

**INFINITE SERIES, STOCHASTIC  
PROCESSES, FUNCTION OPTIMIZATION  
AND THE BAYESIAN PANACEA**

by

**SUCHARITA ROY**

This thesis is submitted in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy (Science) of Jadavpur University



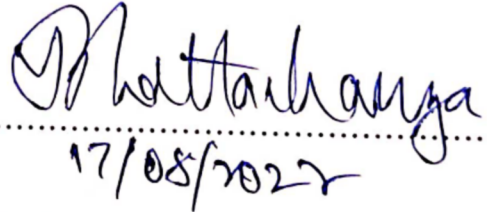
DEPARTMENT OF MATHEMATICS  
JADAVPUR UNIVERSITY  
KOLKATA 700032, WEST BENGAL, INDIA  
2022

To be submitted as per this format

CERTIFICATE FROM THE SUPERVISOR(S)

This is to certify that the thesis entitled "...INFINITE SERIES, STOCHASTIC PROCESSES, FUNCTION OPTIMIZATION AND THE BAYESIAN PANACEA" Submitted by Smt. SUCHARITA ROY who got her name registered on 14.12.2020 for the award of Ph. D. (Science) Degree of Jadavpur University, is absolutely based upon his own work under the supervision of Dr. Sourabh Bhattacharya, Associate Professor, ISRU, ISI, Kolkata and that neither this thesis nor any part of it has been submitted for either any degree / diploma or any other academic award anywhere before.

Dr. Sourabh Bhattacharya

  
17/08/2022

(Signature of the Supervisor(s) date with official seal)

**सौरभ भट्टाचार्य**  
**Sourabh Bhattacharya**  
सह-प्रोफेसर  
Associate Professor  
अंतर्विषयक सांख्यिकीय अनुसंधान युनिट  
Interdisciplinary Statistical Research Unit  
भारतीय सांख्यिकीय संस्थान  
Indian Statistical Institute  
205, डी.टी. रोड, कोलकाता-700108 इंडिया  
205, E. T. Road, Kolkata-700108, INDIA

## ABSTRACT

This thesis aims to solve important problems in topics as varied as deterministic and random infinite series, stochastic processes and function optimization, by embedding the objects in appropriate Bayesian characterization frameworks and then providing the equivalent Bayesian solution. The key philosophy is to view even the deterministic objects as the series elements of deterministic infinite series as realizations of stochastic processes, which facilitates the Bayesian treatment.

Our Bayesian embedding perspective led to Bayesian characterizations of convergence, divergence and oscillations of deterministic and random infinite series; stationarity, nonstationarity, oscillations of general stochastic processes, and also a novel function optimization theory driven by posterior Gaussian derivative process.

Advantages of our Bayesian characterization approach includes equivalent Bayesian solutions to questions of convergence, divergence, oscillations of infinite series where all existing methods fail to provide conclusive answers, equivalent Bayesian assessment of strong and weak stationarity and nonstationarity in time series, spatial and spatio-temporal processes, along with equivalent Bayesian appraisals of complete spatial randomness, strong and weak stationarity and the Poisson assumption in point process analysis. Furthermore, such Bayesian characterization led to method for Bayesian frequency determination in oscillating time series and a reliable method for convergence diagnostics of Markov Chain Monte Carlo algorithms, apart from the novel and accurate function optimization method.

Special mention must be reserved for Bayesian characterization of infinite series, as this attempted to provide solutions to two problems of great importance. One such problem is the celebrated Riemann Hypothesis, the most elusive problem of classical

mathematics, whose solution is the most sought after. The other is related to the global climate change debate, the specific question being the validity of the portentous future global warming projections. The respective results of our Bayesian characterizations of deterministic and random infinite series support neither Riemann Hypothesis, nor future global warming.



# Contents

ABSTRACT	2
<b>1 Introduction</b>	<b>1</b>
<b>2 An Overview of Our Contributions</b>	<b>4</b>
<b>3 Bayes Meets Riemann – Bayesian Characterization of Infinite Series with Application to Riemann Hypothesis</b>	<b>9</b>
3.1 Introduction . . . . .	9
3.2 The key concept . . . . .	14
3.3 A recursive Bayesian procedure for studying infinite series . . . . .	17
3.4 Characterization of convergence properties of the underlying infinite series	22
3.5 Illustrations . . . . .	27
3.6 Application to Riemann Hypothesis . . . . .	43
3.7 Summary and conclusion . . . . .	48
APPENDICES	<b>53</b>
Appendix 3.A1 Proof of Lemma 8 . . . . .	53
Appendix 3.A2 Proof of Lemma 10 . . . . .	54
Appendix 3.A3 Characterization of Riemann Hypothesis based on Bernoulli numbers . . . . .	55
<b>4 Bayesian Characterization of Oscillatory Series with Multiple Limit Points</b>	<b>58</b>

4.1	Introduction . . . . .	58
4.2	Bayesian characterization for finite number of limit points . . . . .	60
4.3	Infinite number of limit points . . . . .	63
4.4	Bayesian characterization of convergence and divergence with our approach on limit points . . . . .	66
4.5	A rule of thumb for diagnosis of convergence, divergence and oscillations	69
4.6	Illustration of our Bayesian theory on oscillation . . . . .	69
4.7	Application of the Bayesian multiple limit points theory to Riemann Hypothesis . . . . .	72
4.8	Summary and conclusion . . . . .	74
<b>5</b>	<b>Bayesian Appraisal of Random Series Convergence with Application to Climate Change</b>	<b>78</b>
5.1	Introduction . . . . .	78
5.2	Random infinite series and parametric upper bound for the partial sums	82
5.3	Simulation experiments with parametric upper bound . . . . .	91
5.4	Nonparametric bounds for the partial sums and simulation experiments	102
5.5	Application of random series convergence diagnostics to global climate change . . . . .	113
5.6	Summary and conclusion . . . . .	119
<b>6</b>	<b>Bayesian Characterizations of Properties of Stochastic Processes with Applications</b>	<b>123</b>
6.1	Introduction . . . . .	123
6.2	Requisite definitions and associated results – prelude to the key concept	126
6.3	The key concept . . . . .	131
6.4	Characterization of stationarity properties of the underlying process . .	133

6.5	Computation of the sup norm between empirical distribution functions associated with $\hat{P}_j$ and $\tilde{P}_K$ . . . . .	137
6.6	Choice of the cardinality of $\mathcal{N}_i$ . . . . .	138
6.7	Stationarity of covariance structure . . . . .	139
6.8	Characterization of stationarity and nonstationarity using non-recursive Bayesian posteriors . . . . .	142
6.9	Summary and conclusion . . . . .	143
<b>7</b>	<b>Application of Bayesian Characterization of Stationarity and Non-stationarity to Time Series and Markov Chain Monte Carlo</b>	<b>145</b>
7.1	Introduction . . . . .	145
7.2	First illustration: AR(1) model . . . . .	149
7.3	Second illustration: AR(2), ARCH(1) and GARCH(1,1) models . . . . .	158
7.4	Third illustration: MCMC convergence diagnostics . . . . .	166
7.5	Summary and conclusion . . . . .	176
<b>8</b>	<b>Applications of Bayesian Characterization of Stochastic Processes to Spatial and Spatio-Temporal Data</b>	<b>177</b>
8.1	Introduction . . . . .	177
8.2	Detection of stationarity and nonstationarity in spatial data . . . . .	180
8.3	Detection of stationarity and nonstationarity in spatio-temporal data . . . . .	193
8.4	Real data analyses for spatial and spatio-temporal data . . . . .	212
8.5	Summary and conclusion . . . . .	217
<b>9</b>	<b>Bayesian Characterization of Point Processes</b>	<b>222</b>
9.1	Introduction . . . . .	222
9.2	A brief overview of the existing CSR test . . . . .	224
9.3	Bayesian characterization of CSR . . . . .	225

9.4	Bayesian characterization of stationarity and nonstationarity of point processes . . . . .	232
9.5	Bayesian characterization of mutual independence among random variables	233
9.6	Bayesian characterization of Poisson point process . . . . .	239
9.7	Simulation experiments . . . . .	242
9.8	Summary and conclusion . . . . .	280
<b>10</b>	<b>Bayesian Determination of Frequencies of Oscillatory Stochastic Processes</b>	<b>282</b>
10.1	Introduction . . . . .	282
10.2	The key idea for Bayesian frequency determination . . . . .	284
10.3	Bayesian theory for finite $M$ . . . . .	285
10.4	Bayesian theory for infinite number of frequencies . . . . .	288
10.5	Simulation experiments . . . . .	290
10.6	Real data example: El Niño and fish population . . . . .	308
10.7	Summary and conclusion . . . . .	314
<b>11</b>	<b>Function Optimization with Posterior Gaussian Derivative Process</b>	<b>316</b>
11.1	Introduction . . . . .	316
11.2	Posterior Gaussian derivative process . . . . .	320
11.3	Posterior distribution of random optima corresponding to the posterior derivative process . . . . .	326
11.4	Almost sure uniform convergence of posterior Gaussian and Gaussian derivative processes . . . . .	328
11.5	Algorithm for optimization with the Gaussian process derivative method	340
11.6	Bayesian characterization of the number of local minima of the objective function with recursive posteriors . . . . .	345
11.7	Experiments . . . . .	348

11.8 Summary and conclusion . . . . .	370
REFERENCES	<b>384</b>

## Listing of figures

3.5.1 Example 1: The series (3.5.1) is divergent. . . . .	29
3.5.2 Example 2: The series (3.5.2) converges for $a > 1$ and diverges for $a \leq 1$ . . . . .	31
3.5.3 Example 3: The series (3.5.6) converges for $a > e$ and diverges for $a \leq e$ . . . . .	33
3.5.4 Example 4: The series (3.1.1) converges for $(a = 3, b = 1)$ , $(a = 1 + 10^{-10}, b = 0)$ , $(a = 1 + 20^{-10}, b = 10^{-10})$ and diverges for $(a = 1/2, b = 1/3)$ . . . . .	35
3.5.5 Example 4: The series (3.1.1) diverges for $(a = \frac{1}{2}(1 - 10^{-11}), b = \frac{1}{2}(1 - 10^{-11}))$ , $(a = 1, b = 0)$ and $(a = 1, b = 1)$ . . . . .	36
3.5.6 Example 5: The series (3.5.13) converges for $(a = 2, b = 1)$ , $(a = 1 + 20^{-10}, b = 10^{-10})$ , $(a = 1 + 30^{-10}, b = 20^{-10})$ and diverges for $(a = 1/2, b = 1/2)$ and $(a = \frac{1}{2}(1 - 10^{-11}), b = \frac{1}{2}(1 - 10^{-11}))$ . . . . .	38
3.5.7 Example 6: The series (3.5.17) converges for $(a = -10^{-10}, b = 2 - 10^{-10})$ , $(a = 10^{-10}, b = 2 - 10^{-10})$ , and diverges for $(a = -10^{-10}, b = 2 + 10^{-10})$ , $(a = 10^{-10}, b = 2 + 10^{-10})$ . . . . .	41
3.5.8 Example 6: The series (3.5.17) converges for $(a = -10^{-10}, b = -10^{-10})$ , $(a = -10^{-10}, b = 10^{-10})$ , $(a = 10^{-10}, b = -10^{-10})$ , $(a = 10^{-10}, b = 10^{-10})$ , and $(a = 0, b = 0)$ . . . . .	42
3.5.9 Example 7: The series (3.5.23) diverges for $(a = \pi^{-1}, b = 1)$ , $(a = 5/7, b = 1)$ . . . . .	43
3.6.1 Plot of the partial sums $S_{1000,1000000}^a$ versus $a$ . Panel (a) shows the plot in the domain $[0, 5]$ while panel (b) magnifies the same in the domain $(0.5, 1)$ . . . . .	46
3.6.2 Riemann Hypothesis: The Möbius function based series diverges for $a = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6$ . . . . .	49

3.6.3 Riemann Hypothesis: The Möbius function based series diverges for $a = 0.7$ but converges for $a = 0.8, 0.9, 1 + 10^{-10}, 2, 3$ . . . . .	50
3.6.4 Riemann Hypothesis: The left panels show the posterior means for the full set of iterations, while the right panels depict the posterior means for the last 500 iterations, for $a = 0.71, 0.715$ and $0.72$ . It is evident that the Möbius function based series diverges for $a = 0.71$ and $0.715$ but converges for $a = 0.72$ . . . . .	51
3.A3. Actual and Stirling-approximated terms $a_m$ of the series $\tilde{S}_1$ and $\tilde{S}_2$ . . . . .	57
4.6.1 Illustration of the Dirichlet process based theory on the first oscillating series: two limit points, each with proportion 0.5, are captured. . . . .	70
4.6.2 Illustration of the Dirichlet process based theory with Example 5: For $(a = 2, b = 1)$ in the series (3.5.13), $\frac{m}{M} = \frac{6}{10} < 0.9$ , indicating convergence, for $(a = 1 + 20^{-10}, b = 10^{-10})$ , $\frac{m}{M} = \frac{6}{10} < 0.9$ , indicating convergence, for $(a = 1 + 30^{-10}, b = 20^{-10})$ , $\frac{m}{M} = \frac{6}{10} < 0.9$ , indicating convergence, for $(a = 1/2, b = 1/2)$ , $\frac{m}{M} = \frac{10}{10} > 0.9$ , indicating divergence, and for $(a = \frac{1}{2}(1 - 10^{-11}), b = \frac{1}{2}(1 - 10^{-11}))$ , $\frac{m}{M} = \frac{10}{10} > 0.9$ , indicating divergence. . . . .	73
4.7.1 Riemann Hypothesis based on Bayesian multiple limit points theory: Divergence for $a = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6$ . . . . .	75
4.7.2 Riemann Hypothesis based on Bayesian multiple limit points theory: Divergence for $a = 0.7$ but convergence for $a = 0.74, 0.8, 0.9, 1, 1 + 10^{-10}$ . . . . .	76
4.7.3 Riemann Hypothesis based on Bayesian multiple limit points theory: Convergence for $a = 2, 3$ . . . . .	77
5.3.1 Example 1: Convergence and divergence for exponential series. . . . .	92
5.3.2 Example 2: Convergence and divergence for normal series. . . . .	94
5.3.3 Example 3: Convergence and divergence for dependent normal series. . . . .	97

5.3.4 Example 4: Convergence and divergence for state-space series. . . . .	99
5.3.5 Example 5: Convergence and divergence for state-space series with hierarchical exponential distribution. . . . .	101
5.3.6 Example 6: Convergence and divergence for RDS. . . . .	103
5.4.1 Example 1 revisited: Convergence and divergence for exponential series with nonparametric bound. . . . .	107
5.4.2 Example 2 revisited: Convergence and divergence for normal series with nonparametric bound. . . . .	108
5.4.3 Example 3 revisited: Convergence and divergence for dependent normal series with nonparametric bound. . . . .	109
5.4.4 Example 4 revisited: Convergence and divergence for state-space series with nonparametric bound. . . . .	111
5.4.5 Example 5 revisited: Convergence and divergence for state-space series with hierarchical exponential distribution. . . . .	112
5.4.6 Example 6 revisited: Convergence and divergence for RDS. . . . .	114
5.5.1 Current, HadCRUT4 global mean temperature data. . . . .	115
5.5.2 Holocene global mean surface temperature reconstructions 12,000 years before present. . . . .	118
7.2.1 Parametric AR(1) example with $K = 20000$ and $n = 10000$ . . . . .	152
7.2.2 Parametric AR(1) example with $K = 50$ and $n = 50$ . . . . .	155
7.2.3 Nonparametric AR(1) example with $K = 50$ and $n = 50$ . . . . .	157
7.3.1 Slow and fast divergence tendencies of AR(2) model for several values of $\alpha$ and $\beta$ . . . . .	160
7.3.2 Nonparametric AR(2) example with $K = 500$ and $n = 5$ . . . . .	162
7.3.3 Nonparametric ARCH(1) example with $K = 500$ and $n = 5$ . . . . .	163
7.3.4 Comparison of ARCH(1) samples for several values of $\alpha$ where our Bayesian method failed. . . . .	165



7.3.5 Nonparametric GARCH(1,1) example with $K = 500$ and $n = 5$ . . . . .	167
7.3.6 GARCH(1,1) sample for $\alpha = 0.5$ and $\beta = 0.5$ where our Bayesian method failed. . . . .	168
7.4.1 Additive TMCMC convergence example, with $K = 1000$ and $n = 1000$ . .	172
7.4.2 Additive TMCMC convergence example, with $K = 1000$ and $n = 1000$ . .	173
7.4.3 Additive TMCMC convergence example, with $K = 1000$ and $n = 1000$ . .	174
7.4.4 Additive TMCMC convergence example for mixture densities. . . . .	175
8.2.1 Detection of strong stationarity and nonstationarity in spatial data drawn from GPs. . . . .	183
8.2.2 Detection of covariance stationarity and nonstationarity in spatial data drawn from GPs. . . . .	186
8.2.3 Detection of strong nonstationarity in spatial data drawn from GP with covariance structure (8.2.4) with $p = 0.99999$ . . . . .	187
8.2.4 Detection of covariance nonstationarity in spatial data drawn from GP with covariance structure (8.2.4) with $p = 0.99999$ . . . . .	189
8.2.5 Detection of strong stationarity and nonstationarity in spatial data of size 1000 drawn from GPs. . . . .	190
8.2.6 Detection of strong nonstationarity in spatial data of size 1000 drawn from GP with covariance structure (8.2.4) with $p = 0.99999$ . . . . .	191
8.3.1 Detection of strong stationarity and nonstationarity in spatio-temporal data drawn from GPs. . . . .	197
8.3.2 Detection of covariance stationarity in spatio-temporal data drawn from GP with covariance structure (8.3.1) with $\rho = 0.99999$ . . . . .	198
8.3.3 Detection of covariance nonstationarity in spatio-temporal data drawn from GP with covariance structure (8.3.3) with $p = 0.99999$ and $\rho = 0.8$ . . . . .	199
8.3.4 Detection of strong stationarity and nonstationarity in spatio-temporal data drawn from GPs with 50 locations and 20 time points. . . . .	201

8.3.5	Detection of strong stationarity in spatio-temporal data drawn from models $S1$ and $S2$ with sample size 100 locations and 200 time points, with $\psi = 1$ and $\lambda = 5$ .	203
8.3.6	Detection of covariance stationarity in spatio-temporal data drawn from model $S1$ with sample size 100 locations and 200 time points, with $\psi = 1$ and $\lambda = 5$ .	204
8.3.7	Detection of covariance stationarity in spatio-temporal data drawn from model $S2$ with sample size 100 locations and 200 time points, with $\psi = 1$ and $\lambda = 5$ .	205
8.3.8	Detection of strong stationarity in spatio-temporal data drawn from models $S1$ and $S2$ with sample size 100 locations and 200 time points, corresponding to Whittle spatial covariance with $\psi = 0.72$ and $\lambda = 5$ .	206
8.3.9	Detection of covariance stationarity in spatio-temporal data drawn from model $S1$ with sample size 100 locations and 200 time points, corresponding to Whittle spatial covariance with $\psi = 0.72$ and $\lambda = 5$ .	207
8.3.10	Detection of covariance stationarity in spatio-temporal data drawn from model $S2$ with sample size 100 locations and 200 time points, corresponding to Whittle spatial covariance with $\psi = 0.72$ and $\lambda = 5$ .	208
8.3.11	Detection of strong nonstationarity in spatio-temporal data drawn from models $NS1$ , $NS2$ and $NS3$ with sample size 100 locations and 200 time points.	210
8.3.12	Detection of covariance nonstationarity in spatio-temporal data drawn from model $NS1$ with sample size 100 locations and 200 time points.	211
8.4.1	Detection of nonstationarity of the ozone data with our Bayesian method.	213
8.4.2	Detection of covariance nonstationarity of the ozone data.	214
8.4.3	GP samples of sizes 10000 and 20000 for Whittle covariance with $\psi = 0.8$ for PM 10 data.	216

8.4.4	Detection of nonstationarity of the PM 10 data with our Bayesian method.	217
8.4.5	Detection of covariance nonstationarity of the PM 10 data. . . . .	218
8.4.6	GP sample of size 17496 for Whittle covariance with $\psi = 0.8$ for PM 2.5 data. . . . .	219
8.4.7	Detection of stationarity of the PM 2.5 data with our Bayesian method.	219
9.7.1	Homogeneous and inhomogeneous Poisson point processes. . . . .	243
9.7.2	Detection of CSR with our Bayesian method and traditional classical method. . . . .	245
9.7.3	Detection of stationarity and nonstationarity of point processes (here HPP and IHPP) with our Bayesian method. . . . .	246
9.7.4	Detection of independence in point patterns (here HPP and IHPP) with our Bayesian method, suggesting that both the point processes are Poisson point processes. . . . .	247
9.7.5	Homogeneous LGCP. . . . .	248
9.7.6	Detection of CSR with our Bayesian method and traditional classical method for LGCP. The Bayesian method correctly identifies that the underlying point process is not CSR, but the classical method falsely indicates CSR. . . . .	249
9.7.7	Detection of stationarity and dependence of homogeneous LGCP with our Bayesian method. . . . .	249
9.7.8	Inhomogeneous LGCP. . . . .	250
9.7.9	Detection of CSR with our Bayesian method and traditional classical method for LGCP. Both the methods correctly identify that the underlying point process is not CSR. . . . .	251
9.7.10	Detection of nonstationarity and dependence of inhomogeneous LGCP with our Bayesian method. . . . .	252
9.7.11	Inhomogeneous LGCP. . . . .	253

9.7.1	Detection of CSR with our Bayesian method and traditional classical method for LGCP. Both the methods correctly identify that the underlying point process is not CSR. . . . .	253
9.7.1	Detection of nonstationarity and dependence of inhomogeneous LGCP with our Bayesian method. . . . .	254
9.7.1	Matérn cluster point process pattern. . . . .	255
9.7.1	Detection of CSR with our Bayesian method and traditional classical method for Matérn cluster process. Both the methods correctly identify that the underlying point process is not CSR. . . . .	256
9.7.1	Detection of stationarity and dependence of Matérn cluster process with our Bayesian method. . . . .	256
9.7.1	Inhomogeneous Matérn cluster point process pattern. . . . .	257
9.7.1	Detection of CSR with our Bayesian method and traditional classical method for inhomogeneous Matérn cluster process. Both the methods correctly identify that the underlying point process is not CSR. . . . .	258
9.7.1	Detection of nonstationarity and dependence of Matérn cluster process with our Bayesian method. . . . .	258
9.7.2	Inhomogeneous Matérn cluster point process pattern. . . . .	259
9.7.2	Detection of CSR with our Bayesian method and traditional classical method for inhomogeneous Matérn cluster process. Both the methods correctly identify that the underlying point process is not CSR. . . . .	260
9.7.2	Detection of nonstationarity and dependence of Matérn cluster process with our Bayesian method. . . . .	260
9.7.2	Homogeneous Thomas point process pattern. . . . .	261
9.7.2	Detection of CSR with our Bayesian method and traditional classical method for homogeneous Thomas point process. Both the methods correctly identify that the underlying point process is not CSR. . . . .	262

9.7.2	Detection of stationarity and dependence of homogeneous Thomas process with our Bayesian method. . . . .	262
9.7.2	Inhomogeneous Thomas point process pattern. . . . .	263
9.7.2	Detection of CSR with our Bayesian method and traditional classical method for Inhomogeneous Thomas point process. Both the methods correctly identify that the underlying point process is not CSR. . . . .	263
9.7.2	Detection of nonstationarity and dependence of inhomogeneous Thomas process with our Bayesian method. . . . .	264
9.7.2	Inhomogeneous Thomas point process pattern. . . . .	265
9.7.3	Detection of CSR with our Bayesian method and traditional classical method for inhomogeneous Thomas point process. Both the methods correctly identify that the underlying point process is not CSR. . . . .	265
9.7.3	Detection of nonstationarity and dependence of inhomogeneous Thomas process with our Bayesian method. . . . .	266
9.7.3	Inhomogeneous Thomas point process pattern. . . . .	267
9.7.3	Detection of CSR with our Bayesian method and traditional classical method for inhomogeneous Thomas point process. Both the methods correctly identify that the underlying point process is not CSR. . . . .	267
9.7.3	Detection of nonstationarity and dependence of inhomogeneous Thomas process with our Bayesian method. . . . .	268
9.7.3	Inhomogeneous Thomas point process pattern. . . . .	269
9.7.3	Detection of CSR with our Bayesian method and traditional classical method for inhomogeneous Thomas point process. Both the methods correctly identify that the underlying point process is not CSR. . . . .	269
9.7.3	Detection of nonstationarity and dependence of inhomogeneous Thomas process with our Bayesian method. . . . .	270
9.7.3	Inhomogeneous Thomas point process pattern. . . . .	271

9.7.3	Detection of CSR with our Bayesian method and traditional classical method for inhomogeneous Thomas point process. Both the methods correctly identify that the underlying point process is not CSR. . . . .	271
9.7.4	Detection of nonstationarity and dependence of Inhomogeneous Thomas process with our Bayesian method. . . . .	272
9.7.4	Homogeneous Neyman-Scott point process pattern. . . . .	273
9.7.4	Detection of CSR with our Bayesian method and traditional classical method for homogeneous Neyman-Scott point process. Both the methods correctly identify that the underlying point process is not CSR. . . . .	274
9.7.4	Detection of stationarity and dependence of homogeneous Neyman-Scott process with our Bayesian method. . . . .	274
9.7.4	Inhomogeneous Neyman-Scott point process pattern. . . . .	275
9.7.4	Detection of CSR with our Bayesian method and traditional classical method for inhomogeneous Neyman-Scott point process. Both the methods correctly identify that the underlying point process is not CSR. . . . .	275
9.7.4	Detection of nonstationarity and dependence of inhomogeneous Neyman-Scott process with our Bayesian method. . . . .	276
9.7.4	Strauss point process pattern. . . . .	277
9.7.4	Detection of CSR with our Bayesian method and traditional classical method for Strauss process. Both the methods correctly identify that the underlying point process is not CSR. . . . .	278
9.7.4	Detection of stationarity and dependence of Strauss process with our Bayesian method. . . . .	278
9.7.5	Strauss Process. . . . .	279
9.7.5	Detection of CSR with our Bayesian method and traditional classical method for Strauss process. Both the methods correctly identify that the underlying point process is not CSR. . . . .	279

9.7.5	Detection of stationarity and dependence of Strauss process with our Bayesian method. . . . .	280
10.5.	Simulated oscillating time series with true frequency 0.02. . . . .	291
10.5.	Illustration of effects of $r$ in $\mathbf{Z}^r$ in determining single frequency in (10.5.1). Here the true frequency is 0.02. . . . .	292
10.5.	Illustration of our Bayesian method for determining single frequency. Here the true frequency is 0.02. . . . .	294
10.5.	Illustration of our Bayesian method for determining single frequency. Here the true frequency is 0.02. . . . .	295
10.5.	Illustration of our Bayesian method for determining single frequency for long enough time series. Here the true frequency is 0.02. . . . .	297
10.5.	Convergence of our Bayesian method to the true frequency 0.02 for long enough time series with $r = 1000$ and $M = 40$ . . . . .	298
10.5.	Simulated oscillating time series with true frequencies 0.4, 0.1 and 0.06. . . . .	298
10.5.	Illustration of effects of $r$ in $\mathbf{Z}^r$ in determining multiple frequencies in (10.5.2). Here the true frequencies are 0.4, 0.1 and 0.06. . . . .	300
10.5.	Illustration of our Bayesian method for determining multiple frequencies. Here the true frequencies are 0.4, 0.1 and 0.06. . . . .	302
10.5.	Illustration of our Bayesian method for determining multiple frequencies. Here the true frequencies are 0.4, 0.1 and 0.06. . . . .	303
10.5.	Illustration of our Bayesian method for determining multiple frequencies. Here the true frequencies are 0.4, 0.1 and 0.06. . . . .	304
10.5.	Illustration of our Bayesian method for determining multiple frequencies. Here the true frequencies are 0.4, 0.1 and 0.06. . . . .	305
10.5.	Illustration of our Bayesian method for determining multiple frequencies. Here the true frequencies are 0.4, 0.1 and 0.06. . . . .	306
10.5.	The original and the transformed signal with 6 harmonics. . . . .	307

10.5.1	Illustration of our Bayesian method for determining multiple frequencies in non-sinusoidal signals. Here the true frequencies are 2, 4, 6, 8, 10 and 12 oscillations per unit time. . . . .	309
10.6.1	The original and the transformed SOI time series. . . . .	311
10.6.2	Bayesian results for frequency determination of the SOI time series. . .	312
10.6.3	The original and the transformed Recruitment time series. . . . .	313
10.6.4	Bayesian results for frequency determination of the Recruitment time series with transformed time series $\exp(25) \times \mathbf{Z}^{50}$ . . . . .	314
11.7.1	EMCMC trace plots for Example 1. . . . .	351
11.7.2	EMCMC trace plots for Example 2. . . . .	352
11.7.3	EMCMC trace plots for Example 3 for finding maxima. . . . .	355
11.7.4	EMCMC trace plots for Example 3 for finding the first saddle point. . .	356
11.7.5	EMCMC trace plots for Example 3 for finding the second saddle point. .	357
11.7.6	EMCMC trace plots for Example 3 for investigating inconclusiveness. . .	359
11.7.7	EMCMC trace plots for Example 4 for finding MLE. . . . .	362
11.7.8	EMCMC trace plots for Example 5 for finding MLE for dimension $d = 2$ . .	365
11.7.9	EMCMC trace plots for Example 5 for finding MLE for dimension $d = 5$ . .	367
11.7.10	EMCMC trace plots for Example 5 for finding MLE for dimension $d = 10$ . .	368
11.7.11	EMCMC trace plots for Example 5 for finding MLE for dimension $d = 50$ . .	369
11.7.12	EMCMC trace plots for Example 5 for finding MLE for dimension $d = 100$ . .	371



## Listing of tables

8.2.1 The performance evaluation of the test statistics $T$ and $V$ of Bandopadhyay and Rao (2017) and Bandopadhyay <i>et al.</i> (2017) applied to our simulated spatial datasets. . . . .	221
---	-----

*I dedicate this thesis to my parents and teachers*

# Acknowledgments

*I thank my advisor Dr. Sourabh Bhattacharya for encouraging my research and for allowing me to grow as a researcher.*

*I would also like to express my gratitude to all the faculty members and the staff of St. Xavier's College, Kolkata, for their continued support.*

*I also thank Arun Kumar Kuchibhotla, Debapratim Banerjee and Satyaki Mazumder for comments on some parts of the thesis.*

# 1

## Introduction

As the thesis title indicates, the goal of this thesis is to offer solutions to important problems in topics, as varied as deterministic and random infinite series, stochastic processes and function optimization, after essentially bringing them under similar Bayesian characterization umbrella. Outwardly, this may seem outrageously uncanny, since the Bayesian premise is a statistical paradigm that deals with random objects, and so even though stochastic processes are appropriate candidates for the Bayesian treatment, it may be extremely difficult to perceive the links of the Bayesian paradigm with the completely deterministic mathematical topics like deterministic infinite series and function optimization. Hitherto, even random infinite series has no connection whatsoever with the Bayesian paradigm.

Nevertheless, however uncanny it might sound, it is not difficult to anticipate that if such a task is at all possible, then sweet might be the fruits of embedding the seeds of important mathematical topics into the fertile Bayesian soil.

In this thesis, we explore such possibilities, as we attempt to provide Bayesian characterizations of convergence, divergence and oscillations of infinite series; stationarity, nonstationarity, oscillations and other important properties of stochastic processes in general, and also function optimization in a novel Bayesian Gaussian derivative process setup. Our efforts led us to deal with some problems of great importance, namely, the celebrated Riemann Hypothesis, the most notorious unsolved problem for more than 150 years, and global warming, that lies at the heart of the most controversial climate change debate. Our Bayesian solutions to the Riemann Hypothesis yielded the very surprising conclusion that its validity can not be supported. The ominous future global warming projections are not upheld either by our Bayesian characterization ideas. Both the solutions are applications of our Bayesian characterization of deterministic and random infinite series, respectively.

Apart from dealing with the aforementioned important problems, other fruits of our investigations include powerful Bayesian characterization theories and methods for testing strong and weak stationarity and nonstationarity in time series analysis, spatial and spatio-temporal analysis; Bayesian characterization based tests for stationarity, nonstationarity, complete spatial randomness and the Poisson assumption in point process analysis, frequency determination of oscillating time series and an effective function optimization theory embedded in a novel Bayesian framework that facilitates more accurate optimization compared to the existing optimization methods. Furthermore, our Bayesian characterization of stationarity and nonstationarity provides a novel convergence assessment method for Markov Chain Monte Carlo algorithms designed to simulate from complicated distributions of interest. The above-mentioned areas are either very little explored or completely unexplored in the literature, and moreover, comparison of our Bayesian results with results of the other methods, whenever existent and relevant, indicates superiority of our ideas in most situations.

It is interesting to note that the key idea of Bayesian characterization emerged as a

response to a simple curiosity of the thesis author, with regard to infinite series, which was even rejected at the first thought on the ground of being too esoteric. The thesis author, who is a professor and the head of the Department of Mathematics at St. Xavier's College, Kolkata, noted with dismay, perhaps like many other mathematics teachers, that although determination of convergence, divergence or oscillation of infinite series is a much-studied problem in classical mathematics, unfortunately there does not yet seem to exist any universal test that can provide conclusive answers regarding convergence of most infinite series. This issue kept preventing her from answering the relevant questions from her students regarding series convergence. Hearing of the powerful Bayesian paradigm from some of her (over-enthusiastic!) colleagues as the panacea to all problems, she was left wondering about answering questions of series convergence by surrendering to the Bayesian power. Her thesis supervisor, a Bayesian, considering this an innocuous banter, did not take it seriously at the first thought. However, importance of the banter dawned on him with an afterthought, and the rest is . . .this thesis!

In the next chapter we provide a brief overview of our contributions.

# 2

## An Overview of Our Contributions

To begin with, in Chapter 3, we propose a novel Bayesian approach that attempts to provide conclusive answers to the question of series convergence even where all the existing tests fail. As can be anticipated from our “Bayesian approach”, the key philosophy is to embed this deterministic problem of classical mathematics in a stochastic framework. In a nutshell, we develop a recursive Bayesian technique and construct Bayesian posteriors at successive stages of the partial sums associated with the infinite series. Interestingly enough our posterior distributions characterized the convergence as well as divergence of the infinite series in the Bayesian framework. The theory that we propose does not even assume independence of the random variables and applies to any arbitrary infinite series. Application of our Bayesian theory to various infinite series, ranging from simple to complicated, has not only shown results that are in complete agreement with the existing literature but also provided conclusion even where all the existing tests of convergence failed. However, the most path-breaking application of our approach turns out to be in

the investigation of the hardest unsolved problem of mathematics today, the celebrated Riemann Hypothesis. Indeed, since Riemann Hypothesis can be characterized in terms of an infinite series based on the Möbius function, our theory and methods are readily applicable for the investigation of the famous conjecture. As already mentioned, we have obtained results that do not support the hypothesis.

In Chapter 4, we extend our theory to encompass infinite series with finite as well as countably infinite number of limit points. We also apply the multiple limit point theory to characterize convergence and divergence of non-oscillating infinite series, using which we further investigate Riemann Hypothesis and obtain identical conclusion as in Chapter 3. These results have strengthened our belief that the conjecture cannot be completely supported. The insights that we obtained about the most challenging problem of classical mathematics is definitely one of the most encouraging parts of our research works presented in this thesis.

In contrast with deterministic series considered in Chapters 3 and 4, in Chapter 5 we take up random infinite series for our investigation. Remarkably, our method does not require any simplifying assumption, such as independence or restrictive dependence among the random variables. Albeit the Bayesian characterization theory for random series is no different from that for the deterministic setup, construction of effective upper bounds for partial sums, required for implementation, turns out to be a challenging undertaking in the random setup. The difficulty steps in as the consequence of non-availability of the functional forms of the random summands of the series, and the problem persists even if the distributions of the summands are assumed to be known. In Chapter 5, we first construct parametric upper bound forms assuming parametric densities of the random summands. But despite their mathematical validity for non-negative summands and good performance in such setups, they are not generally applicable, which leads us to propose a flexible bound for general setups. But even for series driven by normal distributions, the general bound exhibits correct but very inefficient and



less persuasive convergence analysis. Moreover, application to random Dirichlet series yields wrong answers in many cases. Hence, we propose a general nonparametric bound structure for our purpose. Simulation studies demonstrate high accuracy and efficiency of the nonparametric bound in all the setups that we consider. Finally, exploiting the property that the summands tend to zero in the case of series convergence, we consider application of our nonparametric bound driven Bayesian method to global climate change analysis. Specifically, analyzing the global average temperature record over the years 1850 – 2016 and Holocene global average temperature reconstruction data 12,000 years before present, we conclude, in spite of the current global warming situation, that global climate dynamics is subject to temporary variability only, the current global warming being an instance, and long term global warming or cooling either in the past or in the future, are highly unlikely.

We next turn to Bayesian characterizations of properties of stochastic processes. In this regard, in Chapter 6, we primarily propose a novel Bayesian characterization of stationary and nonstationary stochastic processes. In practice, this theory aims to distinguish between global stationarity and nonstationarity for both parametric and nonparametric stochastic processes. Interestingly, our theory builds on our previous work on Bayesian characterization of infinite series, which was applied to verification of the (in)famous Riemann Hypothesis. Thus, there seems to be interesting and important connections between pure mathematics and Bayesian statistics, with respect to our proposed ideas. We validate our proposed method with simulation and real data experiments associated with different setups. In particular, applications of our method include stationarity and nonstationarity determination in various time series models, spatial and spatio-temporal setups, and convergence diagnostics of Markov Chain Monte Carlo. Our results demonstrate very encouraging performance, even in very subtle situations. These applications are considered in Chapters 7 and 8. Using similar principles, in Chapter 9 we also provide a novel Bayesian characterization of mutual

independence among any number of random variables, using which we characterize the properties of point processes, including characterizations of Poisson point processes, complete spatial randomness, stationarity and nonstationarity. Applications to simulation experiments with ample Poisson and non-Poisson point process models again indicate quite encouraging performance of our proposed ideas.

Further, in Chapter 10, we propose a novel recursive Bayesian method for determination of frequencies of oscillatory stochastic processes, based on our general principle. Simulation studies and real data experiments with varieties of time series models consisting of single and multiple frequencies bring out the worth of our method.

Function optimization is a research area that has wide applications in all scientific disciplines. Yet, for any sufficiently large class of optimization problems it is considerably difficult to single out any optimization methodology that can outperform the others in terms of theoretical foundation, accuracy, computational efficiency or robustness. Indeed, given any optimization problem, it is customary to search for methods that might be effective, and in most cases, some heuristic method is ultimately taken into consideration. In Chapter 11, we propose and develop a novel Bayesian algorithm for optimization of functions whose first and second partial derivatives are known. The basic premise is the Gaussian process representation of the function which induces a first derivative process that is also Gaussian. The Bayesian posterior solutions of the derivative process set equal to zero, given data consisting of suitable choices of input points in the function domain and their function values, emulate the stationary points of the function, which can be fine-tuned by setting restrictions on the prior in terms of the first and second derivatives of the objective function. These observations motivate us to propose a general and effective algorithm for function optimization that attempts to get closer to the true optima adaptively with in-built iterative stages. We provide theoretical foundation to this algorithm, proving almost sure convergence to the true optima as the number of iterative stages tends to infinity. The theoretical foundation

hinges upon our proofs of almost sure uniform convergence of the posteriors associated with Gaussian and Gaussian derivative processes to the underlying function and its derivatives in appropriate fixed-domain infill asymptotics setups; rates of convergence are also available. We also provide Bayesian characterization of the number of optima using information inherent in our optimization algorithm. We illustrate our Bayesian optimization algorithm with five different examples involving maxima, minima, saddle points and even inconclusiveness. Our examples range from simple, one-dimensional problems to challenging 50 and 100-dimensional problems. While we obtain encouraging and interesting results in each case, we shed light on various issues regarding computation and accuracy along the way. A general and important issue is that our algorithm is able to capture significantly more accurate solutions than the existing optimization algorithms thanks to the posterior simulation approach embedded in our method.

# 3

## Bayes Meets Riemann – Bayesian Characterization of Infinite Series with Application to Riemann Hypothesis

### 3.1 Introduction

Determination of convergence, divergence or oscillation of infinite series has a very rich tradition in mathematics, and a large number of tests exist for the purpose. Unfortunately, there does not seem to exist any universal test that provides conclusive answers to all infinite series; see, for example, [Ilyin and Poznyak \(1982\)](#), [Knopp \(1990\)](#), [Bourchtein \*et al.\* \(2012\)](#). Attempts to resolve the issue as much as possible using hierarchies of tests, with the successive tests in the hierarchy providing conclusive answers to successively larger ranges of infinite series, are provided by [Knopp \(1990\)](#), [Bromwich \(2005\)](#), [Bourchtein](#)

*et al.* (2011) and Lifyand *et al.* (2011). These tests are based on the Kummer approach for positive series and the chain of the Ermakov tests for positive monotone series. The hierarchy of tests provided in Burchtein *et al.* (2012) are based on Bromwich (2005) and are related to the well-known Cauchy's test (see, for example, Fichtenholz (1970), Rudin (1976), Spivak (1994)). Below we briefly discuss the approach of Burchtein *et al.* (2012), who consider positive series. It is important to remark at the outset that positive series is not a requirement for the approaches that we propose and develop in this work.

### 3.1.1 Hierarchical tests of convergence

The tests of Burchtein *et al.* (2012) are based on the following theorem, which is a refinement of a result of Bromwich (2005).

**Theorem 1 (Burchtein *et al.* (2012))** *Let  $\sum_{i=1}^{\infty} F'(i)$  be a divergent series where  $F(x) > 0$ ,  $F'(x) > 0$  and  $F'(x)$  is decreasing. If  $\sum_{i=1}^{\infty} X_i$  is a positive series, then denoting  $\frac{\log\left\{\frac{F'(i)}{X_i}\right\}}{\log F'(i)} = W_i$ , the following hold:*

$$\begin{aligned} \text{If } \liminf_{i \rightarrow \infty} W_i > 1, \text{ then } \sum_{i=1}^{\infty} X_i \text{ converges;} \\ \text{If } \limsup_{i \rightarrow \infty} W_i < 1, \text{ then } \sum_{i=1}^{\infty} X_i \text{ diverges.} \end{aligned}$$

Letting  $F(z) = z$  in the above theorem, Burchtein *et al.* (2012) obtain their first test, which we provide below.

**Theorem 2 (Test  $T_1$  of Burchtein *et al.* (2012))** *Consider a positive series  $\sum_{i=1}^{\infty} X_i$*

and let  $T_{1,i} = \frac{i}{\log i} \left(1 - X_i^{\frac{1}{i}}\right)$ . Then

If  $\liminf_{i \rightarrow \infty} T_{1,i} > 1$ , then  $\sum_{i=1}^{\infty} X_i$  converges;

If  $\limsup_{i \rightarrow \infty} T_{1,i} < 1$ , then  $\sum_{i=1}^{\infty} X_i$  diverges.

This result is the same as that of [Bromwich \(2005\)](#), but a proof was not supplied in that work.

Now choosing  $F(z) = \log z$ , [Bourchtein et al. \(2012\)](#) form their second test of the hierarchy; we provide the result below. Again, the result has been formulated by [Bromwich \(2005\)](#), but a proof was not given.

**Theorem 3 (Test  $T_2$  of [Bourchtein et al. \(2012\)](#))** Consider a positive series  $\sum_{i=1}^{\infty} X_i$  and let  $T_{2,i} = \frac{\log i}{\log \log i} (T_{1,i} - 1)$ . Then

If  $\liminf_{i \rightarrow \infty} T_{2,i} > 1$ , then  $\sum_{i=1}^{\infty} X_i$  converges;

If  $\limsup_{i \rightarrow \infty} T_{2,i} < 1$ , then  $\sum_{i=1}^{\infty} X_i$  diverges.

Setting  $F(z) = \log \log z$ , the following result has been proved by [Bourchtein et al. \(2012\)](#):

**Theorem 4 (Test  $T_3$  of [Bourchtein et al. \(2012\)](#))** Consider a positive series  $\sum_{i=1}^{\infty} X_i$  and let  $T_{3,i} = \frac{\log i}{\log \log i} (T_{2,i} - 1)$ . Then

If  $\liminf_{i \rightarrow \infty} T_{3,i} > 1$ , then  $\sum_{i=1}^{\infty} X_i$  converges;

If  $\limsup_{i \rightarrow \infty} T_{3,i} < 1$ , then  $\sum_{i=1}^{\infty} X_i$  diverges.

Successively selecting  $F(z) = \log \log \log z$ ,  $F(z) = \log \log \log \log z$ , etc. successively more

refined tests  $T_4, T_5$ , etc. can be constructed, with each test having wider scope compared to the preceding test with regard to obtaining conclusive decision on convergence or divergence of the underlying series.

However, if, say, at stage  $k$ ,  $\liminf_{i \rightarrow \infty} T_{k,i} < 1 < \limsup_{i \rightarrow \infty} T_{k,i}$  so that  $T_k$  is inconclusive, then all the subsequent tests will also fail to provide any conclusion. Thus, in spite of the above developments, conclusion regarding the series can still be elusive. For instance, an example considered in [Bourchtein \*et al.\* \(2012\)](#) is the following series:

$$S_1 = \sum_{i=3}^{\infty} \left( 1 - \frac{\log i}{i} - \frac{\log \log i}{i} \left\{ \cos^2 \left( \frac{1}{i} \right) \right\} (a + (-1)^i b) \right)^i, \quad (3.1.1)$$

where  $a \geq 0$  and  $b \geq 0$ . For  $a = b = 1$ ,  $\liminf_{i \rightarrow \infty} T_{2,i} = 0 < 1 < 2 = \limsup_{i \rightarrow \infty} T_{2,i}$ . Hence, the hierarchy of tests  $\{T_k; k \geq 1\}$  fails to provide definitive answer to the question of convergence of the above series.

In fact, we can generalize the series (3.1.1) such that the hierarchy of tests fails for the general class of series. Indeed, consider

$$S_2 = \sum_{i=3}^{\infty} \left( 1 - \frac{\log i}{i} - \frac{\log \log i}{i} f(i) (a + (-1)^i b) \right)^i, \quad (3.1.2)$$

where  $0 \leq f(i) \leq 1$  for all  $i = 1, 2, 3, \dots$ , and  $f(i) \rightarrow 1$  as  $i \rightarrow \infty$ . Such a function can be easily constructed as follows. Let  $g(i)$  be positive and monotonically increase to  $c$ , where  $c > 0$ . Then let  $f(i) = g(i)/c$ , for  $i = 1, 2, 3, \dots$ . A simple example of such a function  $g$  is  $g(i) = c - \frac{1}{i}$ ;  $g(i) = \cos^2 \left( \frac{1}{i} \right)$  is another example, showing the generality of (3.1.2) compared to (3.1.1).

### 3.1.2 Riemann Hypothesis and series convergence

It is well-known that the famous Riemann Hypothesis is equivalent to convergence of an infinite series on a certain interval. A brief introduction to the problem, along with the necessary background, is provided in Section 3.6. Studying the relevant infinite series,

if at all possible, is then the most challenging problem of mathematics. The existing mathematical literature, however, does not seem to be able to provide any directions in this regard. Hence, innovative theories and methods for analyzing infinite series should be particularly welcome.

Note that direct and successive evaluation of sums of consecutive terms of the deterministic series of interest need not even provide any insight into the convergence behaviour of the series. This is because if the said sum seems to have approximately stabilized after a large number of successive evaluations, a further large number of evaluations may reveal a slow increase of the sums. On the other hand, even though initially the sums might exhibit an increasing nature, eventually they might stabilize. To combat such problems, it would be worthwhile to create some appropriate transformation of the sums such that convergence of the series may be indicated if the transformed sums approach a certain pre-defined value (say, 1), and divergence would be anticipated if the transformed sums approach another pre-defined value (say, 0), in a large number of evaluations. Although these two pre-defined values and the progress of the transformed sums towards these values in a large, but finite number of evaluations do not, in any way, formally settle the question of convergence of the underlying series, strong evidence regarding the convergence behaviour may be gained, when the number of evaluations is considerably large.

In this work, our approach of characterization of convergence properties of infinite series is based on the aforementioned intuition, which we formalize rigorously through a novel Bayesian procedure. We subsequently extend the idea and formalism to infinite series with multiple or even infinite number of limit points. The main motivation and the idea of Bayesian formalism is illustrated in Section 3.2.



### 3.2 The key concept

Let us assume that the terms  $\{x_1, x_2, \dots\}$  of any deterministic infinite series of the form  $\sum_{i=1}^{\infty} x_i$  of interest is a realization of some stochastic process  $\{X_i : i = 1, 2, \dots\}$ , so that  $\sum_{i=1}^{\infty} x_i$  is a realization of the corresponding random infinite series

$$S_{1,\infty} = \sum_{i=1}^{\infty} X_i. \quad (3.2.1)$$

In the above, we do not assume any distributional form for  $\{X_i : i = 1, 2, \dots\}$ , signifying the nonparametric nature of our problem. Let  $p \in [0, 1]$  denote the probability of convergence the sum  $S_{1,\infty}$ . In particular, if  $\{X_i : i = 1, 2, \dots\}$  are independent, then by Kolmogorov's 0-1 law (see, for example, [Stroock \(1999\)](#)),  $p$  is either 0 or 1, where 0 stands for divergence of almost all realizations of  $S_{1,\infty}$  and 1 is associated with convergence of almost all realizations of  $S_{1,\infty}$ . Kolmogorov's three series theorem (see, for example, [Stroock \(1999\)](#)) helps determine in this case if  $p = 0$  or  $p = 1$ . However, the three series theorem requires parametric specification of the distributions of  $\{X_i : i = 1, 2, \dots\}$ , and specific choices of the parameters determine if  $p = 0$  or  $p = 1$ . Since our goal is to determine the convergence behaviour of the deterministic series  $\sum_{i=1}^{\infty} x_i$ , interpreted as a realization of the specified stochastic process, different choices of the parameters would lead to convergence and divergence of the same series, along with almost all other realizations of the stochastic process. In other words, Kolmogorov's three series theorem is inappropriate when it comes to determination of convergence behaviour of deterministic series.

If the random variables are not independent, then it may happen that some of the realizations of  $S_{1,\infty}$  are convergent, some are divergent and the rest are oscillatory. Since the above argument regarding Kolmogorov's three series theorem shows that it is inappropriate to assume parametric forms of the distributions of the random variables, we do not assume any particular distributional form of  $\{X_i : i = 1, 2, \dots\}$ . It then

follows that the value of  $p$  is unknown, so that from the Bayesian perspective, one must acknowledge uncertainty about  $p$  in the form of some appropriate prior.

Now, specifying a prior directly on  $p$  associated with the entire infinite series and computing the posterior given  $\{X_i : i = 1, 2, \dots\}$ , is not a valid proposition, as computing the likelihood would require evaluation of infinite number of terms associated with the infinite series, which amounts to knowing the convergence behaviour of the series of interest. Instead, it makes sense to specify priors on the probabilities associated with the finite partial sums of the form  $\sum_{i=m}^n X_i$ , for  $m \leq n$ . Indeed, let

$$P\left(\left|\sum_{i=m}^n X_i\right| \leq c_{m,n}\right) = p_{m,n},$$

where  $c_{m,n}$  are non-negative quantities satisfying  $c_{m,n} \downarrow 0$  as  $m, n \rightarrow \infty$ . Thus, the probability depends on how large  $m$  and  $n$  are.

Now note that, as  $m, n \rightarrow \infty$ ,

$$\mathbb{I}\{|\sum_{i=m}^n X_i| \leq c_{m,n}\} \rightarrow \mathbb{I}\left\{\lim_{m,n \rightarrow \infty} |\sum_{i=m}^n X_i| = 0\right\}$$

almost surely, so that uniform integrability leads to

$$\begin{aligned} \lim_{m,n \rightarrow \infty} p_{m,n} &= \lim_{m,n \rightarrow \infty} P\left(\left|\sum_{i=m}^n X_i\right| \leq c_{m,n}\right) \\ &= \lim_{m,n \rightarrow \infty} E\left(\mathbb{I}\{|\sum_{i=m}^n X_i| \leq c_{m,n}\}\right) = E\left(\mathbb{I}\left\{\lim_{m,n \rightarrow \infty} |\sum_{i=m}^n X_i| = 0\right\}\right) \\ &= P\left(\lim_{m,n \rightarrow \infty} \left|\sum_{i=m}^n X_i\right| = 0\right) = \lim_{m,n \rightarrow \infty} p_{m,n} = p, \end{aligned} \quad (3.2.2)$$

so that it is sufficient to deal with  $p_{m,n}$  associated with the partial sums rather than  $p$ . It is only required to ensure that the priors on  $p_{m,n}$  are built such that given any realization  $\{x_i : i = 1, 2, \dots\}$  of the stochastic process  $\{X_i : i = 1, 2, \dots\}$  associated with

the corresponding series of interest  $\sum_{i=1}^{\infty} x_i$ , the posterior corresponding to the prior of  $p_{m,n}$ , which we denote by  $\pi_{m,n} \left( \cdot \mid \left| \sum_{i=m}^n x_i \right| \right)$ , converges to  $\pi \left( \cdot \mid \lim_{m,n \rightarrow \infty} \left| \sum_{i=m}^n x_i \right| \right)$ , the posterior corresponding to the prior of  $p$ . Since the latter posterior is based on some given, single realization of the underlying stochastic process, the overall probability of convergence  $p$  is informed with respect to the conditioned single realization only. Consequently, the overall probability of convergence, given the series of interest, admits interpretation as the probability of convergence of the series of interest. Hence, it is reasonable to require that,  $\pi \left( \cdot \mid \lim_{m,n \rightarrow \infty} \left| \sum_{i=m}^n x_i \right| \right) = \delta_{\{z\}}(\cdot)$ , the point mass at  $z$ , where  $z = 1$  or  $z = 0$  accordingly as  $\lim_{m,n \rightarrow \infty} \left| \sum_{i=m}^n x_i \right|$  is zero or positive, that is, accordingly as  $\sum_{i=1}^{\infty} x_i$  is convergent or divergent. Thus, it is required to construct the priors on  $p_{m,n}$  such that  $\pi_{m,n} \left( \cdot \mid \left| \sum_{i=m}^n x_i \right| \right) \rightarrow \delta_{\{z\}}(\cdot)$  in some appropriate sense, as  $m, n \rightarrow \infty$ , for any realization of the stochastic process.

It is important to appreciate that for another realization  $\{\tilde{x}_i : i = 1, 2, \dots\}$  of the underlying stochastic process, the corresponding infinite sum  $\sum_{i=1}^{\infty} \tilde{x}_i$  may have different convergence behaviour than  $\sum_{i=1}^{\infty} x_i$ . For instance,  $\sum_{i=1}^{\infty} \tilde{x}_i$  may be divergent while  $\sum_{i=1}^{\infty} x_i$  may be convergent. Hence, the corresponding posteriors based on the partial sums of  $\sum_{i=1}^{\infty} \tilde{x}_i$  will converge to 0, while those associated with  $\sum_{i=1}^{\infty} x_i$  will converge to 1. Since  $p$  is the probability that  $S_{1,\infty}$  converges, at first glance such discrepant posteriors may create the impression that the Bayesian inference procedure regarding  $p$  is inconsistent. However, as discussed above, given only the series of interest, the overall probability of convergence  $p$  admits interpretability as the probability of convergence of the series at hand. This is exactly what is desired, since our goal is to study the convergence properties of the series of our interest only, not to learn about  $p$ . As an aside, note that it is of course possible to learn about  $p$  via its posterior distribution which may be obtained by conditioning on adequate number of realizations (instead of a single realization) of the stochastic process as in the usual Bayesian inference problems of learning about unknown parameters.

In Section 3.3 we devise a recursive Bayesian methodology that achieves the goal discussed above. It is important to remark that no restrictive assumption is necessary for the development of our ideas, not even independence of  $X_i$ . With this methodology, we then characterize convergence and divergence of infinite series in Section 3.4, illustrating in Section 3.5 our theory and methods with seven examples. In Section 3.6 we apply our ideas to Riemann Hypothesis, obtaining results that are not in complete favour of the conjecture. Finally, we make concluding remarks in Section 3.7.

### 3.3 A recursive Bayesian procedure for studying infinite series

Since we view  $X_i$  as realizations from some random process, we first formalize the notion in terms of the relevant probability space. Let  $(\Omega, \mathcal{A}, \mu)$  be a probability space, where  $\Omega$  is the sample space,  $\mathcal{A}$  is the Borel  $\sigma$ -field on  $\Omega$ , and  $\mu$  is some probability measure. Let, for  $i = 1, 2, 3, \dots$ ,  $X_i : \Omega \mapsto \mathbb{R}$  be real valued random variables measurable with respect to the Borel  $\sigma$ -field  $\mathcal{B}$  on  $\mathbb{R}$ . As in [Schervish \(1995\)](#), we can then define a  $\sigma$ -field of subsets of  $\mathbb{R}^\infty$  with respect to which  $X = (X_1, X_2, \dots)$  is measurable. Indeed, let us define  $\mathbb{B}^\infty$  to be the smallest  $\sigma$ -field containing sets of the form

$$B = \{X : X_{i_1} \leq r_1, X_{i_2} \leq r_2, \dots, X_{i_p} \leq r_p, \text{ for some } p \geq 1, \\ \text{some integers } i_1, i_2, \dots, i_p, \text{ and some real numbers } r_1, r_2, \dots, r_p\}.$$

Since  $B$  is an intersection of finite number of sets of the form  $\{X : X_{i_j} \leq r_j\}; j = 1, \dots, p$ , all of which belong to  $\mathcal{A}$  (since  $X_{i_j}$  are measurable) it follows that  $X^{-1}(B) \in \mathcal{A}$ , so that  $X$  is measurable with respect to  $(\mathbb{R}^\infty, \mathbb{B}^\infty, P)$ , where  $P$  is the probability measure induced by  $\mu$ .

Alternatively, note that it is possible to represent any stochastic process  $\{X_i : i \in \mathcal{J}\}$ , for fixed  $i$  as a random variable  $\omega \mapsto X_i(\omega)$ , where  $\omega \in \Omega$ ;  $\Omega$  being the set of all functions

from  $\mathfrak{J}$  into  $\mathbb{R}$ . Also, fixing  $\omega \in \Omega$ , the function  $i \mapsto X_i(\omega)$ ;  $i \in \mathfrak{J}$ , represents a path of  $X_i$ ;  $i \in \mathfrak{J}$ . Indeed, we can identify  $\omega$  with the function  $i \mapsto X_i(\omega)$  from  $\mathfrak{J}$  to  $\mathbb{R}$ ; see, for example, Øksendal (2000), for a lucid discussion.

This latter identification will be convenient for our purpose, and we adopt this here. Note that the  $\sigma$ -algebra  $\mathcal{F}$  induced by  $X$  is generated by sets of the form

$$\{\omega : \omega(i_1) \in B_1, \omega(i_2) \in B_2, \dots, \omega(i_k) \in B_k\},$$

where  $B_j \subset \mathbb{R}$ ;  $j = 1, \dots, k$ , are Borel sets in  $\mathbb{R}$ .

### 3.3.1 Development of the stage-wise likelihoods

For  $j = 1, 2, 3, \dots$ , let

$$S_{j,n_j} = \sum_{i=\sum_{k=0}^{j-1} n_k+1}^{\sum_{k=0}^j n_k} X_i, \tag{3.3.1}$$

where  $n_0 = 0$  and  $n_j \geq 1$  for all  $j \geq 1$ . Also let  $\{c_j\}_{j=1}^\infty$  be a non-negative decreasing sequence and

$$Y_{j,n_j} = \mathbb{I}\{|S_{j,n_j}| \leq c_j\}. \tag{3.3.2}$$

Let, for  $j \geq 1$ ,

$$P(Y_{j,n_j} = 1) = p_{j,n_j}. \tag{3.3.3}$$

Hence, the likelihood of  $p_{j,n_j}$ , given  $y_{j,n_j}$ , is given by

$$L(p_{j,n_j}) = p_{j,n_j}^{y_{j,n_j}} (1 - p_{j,n_j})^{1-y_{j,n_j}} \tag{3.3.4}$$

It is important to relate  $p_{j,n_j}$  to convergence or divergence of the underlying series. Note that  $p_{j,n_j}$  is the probability that  $|S_{j,n_j}|$  falls below  $c_j$ . Thus,  $p_{j,n_j}$  can be interpreted as the probability that the series  $S_{1,\infty}$  is convergent when the data observed is  $S_{j,n_j}$ . If

$S_{1,\infty}$  is convergent, then it is to be expected *a posteriori*, that

$$p_{j,n_j} \rightarrow 1 \quad \text{as } j \rightarrow \infty. \quad (3.3.5)$$

Note that the above is expected to hold even for  $n_j = n$  for all  $j \geq 1$ , and for all  $n \geq 1$ . This is related to Cauchy's criterion of convergence of partial sums: for every  $\epsilon > 0$  there exists a positive integer  $N$  such that for all  $n \geq m \geq N$ ,  $|\sum_{i=m}^n X_i| < \epsilon$ . Indeed, as we will formally show, condition (3.3.5) is both necessary and sufficient for convergence of the series.

On the other hand, if the series is divergent, then there exist  $j_0 \geq 1$  such that for every  $j > j_0$  there exists  $n_j \geq 1$  satisfying  $|S_{j,n_j}| > c_j$ . Here we expect, *a posteriori*, that

$$p_{j,n_j} \rightarrow 0 \quad \text{as } j \rightarrow \infty. \quad (3.3.6)$$

Again, we will prove formally that the above condition is both necessary and sufficient for divergence.

In this work we call the series  $S_{1,\infty}$  oscillating if the sequence  $\{S_{1,n}; n = 1, 2, \dots\}$  has more than one limit points. Thus, these are non-convergent series, and so, the probability of convergence of these series must tend to zero in our Bayesian framework, which is in fact ensured by our theoretical developments. But it is also important to be able to categorize and learn about the limit points. A general theory, which encompasses finite as well as infinite number of limit points, with perhaps unequal frequencies of occurrences, is developed in Chapter 4.

In what follows we shall first construct a recursive Bayesian methodology that formally characterizes convergence and divergence in terms of formal posterior convergence related to (3.3.5) and (3.3.6).

### 3.3.2 Development of recursive Bayesian posteriors

We assume that  $\{y_{j,n_j}; j = 1, 2, \dots\}$  is observed successively at stages indexed by  $j$ . That is, we first observe  $y_{1,n_1}$ , and based on our prior belief regarding the first stage probability,  $p_{1,n_1}$ , compute the posterior distribution of  $p_{1,n_1}$  given  $y_{1,n_1}$ , which we denote by  $\pi(p_{1,n_1}|y_{1,n_1})$ . Based on this posterior we construct a prior for the second stage, and compute the posterior  $\pi(p_{2,n_2}|y_{1,n_1}, y_{2,n_2})$ . We continue this procedure for as many stages as we desire. Details follow.

Consider the sequences  $\{\alpha_j\}_{j=1}^{\infty}$  and  $\{\beta_j\}_{j=1}^{\infty}$ , where  $\alpha_j = \beta_j = 1/j^2$  for  $j = 1, 2, \dots$ . At the first stage of our recursive Bayesian algorithm, that is, when  $j = 1$ , let us assume that the prior is given by

$$\pi(p_{1,n_1}) \equiv \text{Beta}(\alpha_1, \beta_1), \quad (3.3.7)$$

where, for  $a > 0$  and  $b > 0$ ,  $\text{Beta}(a, b)$  denotes the Beta distribution with mean  $a/(a+b)$  and variance  $(ab)/\{(a+b)^2(a+b+1)\}$ . Combining this prior with the likelihood (3.3.4) (with  $j = 1$ ), we obtain the following posterior of  $p_{1,n_1}$  given  $y_{1,n_1}$ :

$$\pi(p_{1,n_1}|y_{1,n_1}) \equiv \text{Beta}(\alpha_1 + y_{1,n_1}, \beta_1 + 1 - y_{1,n_1}). \quad (3.3.8)$$

At the second stage (that is, for  $j = 2$ ), for the prior of  $p_{2,n_2}$  we consider the posterior of  $p_{1,n_1}$  given  $y_{1,n_1}$  associated with the  $\text{Beta}(\alpha_1 + \alpha_2, \beta_1 + \beta_2)$  prior. That is, our prior on  $p_{2,n_2}$  is given by:

$$\pi(p_{2,n_2}) \equiv \text{Beta}(\alpha_1 + \alpha_2 + y_{1,n_1}, \beta_1 + \beta_2 + 1 - y_{1,n_1}). \quad (3.3.9)$$

The reason for such a prior choice is that the uncertainty regarding convergence of the series is reduced once we obtain the posterior at the first stage, so that at the second stage the uncertainty regarding the prior is expected to be lesser compared to the first stage posterior. With our choice, it is easy to see that the prior variance at the second

### 3.13. A RECURSIVE BAYESIAN PROCEDURE FOR STUDYING INFINITE SERIES

stage, given by

$$\{(\alpha_1 + \alpha_2 + y_{1,n_1})(\beta_1 + \beta_2 + 1 - y_{1,n_1})\} / \{(\alpha_1 + \alpha_2 + \beta_1 + \beta_2 + 1)^2(\alpha_1 + \alpha_2 + \beta_1 + \beta_2 + 2)\},$$

is smaller than the first stage posterior variance, given by

$$\{(\alpha_1 + y_{1,n_1})(\beta_1 + 1 - y_{1,n_1})\} / \{(\alpha_1 + \beta_1 + 1)^2(\alpha_1 + \beta_1 + 2)\}.$$

The posterior of  $p_{2,n_2}$  given  $y_{2,n_2}$  is then obtained by combining the second stage prior (3.3.9) with (3.3.4) (with  $j = 2$ ). The form of the posterior at the second stage is thus given by

$$\pi(p_{2,n_2}|y_{2,n_2}) \equiv \text{Beta}(\alpha_1 + \alpha_2 + y_{1,n_1} + y_{2,n_2}, \beta_1 + \beta_2 + 2 - y_{1,n_1} - y_{2,n_2}). \quad (3.3.10)$$

Continuing this way, at the  $k$ -th stage, where  $k > 1$ , we obtain the following posterior of  $p_{k,n_k}$ :

$$\pi(p_{k,n_k}|y_{k,n_k}) \equiv \text{Beta}\left(\sum_{j=1}^k \alpha_j + \sum_{j=1}^k y_{j,n_j}, k + \sum_{j=1}^k \beta_j - \sum_{j=1}^k y_{j,n_j}\right). \quad (3.3.11)$$

It follows from (3.3.11) that

$$E(p_{k,n_k}|y_{k,n_k}) = \frac{\sum_{j=1}^k \alpha_j + \sum_{j=1}^k y_{j,n_j}}{k + \sum_{j=1}^k \alpha_j + \sum_{j=1}^k \beta_j}; \quad (3.3.12)$$

$$\text{Var}(p_{k,n_k}|y_{k,n_k}) = \frac{(\sum_{j=1}^k \alpha_j + \sum_{j=1}^k y_{j,n_j})(k + \sum_{j=1}^k \beta_j - \sum_{j=1}^k y_{j,n_j})}{(k + \sum_{j=1}^k \alpha_j + \sum_{j=1}^k \beta_j)^2(1 + k + \sum_{j=1}^k \alpha_j + \sum_{j=1}^k \beta_j)}. \quad (3.3.13)$$

Since  $\sum_{j=1}^k \alpha_j = \sum_{j=1}^k \beta_j = \sum_{j=1}^k \frac{1}{j^2}$ , (3.3.12) and (3.3.13) admit the following simplifi-



cations:

$$E(p_{k,n_k} | y_{k,n_k}) = \frac{\sum_{j=1}^k \frac{1}{j^2} + \sum_{j=1}^k y_{j,n_j}}{k + 2 \sum_{j=1}^k \frac{1}{j^2}}; \quad (3.3.14)$$

$$\text{Var}(p_{k,n_k} | y_{k,n_k}) = \frac{(\sum_{j=1}^k \frac{1}{j^2} + \sum_{j=1}^k y_{j,n_j})(k + \sum_{j=1}^k \frac{1}{j^2} - \sum_{j=1}^k y_{j,n_j})}{(k + 2 \sum_{j=1}^k \frac{1}{j^2})^2 (1 + k + 2 \sum_{j=1}^k \frac{1}{j^2})}. \quad (3.3.15)$$

### 3.4 Characterization of convergence properties of the underlying infinite series

Based on our recursive Bayesian theory we have the following theorem that characterizes convergence of  $S_{1,\infty}$  in terms of the limit of the posterior probability of  $p_{k,n_k}$ , as  $k \rightarrow \infty$ . Note that the sample space of  $S_{1,\infty}$  is also given by  $\mathfrak{S}$ . We also assume, for the sake of generality, that for any  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ , where  $\mathfrak{N} (\subset \mathfrak{S})$  has zero probability measure, the non-negative monotonically decreasing sequence  $\{c_j\}_{j=1}^\infty$  depends upon  $\omega$ , so that we shall denote the sequence by  $\{c_j(\omega)\}_{j=1}^\infty$ . In other words, we allow  $\{c_j(\omega)\}_{j=1}^\infty$  to depend upon the corresponding series  $S_{1,\infty}(\omega)$ . Note that if  $S_{1,\infty}(\omega) < \infty$ , then the sequence  $\{|S_{j,n_j}(\omega)|\}_{j=1}^\infty$  is uniformly bounded, for all sequences  $\{n_j\}_{j=1}^\infty$ , and converges to zero for all sequences  $\{n_j\}_{j=1}^\infty$ , which implies that there exists a monotonically decreasing sequence  $\{c_j(\omega)\}_{j=1}^\infty$  independent of the choice of  $\{n_j\}_{j=1}^\infty$  such that for some  $j_0(\omega) \geq 1$ ,

$$|S_{j,n_j}(\omega)| \leq c_j(\omega), \text{ for } j \geq j_0(\omega). \quad (3.4.1)$$

Indeed, in most of our illustrations presented in this chapter, including the Riemann Hypothesis, we choose  $\{c_j(\omega)\}_{j=1}^\infty$  in a way that depends upon the infinite series at hand.

**Theorem 5** *For any  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ , where  $\mathfrak{N}$  is some null set having probability measure zero,  $S_{1,\infty}(\omega) < \infty$  if and only if there exists a non-negative monotonically decreasing*

sequence  $\{c_j(\omega)\}_{j=1}^{\infty}$  such that for any choice of the sequence  $\{n_j\}_{j=1}^{\infty}$ ,

$$\pi(\mathcal{N}_1 | y_{k, n_k}(\omega)) \rightarrow 1, \quad (3.4.2)$$

as  $k \rightarrow \infty$ , where  $\mathcal{N}_1$  is any neighborhood of 1 (one).

**Proof.** Let, for  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ ,  $S_{1, \infty}(\omega)$  be convergent. Then, by (3.4.1),  $|S_{j, n_j}(\omega)| \leq c_j(\omega)$  for all  $n_j$ , so that  $y_{j, n_j}(\omega) = 1$  for all  $j > j_0(\omega)$ , for all  $n_j$ . Hence, in this case,  $\sum_{j=1}^k y_{j, n_j}(\omega) = k - k_0(\omega)$ , where  $k_0(\omega) \geq 0$ . Also,  $\sum_{j=1}^k \frac{1}{j^2} \rightarrow \frac{\pi^2}{6}$ , as  $k \rightarrow \infty$ . Consequently, it is easy to see that

$$\mu_k = E(p_{k, n_k} | y_{k, n_k}(\omega)) \sim \frac{\frac{\pi^2}{6} + k - k_0(\omega)}{k + \frac{\pi^2}{3}} \rightarrow 1, \text{ as } k \rightarrow \infty, \text{ and,} \quad (3.4.3)$$

$$\sigma_k^2 = Var(p_{k, n_k} | y_{k, n_k}(\omega)) \sim \frac{(\frac{\pi^2}{6} + k)(\frac{\pi^2}{6})}{(k + \frac{\pi^2}{3})^2(1 + k + \frac{\pi^2}{3})} \rightarrow 0 \text{ as } k \rightarrow \infty. \quad (3.4.4)$$

In the above, for any two sequences  $\{a_k\}_{k=1}^{\infty}$  and  $\{b_k\}_{k=1}^{\infty}$ ,  $a_k \sim b_k$  indicates  $\frac{a_k}{b_k} \rightarrow 1$ , as  $k \rightarrow \infty$ . Now let  $\mathcal{N}_1$  denote any neighborhood of 1, and let  $\epsilon > 0$  be sufficiently small such that  $\mathcal{N}_1 \supseteq \{1 - p_{k, n_k} < \epsilon\}$ . Combining (3.4.3) and (3.4.4) with Chebychev's inequality ensures that (3.4.2) holds.

Now assume that (3.4.2) holds. Then for any given  $\epsilon > 0$ ,

$$\pi(p_{k, n_k} > 1 - \epsilon | y_{k, n_k}(\omega)) \rightarrow 1, \text{ as } k \rightarrow \infty. \quad (3.4.5)$$

Hence, it can be seen, using Markov's inequality, that

$$E(p_{k, n_k} | y_{k, n_k}(\omega)) \rightarrow 1; \quad (3.4.6)$$

$$Var(p_{k, n_k} | y_{k, n_k}(\omega)) \rightarrow 0, \quad (3.4.7)$$

as  $k \rightarrow \infty$ . If  $S_{1, \infty}(\omega)$  does not converge then there exists  $j_0(\omega)$  such that for each

$j \geq j_0(\omega)$ , there exists  $n_j(\omega)$  satisfying  $|S_{j,n_j(\omega)}(\omega)| > c_j(\omega)$ , for any choice of non-negative sequence  $\{c_j(\omega)\}_{j=1}^{\infty}$  monotonically converging to zero. Hence, in this situation,  $0 \leq \sum_{j=1}^k y_{j,n_j(\omega)}(\omega) \leq j_0(\omega)$ . Substituting this in (3.3.14) and (3.3.15), it is easy to see that, as  $k \rightarrow \infty$ ,

$$E(p_{k,n_k(\omega)}|y_{k,n_k(\omega)}(\omega)) \rightarrow 0; \quad (3.4.8)$$

$$Var(p_{k,n_k(\omega)}|y_{k,n_k(\omega)}(\omega)) \rightarrow 0, \quad (3.4.9)$$

so that (3.4.6) is contradicted.

■

We now prove the following theorem that provides necessary and sufficient conditions for divergence of  $S_{1,\infty}(\omega)$  in terms of the limit of the posterior probability of  $p_{k,n_k(\omega)}$ , as  $k \rightarrow \infty$ .

**Theorem 6** *For any  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ , where  $\mathfrak{N}$  is some null set having probability measure zero,  $S_{1,\infty}(\omega)$  is divergent if and only if there exists a sequence  $\{n_j(\omega)\}_{j=1}^{\infty}$  such that*

$$\pi(\mathcal{N}_0|y_{k,n_k(\omega)}(\omega)) \rightarrow 1, \quad (3.4.10)$$

$k \rightarrow \infty$ , where  $\mathcal{N}_0$  is any neighborhood of 0 (zero).

**Proof.** Assume that  $S_{1,\infty}(\omega)$  is divergent. Then there exist  $j_0(\omega) \geq 1$  such that for every  $j \geq j_0(\omega)$ , one can find  $n_j(\omega)$  satisfying  $|S_{j,n_j(\omega)}(\omega)| > c_j(\omega)$ , for any choice of non-negative sequence  $\{c_j(\omega)\}_{j=1}^{\infty}$  monotonically converging to zero. From the proof of the sufficient condition of Theorem 5 it follows that (3.4.8) and (3.4.9) hold. Let  $\epsilon > 0$  be small enough so that  $\mathcal{N}_0 \supseteq \{p_{k,n_k(\omega)} < \epsilon\}$ . Then combining Chebychev's inequality with (3.4.8) and (3.4.9) it is easy to see that (3.4.10) holds.

Now assume that (3.4.10) holds. Then for any given  $\epsilon > 0$ ,

$$\pi(p_{k,n_k(\omega)} < \epsilon|y_{k,n_k(\omega)}(\omega)) \rightarrow 1, \text{ as } k \rightarrow \infty. \quad (3.4.11)$$

It can be seen, now using Markov's inequality with respect to  $1 - p_{k,n_k(\omega)}$ , that

$$E(p_{k,n_k(\omega)} | y_{k,n_k(\omega)}(\omega)) \rightarrow 0; \quad (3.4.12)$$

$$\text{Var}(p_{k,n_k(\omega)} | y_{k,n_k(\omega)}(\omega)) \rightarrow 0, \quad (3.4.13)$$

as  $k \rightarrow \infty$ .

If  $S_{1,\infty}(\omega)$  is convergent, then by Theorem 5,  $\pi(\mathcal{N}_1 | y_{k,n_k}(\omega)) \rightarrow 1$  as  $k \rightarrow \infty$ , for all sequences  $\{n_j\}_{j=1}^{\infty}$ , so that  $E(p_{k,n_k(\omega)} | y_{k,n_k(\omega)}(\omega)) \rightarrow 1$ , which is a contradiction to (3.4.12).

■

Note that Theorem 6 encompasses even oscillatory series. For instance, if for some  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ ,  $S_{1,\infty}(\omega) = \sum_{i=1}^{\infty} (-1)^{i-1}$ , then the sequence  $n_j(\omega) = 1 + 2(j-1)$  ensures that  $|S_{j,n_j}(\omega)| > c_j(\omega)$  for all  $j \geq j_0(\omega)$ , for some  $j_0(\omega) \geq 1$ , for any monotonically decreasing non-negative sequence  $\{c_j(\omega)\}_{j=1}^{\infty}$ . This of course forces declaration of divergence of this particular series, as per Theorem 6. We show in Section 4.6.1, with the help of our Bayesian idea of studying oscillatory series, how to identify the number and proportions of the limit points of this oscillatory series.

### 3.4.1 Characterization of infinite series using non-recursive Bayesian posteriors

Observe that it is not strictly necessary for the prior at any stage to depend upon the previous stage. Indeed, we may simply assume that  $\pi(p_{j,n_j}) \equiv \text{Beta}(\alpha_j, \beta_j)$ , for  $j = 1, 2, \dots$ . In this case, the posterior of  $p_{k,n_k}$  given  $y_{k,n_k}$  is simply  $\text{Beta}(\alpha_k + y_{k,n_k}, 1 + \beta_k - y_{k,n_k})$ . The posterior mean and variance are then given by

$$E(p_{k,n_k} | y_{k,n_k}(\omega)) = \frac{\alpha_k + y_{k,n_k}(\omega)}{1 + \alpha_k + \beta_k}; \quad (3.4.14)$$

$$\text{Var}(p_{k,n_k} | y_{k,n_k}(\omega)) = \frac{(\alpha_k + y_{k,n_k}(\omega))(1 + \beta_k - y_{k,n_k}(\omega))}{(1 + \alpha_k + \beta_k)^2(2 + \alpha_k + \beta_k)}. \quad (3.4.15)$$

Since  $y_{k,n_k}(\omega)$  converges to 1 or 0 as  $k \rightarrow \infty$ , accordingly as  $S_{1,\infty}(\omega)$  is convergent or divergent, it is easily seen, provided that  $\alpha_k \rightarrow 0$  and  $\beta_k \rightarrow 0$  as  $k \rightarrow \infty$ , that (3.4.14) converges to 1 (respectively, 0) if and only if  $S_{1,\infty}(\omega)$  is convergent (respectively, divergent).

Thus, characterization of convergence or divergence of infinite series is possible even with the non-recursive approach. Indeed, note that the prior parameters  $\alpha_k$  and  $\beta_k$  are more flexible compared to those associated with the recursive approach. This is because, in the non-recursive approach we only require  $\alpha_k \rightarrow 0$  and  $\beta_k \rightarrow 0$  as  $k \rightarrow \infty$ , so that convergence of the series  $\sum_{j=1}^{\infty} \alpha_j$  and  $\sum_{j=1}^{\infty} \beta_j$  are not necessary, unlike the recursive approach. However, choosing  $\alpha_k$  and  $\beta_k$  to be of sufficiently small order ensures much faster convergence of the posterior mean and variance as compared to the recursive approach.

Unfortunately, an important drawback of the non-recursive approach is that it does not admit extension to the case of general oscillatory series with multiple limit points, where blocks of partial sums can not be used; see Chapter 4. On the other hand, as we show in Chapter 4, the principles of our recursive theory can be easily adopted to develop a Bayesian characterization of oscillating series, which also includes the characterization of non-oscillating series as a special case. In other words, the recursive approach seems to be more powerful from the perspective of development of a general characterization theory. Moreover, as our examples on convergent and divergent series demonstrate, the recursive posteriors converge sufficiently fast to the correct degenerate distributions, obviating the need to consider the non-recursive approach. Consequently, we do not further pursue the non-recursive approach.

**Remark 7** *An important issue associated with our characterization results is that the terms  $\{x_1, x_2, \dots\}$  of the underlying deterministic series of interest  $\sum_{i=1}^{\infty} x_i$  is assumed to lie in the complement of the null set. For appropriately specified stochastic processes this need not be difficult to verify. However, for the sake of sufficient generality we have*

not assumed any specific form of the underlying stochastic process, which makes the question of null sets relevant in our case. The solution is that, even if  $\{x_1, x_2, \dots\}$  falls in some null set, we can still compute a pseudo-posterior distribution of  $p_{k, n_k}$  conditional on  $\{x_1, x_2, \dots\}$ , which has exactly the same form as before. This pseudo-posterior may not admit interpretability as a bona fide posterior distribution, but characterizes the convergence property of  $\sum_{i=1}^{\infty} x_i$  in exactly the same way as before. In other words, interestingly and very importantly, all our results of characterization hold for all  $\omega \in \mathfrak{S}$ .

### 3.5 Illustrations

We now illustrate our ideas with seven examples. These seven examples can be categorized into three categories in terms of construction of the upper bound  $c_j$ . With the first example we demonstrate that it may sometimes be easy to devise an appropriate upper bound. In Examples 2 – 5, we show that usually simple bounds such as that in Example 1, are not adequate in practice, but appropriate bounds may be constructed if convergence and divergence of the series in question is known for some values of the parameters; the resultant bounds can be utilized to learn about convergence or divergence of the series for the remaining values of the parameters. In Examples 6 and 7, the series in question are stand-alone in the sense they are not defined by parameters with known convergence/divergence for some of their values which might have aided our construction of  $c_j$ . However, we show that these series can be embedded into appropriately parameterized series, facilitating similar analysis as Examples 2 – 5.

For these examples, we consider  $n_j = n$  for  $j = 1, \dots, K$ , with  $n = 10^6$  and  $K = 10^5$ . Since  $n$  seems to be sufficiently large, in the case of divergence we expect  $|S_{j, n}|$  to exceed the monotonically decreasing  $c_j$  for all  $j \geq j_0$ , for sufficiently large  $j_0$ . Our experiments demonstrate that this is indeed the case. For further justification we conducted some experiments with larger values of  $n$ , but the results remained unchanged. Hence, for relative computational ease we set  $n = 10^6$  for the illustrations in this work.

Since we needed to sum  $10^6$  terms at each step of  $10^5$  stages, the associated computation is extremely demanding. For the purpose of efficiency, we parallelized the computation of the sums of  $10^6$  terms, splitting the job on many processors, using the Message Passing Interface (MPI) protocol. In more details, we implemented our parallelized codes, written in C, in VMware consisting of 60 double-threaded, 64-bit physical cores, each running at 2793.269 MHz. Parallel computation of our methods associated with Examples 1 to 5 take, respectively, 1 minute, 4 minutes, 7 minutes, 6 minutes, and 9 minutes. Examples 6 and 7 require about 6 minutes and 4 minutes of computational time.

### 3.5.1 Example 1

In their first example [Bourchtein et al. \(2012\)](#) study the following divergent series with their methods:

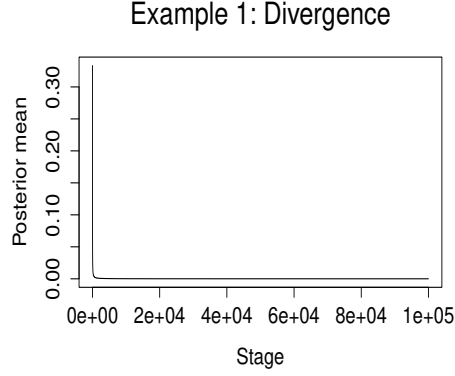
$$S = \sum_{i=2}^{\infty} \frac{1}{\log(i)}. \quad (3.5.1)$$

We test our Bayesian idea on this series choosing the monotonically decreasing sequence as  $c_{j,n} = 1/\sqrt{nj}$ , where we represent  $c_j$  as  $c_{j,n}$  to reflect dependence on  $n$ . Figure 3.5.1, a plot of the posterior means of  $\{p_{k,n}; k = 1, \dots, 10^5\}$ , clearly and correctly indicates that the series is divergent. We also constructed approximate 95% highest posterior density credible intervals at each recursive step; however, thanks to very less variances at each stage, the intervals turned out to be too small to be clearly distinguishable from the plot of the stage-wise posterior means.

### 3.5.2 Example 2

Example 2 of [Bourchtein et al. \(2012\)](#) deals with the following series:

$$S^a = \sum_{i=2}^{\infty} \left( 1 - \left\{ \frac{\log(i)}{i} \right\} - a \frac{\log \log(i)}{i} \right)^i, \quad (3.5.2)$$



**Figure 3.5.1:** Example 1: The series (3.5.1) is divergent.

where  $a \in \mathbb{R}$ . [Bouchtein \*et al.\* \(2012\)](#) prove that the series converges for  $a > 1$  and diverges for  $a \leq 1$ .

### Choice of $c_{j,n}$

Now, however, selecting the monotone sequence as  $c_{j,n} = 1/\sqrt{nj}$  turn out to be inappropriate for this series, the behaviour of which is quite sensitive to the parameter  $a$ , particularly around  $a = 1$ . Hence, any appropriate sequence  $\{c_{j,n}\}_{j=1}^{\infty}$  must depend on the parameter  $a$  of the series (3.5.2).

Denoting  $c_{j,n}$  by  $c_{j,n}^a$  to reflect the dependence on  $a$  as well, we first set

$$u_{j,n}^a = S_{j,n}^{a_0} + \frac{(a - 1 - 9 \times 10^{-11})}{\log(j + 1)}, \quad (3.5.3)$$

and then let

$$c_{j,n}^a = \begin{cases} u_{j,n}^a, & \text{if } u_{j,n}^a > 0; \\ S_{j,n}^{a_0}, & \text{otherwise.} \end{cases} \quad (3.5.4)$$

where  $a_0 = 1 + 10^{-10}$ . The reason behind such a choice of  $c_{j,n}^a$  is provided below.

Let, for  $\epsilon > 0$ ,

$$\tilde{S} = \sup \{S^a : a \geq 1 + \epsilon\}. \quad (3.5.5)$$



Thus,  $\tilde{S}$  may be interpreted as the convergent series which is closest to divergence given the convergence criterion  $a \geq 1 + \epsilon$ . Since  $S^a$  is decreasing in  $a$ , it easily follows that equality of (3.5.5) is attained at  $a_0 = 1 + \epsilon$ .

Since the terms of the series  $S^a$  are decreasing in  $i$ , it follows that  $S_{j,n}^{a_0}$  in (3.5.4) is decreasing in  $j$ . We assume that  $\epsilon$  is chosen to be so small that convergence properties of the series for  $\{a \leq 1\} \cup \{a \geq 1 + \epsilon\}$  are only desired. Indeed, since  $\left(1 - \left\{\frac{\log(i)}{i}\right\} - a \frac{\log \log(i)}{i}\right)^i$  is decreasing in  $a$  for any given  $i \geq 3$ , our method of constructing  $c_{j,n}^a$  need not be able to correctly identify the convergence properties of the series for  $1 < a < 1 + \epsilon$ .

For the purpose of illustrations we choose  $\epsilon = 10^{-10}$ . Note that for  $a > 1$  the term  $\frac{(a-1-9 \times 10^{-11})}{\log(j+1)}$  inflates  $c_{j,n}^a$  making  $S_{j,n}^a$  more likely to fall below  $c_{j,n}^a$  for increasing  $a$ , thus paving the way for diagnosing convergence. The same term also ensures that for  $a \leq 1$ ,  $c_{j,n}^a < S_{j,n}^{a_0}$ , so that  $S_{j,n}^a$  is likely to exceed  $c_{j,n}^a$ , thus providing an inclination towards divergence. The term  $-9 \times 10^{-11}$  is an adjustment for the case  $a = 1 + 10^{-10}$ , ensuring that  $c_{j,n}^a$  marginally exceeds  $S_{j,n}^a$  to ensure convergence. The scaling factor  $\log(j+1)$  ensures that the part  $\frac{(a-1-9 \times 10^{-11})}{\log(j+1)}$  of (3.5.4) tends to zero at a slow rate so that  $c_{j,n}^a$  is decreasing with  $j$  and  $n$  even if  $a - 1 - 9 \times 10^{-11}$  is negative.

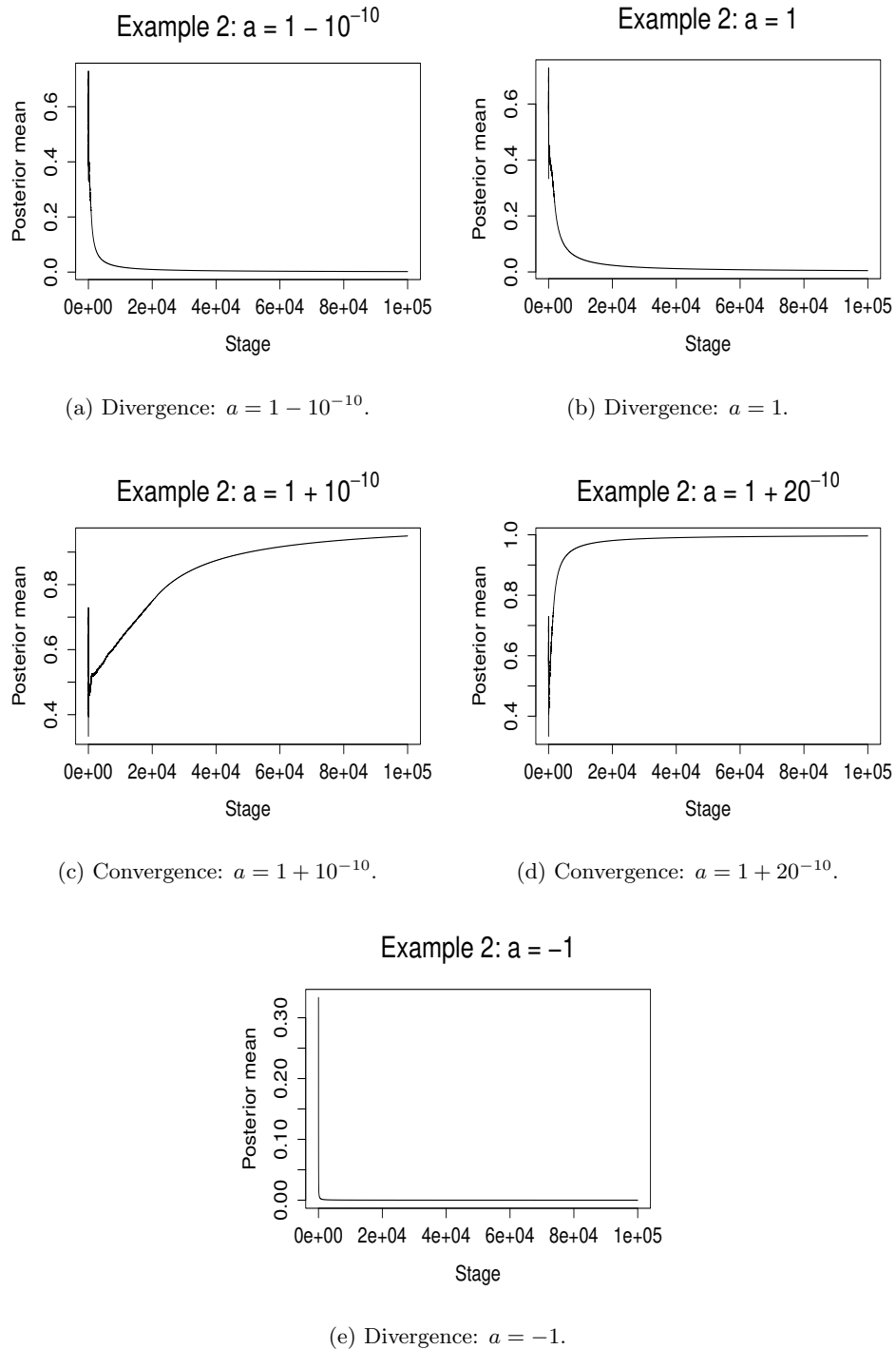
Figure 3.5.2, depicting our Bayesian results for this series, is in agreement with the results of [Bourchtein et al. \(2012\)](#). In fact, we have applied our methods to many more values of  $a \in A_\epsilon$  with  $\epsilon = 10^{-10}$ , and in every case the correct result is vindicated.

### 3.5.3 Example 3

Let us now consider the following series analysed by [Bourchtein et al. \(2012\)](#):

$$S = \sum_{i=3}^{\infty} \left(1 - \left(\frac{\log(i)}{i}\right) a \frac{\log \log(i)}{\log(i)}\right)^i, \quad (3.5.6)$$

where  $a > 0$ . As is shown by [Bourchtein et al. \(2012\)](#), the series converges for  $a > e$  and diverges for  $a \leq e$ .



**Figure 3.5.2:** Example 2: The series (3.5.2) converges for  $a > 1$  and diverges for  $a \leq 1$ .

### Choice of $c_{j,n}$

Here we first set

$$u_{j,n}^a = S_{j,n}^{a_0} + \frac{(a - e - 9 \times 10^{-11})}{\log(j+1)}, \quad (3.5.7)$$

and then let  $c_{j,n}^a$  defined by (3.5.4). Again, it is easily seen that  $S_{j,n}^{a_0}$  is decreasing in  $j$ . In this example we set  $a_0 = e + 10^{-10}$ . The rationale behind the choice remains the same as detailed in Section 3.5.2.

As before, the results obtained by our Bayesian theory, as displayed in Figure 3.5.3, are in complete agreement with the results obtained by [Bourchtein et al. \(2012\)](#).

### 3.5.4 Example 4

We now consider series (3.1.1). It has been proved by [Bourchtein et al. \(2012\)](#) that the series is convergent for  $a - b > 1$  and divergent for  $a + b < 1$ . As mentioned before, the hierarchy of tests of [Bourchtein et al. \(2012\)](#) are inconclusive for  $a = b = 1$ .

In this example we denote the partial sums by  $S_{j,n}^{a,b}$  and the actual series  $S$  by  $S^{a,b}$  to reflect the dependence on both the parameters  $a$  and  $b$ .

$$S_{j,n}^{a,b} = \sum_{i=3+n(j-1)}^{3+nj-1} \left( 1 - \frac{\log i}{i} - \frac{\log \log i}{i} \left\{ \cos^2 \left( \frac{1}{i} \right) \right\} (a + (-1)^i b) \right)^i, \quad (3.5.8)$$

We then have the following lemma, the proof of which is presented in Appendix 3.A1

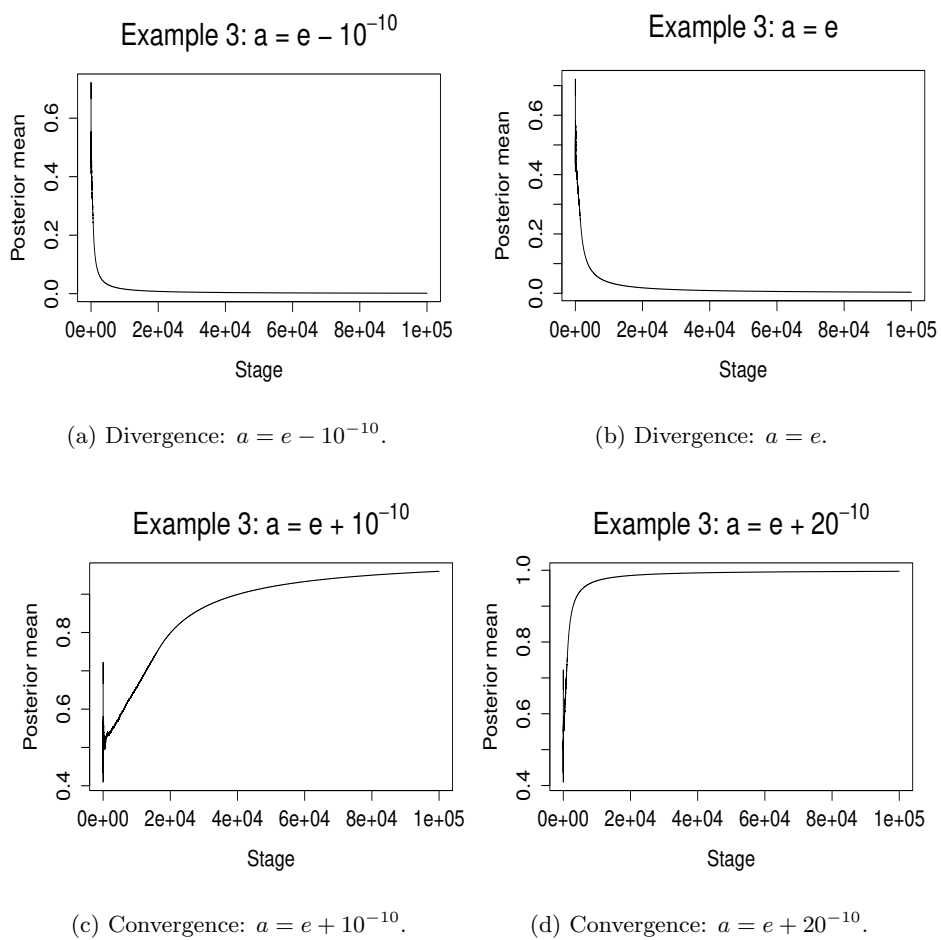
**Lemma 8** *For series (3.1.1), for  $j \geq 1$  and  $n$  even,  $S_{j,n}^{a,b}$  given by (3.5.8) is decreasing in  $a$  but increasing in  $b$ .*

Since  $S^{a,b}$  is just summation of the partial sums, it follows that

**Corollary 9**  *$S^{a,b}$  is decreasing in  $a$  and increasing in  $b$ .*

We let

$$A_\epsilon = \{a : 0 \leq a \leq 1\} \cup \{a : a \geq 1 + \epsilon\}, \quad (3.5.9)$$



**Figure 3.5.3:** Example 3: The series (3.5.6) converges for  $a > e$  and diverges for  $a \leq e$ .

and

$$\tilde{S} = \inf_{a \in A_\epsilon} \sup_{b \geq 0} \left\{ S^{a,b} : a - b > 1 \right\}. \quad (3.5.10)$$

It is easy to see in this case, due to Corollary 9 and the convergence criterion  $a - b > 1$ , that  $\tilde{S}$  is attained at  $a_0 = 1 + \epsilon$  and  $b_0 = 0$ . As before, we set  $\epsilon = 10^{-10}$ . Hence, arguments similar to those in Section 3.5.2 lead to the following choice of the upper bound for  $S_{j,n}^{a,b}$ , which we denote in this example by  $c_{j,n}^{a,b}$ :

$$c_{j,n}^{a,b} = \begin{cases} u_{j,n}^{a,b}, & \text{if } u_{j,n}^{a,b} > 0; \\ S_{j,n}^{a_0,b_0}, & \text{otherwise,} \end{cases} \quad (3.5.11)$$

where  $a_0 = 1 + 10^{-10}$ ,  $b_0 = 0$ , and

$$u_{j,n}^{a,b} = S_{j,n}^{a_0,b_0} + \frac{(a - 1 - b - 9 \times 10^{-11})}{\log(j + 1)}. \quad (3.5.12)$$

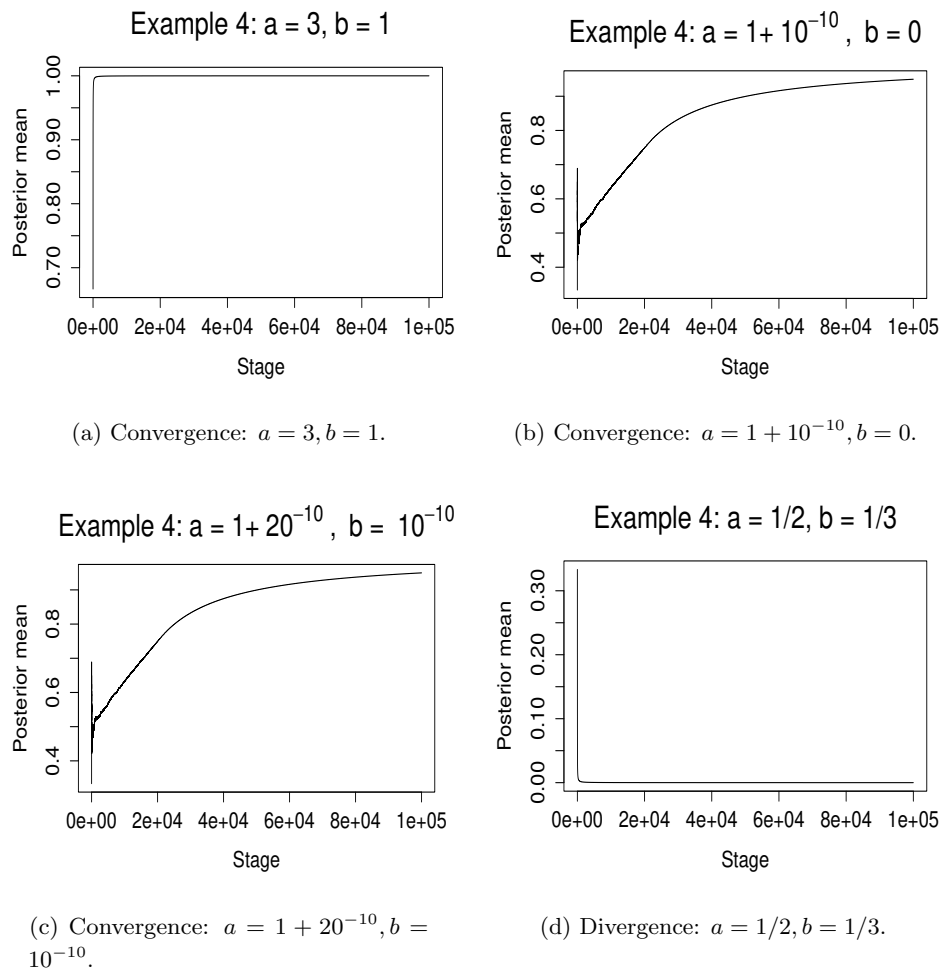
As before, it is easily seen that  $S_{j,n}^{a_0,b_0}$  is decreasing in  $j$ . Also note that  $-b$  in (3.5.12) takes account of the fact that the partial sums are increasing in  $b$ , thus favouring divergence for increasing  $b$ .

Setting aside panel (c) of Figure 3.5.5, observe that the remaining panels of Figures 3.5.4 and 3.5.5 are in agreement with the results of [Bourchtein \*et al.\* \(2012\)](#), but in the case  $a = b = 1$ , the tests of [Bourchtein \*et al.\* \(2012\)](#) turned out to be inconclusive. Panel (c) of Figure 3.5.5 demonstrates that the series is divergent for  $a = b = 1$ .

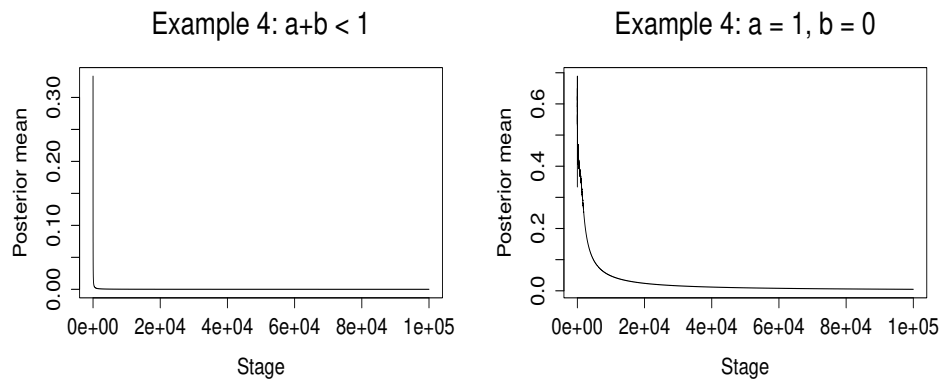
### 3.5.5 Example 5

Now consider the following series presented and analysed in [Bourchtein \*et al.\* \(2012\)](#):

$$S = \sum_{i=3}^{\infty} \left( 1 - \left( \frac{\log(i)}{i} \right) \left( a \left( 1 + \sin^2 \left( \sqrt{\left( \frac{\log(\log(i))}{\log(i)} \right)} \right) \right) + b \sin \left( \frac{i\pi}{4} \right) \right) \right)^i; \quad a > 0, b > 0. \quad (3.5.13)$$

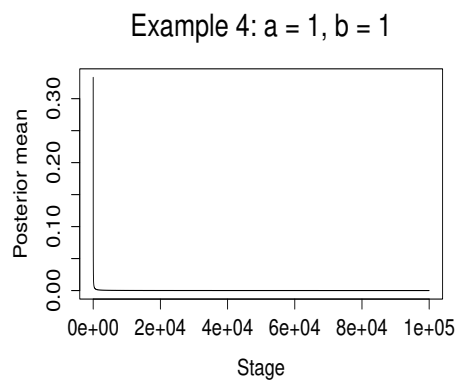


**Figure 3.5.4:** Example 4: The series (3.1.1) converges for  $(a = 3, b = 1)$ ,  $(a = 1 + 10^{-10}, b = 0)$ ,  $(a = 1 + 20^{-10}, b = 10^{-10})$  and diverges for  $(a = 1/2, b = 1/3)$ .



(a) Divergence:  $a = \frac{1}{2} (1 - 10^{-11}), b = \frac{1}{2} (1 - 10^{-11})$ .

(b) Divergence:  $a = 1, b = 0$ .



(c) Divergence:  $a = 1, b = 1$ .

**Figure 3.5.5:** Example 4: The series (3.1.1) diverges for  $(a = \frac{1}{2} (1 - 10^{-11}), b = \frac{1}{2} (1 - 10^{-11}))$ ,  $(a = 1, b = 0)$  and  $(a = 1, b = 1)$ .

[Bourchtein et al. \(2012\)](#) show that the series converges when  $a - b > 1$  and diverges when  $a + b < 1$ . Again, as in the case of Example 4, the following lemma holds in Example 5, the proof of which is provided in Appendix 3.A2. Note that for mathematical convenience we consider partial sums from the 5-th term onwards. We also assume  $n$  to be a multiple of 4.

**Lemma 10** *For the series (3.5.13), let*

$$S_{j,n}^{a,b} = \sum_{i=5+n(j-1)}^{5+nj-1} \left( 1 - \left( \frac{\log(i)}{i} \right) \left( a \left( 1 + \sin^2 \left( \sqrt{\left( \frac{\log(\log(i))}{\log(i)} \right)} \right) \right) + b \sin \left( \frac{i\pi}{4} \right) \right) \right)^i, \quad (3.5.14)$$

for  $j \geq 1$  and  $n$ , a multiple of 4. Then  $S_{j,n}^{a,b}$  is decreasing in  $a$  and increasing in  $b$ .

The following corollary with respect to  $S^{a,b}$  again holds:

**Corollary 11**  *$S^{a,b}$  is decreasing in  $a$  and increasing in  $b$ .*

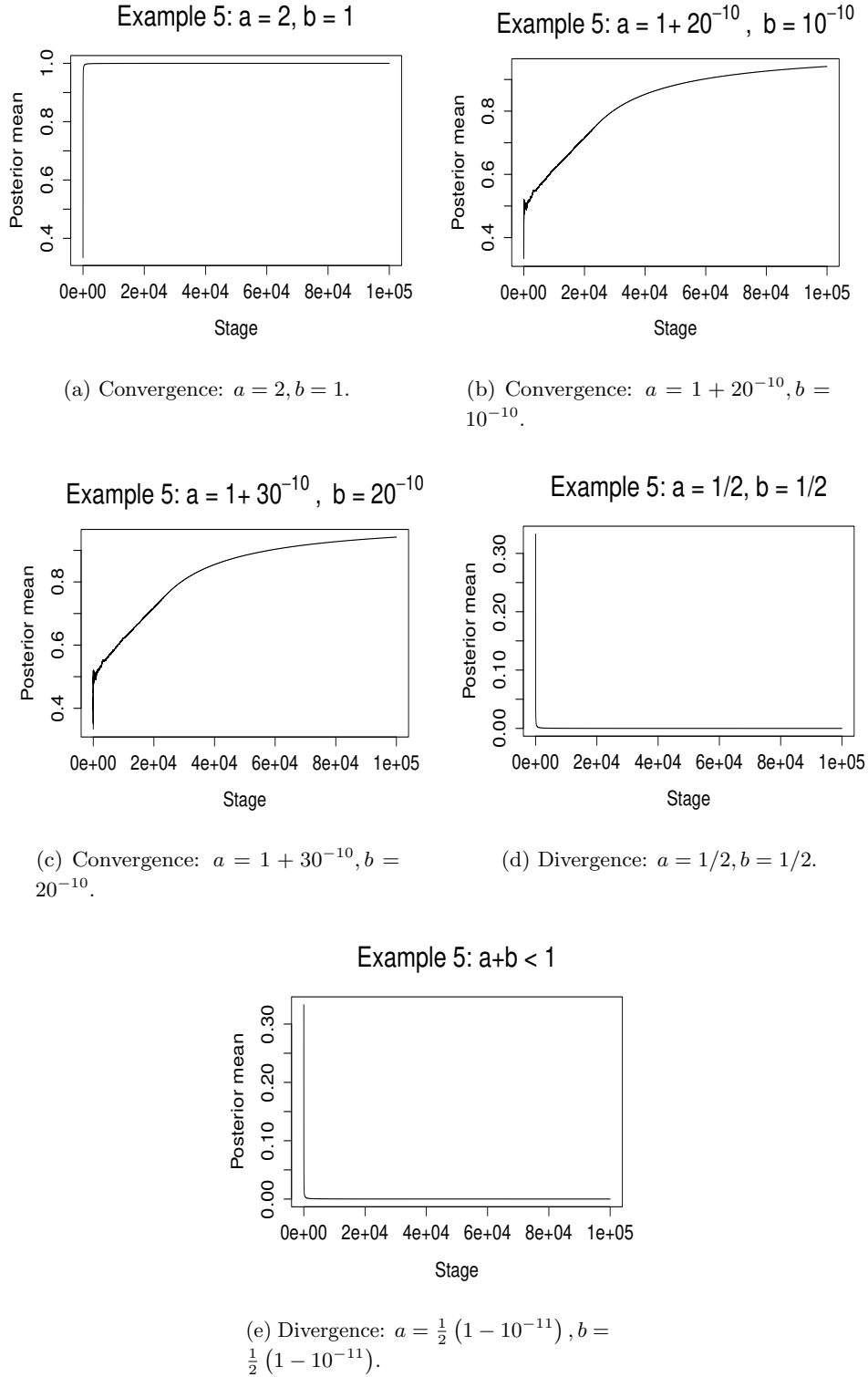
Thus, we follow the same method as in Example 4 to determine  $c_{j,n}^{a,b}$ , but we need to note that in this example  $a > 0$  and  $b > 0$  instead of  $a \geq 0$  and  $b \geq 0$  of Example 4. Consequently, here we define  $b \geq \epsilon$ , for  $\epsilon > 0$ , the set  $A_\epsilon$  given by (3.5.9) and

$$\tilde{S} = \inf_{a \in A_\epsilon} \sup_{b \geq \epsilon} \left\{ S^{a,b} : a - b > 1 \right\}. \quad (3.5.15)$$

In this case, Corollary 11 and the convergence criterion  $a - b > 1$  ensure that  $\tilde{S}$  is attained at  $a_0 = 1 + \epsilon$  and  $b_0 = \epsilon$ . As before, we set  $\epsilon = 10^{-10}$ . The rest of the arguments leading to the choice of  $c_{j,n}^{a,b}$  remains the same as in Example 4, and hence in this example  $c_{j,n}^{a,b}$  has the same form as (3.5.11), with  $a_0 = 1 + 10^{-10}$ ,  $b_0 = 10^{-10}$ , where  $S_{j,n}^{a_0,b_0}$  is decreasing in  $j$  as before.

Figure 3.5.6 depicts the results of our Bayesian analysis of the series (3.5.13) for various values of  $a$  and  $b$ . All the results are in accordance with those of [Bourchtein et al. \(2012\)](#).





**Figure 3.5.6:** Example 5: The series (3.5.13) converges for  $(a = 2, b = 1)$ ,  $(a = 1 + 20^{-10}, b = 10^{-10})$ ,  $(a = 1 + 30^{-10}, b = 20^{-10})$  and diverges for  $(a = 1/2, b = 1/2)$  and  $(a = \frac{1}{2}(1 - 10^{-11}), b = \frac{1}{2}(1 - 10^{-11}))$ .

### 3.5.6 Example 6

We now investigate whether or not the following series converges:

$$S = \sum_{i=1}^{\infty} \frac{1}{i^3 |\sin i|}. \quad (3.5.16)$$

This series is a special case of the generalized form of the Flint Hills series (see [Pickover \(2002\)](#) and [Alekseyev \(2011\)](#)).

For our purpose, we first embed the above series into

$$S^{a,b} = \sum_{i=1}^{\infty} \frac{i^{b-3}}{a + |\sin i|}, \quad (3.5.17)$$

where  $b \in \mathbb{R}$  and  $|a| \leq \eta$ , for some  $\eta > 0$ , specified according to our purpose. Note that,  $S = S^{0,0}$ , and we set  $\eta = 10^{-10}$  for our investigation of (3.5.16).

Note that for any fixed  $a \neq 0$ ,  $S^{a,b}$  converges if  $b < 2$  and diverges if  $b \geq 2$ . Since  $S^{a,b}$  increases in  $b$  it follows that the equality in

$$\tilde{S} = \sup \left\{ S^{a,b} : a = \epsilon, b \leq 2 - \epsilon \right\} \quad (3.5.18)$$

is attained at  $(a_0, b_0) = (\epsilon, 2 - \epsilon)$ .

Arguments in keeping with those in the previous examples lead to the following choice of the upper bound for  $S_{j,n}^{a,b}$ , which we again denote by  $c_{j,n}^{a,b}$ :

$$c_{j,n}^{a,b} = \begin{cases} u_{j,n}^{a,b}, & \text{if } b < 2; \\ v_{j,n}^{a,b}, & \text{otherwise,} \end{cases} \quad (3.5.19)$$

where

$$u_{j,n}^{a,b} = S_{j,n}^{a_0,b_0} + \frac{(|a| - b + 2 - 2\epsilon + 10^{-5})}{\log(j+1)}; \quad (3.5.20)$$

$$v_{j,n}^{a,b} = S_{j,n}^{a_0,b_0} + \frac{(|a| - b + 2 - 2\epsilon - 10^{-5})}{\log(j+1)}. \quad (3.5.21)$$

It can be easily verified that the upper bound is decreasing in  $j$ . Notice that we add the term  $10^{-5}$  when  $b < 2$  so that our Bayesian method favours convergence and subtract the same when  $b \geq 2$  to facilitate detection of divergence. Since convergence or divergence of  $S^{a,b}$  does not depend upon  $a \in [-\eta, \eta] \setminus \{0\}$ , we use  $|a|$  in (3.5.20) and (3.5.21).

Setting  $\epsilon = 10^{-10}$ , Figures 3.5.7 and 3.5.8 depict convergence and divergence of  $S^{a,b}$  for various values of  $a$  and  $b$ . In particular, panel (e) of Figure 3.5.8 shows that our main interest, the series  $S$ , given by (3.5.16), converges.

### 3.5.7 Example 7

We now consider

$$S = \sum_{i=1}^{\infty} \frac{|\sin i|^i}{i}. \quad (3.5.22)$$

We embed this series into

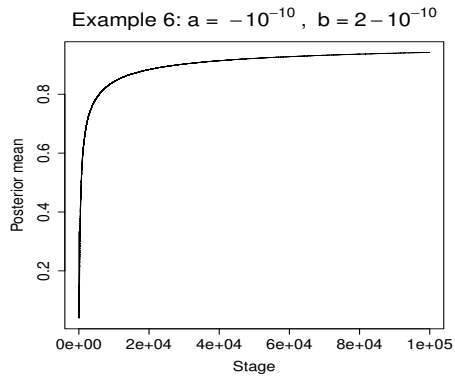
$$S^{a,b} = \sum_{i=1}^{\infty} \frac{|\sin a\pi i|^i}{i^b}, \quad (3.5.23)$$

where  $a \in \mathbb{R}$  and  $b \geq 1$ . The above series converges if  $b > 1$ , for all  $a \in \mathbb{R}$ . But for  $b = 1$ , it is easy to see that the series diverges if  $a = \ell/2m$ , where  $\ell$  and  $m$  are odd integers.

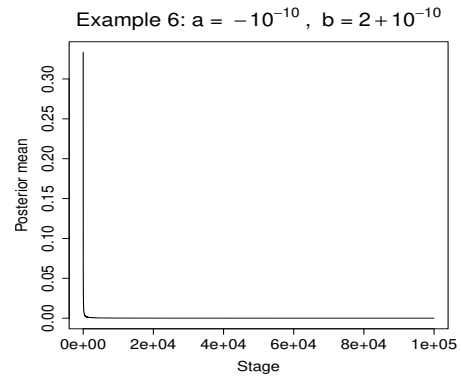
Letting  $a_0 = \pi^{-1}$  and  $b_0 = 1 + \epsilon$ , with  $\epsilon = 10^{-10}$ , we set the following upper bound that is decreasing in  $j$ :

$$c_{j,n}^{a,b} = S_{j,n}^{a_0,b_0} + \frac{\epsilon}{j}. \quad (3.5.24)$$

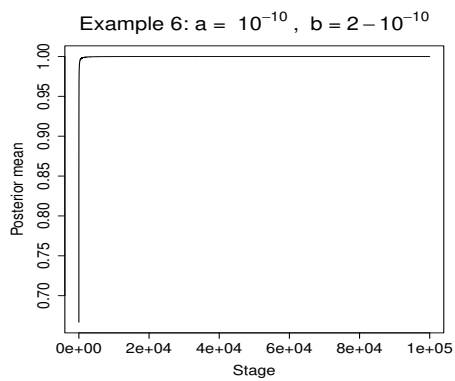
Thus,  $c_{j,n}^{a,b}$  corresponds to a convergent series which is also sufficiently close to divergence. Addition of the term  $\frac{\epsilon}{j}$  provides further protection from erroneous conclusions regarding



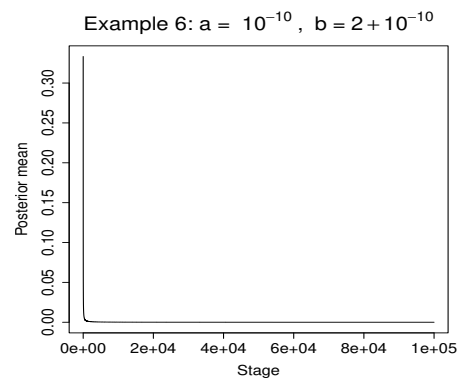
(a) Convergence:  $a = -10^{-10}$ ,  $b = 2 - 10^{-10}$ .



(b) Divergence:  $a = -10^{-10}$ ,  $b = 2 + 10^{-10}$ .

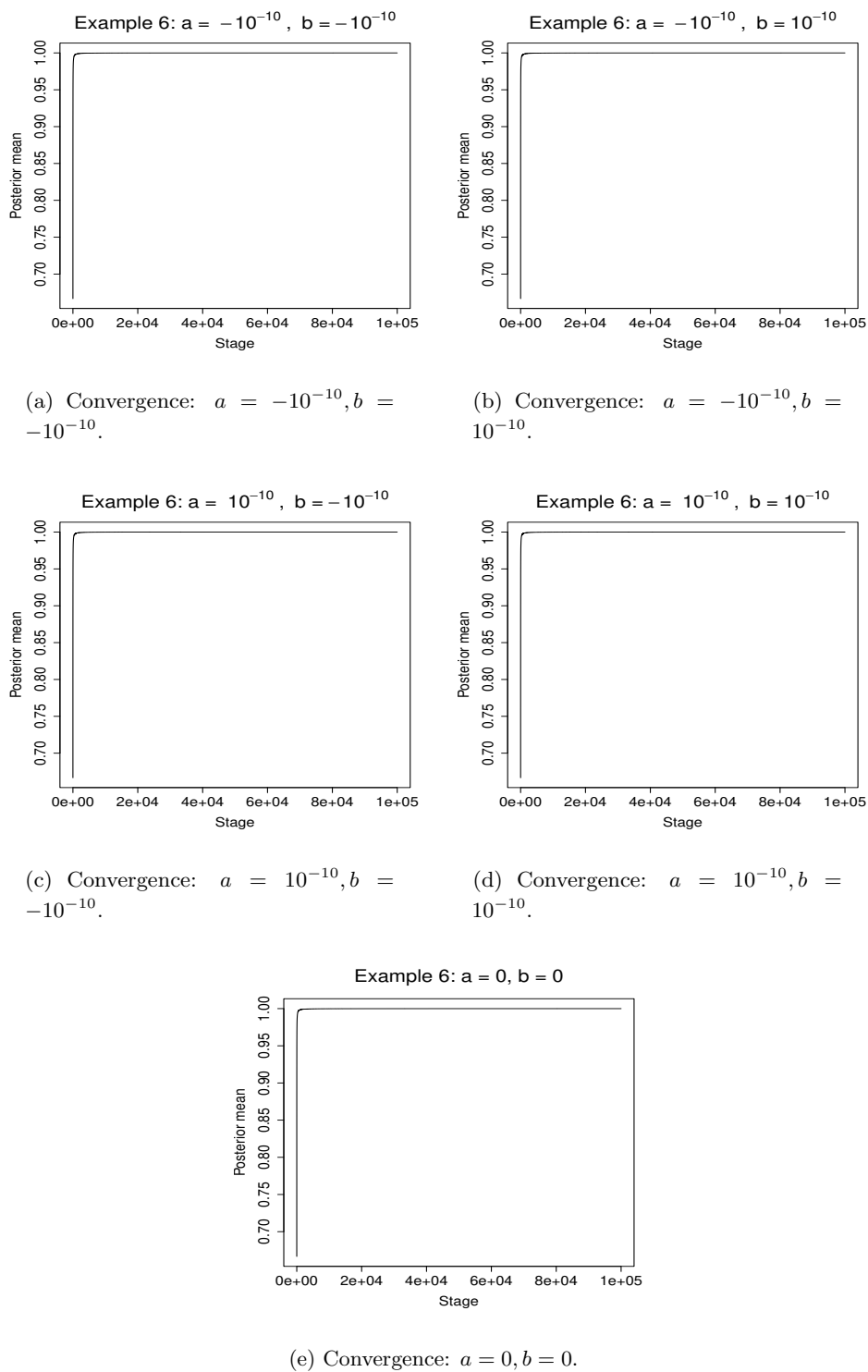


(c) Convergence:  $a = 10^{-10}$ ,  $b = 2 - 10^{-10}$ .

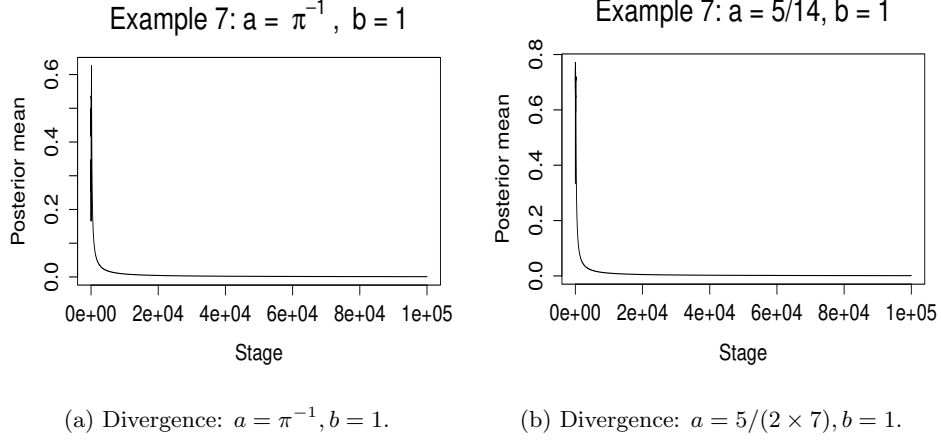


(d) Divergence:  $a = 10^{-10}$ ,  $b = 2 + 10^{-10}$ .

**Figure 3.5.7:** Example 6: The series (3.5.17) converges for  $(a = -10^{-10}, b = 2 - 10^{-10})$ ,  $(a = 10^{-10}, b = 2 - 10^{-10})$ , and diverges for  $(a = -10^{-10}, b = 2 + 10^{-10})$ ,  $(a = 10^{-10}, b = 2 + 10^{-10})$ .



**Figure 3.5.8:** Example 6: The series (3.5.17) converges for  $(a = -10^{-10}, b = -10^{-10})$ ,  $(a = -10^{-10}, b = 10^{-10})$ ,  $(a = 10^{-10}, b = -10^{-10})$ ,  $(a = 10^{-10}, b = 10^{-10})$ , and  $(a = 0, b = 0)$ .



**Figure 3.5.9:** Example 7: The series (3.5.23) diverges for  $(a = \pi^{-1}, b = 1)$ ,  $(a = 5/7, b = 1)$ .

divergence.

Panel(a) of Figure 3.5.9 demonstrates that the series of our interest, given by (3.5.22), diverges. Panel (b) confirms that for  $a = 5/(2 \times 7)$  and  $b = 1$ , the series indeed diverges, as it should.

## 3.6 Application to Riemann Hypothesis

### 3.6.1 Brief background

Consider the Riemann zeta function given by

$$\zeta(a) = \frac{1}{1 - 2^{1-a}} \sum_{n=0}^{\infty} \frac{1}{2^{n+1}} \sum_{k=0}^n (-1)^k \frac{n!}{k!(n-k)!} (k+1)^{-a}, \quad (3.6.1)$$

where  $a$  is complex. The above function is formed by first considering Euler's function

$$Z(a) = \sum_{n=1}^{\infty} \frac{1}{n^a}, \quad (3.6.2)$$

then by multiplying both sides of (3.6.2) by  $(1 - \frac{2}{2^a})$  to obtain

$$\left(1 - \frac{2}{2^a}\right) Z(a) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^a}, \quad (3.6.3)$$

and then dividing the right hand side of (3.6.3) by  $(1 - \frac{2}{2^a})$ . The advantage of the function  $\zeta(a)$  in comparison with the parent function  $Z(a)$  is that,  $Z(a)$  is divergent if the real part of  $a$ , which we denote by  $Re(a)$ , is less than or equal to 1, while  $\zeta(a)$  is convergent for all  $a$  with  $Re(a) > 0$ . Importantly,  $\zeta(a) = Z(a)$  whenever  $Z(a)$  is convergent.

Whenever  $0 < Re(a) < 1$ ,  $\zeta(a)$  satisfies the following identity:

$$\zeta(a) = 2^a \pi^{a-1} \sin\left(\frac{\pi a}{2}\right) \Gamma(1-a) \zeta(1-a), \quad (3.6.4)$$

where  $\Gamma(\cdot)$  is the gamma function. This can be extended to the set of complex numbers by defining a function with non-positive real part by the right hand side of (3.6.4); abusing notation, we denote the new function by  $\zeta(a)$ . Because of the sine function, it follows that the trivial zeros of the above function occur when the values of  $a$  are negative even integers. Hence, the non-trivial zeros must satisfy  $0 < Re(a) < 1$ .

[Riemann \(1859\)](#) conjectured that all the non-trivial zeros have the real part  $1/2$ , which is the famous Riemann Hypothesis. For accessible account of the Riemann Hypothesis, see [Borwein \*et al.\* \(2006\)](#), [Derbyshire \(2004\)](#).

One equivalent condition for the Riemann Hypothesis is related to sums of the Möbius function, given by

$$\mu(n) = \begin{cases} -1 & \text{if } n \text{ is a square-free positive integer with an odd number of prime factors;} \\ 0 & \text{if } n \text{ has a squared prime factor;} \\ 1 & \text{if } n \text{ is a square-free positive integer with an even number of prime factors,} \end{cases} \quad (3.6.5)$$

where, by square-free integer we mean that the integer is not divisible by any perfect square other than 1. Specifically, the condition

$$\sum_{n=1}^x \mu(n) = O\left(x^{\frac{1}{2}+\epsilon}\right) \quad (3.6.6)$$

for any  $\epsilon > 0$ , is equivalent to Riemann Hypothesis. This condition implies that the Dirichlet series for the Möbius function, given by

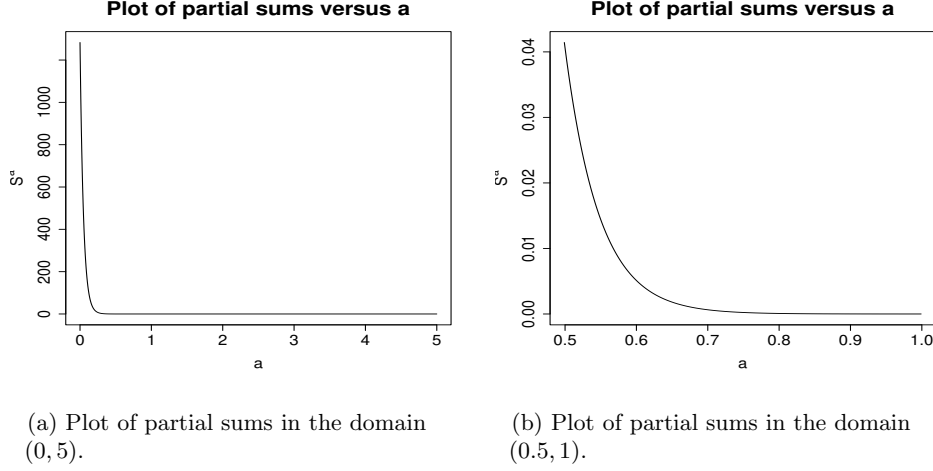
$$M(a) = \sum_{n=1}^{\infty} \frac{\mu(n)}{n^a} = \frac{1}{\zeta(a)}, \quad (3.6.7)$$

is analytic in  $Re(a) > 1/2$ . This again ensures that  $\zeta(a)$  is meromorphic in  $Re(a) > 1/2$  and that it has no zeros in this region. Using the functional equation (3.6.4) it follows that there are no zeros of  $\zeta(a)$  in  $0 < Re(a) < 1/2$  either. Hence, (3.6.6) implies Riemann Hypothesis. The converse is also certainly true.

The above arguments also imply that convergence of  $M(a)$  in (3.6.7) for  $Re(a) > 1/2$  is equivalent to Riemann Hypothesis, and it is this criterion that is of our interest here. Now,  $M(a)$  converges absolutely for  $Re(a) > 1$ ; moreover,  $M(1) = 0$ . The latter is equivalent to the prime number theorem stating that the number of primes below  $x$  is asymptotically  $x/\log(x)$ , as  $x \rightarrow \infty$  (Landau (1906)). Thus,  $M(a)$  converges for  $Re(a) \geq 1$ . That  $M(a)$  diverges for  $Re(a) \leq 1/2$  can be seen as follows. Note that if  $M(a)$  converged for any  $a^*$  such that  $Re(a^*) \leq 1/2$ , then analytic continuation for Dirichlet series of the form  $M(a)$  would guarantee convergence of  $M(a)$  for all  $a$  with  $Re(a) > Re(a^*)$ . But  $\zeta(a)$  is not analytic on  $0 < Re(a) < 1$  because of its non-trivial zeros on the strip. This would contradict the analytic continuation leading to the identity  $M(a) = 1/\zeta(a)$  on the entire set of complex numbers. Hence,  $M(a)$  must be divergent for  $Re(a) \leq 1/2$ .

In this work, we apply our ideas to particularly investigate convergence of  $M(a)$  when  $1/2 < a < 1$ .





**Figure 3.6.1:** Plot of the partial sums  $S_{1000,1000000}^a$  versus  $a$ . Panel (a) shows the plot in the domain  $[0, 5]$  while panel (b) magnifies the same in the domain  $(0.5, 1)$ .

### 3.6.2 Choice of the upper bound and implementation details

To form an idea of the upper bound we first plot the partial sums  $S_{j,n}^a$ , for  $j = 1000$  and  $n = 10^6$ , with respect to  $a$ . In this regard, panel (a) of Figure 3.6.1 shows the decreasing nature of the partial sums with respect to  $a$ , and panel (b) magnifies the plot in the domain  $1/2 < a < 1$  that we are particularly interested in. The latter shows that the partial sums decrease sharply till about 0.7, getting appreciably close to zero around that point, after which the rate of decrease diminishes. Thus, one may expect a change point around 0.7 regarding convergence. Specifically, divergence may be expected below a point slightly larger than 0.7 and convergence above it.

Since  $M(1) < \infty$ , we consider this series as the basis for our upper bound, with the value of  $a$  also taken into account. Specifically, we choose the upper bound as

$$c_{j,n} = \left| S_{j,n}^1 + \frac{a}{j+1} \right|. \quad (3.6.8)$$

Since Figure 3.6.1 shows that the partial sums are of monotonically decreasing nature,

the above choice of upper bound facilitates detection of convergence for relatively large values of  $a$ . The part  $\frac{a}{j+1}$ , which tends to zero as  $j \rightarrow \infty$ , takes care of the fact that the series may be convergent if  $a < 1$ , by slightly inflating  $S_{j,n}^1$ .

For our purpose, we compute the first  $10^9$  values of the Möbius function using an efficient algorithm proposed in [Lioen and van de Lune \(1994\)](#), which is based on the Sieve of Eratosthenes ([Horsley \(1772\)](#)). We set  $K = 1000$  and  $n = 10^6$ . A complete analysis with our VMware with our parallel implementation takes about 2 minutes.

### 3.6.3 Results of our Bayesian analysis

Panels (a)–(e) of Figure 3.6.2 and panels (d)–(f) of Figure 3.6.3 show the  $M(a)$  diverges for  $a = 0.1, 0.2, 0.3, 0.4, 0.5$ , but converges for  $a = 1 + 10^{-10}, 2$  and  $3$ . In fact, for many other values that we experimented with,  $M(a)$  converged for  $a > 1$  and diverged for  $a < 1/2$ , demonstrating remarkable consistency with the known, existing results.

Certainly far more important are the results for  $1/2 < a < 1$ . Indeed, panel (f) of Figure 3.6.2 and panels (a)–(c) of Figure 3.6.3 show that  $M(a)$  diverged for  $a = 0.6$  and  $0.7$  and converged for  $a = 0.8$  and  $0.9$ . It thus appears that  $M(a)$  diverges for  $a < a^*$  and converges for  $a \geq a^*$ , for some  $a^* \in (0.7, 0.8)$ . Figure 3.6.4 displays results of our further experiments in this regard. Panels (a) and (b) of Figure 3.6.4 show the posterior means for the full set of iterations and the last 500 iterations, respectively, for  $a = 0.71$ . Note that from panel (a), convergence seems to be attained, although towards the end, the plot seems to be slightly tilted downwards. Panel (b) magnifies this, clearly showing divergence. Panels (c) and (d) of Figure 3.6.4 depict similar phenomenon for  $a = 0.715$ , but as per panel (d), divergence seems to ensue all of a sudden, even after showing signs of convergence for the major number of iterative stages. Convergence of  $M(a)$  begins at  $a = 0.72$  (approximately); panels (e) and (f) of Figure 3.6.4 take clear note of this.

Thus, as per our methods,  $M(a)$  diverges for  $a < 0.72$  and converges for  $a \geq 0.72$ . This is remarkably in keeping with the wisdom gained from panel (b) of Figure 3.6.1

that convergence is expected to occur for values of  $a$  exceeding 0.7. Note that neither the upper bound (3.6.8), nor our methodology, is in any way biased towards  $a \approx 0.7$ ; hence, our result is perhaps not implausible.

#### 3.6.4 Implications of our result

As per our results,  $M(a)$  does not converge for all  $a > 1/2$ , and hence does not completely support Riemann Hypothesis. However, convergence of  $M(a)$  fails only for the relatively small region  $0.5 < a < 0.72$ , which perhaps is the reason why there exists much evidence in favour of Riemann Hypothesis.

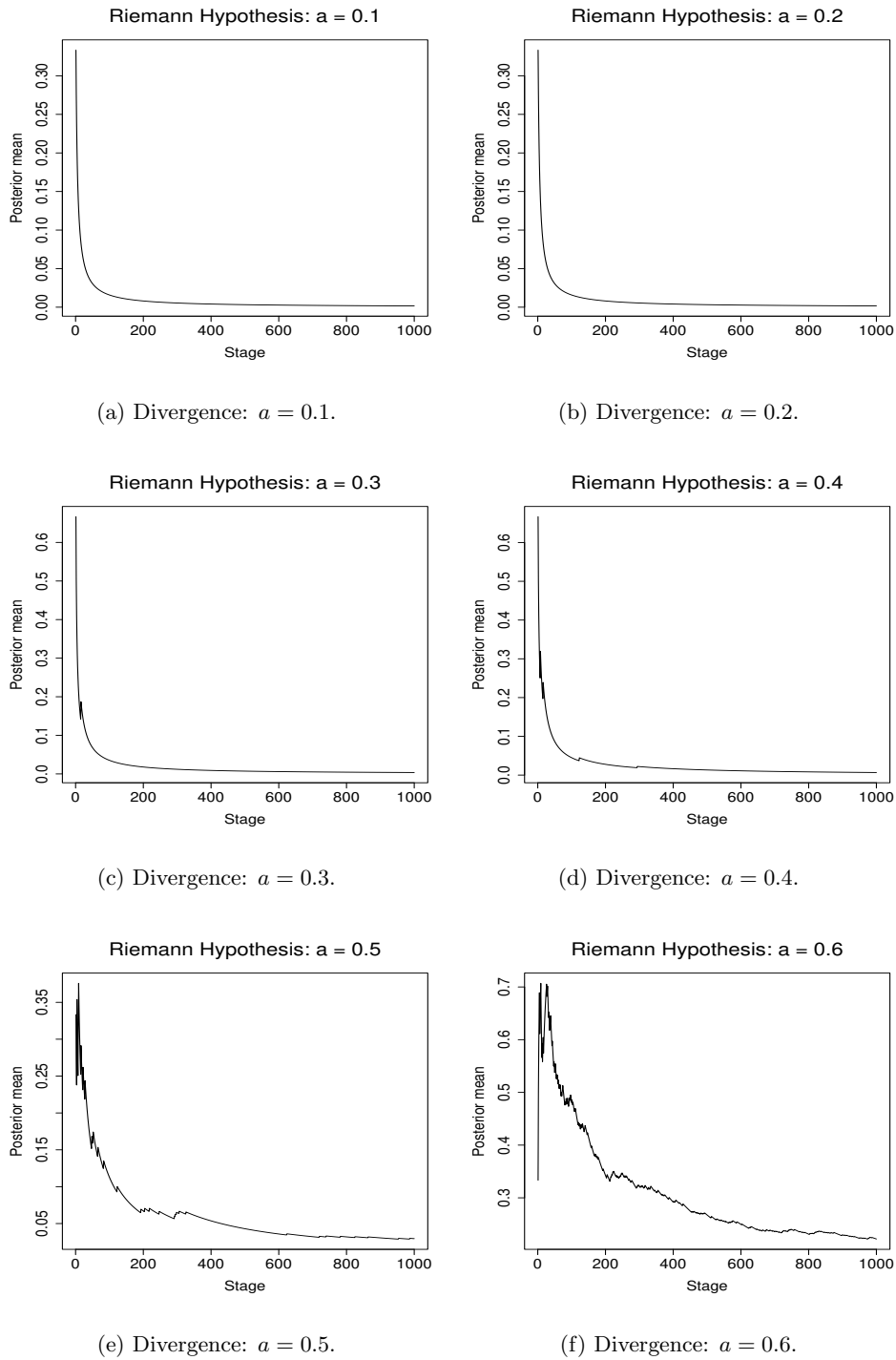
### 3.7 Summary and conclusion

In this chapter, we proposed and developed a novel Bayesian methodology for assessment of convergence of infinite series. Our developments do not require any restrictive assumption, not even independence of the elements  $X_i$  of the infinite series.

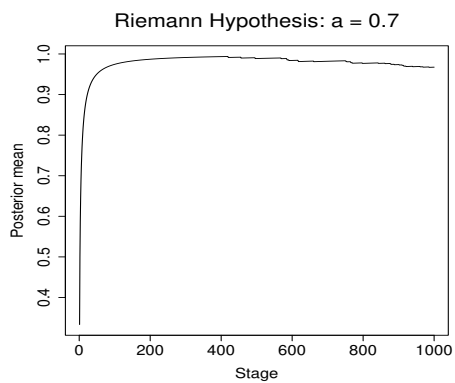
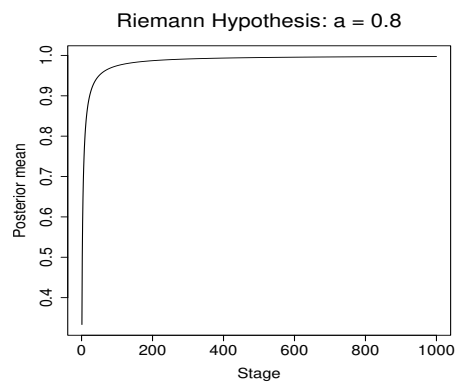
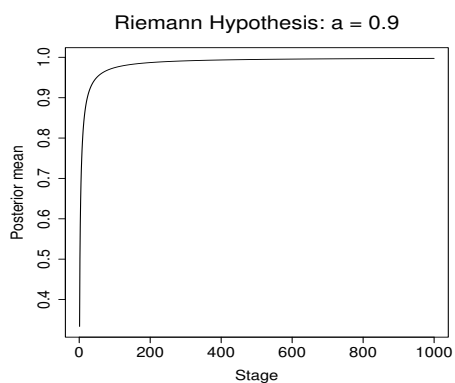
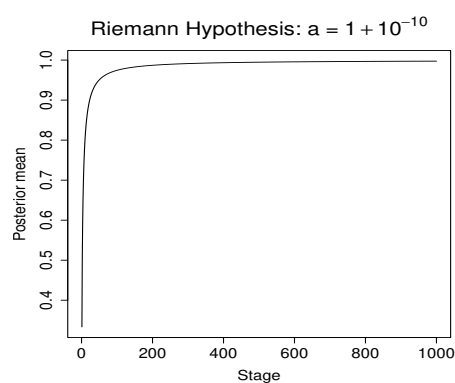
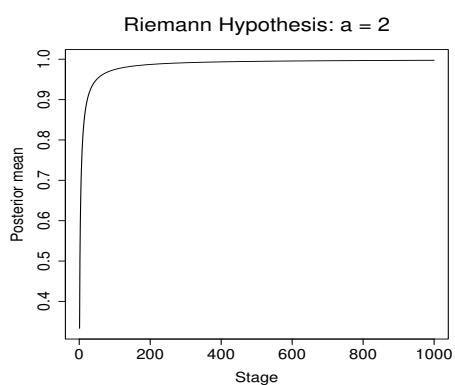
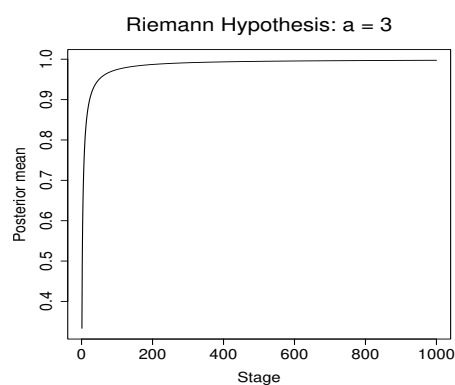
We demonstrated the reliability and efficiency of our methods with varieties of examples, the most important one being associated with Riemann Hypothesis.

The results of our Bayesian characterization are not in support of the Riemann Hypothesis, and this is upheld by informal plots of the partial sums depicted in Figure 3.6.1. Further support of our Riemann hypothesis results can be obtained by exploiting the characterization of Riemann hypothesis by convergence of certain infinite series based on Bernoulli numbers; the details are presented in Section 3.A3.

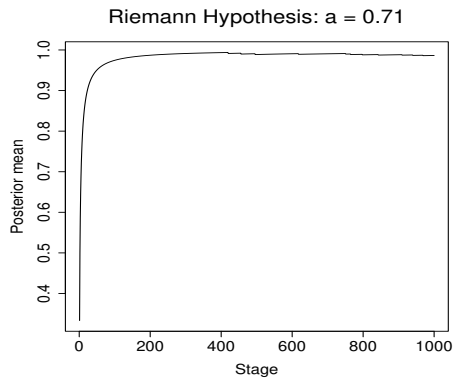
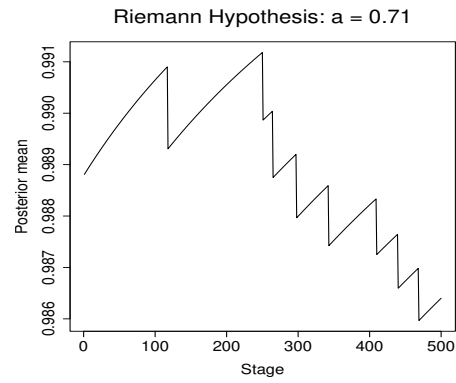
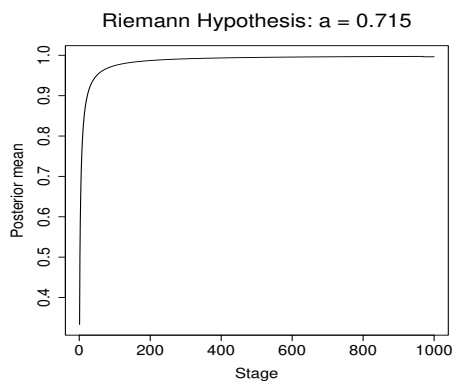
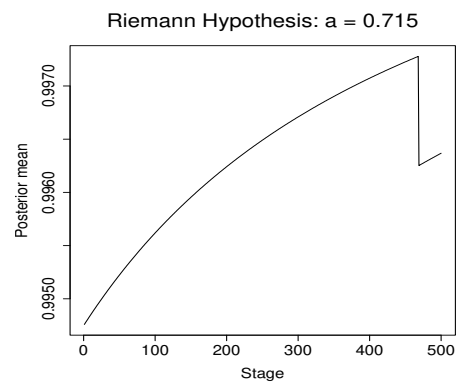
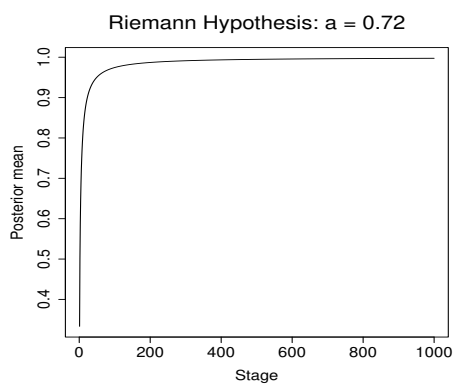
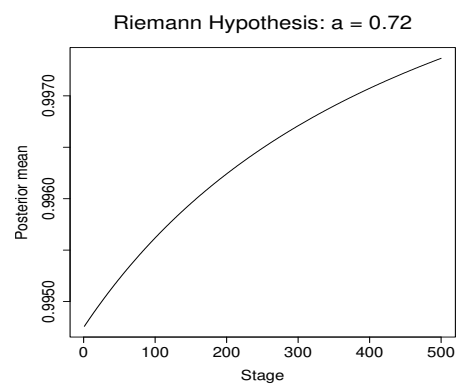
In fine, it is worth reminding the reader that although our work attempts to provide insights regarding Riemann hypothesis, we did not develop our Bayesian approach keeping Riemann hypothesis in mind. Indeed, our primary objective is to develop Bayesian approaches to studying convergence properties of infinite series in general. From this perspective, Riemann hypothesis is just an example where it makes sense to learn about convergence properties of a certain class of infinite series. Further development



**Figure 3.6.2:** Riemann Hypothesis: The Möbius function based series diverges for  $a = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6$ .

(a) Divergence:  $a = 0.7$ .(b) Convergence:  $a = 0.8$ .(c) Convergence:  $a = 0.9$ .(d) Convergence:  $a = 1 + 10^{-10}$ .(e) Convergence:  $a = 2$ .(f) Convergence:  $a = 3$ .

**Figure 3.6.3:** Riemann Hypothesis: The Möbius function based series diverges for  $a = 0.7$  but converges for  $a = 0.8, 0.9, 1 + 10^{-10}, 2, 3$ .

(a) Divergence:  $a = 0.71$ .(b) Divergence:  $a = 0.71$ .(c) Divergence:  $a = 0.715$ .(d) Divergence:  $a = 0.715$ .(e) Convergence:  $a = 0.72$ .(f) Convergence:  $a = 0.72$ .

**Figure 3.6.4:** Riemann Hypothesis: The left panels show the posterior means for the full set of iterations, while the right panels depict the posterior means for the last 500 iterations, for  $a = 0.71$ ,  $0.715$  and  $0.72$ . It is evident that the Möbius function based series diverges for  $a = 0.71$  and  $0.715$  but converges for  $a = 0.72$ .

of our approach is of course in the cards. Note that the theory that we developed for deterministic series remains valid for random series as well, but since the forms of the terms of random series are unknown, direct application of our methods is not possible. In Chapter 5 we develop the detailed theory and methods for Bayesian characterization of random series, with important applications to climate change.

# Appendix

## 3.A1 Proof of Lemma 8

Since each term of the series (3.1.1) is decreasing in  $a$ , it is clear that  $S_{j,n}^{a,b}$  is decreasing in  $a$ . We need to show that  $S_{j,n}^{a,b}$  is increasing in  $b$ .

Let, for  $i \geq 3$ ,

$$g(i) = \left(1 - \frac{\log i}{i} - \frac{\log \log i}{i} \left\{ \cos^2 \left( \frac{1}{i} \right) \right\} (a + (-1)^i b) \right)^i. \quad (3.A1.1)$$

Observe that all our partial sums of the form  $S_{j,n}^{a,b}$  for  $j \geq 3$  admit the form

$$S_{j,n}^{a,b} = \sum_{i=r}^{r+n-1} g(i), \quad (3.A1.2)$$

where  $r = 3 + n(j - 1)$ , which is clearly odd because  $n$  is even. Now,

$$\sum_{i=r}^{r+n-1} g(i) = \{g(r) + g(r+1)\} + \{g(r+2) + g(r+3)\} + \cdots + \{g(r+n-2) + g(r+n-1)\}, \quad (3.A1.3)$$

where the sums of the consecutive terms within the parentheses have the form

$$\begin{aligned} & g(r+\ell) + g(r+\ell+1) \\ &= \left(1 - \frac{\log(r+\ell)}{r+\ell} - \frac{\log \log(r+\ell)}{r+\ell} \left\{ \cos^2 \left( \frac{1}{r+\ell} \right) \right\} (a + (-1)^{(r+\ell)} b) \right)^{(r+\ell)} \\ & \quad + \left(1 - \frac{\log(r+\ell+1)}{r+\ell+1} - \frac{\log \log(r+\ell+1)}{r+\ell+1} \left\{ \cos^2 \left( \frac{1}{r+\ell+1} \right) \right\} (a + (-1)^{(r+\ell+1)} b) \right)^{(r+\ell+1)}. \end{aligned} \quad (3.A1.4)$$



Since  $r$  is odd, and since the terms are represented pairwise in (3.A1.3) it follows that in (3.A1.4),  $r + \ell$  is odd and  $r + \ell + 1$  is even. That is, in (3.A1.4),  $a + (-1)^{(r+\ell)}b = a - b$  and  $a + (-1)^{(r+\ell+1)}b = a + b$ . Since  $\cos^2(\theta)$  is decreasing on  $[0, \frac{\pi}{2}]$ , and since  $\frac{1}{i} \leq \frac{\pi}{2}$  for  $i \geq 3$ , it follows that  $\cos^2\left(\frac{1}{i}\right)$  is increasing in  $i$ . Moreover,  $\frac{\log \log i}{i}$  decreases in  $i$  at a rate faster than  $\cos^2\left(\frac{1}{i}\right)$  increases, so that  $\frac{\log \log i}{i} \times \cos^2\left(\frac{1}{i}\right)$  decreases in  $i$ . It follows that

$$\frac{\log \log(r + \ell)}{r + \ell} \cos^2\left(\frac{1}{r + \ell}\right) > \frac{\log \log(r + \ell + 1)}{r + \ell + 1} \cos^2\left(\frac{1}{r + \ell + 1}\right). \quad (3.A1.5)$$

Note that in  $g(r + \ell) + g(r + \ell + 1)$ ,  $\frac{\log \log(r+\ell)}{r+\ell} \cos^2\left(\frac{1}{r+\ell}\right)$  is associated with  $-b$  while  $\frac{\log \log(r+\ell+1)}{r+\ell+1} \cos^2\left(\frac{1}{r+\ell+1}\right)$  involves  $b$ . Hence, increasing  $b$  increases  $g(r + \ell)$  but decreases  $g(r + \ell + 1)$ , and because of (3.A1.5),  $g(r + \ell) + g(r + \ell + 1)$  increases in  $b$ . This ensures that  $\sum_{i=r}^{r+n-1} g(i)$ , given by (3.A1.3), is increasing in  $b$ . In other words, partial sums of the form (3.A1.2) are increasing in  $b$ , proving Lemma 8 when  $n$  is even.

### 3.A2 Proof of Lemma 10

That  $S_{j,n}^{a,b}$  is decreasing in  $a$  follows trivially since each term of (3.5.13) is decreasing in  $a$ . We need to show that  $S_{j,n}^{a,b}$  is increasing in  $b$ .

Let, for  $i \geq 5$ ,

$$g(i) = \left(1 - \left(\frac{\log(i)}{i}\right) \left(a \left(1 + \sin^2\left(\sqrt{\left(\frac{\log(\log(i))}{\log(i)}\right)}\right)\right) + b \sin\left(\frac{i\pi}{4}\right)\right)\right)^i. \quad (3.A2.1)$$

Now note that, with  $r = 5 + n(j - 1)$ ,

$$\begin{aligned} \sum_{i=r}^{r+n-1} g(i) &= \sum_{m=1}^{\frac{n}{4}} Z_{r,m} \\ &= \{Z_{r,1} + Z_{r,2}\} + \{Z_{r,3} + Z_{r,4}\} + \cdots + \left\{Z_{r, \frac{n}{4}-1} + Z_{r, \frac{n}{4}}\right\}, \end{aligned} \quad (3.A2.2)$$

where

$$Z_{r,m} = \sum_{\ell=5+4(m-1)}^{5+4(m-1)+3} g(r + \ell). \quad (3.A2.3)$$

Now, for any  $\ell \geq 1$ , observe that in  $\{Z_{r,\ell} + Z_{r,\ell+1}\}$ , the term  $Z_{r,\ell}$  consists of only negative signs of the sine-values, while in  $Z_{r,\ell+1}$  the corresponding signs are positive, although the magnitudes are the same. Since  $\log(i)/i$  is decreasing in  $i$ , it follows that  $\{Z_{r,\ell} + Z_{r,\ell+1}\}$  is increasing in  $b$  for  $\ell \geq 1$ . Hence, it follows that (3.A2.2), and  $S_{j,n}^{a,b}$ , defined by (3.5.14), are increasing in  $b$  for  $j \geq 1$  and  $n$ , a multiple of 4, proving Lemma 10.

### 3.A3 Characterization of Riemann Hypothesis based on Bernoulli numbers

Characterization of Riemann Hypothesis by convergence of infinite sums associated with Bernoulli numbers are provided in [Carey \(2003\)](#) (unpublished, according to our knowledge). In particular, it has been shown that Riemann hypothesis is true if and only if the following series is convergent:

$$\tilde{S}_1 = \sum_{m=1}^{\infty} \frac{\pi(4m+3)}{2^{4m+1}} \sum_{k=0}^m (-1)^k \frac{\binom{2m+1}{k} \binom{4m+2-2k}{2m+1}}{2m+2-2k} \log \left( \frac{(2\pi)^{2m+2-2k} |B_{2m+2-2k}|}{2(2m+2-2k)^2 (2m-2k)!} \right), \quad (3.A3.1)$$

where  $\{B_n; n = 0, 1, \dots\}$  are Bernoulli numbers characterized by their generating function  $\sum_{n=0}^{\infty} B_n x^n / n! = x / (\exp(x) - 1)$ . The Bernoulli numbers are related to the Riemann zeta function by (see, for example [Sury \(2003\)](#))

$$B_{2m} = (-1)^{m-1} \frac{2(2m)!}{(2\pi)^{2m}} \zeta(2m). \quad (3.A3.2)$$

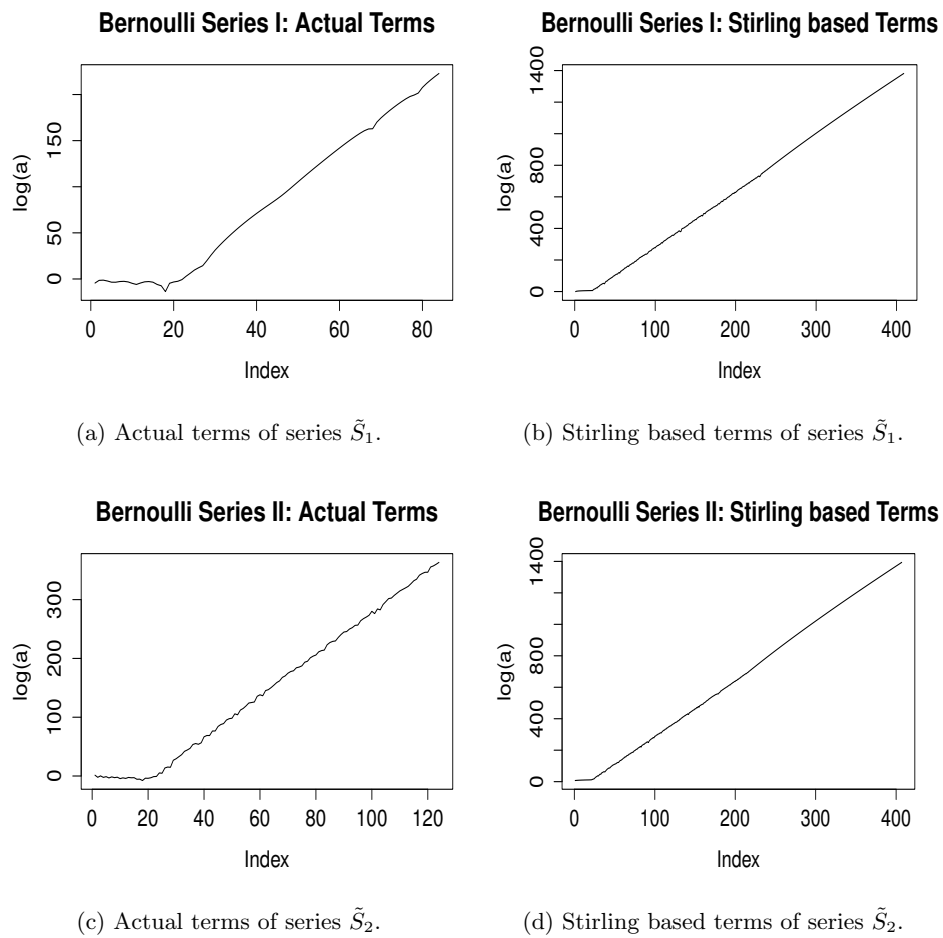
Carey (2003) further showed that convergence of the related series

$$\tilde{S}_2 = \sum_{m=1}^{\infty} \frac{\pi(4m+3)}{2^{4m+1}} \sum_{k=0}^m (-1)^k \frac{\binom{2m+1}{k} \binom{4m+2-2k}{2m+1}}{2m+2-2k} \log \left( (2m+1-2k) \frac{|B_{2m+2-2k}|}{|B_{2m+4-2k}|} \right), \quad (3.A3.3)$$

is also equivalent to the assertion that Riemann hypothesis is correct. However, the terms of both the series (3.A3.1) and (3.A3.3) tend to explode very quickly. Stirlings's approximation of the factorials involved in the summands facilitates computation of larger number of summands compared to the original terms. In this context, note that Stirling's approximation applied to the factorials in (3.A3.2), along with the approximation  $\zeta(2m) \sim 1$ , as  $m \rightarrow \infty$ , lead the following asymptotic form of  $B_{2m}$  as  $m \rightarrow \infty$ :

$$B_{2m} \sim (-1)^{m-1} 4\sqrt{\pi m} \left( \frac{m}{\pi e} \right)^{2m}. \quad (3.A3.4)$$

Figure 3.A3.1 shows the logarithms of the first few terms  $a_m$  of the above two series, based on the actual terms  $a_m$  and the Stirling-approximated  $a_m$  (ignoring a multiplicative constant); the rest of the terms become too large to be reliably computed, even with Stirling's approximation. The bottomline that emerges from (3.A3.1) is that the series  $\tilde{S}_1$  and  $\tilde{S}_2$  appear to be clearly divergent, providing some support to our result on Riemann hypothesis.



**Figure 3.A3.1:** Actual and Stirling-approximated terms  $a_m$  of the series  $\tilde{S}_1$  and  $\tilde{S}_2$ .

# 4

## Bayesian Characterization of Oscillatory Series with Multiple Limit Points

### 4.1 Introduction

As a follow-up of Chapter 3, in this chapter we assume that the sequence  $\{S_{1,n}\}_{n=1}^{\infty}$  has multiple limit points, including the possibility that the number of limit points is countably infinite, and develop Bayesian characterizations of the number of limit points. The multiple limit point premise naturally suggests extension of our characterization theory with Bernoulli and Beta distributions in Chapter 3 to characterization with Multinomial and Dirichlet distributions, when the number of limit points is finite. However, as we shall elucidate, there are important differences in the characterization theory for multiple limit points, in the conceptualization procedure and the mathematical treatise, as well as in the computational aspect.

For infinite number of limit points, we develop the Bayesian characterization theory after extending the Multinomial and Dirichlet distributions to infinite-dimensional Multinomial distribution and the Dirichlet process, the well-known prior for Bayesian nonparametric problems (Ferguson (1973)). In fact, our Bayesian characterization for finite number of limit points becomes a special case of this infinite-dimensional situation.

Moreover, as is intuitively expected, convergence and divergence of non-oscillating infinite series can also be characterized with the Bayesian characterization concepts for multiple limit points. As such, we also provide a formal theory in this regard.

We illustrate the effectiveness of our multiple limit point characterization theories with several examples, consisting of both oscillating and non-oscillating series. Finally, we apply our multiple limit point characterization strategies to the Riemann Hypothesis problem and obtain results that again negate the validity of the most (in)famous mathematical conjecture.

The rest of this chapter is structured as follows. In Section 4.2 we develop the Bayesian characterization when the number of limit points is finite, whereas the case of infinite number of limit points is undertaken in Section 4.3. Bayesian characterization of convergence and divergence of non-oscillating series using the generalized concepts of oscillating series, is developed in Section 4.4. A rule of thumb for implementation of our theories and methods is provided in Section 4.5. Section 4.6 presents illustrations of our theory with an oscillating example and a non-oscillating example. The details of the application of our Bayesian multiple limit point characterization theory are presented in Section 4.7. Finally, we make concluding remarks in Section 4.8.

## 4.2 Bayesian characterization for finite number of limit points

Let us assume that there are  $M$  ( $> 1$ ) limit points of the sequence  $\{S_{1,n}\}_{n=1}^{\infty}$ . Then there exist sequences  $\{c_{m,j}\}_{j=1}^{\infty}$ ;  $m = 0, \dots, M$ , such that  $\{(c_{m-1,j}, c_{m,j}]; m = 1, \dots, M\}$  partition the real line  $\mathbb{R}$  for every  $j \geq 1$  and that there exists  $j_0 \geq 1$  such that for all  $j \geq j_0$ , the interval  $(c_{m-1,j}, c_{m,j}]$  contains at most one limit point of the sequence  $\{S_{1,n}\}_{n=1}^{\infty}$ , for every  $m = 1, \dots, M$ . With these sequences we define

$$Y_j = m \text{ if } c_{m-1,j} < S_{1,j} \leq c_{m,j}; m = 1, 2, \dots, M, \quad (4.2.1)$$

Recall that in Section 3.4, we allowed the sequence  $\{c_j\}_{j=1}^{\infty}$  to depend upon the underlying series  $S_{1,\infty}$ . Likewise, here also we allow the quantities  $c_{0,j}, c_{1,j}, \dots, c_{M,j}$  to depend upon  $S_{1,\infty}$ . In other words, for  $\omega \in \mathfrak{S}$ , for  $m = 0, 1, 2, \dots, M$ , and  $j = 1, 2, 3, \dots$ ,  $c_{m,j} = c_{m,j}(\omega)$  corresponds to  $S_{1,\infty}(\omega)$ .

Note that unlike our ideas appropriate for non-oscillating series, here do not consider blocks of partial sums,  $S_{j,n_j} = \sum_{i=\sum_{k=0}^{j-1} n_{k+1}}^{\sum_{k=0}^j n_k} X_i$ , but  $S_{1j} = \sum_{i=1}^j X_i$ . In other words, for Bayesian analysis of non-oscillating series we compute sums of  $n_j$  terms in each iteration, whereas for oscillating series we keep adding a single term at every iteration. Thus, computationally, the latter is a lot simpler.

We assume that

$$(\mathbb{I}(Y_j = 1), \dots, \mathbb{I}(Y_j = M)) \sim \text{Multinomial}(1, p_{1,j}, \dots, p_{M,j}), \quad (4.2.2)$$

where  $p_{m,j}$  can be interpreted as the probability that  $S_{1,j} \in (c_{m-1,j}, c_{m,j}]$ . As  $j \rightarrow \infty$  it is expected that  $c_{m-1,j}$  and  $c_{m,j}$  will converge to appropriate constants depending upon  $m$ , and that  $p_{m,j}$  will tend to the correct proportion of the limit point indexed by  $m$ . Indeed, let  $\{p_{m,0}; m = 1, \dots, M\}$  denote the actual proportions of the limit points

indexed by  $\{1, \dots, M\}$ , as  $j \rightarrow \infty$ .

Following the same principle discussed in Section 3.3, and extending the Beta prior to the Dirichlet prior, at the  $k$ -th stage we arrive at the following posterior of  $\{p_{m,k} : m = 1, \dots, M\}$ :

$$\pi(p_{1,k}, \dots, p_{M,k} | y_k) \equiv \text{Dirichlet} \left( \sum_{j=1}^k \frac{1}{j^2} + \sum_{j=1}^k \mathbb{I}(y_j = 1), \dots, \sum_{j=1}^k \frac{1}{j^2} + \sum_{j=1}^k \mathbb{I}(y_j = M) \right). \quad (4.2.3)$$

The posterior mean and posterior variance of  $p_{m,k}$ , for  $m = 1, \dots, M$ , are given by:

$$E(p_{m,k} | y_k) = \frac{\sum_{j=1}^k \frac{1}{j^2} + \sum_{j=1}^k \mathbb{I}(y_j = m)}{M \sum_{j=1}^k \frac{1}{j^2} + k}; \quad (4.2.4)$$

$$\text{Var}(p_{m,k} | y_k) = \frac{\left( \sum_{j=1}^k \frac{1}{j^2} + \sum_{j=1}^k \mathbb{I}(y_j = m) \right) \left( (M-1) \sum_{j=1}^k \frac{1}{j^2} + k - \sum_{j=1}^k \mathbb{I}(y_j = m) \right)}{\left( M \sum_{j=1}^k \frac{1}{j^2} + k \right)^2 \left( M \sum_{j=1}^k \frac{1}{j^2} + k + 1 \right)}. \quad (4.2.5)$$

Let  $k = M\tilde{k}$ , where  $\tilde{k} \rightarrow \infty$ . Then, from (4.2.4) and (4.2.5) it is easily seen, using  $\frac{\sum_{j=1}^k \mathbb{I}(y_j(\omega) = m)}{k} \rightarrow p_{m,0}$  as  $k \rightarrow \infty$ , that,

$$E(p_{m,k} | y_k) \rightarrow p_{m,0}, \quad \text{and} \quad (4.2.6)$$

$$\text{Var}(p_{m,k} | y_k) = O\left(\frac{1}{k}\right) \rightarrow 0, \quad (4.2.7)$$

as  $k \rightarrow \infty$ .

We can now characterize the  $m$  limit points of  $S_{1,\infty}(\omega)$  in terms of the limits of the marginal posterior probabilities of  $p_{m,k}$ , denoted by  $\pi_m(\cdot | y_k(\omega))$ , as  $k \rightarrow \infty$ .

**Theorem 12** For  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ , where  $\mathfrak{N}$  has zero probability measure,  $\{S_{1,n}(\omega)\}_{n=1}^{\infty}$  has  $M (> 1)$  limit points almost surely if and only if

- (1) There exist sequences  $\{c_{m,j}(\omega)\}_{j=1}^{\infty}$ ;  $m = 0, \dots, M$ , such that  $(c_{m-1,j}(\omega), c_{m,j}(\omega))$



partition the real line  $\mathbb{R}$  for every  $j \geq 1$  and  $m = 1, \dots, M$ .

(2) There exists  $j_0(\omega) \geq 1$  such that for all  $j \geq j_0(\omega)$ , for  $m = 1, \dots, M$ ,  $(c_{m-1,j}(\omega), c_{m,j}(\omega)]$  contains at most one limit point of  $\{S_{1,n}(\omega)\}_{n=1}^{\infty}$ .

(3) With  $Y_j$  defined as in (4.2.1),

$$\pi_m(\mathcal{N}_{p_{m,0}}|y_k(\omega)) \rightarrow 1, \quad (4.2.8)$$

as  $k \rightarrow \infty$ . In the above,  $\mathcal{N}_{p_{m,0}}$  is any neighborhood of  $p_{m,0}$ , with  $p_{m,0}$  satisfying  $0 < p_{m,0} < 1$  for  $m = 1, \dots, M$  such that  $\sum_{m=1}^M p_{m,0} = 1$ .

**Proof.** For  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ , where  $\mathfrak{N}$  has zero probability measure, let  $S_{1,\infty}(\omega)$  be oscillatory with  $M$  limit points having proportions  $\{p_{m,0}; m = 1, \dots, M\}$ . Conditions (1) and (2) then clearly hold. Then with our definition of  $Y_j$  provided in (4.2.1), the results (4.2.6) and (4.2.7) hold with  $k = M\tilde{k}$ , where  $\tilde{k} \rightarrow \infty$ . Now let  $\mathcal{N}_{p_{m,0}}$  be any neighborhood of  $p_{m,0}$ . Let  $\epsilon > 0$  be sufficiently small so that  $\mathcal{N}_{p_{m,0}} \supseteq \{|p_{m,k} - p_{m,0}| < \epsilon\}$ . Then by Chebychev's inequality, using (4.2.6) and (4.2.7), it is seen that  $\pi_m(\mathcal{N}_{p_{m,0}}|y_k(\omega)) \rightarrow 1$ , as  $k \rightarrow \infty$ . Thus, (4.2.8) holds. In fact, more generally, condition (3) holds.

Now assume that conditions (1), (2), (3) hold. Then  $\pi_m(|p_{m,k} - p_{m,0}| < \epsilon|y_k(\omega)) \rightarrow 1$ , as  $k \rightarrow \infty$ . Combining this with Chebychev's inequality it follows that (4.2.6) and (4.2.7) hold with  $0 < p_{m,0} < 1$  for  $m = 1, \dots, M$  such that  $\sum_{m=1}^M p_{m,0} = 1$ . If  $\{S_{1,n}(\omega)\}_{n=1}^{\infty}$  has less than  $M$  limit points, then at least one  $p_{m,0} = 0$ , providing a contradiction. Hence  $\{S_{1,n}(\omega)\}_{n=1}^{\infty}$  must have  $M$  limit points. ■

#### 4.2.1 Choice of $c_{0,j}, \dots, c_{M,j}$ for a given series

Let us define, for  $j = 1, 2, \dots, k$ ,

$$\tilde{p}_{\ell,j} = \begin{cases} 0 & \text{if } \ell = 0; \\ E(p_{\ell,j}|y_j) & \text{if } \ell = 1, 2, \dots, M. \end{cases} \quad (4.2.9)$$

We also define, for  $\ell = 1, 2, \dots, M$ ,

$$\tilde{p}_{\ell,0} = E(p_{\ell,1}), \quad (4.2.10)$$

the prior mean at the first stage, before observing any data.

We then set  $c_{0,j} \equiv 0$  for all  $j = 1, 2, \dots, k$ , and, for  $m \geq 1$ , define

$$c_{m,j} = \log \left[ \frac{(\sum_{\ell=1}^m \tilde{p}_{\ell,j-1})^{1/\rho(\theta)}}{1 - (\sum_{\ell=1}^m \tilde{p}_{\ell,j-1})^{1/\rho(\theta)}} \right], \quad (4.2.11)$$

for  $j = 1, 2, \dots, k$ . Thus, the inequality  $c_{m-1,j} < S_{1,j} \leq c_{m,j}$  in (4.2.1) is equivalent to

$$\sum_{\ell=1}^{m-1} \tilde{p}_{\ell,k} < \left( \frac{\exp(S_{1,j})}{1 + \exp(S_{1,j})} \right)^{\rho(\theta)} \leq \sum_{\ell=1}^m \tilde{p}_{\ell,k}, \quad (4.2.12)$$

where  $\rho(\theta)$  is some relevant power depending upon the set of parameters  $\theta$  of the given series, responsible for appropriately inflating or contracting the quantity  $\frac{\exp(S_{1,j})}{1 + \exp(S_{1,j})}$  for properly diagnosing the limit points. Thus, given the series  $S_{1,\infty}(\omega)$ ,  $\theta = \theta(\omega)$  is allowed to depend upon the underlying series. If  $\left( \frac{\exp(S_{1,j})}{1 + \exp(S_{1,j})} \right)^{\rho(\theta)} \geq 1$ , we set  $Y_j = M$ . By (4.2.8), for large  $k$ ,  $\tilde{p}_{\ell,k}$  and  $S_{1,j}$  adaptively adjust themselves so that the correct proportions of the limit points are achieved in the long run.

### 4.3 Infinite number of limit points

We now assume that the number of limits points of  $\{S_{1,n}(\omega)\}_{n=1}^{\infty}$  is countably infinite, and that  $\{p_{m,0}; m = 1, 2, 3, \dots\}$ , where  $0 \leq p_{m,0} \leq 1$  and  $\sum_{m=1}^{\infty} p_{m,0} = 1$ , are the true proportions of the limit points.

Now we define

$$Y_j = m \text{ if } c_{m-1,j} < S_{1,j} \leq c_{m,j}; \quad m = 1, 2, \dots, \infty, \quad (4.3.1)$$

where the sequences  $\{c_{m,j}\}_{j=1}^{\infty}$ ;  $m \geq 1$ , are such that  $(c_{m-1,j}, c_{m,j}]$ ;  $m \geq 1$ , partition  $\mathbb{R}$  for every  $j \geq 1$ , and that there exists  $j_0 \geq 1$  such that for all  $j \geq j_0$ , these intervals contain at most one limit point of  $\{S_{1,n}\}_{n=1}^{\infty}$ .

Let  $\mathcal{X} = \{1, 2, \dots\}$  and let  $\mathcal{B}(\mathcal{X})$  denote the Borel  $\sigma$ -field on  $\mathcal{X}$  (assuming every singleton of  $\mathcal{X}$  is an open set). Let  $\mathcal{P}$  denote the set of probability measures on  $\mathcal{X}$ . Then, at the first stage,

$$\pi(Y_1|P_1) \equiv P_1, \quad (4.3.2)$$

where  $P_1 \in \mathcal{P}$ . We assume that  $P_1$  is the following Dirichlet process:

$$\pi(P_1) \equiv DP(G), \quad (4.3.3)$$

where, the probability measure  $G$  is such that, for every  $j \geq 1$ ,

$$G(Y_j = m) = \frac{1}{2^m}. \quad (4.3.4)$$

Then

$$\pi(P_1|y_1) \equiv DP(G + \delta_{y_1}),$$

where, for any  $x$ ,  $\delta_x$  denotes point mass at  $x$ .

At the second stage, we set the prior for  $P_2$  to be the posterior of  $y_1$  corresponding to  $DP\left(\left(1 + \frac{1}{2^2}\right)G\right)$  prior for  $P_1$ . That is,  $\pi(P_2) \equiv DP\left(\left(1 + \frac{1}{2^2}\right)G + \delta_{y_1}\right)$ . Then, with respect to this prior for  $P_2$ , the posterior of  $P_2$  is given by

$$\pi(P_2|y_2) \equiv DP\left(\left(\left(1 + \frac{1}{2^2}\right)G + \delta_{y_1} + \delta_{y_2}\right)\right).$$

Continuing this recursive process as obtain that, at the  $k$ -th stage, the posterior of  $P_k$

is a Dirichlet process, given by

$$\pi(P_k|y_k) \sim DP\left(\sum_{j=1}^k \frac{1}{j^2} G + \sum_{j=1}^k \delta_{y_j}\right). \quad (4.3.5)$$

It follows from (4.3.5) that

$$E(p_{mk}|y_k) = \frac{\frac{1}{2^m} \sum_{j=1}^k \frac{1}{j^2} + \sum_{j=1}^k I(y_j = m)}{\sum_{j=1}^k \frac{1}{j^2} + k}; \quad (4.3.6)$$

$$\text{Var}(p_{mk}|y_k) = \frac{\left(\sum_{j=1}^k \frac{1}{j^2} + \sum_{j=1}^k I(y_j = m)\right) \left(\left(1 - \frac{1}{2^m}\right) \sum_{j=1}^k \frac{1}{j^2} + k - \sum_{j=1}^k I(y_j = m)\right)}{\left(\sum_{j=1}^k \frac{1}{j^2} + k\right)^2 \left(\sum_{j=1}^k \frac{1}{j^2} + k + 1\right)}. \quad (4.3.7)$$

As before, it easily follows from (4.3.6) and (4.3.7) that for  $m = 1, 2, 3, \dots$ ,

$$E(p_{m,k}|y_k) \rightarrow p_{m,0}, \quad \text{and} \quad (4.3.8)$$

$$\text{Var}(p_{m,k}|y_k) = O\left(\frac{1}{k}\right) \rightarrow 0, \quad (4.3.9)$$

almost surely, as  $k \rightarrow \infty$ .

The theorem below characterizes countable number of limit points of  $S_{1,\infty}$  in terms of the limit of the marginal posterior probabilities of  $p_{m,k}$ , as  $k \rightarrow \infty$ .

**Theorem 13** *For  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ , where  $\mathfrak{N}$  has zero probability measure,  $\{S_{1,n}(\omega)\}_{n=1}^{\infty}$  has countable limit points almost surely if and only if*

- (1) *There exist sequences  $\{c_{m,j}(\omega)\}_{j=1}^{\infty}$ ;  $m = 0, 1, 2, \dots$ , such that  $(c_{m-1,j}(\omega), c_{m,j}(\omega)]$  partition the real line  $\mathbb{R}$  for every  $j \geq 1$  and  $m \geq 1$ .*
- (2) *There exists  $j_0(\omega) \geq 1$  such that for all  $j \geq j_0(\omega)$ ,  $(c_{m-1,j}(\omega), c_{m,j}(\omega)]$  contains at most one limit point of  $\{S_{1,n}(\omega)\}_{n=1}^{\infty}$ , for every  $m \geq 1$ .*

(3) With  $Y_j$  defined as in (4.3.1),

$$\pi_m(\mathcal{N}_{p_{m,0}}|y_k(\omega)) \rightarrow 1, \quad (4.3.10)$$

as  $k \rightarrow \infty$ . In the above,  $\mathcal{N}_{p_{m,0}}$  is any neighborhood of  $p_{m,0}$ , with  $p_{m,0}$  satisfying  $0 \leq p_{m,0} \leq 1$  for  $m = 1, 2, \dots$  such that  $\sum_{m=1}^{\infty} p_{m,0} = 1$ , with at most finite number of  $m$  such that  $p_{m,0} = 0$ .

**Proof.** Follows using the same ideas as the proof of Theorem 12. ■

As regards the choice of the quantities  $c_{m,j}$ , we simply extend the construction detailed in Section 4.2.1 by only letting  $M \rightarrow \infty$ , and with obvious replacement of the posterior means with those associated with the posterior Dirichlet process.

It is useful to remark that our theory with countably infinite number of limit points is readily applicable to situations where the number of limit points is finite but unknown. In such cases, only a finite number of the probabilities  $\{p_{m,j}; m = 1, 2, 3 \dots\}$  will have posterior probabilities around positive quantities, while the rest will concentrate around zero. For known finite number of limit points, it is only required to specify  $G$  such that it gives positive mass to only a specific finite set.

## 4.4 Bayesian characterization of convergence and divergence with our approach on limit points

Note that for convergent series,  $\pi_m(\mathcal{N}_1|y_k(\omega)) \rightarrow 1$  as  $k \rightarrow \infty$  for smaller values of  $m$ , while for divergent series with  $S_{1,\infty}(\omega) = \infty$  or  $S_{1,\infty}(\omega) = -\infty$ ,  $\pi_m(\mathcal{N}_1|y_k(\omega)) \rightarrow 1$  as  $k \rightarrow \infty$  for much larger values of  $m$  and the smallest value of  $m$ , respectively. We formalize these statements below as the following theorems.

**Theorem 14** *Let there be  $M$  number of possible limit points of  $S_{1,\infty}(\omega)$ , where  $M$  may be infinite. Then for any  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ , where  $\mathfrak{N}$  has zero probability measure,*

$S_{1,\infty}(\omega) = \infty$  if and only if, for any sequences  $\{c_{m,j}(\omega)\}_{j=1}^{\infty}$ ;  $m = 1, 2, \dots, M$ , such that  $(c_{m-1,j}(\omega), c_{m,j}(\omega))$ ;  $m = 1, \dots, M$ , partitions the real line  $\mathbb{R}$  for every  $j \geq 1$ , it holds that

$$\pi_{m,k}(\mathcal{N}_1|y_k(\omega)) \rightarrow 1, \quad (4.4.1)$$

as  $k \rightarrow \infty$  and  $m \rightarrow M$ .

**Proof.** For  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ , where  $\mathfrak{N}$  has zero probability measure, let  $S_{1,\infty}(\omega) = \infty$ . Then as  $k \rightarrow \infty$ ,

$$\left( \frac{\exp(S_{1,k}(\omega))}{1 + \exp(S_{1,k}(\omega))} \right)^{\rho(\theta(\omega))} \rightarrow 1. \quad (4.4.2)$$

In other words, for any fixed  $M (> 1)$ ,  $y_k(\omega) \rightarrow M$ , as  $k \rightarrow \infty$ . Hence, as  $k \rightarrow \infty$  and  $m \rightarrow M$ , it easily follows using the same techniques as before, that (4.4.1) holds. Consequently, for infinite number of limit points, (4.4.1) holds as  $m \rightarrow \infty$ .

Now assume that (4.4.1) holds. It then follows from the formula of the posterior mean that  $y_k(\omega) \rightarrow M$ , as  $k \rightarrow \infty$ , for fixed  $M$ . Hence, (4.4.2) holds, from which it follows that  $S_{1,\infty}(\omega) = \infty$ . ■

**Theorem 15** *Let there be  $M$  number of possible limit points of  $S_{1,\infty}(\omega)$ , where  $M$  may be infinite. Then for any  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ , where  $\mathfrak{N}$  has zero probability measure,  $S_{1,\infty}(\omega) = -\infty$  almost surely if and only if for any sequences  $\{c_{m,j}(\omega)\}_{j=1}^{\infty}$ ;  $m = 1, 2, \dots, M$ , such that  $(c_{m-1,j}(\omega), c_{m,j}(\omega))$ ;  $m = 1, \dots, M$ , partitions the real line  $\mathbb{R}$  for every  $j \geq 1$ , it holds that*

$$\pi_{m,k}(\mathcal{N}_1|y_k(\omega)) \rightarrow 1, \quad (4.4.3)$$

as  $k \rightarrow \infty$  and  $m \rightarrow 1$ .

**Proof.** For  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ , where  $\mathfrak{N}$  has zero probability measure, let  $S_{1,\infty}(\omega) = -\infty$ . Then as  $k \rightarrow \infty$ ,

$$\left( \frac{\exp(S_{1,k}(\omega))}{1 + \exp(S_{1,k}(\omega))} \right)^{\rho(\theta(\omega))} \rightarrow 0. \quad (4.4.4)$$

In other words, for any fixed  $M (> 1)$ ,  $y_k(\omega) \rightarrow 1$ , as  $k \rightarrow \infty$ . Hence, as  $k \rightarrow \infty$  and  $m \rightarrow 1$ , it is easily seen that (4.4.3) holds.

Also, if (4.4.3) holds, then it follows from the formula of the posterior mean that  $y_k(\omega) \rightarrow 1$ , as  $k \rightarrow \infty$ . Hence, (4.4.4) holds, from which it follows that  $S_{1,\infty}(\omega) = -\infty$ .

■

**Theorem 16** For  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ , where  $\mathfrak{N}$  has zero probability measure,  $S_{1,\infty}(\omega)$  is convergent if and only if for any sequences  $\{c_{m,j}(\omega)\}_{j=1}^{\infty}$ ;  $m = 1, 2, \dots, M$ , such that  $(c_{m-1,j}(\omega), c_{m,j}(\omega))$ ;  $m = 1, \dots, M$ , partitions the real line  $\mathbb{R}$  for every  $j \geq 1$ , it holds for some finite  $m_0(\omega) \geq 1$ , that

$$\pi_{m_0(\omega),k}(\mathcal{N}_1|y_k(\omega)) \rightarrow 1, \quad (4.4.5)$$

as  $k \rightarrow \infty$ .

**Proof.** Let  $S_{1,\infty}(\omega)$  be convergent. Then as  $k \rightarrow \infty$ ,

$$\left( \frac{\exp(S_{1,k}(\omega))}{1 + \exp(S_{1,k}(\omega))} \right)^{\rho(\theta(\omega))} \rightarrow c(\omega), \quad (4.4.6)$$

for some constant  $0 \leq c(\omega) < 1$ . Hence, there exists some finite  $m_0(\omega) \geq 1$  such that  $y_k(\omega) \rightarrow m_0(\omega)$ , as  $k \rightarrow \infty$ . Using the same techniques as before, it is seen that that (4.4.5) holds.

Now assume that (4.4.5) holds. It then follows from the formula of the posterior mean, that  $y_k(\omega) \rightarrow m_0(\omega)$ , as  $k \rightarrow \infty$ . Hence, (4.4.6) holds, from which it follows that  $S_{1,\infty}(\omega)$  is convergent. ■

According to Theorems 15 and 16,  $m$  tends to 1 and a finite quantity greater than or equal to 1, accordingly as the series diverges to  $-\infty$  or converges. If the finite quantity in the latter case turns out to be 1, then it is not possible to distinguish between convergence and divergence to  $-\infty$  by this method. However, Theorem 5 can be usefully exploited

in this case. If this method based on oscillating series yields  $m = 1$ , then we suggest checking for convergence using Theorem 4.1, which would then help us confirm if the series is truly convergent.

## 4.5 A rule of thumb for diagnosis of convergence, divergence and oscillations

Based on the above theorems we propose the following rule of thumb for detecting convergence and divergence when  $M$  is finite: if  $\frac{m}{M} > 0.9$  such that  $\pi_{m,k}(\mathcal{N}_1|y_k(\omega)) \rightarrow 1$  as  $k \rightarrow \infty$ , then declare the series as divergent to  $\infty$ . If  $0.1 < \frac{m}{M} \leq 0.9$  such that  $\pi_{m,k}(\mathcal{N}_1|y_k(\omega)) \rightarrow 1$ , then declare the series as convergent. On the other hand, if  $\frac{m}{M} \leq 0.1$ , use Theorem 4.1 to check for convergence; in the case of negative result, declare the series as divergent to  $-\infty$ .

If, instead, there exist  $m_\ell$ ;  $\ell = 1, \dots, L$  ( $L > 1$ ) such that  $\pi_{m_\ell,k}(\mathcal{N}_{p_{m_\ell,0}}|y_k(\omega)) \rightarrow 1$  as  $k \rightarrow \infty$ , where  $0 < p_{m_\ell,0} < 1$  for  $\ell = 1, \dots, L$  and  $\sum_{\ell=1}^L p_{m_\ell,0} = 1$ , then say that the sequence  $\{S_{1,m}(\omega)\}_{n=1}^\infty$  has  $L$  limit points.

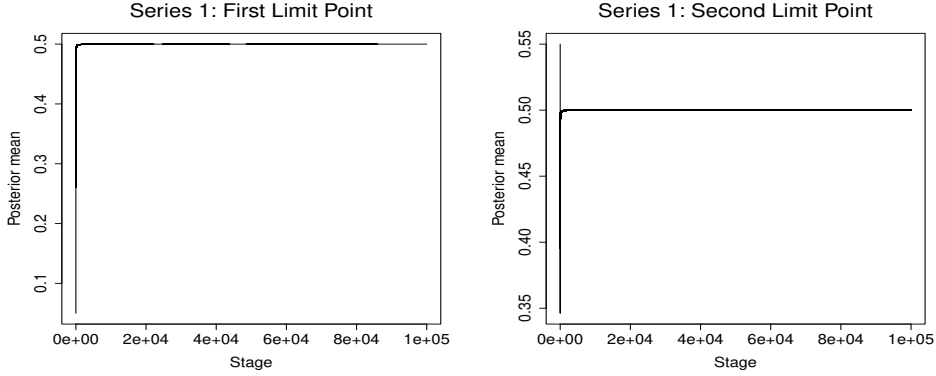
## 4.6 Illustration of our Bayesian theory on oscillation

We first consider a simple oscillatory series to illustrate our Bayesian idea on detection of limit points (Section 4.6.1). Next, in Section 4.6.2, we illustrate our theory on limit points with Example 5, arguably the most complex series in our set of examples (other than Riemann Hypothesis) and in Section 4.7, validate our result on Riemann Hypothesis with our Bayesian limit point theory.

### 4.6.1 Illustration with a simple oscillatory series

Let us re-consider the series  $S_{1,\infty}(\omega) = \sum_{i=1}^\infty (-1)^{i-1}$ , which we already introduced after Theorem 6. We consider the theory based on Dirichlet process developed in Section





(a) First limit point: The posterior of  $p_{5,k}$  converges to 0.5 as  $k \rightarrow \infty$

(b) Second limit point: The posterior of  $p_{6,k}$  converges to 0.5 as  $k \rightarrow \infty$ .

**Figure 4.6.1:** Illustration of the Dirichlet process based theory on the first oscillating series: two limit points, each with proportion 0.5, are captured.

4.3, assuming for the sake of illustrations that  $G$  is concentrated on  $M$  values, with  $G(Y_j = m) = \frac{1}{M}$ ;  $m = 1, 2, \dots, M$ . We set  $M = 10$  and  $K = 10^5$  for our experiments. With  $\rho(\theta) = 2$ , the results are depicted in Figure 4.6.1. Two explicit limit points, with proportions 0.5 each, are correctly recognized. The limit points are obviously 0 and 1 for this example. Implementation takes just a fraction of a second, even on an ordinary 32-bit laptop.

#### 4.6.2 Illustration of the Bayesian limit point theory with Example 5

Since there is at most one limit point in the cases that we investigated, application of our ideas to these cases must be able to re-confirm this. As before we consider the theory based on Dirichlet process with  $G(Y_j = m) = \frac{1}{M}$ ;  $m = 1, 2, \dots, M$ , where we set  $M = 10$ . Thus, by our rule of thumb, divergence is to be declared only if  $\pi_{m=10,k}(\mathcal{N}_1|y_k) \rightarrow 1$ , as  $k \rightarrow \infty$ .

As regards implementation, notice that here there is no scope for parallelization since at the  $j$ -th step only  $y_j$  is added to the existing  $S_{1,j-1}$  to form  $S_{1,j} = S_{1,j-1} + y_j$ . As

such, on our VMware, using a single processor, only about two seconds are required for  $10^5$  iterations associated with the series (3.5.13), for various values of  $a$  ( $> 0$ ) and  $b$  ( $> 0$ ).

**Choice of  $\rho(\theta)$  in  $\left(\frac{\exp(S_{1,k})}{1+\exp(S_{1,k})}\right)^{\rho(\theta)}$**

In our example,  $\theta = (a, b)$ . We choose, for  $j \geq 1$ ,

$$\tilde{\rho}(\theta) = a - b + \epsilon, \quad (4.6.1)$$

and set

$$\left(\frac{\exp(S_{1,j})}{1+\exp(S_{1,j})}\right)^{\rho(\theta)} = \min \left\{ 1, \left(\frac{\exp(S_{1,j})}{1+\exp(S_{1,j})}\right)^{\tilde{\rho}(\theta)} \right\} \quad (4.6.2)$$

Recall that the series (3.5.13), defined for  $a > 0$  and  $b > 0$ , converges for  $a - b > 1$  and diverges for  $a + b < 1$ . In keeping with this result, (4.6.2) decreases as  $(a - b)$  increases, so that the chance of correctly diagnosing convergence increases. Moreover, if both  $a$  and  $b$  are between 0 and 1 such that  $a + b < 1$ , then (4.6.2) tends to be inflated, thereby increasing the chance of correctly detecting divergence. The term  $\epsilon$  in (4.6.2) prevents the power from becoming zero when  $a = b$ . It is important to note here that for  $a + b = 1$  convergence or divergence is not guaranteed, but if  $\epsilon = 0$  in (4.6.2), then  $a = b$  would trivially indicate divergence, even if the series is actually convergent. A positive value of  $\epsilon$  provides protection from such erroneous decision. Note that if  $a < b - \epsilon$ , the convergence criterion  $a - b > 1$  is not met but the divergence criterion  $a + b < 1$  may still be satisfied. Thus, for such instances, greater weight in favour of divergence is indicated. In our illustration, we set  $\epsilon = 10^{-10}$ .

## Results

Figure 4.6.2 shows the results of our Bayesian analysis of the series (3.5.13) based on our Dirichlet process model. Based on the rule of thumb proposed in Section 4.5 all the

results are in agreement with the results based on Figure 3.5.6.

## 4.7 Application of the Bayesian multiple limit points theory to Riemann Hypothesis

To strengthen our result on Riemann Hypothesis presented in Section 3.6 we consider application of our Bayesian multiple limit points theory to Riemann Hypothesis.

### 4.7.1 Choice of $\rho(\theta)$ in $\left(\frac{\exp(S_{1,k})}{1+\exp(S_{1,k})}\right)^{\rho(\theta)}$

For Riemann Hypothesis,  $\theta = a$ ; we choose, for  $j \geq 1$ ,

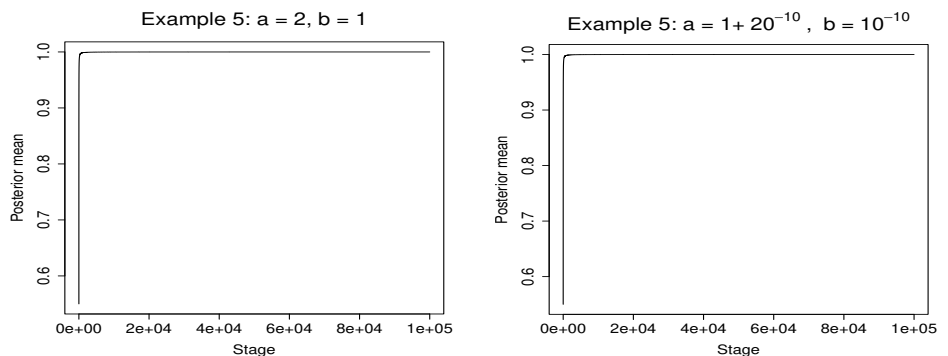
$$\tilde{\rho}(\theta) = a^6. \tag{4.7.1}$$

The reason for such choice with a relatively large power is to allow discrimination between  $\left(\frac{\exp(S_{1,k})}{1+\exp(S_{1,k})}\right)^{\rho(\theta)}$  for close values of  $a$ . However, substantially large powers of  $a$  are not appropriate because that would make the aforementioned term too small to enable detection of divergence. In fact, we have chosen the power after much experimentation. Implementation of our methods takes about 2 seconds on our VMWare, with  $10^5$  iterations.

### 4.7.2 Results

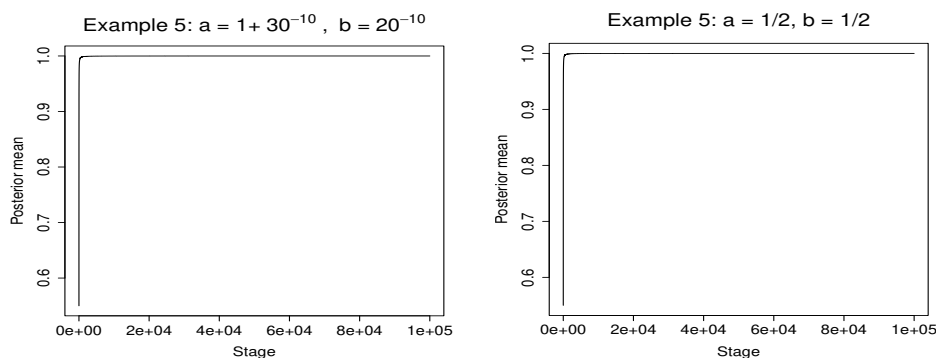
The results of application of our ideas on multiple limit points are depicted in Figures 4.7.1, 4.7.2 and 4.7.3. The values of  $m/M$  and the thumb rule proposed in Section 4.5 show that all the results are consistent with those obtained in Section 3.6. For  $a = 2$  and  $a = 3$  we obtained  $m/M = 0.1$ , but the existing theory and our results reported in Section 3.6 confirm that the series is convergent, and not oscillating, for these values. There seems to be a slight discrepancy only regarding the location of the change point

4.7. APPLICATION OF THE BAYESIAN MULTIPLE LIMIT POINTS THEORY TO  
73 RIEMANN HYPOTHESIS



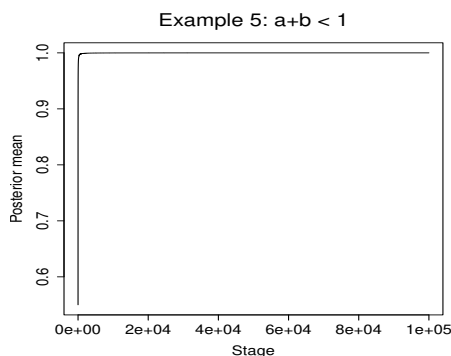
(a) Convergence:  $a = 2, b = 1$ . The posterior of  $p_{6,k}$  converges to 1 as  $k \rightarrow \infty$

(b) Convergence:  $a = 1 + 20^{-10}, b = 10^{-10}$ . The posterior of  $p_{6,k}$  converges to 1 as  $k \rightarrow \infty$ .



(c) Convergence:  $a = 1 + 30^{-10}, b = 20^{-10}$ . The posterior of  $p_{6,k}$  converges to 1 as  $k \rightarrow \infty$ .

(d) Divergence:  $a = 1/2, b = 1/2$ . The posterior of  $p_{10,k}$  converges to 1 as  $k \rightarrow \infty$ .



(e) Divergence:  $a = \frac{1}{2}(1 - 10^{-11}), b = \frac{1}{2}(1 - 10^{-11})$ . The posterior of  $p_{10,k}$  converges to 1 as  $k \rightarrow \infty$ .

**Figure 4.6.2:** Illustration of the Dirichlet process based theory with Example 5: For  $(a = 2, b = 1)$  in the series (3.5.13),  $\frac{m}{M} = \frac{6}{10} < 0.9$ , indicating convergence, for  $(a = 1 + 20^{-10}, b = 10^{-10})$ ,  $\frac{m}{M} = \frac{6}{10} < 0.9$ , indicating convergence, for  $(a = 1 + 30^{-10}, b = 20^{-10})$ ,  $\frac{m}{M} = \frac{6}{10} < 0.9$ , indicating convergence, for  $(a = 1/2, b = 1/2)$ ,  $\frac{m}{M} = \frac{10}{10} > 0.9$ , indicating divergence, and for  $(a = \frac{1}{2}(1 - 10^{-11}), b = \frac{1}{2}(1 - 10^{-11}))$ ,  $\frac{m}{M} = \frac{10}{10} > 0.9$ , indicating divergence.

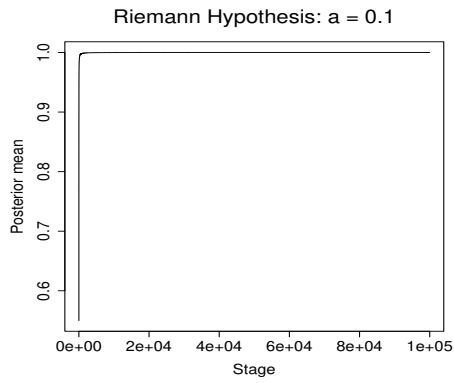
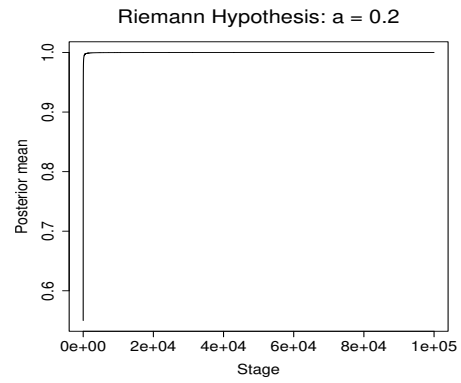
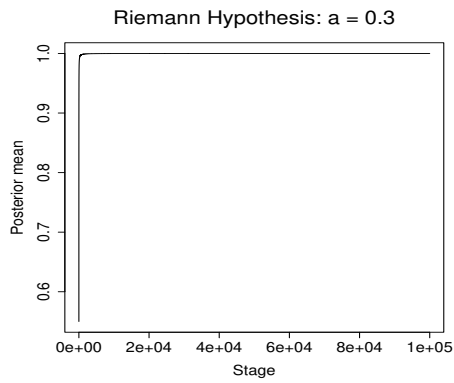
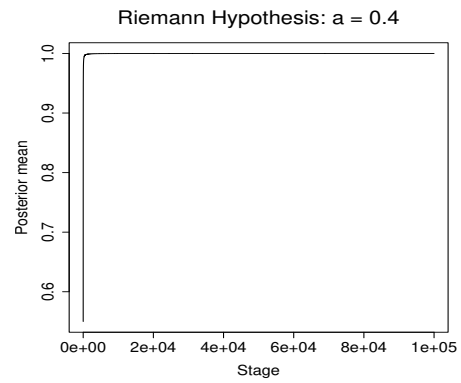
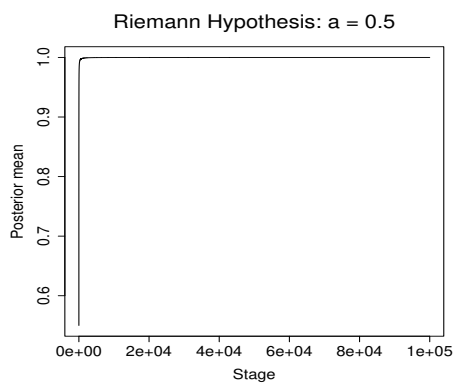
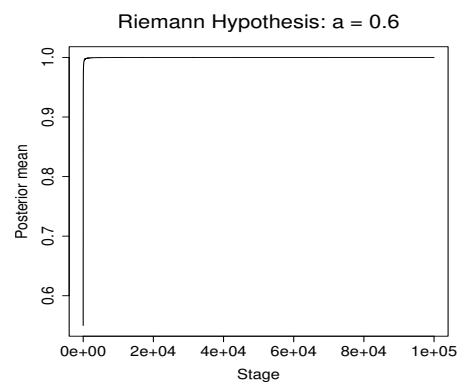
of convergence. In this case, unlike  $a = 0.72$  as obtained in Section 3.6, we obtained  $a = 0.7$  as the change point (see panel (b) of Figure 4.7.2).

This (perhaps) negligible difference notwithstanding, both of our methods are remarkably in agreement with each other, emphasizing our point that Riemann Hypothesis can not be completely supported.

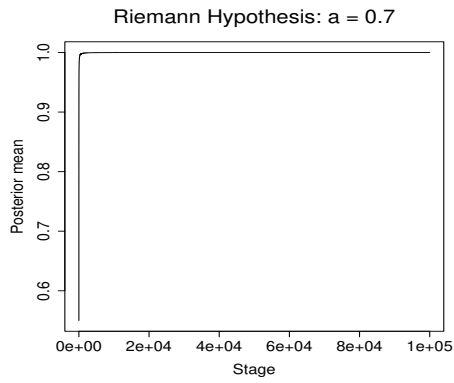
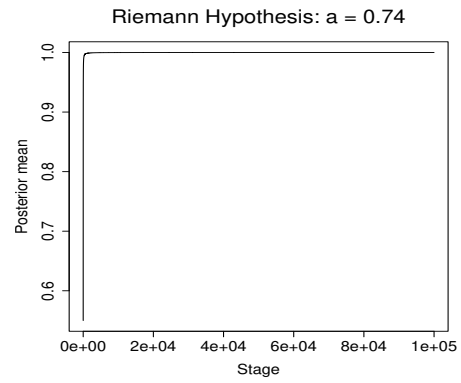
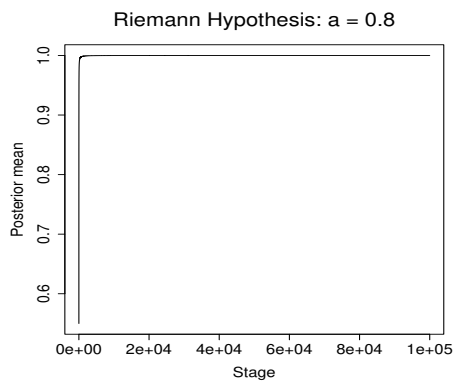
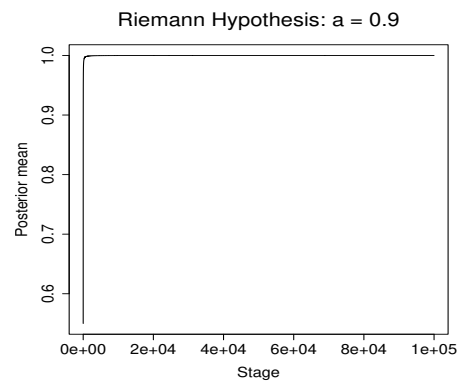
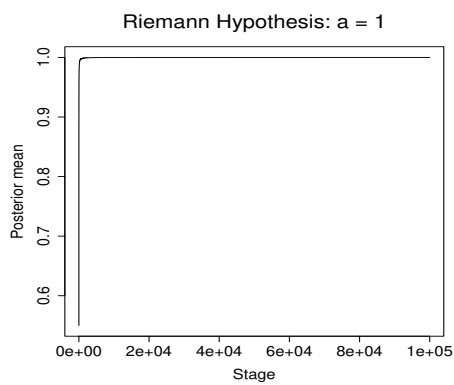
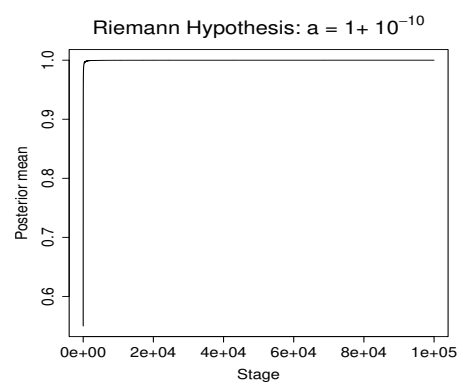
## 4.8 Summary and conclusion

In this chapter, we generalized the Bayesian characterization theory of convergence and divergence of infinite series to Bayesian characterization of multiple limit points in the case of oscillating series, where the number of limit points is allowed to be even countably infinite. The generalization is achieved by extending the Bernoulli-Beta setup of Chapter 3 to Multinomial-Dirichlet and infinite-dimensional Multinomial-Dirichlet process. In the generalization procedure, blocks of partial sums are no longer considered as in Chapter 3; rather, the terms are considered individually in the iterative procedure. This also precludes the idea of parallelization of Chapter 3, but as we demonstrated, does not compromise with computational speed and efficiency. We are also able to characterize convergence and divergence of non-oscillating series using the concepts for oscillating series characterization.

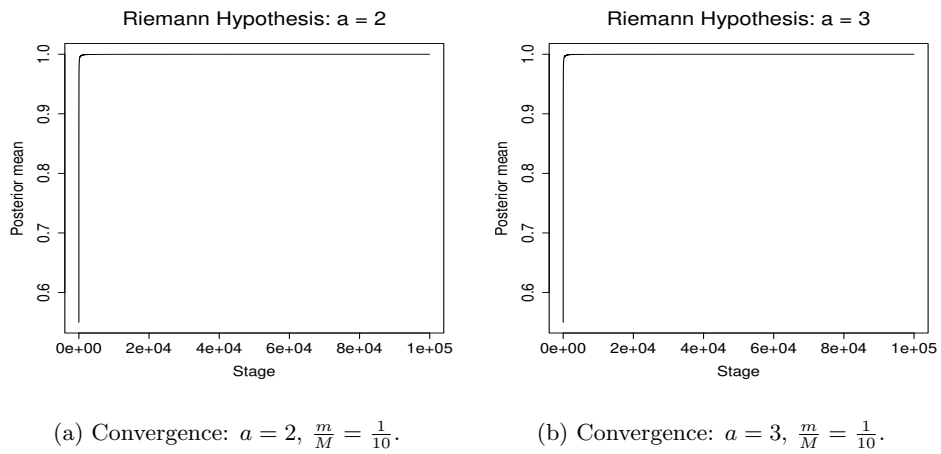
Applications of our developments to oscillating and non-oscillating series vindicate the validity and usefulness of the methods developed in this chapter. Most importantly, our application of the theory and methods developed in this chapter to Riemann Hypothesis again brought out results that are not in support of the most celebrated mathematical conjecture. Thus, the results on Riemann Hypothesis obtained in this chapter are in complete agreement with the results of the theories and methods developed in Chapter 3. Indeed, both the methods agree that there exists some  $a^*$  in the neighborhood of 0.7 such that the infinite series based on the Möbius function diverges for  $a < a^*$  and converges for  $a \geq a^*$ .

(a) Divergence:  $a = 0.1$ ,  $\frac{m}{M} = \frac{10}{10}$ .(b) Divergence:  $a = 0.2$ ,  $\frac{m}{M} = \frac{10}{10}$ .(c) Divergence:  $a = 0.3$ ,  $\frac{m}{M} = \frac{10}{10}$ .(d) Divergence:  $a = 0.4$ ,  $\frac{m}{M} = \frac{10}{10}$ .(e) Divergence:  $a = 0.5$ ,  $\frac{m}{M} = \frac{10}{10}$ .(f) Divergence:  $a = 0.6$ ,  $\frac{m}{M} = \frac{10}{10}$ .

**Figure 4.7.1:** Riemann Hypothesis based on Bayesian multiple limit points theory: Divergence for  $a = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6$ .

(a) Convergence:  $a = 0.7$ ,  $\frac{m}{M} = \frac{9}{10}$ .(b) Convergence:  $a = 0.74$ ,  $\frac{m}{M} = \frac{9}{10}$ .(c) Convergence:  $a = 0.8$ ,  $\frac{m}{M} = \frac{8}{10}$ .(d) Convergence:  $a = 0.9$ ,  $\frac{m}{M} = \frac{7}{10}$ .(e) Convergence:  $a = 1.0$ ,  $\frac{m}{M} = \frac{5}{10}$ .(f) Convergence:  $a = 1 + 10^{-10}$ ,  $\frac{m}{M} = \frac{5}{10}$ .

**Figure 4.7.2:** Riemann Hypothesis based on Bayesian multiple limit points theory: Divergence for  $a = 0.7$  but convergence for  $a = 0.74, 0.8, 0.9, 1, 1 + 10^{-10}$ .



**Figure 4.7.3:** Riemann Hypothesis based on Bayesian multiple limit points theory: Convergence for  $a = 2$ , 3.



# 5

## Bayesian Appraisal of Random Series Convergence with Application to Climate Change

### 5.1 Introduction

Convergence assessment of deterministic infinite series is a part of basic mathematical analysis and is included in the curriculum of almost all schools and colleges. Yet, for most infinite series there still does not exist any test of convergence that can provide conclusive answers, an issue that has concerned among many, the author of this thesis. Obtaining the knowledge that the Bayesian paradigm is a powerful premise for solving problems even of uncanny nature, we began investigation of a Bayesian solution to the infinite series problem and was indeed able to come up with a novel Bayesian procedure

to address questions of series convergence, addressed in Chapters 3 and 4. Our key idea, presented in Chapter 3, is to embed the underlying infinite series, even if deterministic, in a random, stochastic process framework, and then to build a recursive Bayesian algorithm for inference regarding the probability of convergence. We proved that the Bayesian algorithm converges to 1 if and only if the underlying series converges and to 0 if and only if the series diverges. Oscillatory series with multiple limit points, including infinite number of limit points, are also treated under similar Bayesian recursive frameworks in Chapter 4, with proper Bayesian characterizations of their properties. Applications of the Bayesian methods to a variety of infinite series yielded very encouraging results, and answers were obtained even where all existing methods of convergence assessment failed.

Although convergence assessment of infinite series constitutes a part of elementary mathematical analysis, it also holds the key to the solution of the most notorious unsolved problem of mathematics, namely, the Riemann Hypothesis. Establishment of convergence of the Dirichlet series for the Möbius function, for the real part of a complex-valued parameter of the series exceeding  $1/2$ , would establish truth of Riemann Hypothesis. On the other hand, divergence of the series for even any particular value of the real part exceeding  $1/2$  would negate the famous conjecture. On careful application of their Bayesian method to the Dirichlet series, we found, to our utter surprise, that the truth of Riemann Hypothesis is not supported by our Bayesian procedures.

In this chapter, we shall concern ourselves with random series of the form  $\sum_{i=1}^{\infty} X_i$ , where  $X_i$  are random, not deterministic quantities as in the examples of Chapters 3 and 4. Now recall that our Bayesian procedures treat even the deterministic elements of the series as realizations of some stochastic process. Hence, when the elements of the infinite series are random themselves, then there is certainly no need for any new theory for studying random series convergence. But although no new general theory is required, there are important details to pay attention to. The main issue is that, in the case of deterministic infinite series, the functional forms of the series elements are known, which

we usefully exploited to construct bounds for the partial sums associated with the series. However, in the case of random series elements, the functional forms are unavailable. In fact, even the distributional forms of the series elements are not available in reality, and if they are assumed to be available for the sake of theoretical development, construction of bounds for the partial sums in general, is still highly non-trivial.

Our main contribution in this chapter is to create appropriate bounds for the partial sums in the context of random infinite series. We begin with creation of upper bounds in parametric setups, whose mathematical validity is ensured for summands with non-negative supports. Simulation experiments under several such setups corroborate much accuracy and efficiency of such upper bounds when employed in our Bayesian procedure. However, since these bounds are not generally applicable, we propose a flexible parametric upper bound structure, although its mathematical validity in general situations can not be guaranteed. Although the general bound works well in several setups with non-negatively supported summands, its performance in random series driven by hierarchical normal distributions has been very inefficient and less persuasive, in spite of correct indications of convergence and divergence. Furthermore, in the case of random Dirichlet series, the general parametric bound yields wrong answers in many cases. Hence, we propose a nonparametric upper bound for the partial sums. The bound does not require any distributional assumption or non-negativity and improves itself adaptively with the iterations of the recursive Bayesian procedure. Simulation experiments demonstrate that not only is this bound far more accurate and efficient than the general parametric bound, but is also very much comparable in performance with the mathematically valid parametric bounds in the relevant non-negative setups.

Now, investigation of general series convergence, either deterministic or random, may be mathematically or probabilistically extremely challenging and hence makes for commendable undertaking, but such efforts would be more fruitful if determination of series convergence properties can be related to solutions of scientific problems of much

broader interest and importance. In this regard, our efforts that culminated in Chapters 3 and 4 did not seem to go in vain, as our novel Bayesian procedure for general deterministic series convergence assessment led to surprisingly important insights regarding the most challenging but influential unsolved problem of mathematics, the Riemann Hypothesis. Random infinite series seems to be more abstruse compared to deterministic ones as it is not immediately clear if they can be related to scientific problems of broad importance. In this chapter, we attempt to relate investigation of convergence properties of random infinite series to important scientific questions on climate change. Specifically, we attempt to address if global warming will continue or if global temperature will stabilize in the future. We also attempt to learn if global temperature was stable in the past or if there were instances of long periods of global warming or cooling. Based on records of current global temperature data and palaeoclimate reconstruction data, we infer with our Bayesian recursive procedure in conjunction with the nonparametric bound for the partial sums that we propose, that climate dynamics is subject to temporary variations, and long-term global warming or cooling is unlikely in the past as well as in the future.

The rest of this chapter is structured as follows. In Section 5.2, we put in our efforts towards building parametric upper bounds for partial sums of random series and in Section 5.3 assess the performance of such parametric bound structure with simulation experiments. We propose the nonparametric bound structure in Section 5.4 and evaluate its performance with simulation studies in the same section. Using the proposed nonparametric bound structure we analyze past and future global climate change in Section 5.5. Finally, in Section 5.6 we summarize our contributions and provide relevant discussions.

## 5.2 Random infinite series and parametric upper bound for the partial sums

Let us assume that  $\{X_i(\omega)\}_{i=1}^{\infty}$ , for  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$  is a given sequence of random variables (not necessarily independent) such that the marginal distribution of  $X_i$  is  $f_{\theta_i}(\cdot)$ , and that we wish to learn if  $S_{1,\infty}(\omega) = \sum_{i=1}^{\infty} X_i(\omega)$  converges for  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ . In this regard, we assume that the form of the density  $f_{\theta_i}$  is known. We shall consider both known and unknown  $\theta_i$ .

In fact, for our Bayesian theory for characterizing infinite series, it is not strictly necessary to assume that the form of  $f_{\theta_i}$  is known. However, we need to be able to obtain appropriate  $c_j(\omega)$  such that  $|S_{j,n_j}(\omega)| \leq c_j(\omega)$  for  $j \geq j_0(\omega)$  whenever  $S_{1,\infty}(\omega) < \infty$ . Although  $c_j(\omega)$  has been referred to as a non-negative monotonically decreasing sequence in Chapter 3, it is sufficient for  $c_j(\omega)$  to be a non-negative sequence that converges to zero. All the results of Chapter 3, including Theorems 5 and 6 continue to hold with this more flexible condition. This extra flexibility is valuable in our random series context where  $c_j(\omega)$  are non-negative and converge to zero but can not be guaranteed to be monotonically decreasing.

In the case of deterministic series, the functional forms of the series elements are known. Embedding the series in question in a class of series most of whose convergence properties are related to the values of some (set of) parameter(s)  $a$ , in Chapter 3 we could obtain suitable  $c_j(\omega)$  for the series of interest by exploiting the convergence properties of the parameterized class of series. For the current random series scenario, availability of information regarding some suitable class of series in which we can embed our given random series of interest will be useful for our purpose. In this regard, assuming a known form of the density  $f_{\theta_i}$  will be useful for constructing parametric upper bounds for the partial sums. However, we shall also construct a general and effective nonparametric upper bound form that does not require any such information but improves itself

adaptively with the recursive Bayesian steps.

### 5.2.1 Construction of parametric upper bound for the partial sums

It will be convenient for our purpose to build the theory with unknown  $\theta_i$  and to view known  $\theta_i$  situations as special cases.

#### Unknown $\theta_i$

Let us begin with the assumption that  $\{\theta_i\}_{i=1}^{\infty}$  is a stochastic process (again, not necessarily independent) with marginal density  $g_{\psi_i}$  where the density form as well as  $\psi_i$  will be assumed to be known in this parametric bound construction setup.

For  $i \geq 1$ , let us introduce spaces for convergence and divergence, which we denote by  $\Psi_i^{(c)}$  and  $\Psi_i^{(d)}$ , respectively, such that  $\sum_{i=1}^{\infty} \varphi_i$  is convergent and divergent, respectively, for  $\varphi_i \in \Psi_i^{(c)}$  and  $\varphi_i \in \Psi_i^{(d)}$ , for  $i \geq 1$ . In the above infinite sum, we assume that  $\varphi_i$  varies only with respect to  $i$  and is constant with respect to all other possible parameters.

To illustrate, let for any  $\epsilon > 0$ ,  $\Psi_i^{(c)} = \{i^{-p} : p \in [1 + \epsilon, \infty)\}$  and  $\Psi_i^{(d)} = \{i^{-p} : p \in (-\infty, 1]\}$ , or  $\Psi_i^{(c)} = \{q^{-i} : q \in [1 + \epsilon, \infty)\}$  and  $\Psi_i^{(d)} = \{q^{-i} : q \in [0, 1]\}$ . Thus, a typical element of  $\Psi_i^{(c)} = \{i^{-p} : p \in [1 + \epsilon, \infty)\}$  is  $\varphi_i = i^{-p}$ , where  $p \in [1 + \epsilon, \infty)$ . Hence, if  $p \in [1 + \epsilon, \infty)$  is held fixed, then  $\varphi_i$  changes only with respect to  $i$ . Hence,  $\sum_{i=1}^{\infty} \varphi_i < \infty$  for any fixed  $p \in [1 + \epsilon, \infty)$ . On the other hand,  $\sum_{i=1}^{\infty} \varphi_i = \infty$  for  $\varphi_i = i^{-p} \in \Psi_i^{(d)} = \{i^{-p} : p \in (-\infty, 1]\}$ , with  $p$  held fixed.

However, the provision of allowing  $\varphi_i$  to vary only with respect to  $i \geq 1$ , will be restricted to infinite sums only, not elsewhere.

To proceed, we assume that  $E(|\theta_i|) = h_i(\psi_i)$ , where  $h_i : \Psi_i^{(c)} \cup \Psi_i^{(d)} \mapsto \mathbb{R}^+$  (where  $\mathbb{R}^+ = [0, \infty)$ ) is such that  $\sum_{i=1}^{\infty} h_i(\varphi_i) < \infty$  for  $\varphi_i \in \Psi_i^{(c)}$ ;  $i \geq 1$  and  $\sum_{i=1}^{\infty} h_i(\varphi_i) = \infty$  for  $\varphi_i \in \Psi_i^{(d)}$ ;  $i \geq 1$ .

For any  $\varphi_i \in \Psi_i^{(c)} \cup \Psi_i^{(d)}$ , let  $G_{\varphi_i}$  denote the cumulative distribution function (cdf) of  $g_{\varphi_i}$ . Now let, for each  $x \in \mathbb{R}$ ,  $G_i(x) = \inf_{\varphi_i \in \Psi_i^{(c)}} G_{\varphi_i}(x)$ . Assume that  $G_i(\cdot)$  is continuous

for  $i \geq 1$ . Then it follows that  $\lim_{x \rightarrow -\infty} G_i(x) = 0$ ,  $\lim_{x \rightarrow \infty} G_i(x) = 1$ . Also, if  $x_1 < x_2$ ,  $G_i(x_1) \leq G_{\psi_i}(x_1) \leq G_{\psi_i}(x_2)$  for all  $\psi_i \in \Psi_i^{(c)}$ , so that  $G_i(x_1) \leq G_i(x_2)$ , satisfying the monotonicity property. Hence,  $G_i(\cdot)$  is a continuous distribution function for  $i \geq 1$ . Let  $g_i$  denote the corresponding density function.

Let  $\tilde{\theta}_i \sim g_i$ . Then  $\sum_{i=1}^{\infty} E(|\tilde{\theta}_i|) < \infty$ . By Theorem 1 of Kawata (1972) (see also Pakes (2004)) it follows that the series  $\sum_{i=1}^{\infty} \tilde{\theta}_i$  is absolutely convergent almost surely, irrespective of any dependence structure among the  $\tilde{\theta}_i$ 's.

Hence, it follows that if  $G_i(\cdot)$  is continuous for  $i \geq 1$ , then it is a distribution function satisfying  $G_i(x) \leq G_{\psi_i}(x)$  for all  $x$  and  $\psi_i \in \Psi_i^{(c)}$ . Consequently, for any *fixed* random number  $U_i$ , where  $U_i \sim U(0, 1)$ , the uniform distribution on  $(0, 1)$  (this means that we first draw  $U_i \sim U(0, 1)$  and then fix this  $U_i$  to invert the distribution functions  $G_{\psi_i}$  and  $G_i$ , as below), it holds that for all  $\psi_i \in \Psi_i^{(c)}$ ,

$$G_{\psi_i}^{-}(U_i) \leq G_i^{-}(U_i), \quad (5.2.1)$$

where, for any distribution function  $G$ ,  $G^{-}(x) = \inf\{y : G(y) \geq x\}$ , is the inverse of  $G$ .

The inversions in (5.2.1) are nothing but simulations from the distributions corresponding to  $G_{\psi_i}$  and  $G_i$ , respectively. We thus set  $\theta_{\psi_i} = G_{\psi_i}^{-}(U_i)$  and  $\tilde{\theta}_i = G_i^{-}(U_i)$ .

Since inequality (5.2.1) holds for all  $\psi_i \in \Psi_i^{(c)}$ , this implies that for *fixed*  $U_i$ , whatever value of  $\theta_{\psi_i}$  is simulated using the relation  $\theta_{\psi_i} = G_{\psi_i}^{-}(U_i)$ , whatever may be the values of  $\psi_i \in \Psi_i^{(c)}$ , it must always hold that

$$\theta_{\psi_i} \leq \tilde{\theta}_i. \quad (5.2.2)$$

Now suppose that  $X_i$  are non-negative and admits the form  $X_i = F_{\theta_i}^{-}(U_i)$ , where  $F_{\theta_i}$  is the distribution function of  $X_i$  conditional on  $\theta_i$ , and assume that (5.2.2) ensures the inequality  $X_{\theta_{\psi_i}} = F_{\theta_{\psi_i}}^{-}(U_i) \leq F_{\tilde{\theta}_i}^{-}(U_i) = X_{\tilde{\theta}_i}$ . Then, setting  $X_{\theta_{\psi_i}} = X_i$  so that  $F_{\theta_{\psi_i}}^{-}(U_i) = X_i$ , would enable us to obtain  $U_i$  in terms of  $X_i$  and  $\theta_{\psi_i}$ , for given  $\theta_{\psi_i}$ . This

$U_i$  will then be used in  $F_{\tilde{\theta}_i}^-(U_i)$  to form  $X_{\tilde{\theta}_i} = F_{\tilde{\theta}_i}^-(U_i)$ , for given  $\tilde{\theta}_i$ . The partial sums associated with  $\{X_{\tilde{\theta}_i}\}_{i=1}^\infty$  will then constitute valid upper bounds for the partial sums corresponding to the underlying random series summands  $\{X_i\}_{i=1}^\infty$ .

Note that the above assumption of non-negative support of  $X_i$  is crucial, since for general supports, upper bounds for the partial sums can not ensure that the absolute values of the partial sums are bounded above by the absolute values of the corresponding upper bounds.

All the above results and discussions continue to hold if  $X_i$  are discrete random variables with finite support. The proof that  $G_i$  are valid distribution functions in such cases is the same as that presented in Section S-1 of [Mukhopadhyay and Bhattacharya \(2012\)](#). Indeed, the principle of constructing upper bounds in the method described so far has some parallel in [Mukhopadhyay and Bhattacharya \(2012\)](#), although in a very different, perfect sampling context.

**Known  $\theta_i$**

Now, if  $\theta_i$  are known, then we can apply the same procedure to  $f_{\theta_i}$  instead of  $g_{\psi_i}$ . In that case, letting  $F_{\theta_i}$  denote the distribution function associated with  $f_{\theta_i}$  and  $F_i(x) =$

$\inf_{\varphi_i \in \Psi_i^{(c)}} F_{\varphi_i}(x)$  for  $x \in \mathbb{R}$ , we shall then have

$$X_i = F_{\theta_i}^-(U_i) \leq F_i^-(U_i) = \tilde{X}_i, \tag{5.2.3}$$

which ensures  $S_{j,n_j} \leq \tilde{S}_{j,n_j}$ , where  $\tilde{S}_{j,n_j}$  are the partial sums associated with  $\{\tilde{X}_i\}_{i=1}^\infty$ . This would enable us to set  $c_j = \tilde{S}_{j,n_j}$  as the upper bound for the partial sums of  $\{X_i\}_{i=1}^\infty$ . For known  $\theta_i$ , given  $X_i$ ,  $U_i$  is available from the first equality of (5.2.3), which can be used in the second equality of (5.2.3) to form  $\tilde{X}_i$ .



### 5.2.2 Upper bound for partial sums for hierarchical scale families on non-negative supports

To see the utility of (5.2.2), let us assume that the distribution of  $X_i$  given  $\theta_i$  is a scale family on  $[0, \infty)$ , that is,

$$f_{\theta_i}(x_i) = \frac{1}{\theta_i} f\left(\frac{x_i}{\theta_i}\right) \mathbb{I}_{\{x_i > 0\}}, \quad (5.2.4)$$

where  $\theta_i > 0$ , and  $f(\cdot)$  is a density function supported on  $[0, \infty)$ . Let us assume that  $\theta_i$  are random and have densities  $g_{\psi_i}$  with the same details as in Section 5.2.1. Since  $\theta_i$  are also random variables, the model pertains to a hierarchical scale family.

The distribution function corresponding to (5.2.4) is of the form  $F\left(\frac{x_i}{\theta_i}\right)$ , where  $F$  is the cdf corresponding to the density function  $f$ . Hence,  $X_i = \theta_i F^{-}(U_i)$ . Let  $X_{\theta_{\psi_i}} = \theta_{\psi_i} F^{-}(U_i)$  and  $X_{\tilde{\theta}_i} = \tilde{\theta}_i F^{-}(U_i)$ . Here  $U_i$  are *iid*  $U(0, 1)$  random variables assumed to be independent of the uniform random variables used to draw  $\theta_i$  and  $\theta_{\psi_i}$ . Since  $F^{-}(U_i) > 0$ , (5.2.2) ensures

$$X_{\theta_{\psi_i}} \leq X_{\tilde{\theta}_i}. \quad (5.2.5)$$

It follows from (5.2.5) that

$$S_{j, n_j}^{\theta_{\psi}} \leq S_{j, n_j}^{\tilde{\theta}}, \quad (5.2.6)$$

where  $S_{j, n_j}^{\theta_{\psi}}$  and  $S_{j, n_j}^{\tilde{\theta}}$  are the partial sums associated with the series  $\left\{X_{\theta_{\psi_i}}\right\}_{i=1}^{\infty}$  and  $\left\{X_{\tilde{\theta}_i}\right\}_{i=1}^{\infty}$ , respectively. The relation (5.2.6) enables us to set  $c_j = S_{j, n_j}^{\tilde{\theta}}$ . Note that since  $X_{\theta_{\psi_i}}$  and  $\theta_{\psi_i}$  are known in the relation  $X_{\theta_{\psi_i}} = \theta_{\psi_i} F^{-}(U_i)$ ,  $U_i$  can be obtained from this equality, and can be used to form  $X_{\tilde{\theta}_i} = \tilde{\theta}_i F^{-}(U_i)$ . In fact, for given  $X_i$  and  $\theta_{\psi_i}$ , we set  $X_{\theta_{\psi_i}} = \theta_{\psi_i} F^{-}(U_i) = X_i$ , and solve for  $U_i$  from the last equality, which we then use for construct  $X_{\tilde{\theta}_i} = \tilde{\theta}_i F^{-}(U_i)$ .

**Illustration with hierarchical exponential distribution**

Let  $f_{\theta_i}(x) = \frac{1}{\theta_i} \exp\left(-\frac{x}{\theta_i}\right)$ ;  $x > 0$ ,  $\theta_i > 0$ . Also, let  $g_{\psi_i}(\theta) = \frac{1}{\psi_i} \exp\left(-\frac{\theta}{\psi_i}\right)$ ;  $\theta > 0$ ,  $\psi_i > 0$ . Here  $G_{\psi_i}(\theta) = 1 - \exp\left(-\frac{\theta}{\psi_i}\right)$ . Let  $r_i(\epsilon) = \min\{i^{(1+\epsilon)}, (1+\epsilon)^i\}$ . Then  $G_{\tilde{\theta}_i}(\theta) = 1 - \exp(-\theta r_i(\epsilon))$ .

The upper bounds for the partial sums in this case can be constructed in the following manner. Note that here  $X_{\theta_{\psi_i}} = \theta_{\psi_i} F^-(U_i) = -\theta_{\psi_i} \log U_i$  and  $X_{\tilde{\theta}_i} = \tilde{\theta}_i F^-(U_i) = -\tilde{\theta}_i \log U_i$ , where, for  $i \geq 1$ ,  $U_i \stackrel{iid}{\sim} U(0, 1)$ . Also,  $\theta_{\psi_i} = -\psi_i \log U_i^*$  and  $\tilde{\theta}_i = -r_i^{-1}(\epsilon) \log U_i^*$ , where  $U_i^* \stackrel{iid}{\sim} U(0, 1)$  and are independent of  $U_i$ , for  $i \geq 1$ . For theoretically sound bound construction in practice, we shall first simulate  $\theta_{\psi_i}$  and  $\tilde{\theta}_i$  using the same  $U_i^*$ . Then, we shall obtain  $U_i$  from the equality  $X_{\theta_{\psi_i}} = -\theta_{\psi_i} \log U_i = X_i$ , which we shall use to construct  $X_{\tilde{\theta}_i} = -\tilde{\theta}_i \log U_i$ . These, in turn, lead to (5.2.5) and (5.2.6).

To obtain the relevant result regarding upper bounds for the partial sums we begin with the following theorem.

**Theorem 17** *Let  $\theta_i$  be independent. Then  $\sum_{i=1}^{\infty} \theta_i < \infty$  almost surely if and only if  $\sum_{i=1}^{\infty} \psi_i < \infty$ .*

**Proof.** By Kolmogorov's three series theorem (see, for example, [Resnick \(2014\)](#)), it is easy to see that  $\sum_{i=1}^{\infty} \psi_i < \infty$  implies  $\sum_{i=1}^{\infty} \theta_i < \infty$  almost surely. We now show that for any  $R > 0$ ,  $\sum_{i=1}^{\infty} E(\theta_i \mathbb{I}_{\{\theta_i < R\}}) = \infty$  if  $\sum_{i=1}^{\infty} \psi_i = \infty$ . This would then ensure, by Kolmogorov's three series theorem, that  $\sum_{i=1}^{\infty} \theta_i = \infty$  almost surely.

Note that

$$E(\theta_i \mathbb{I}_{\{\theta_i < R\}}) = \psi_i \times \left[ 1 - \exp\left(-\frac{R}{\psi_i}\right) \left(1 + \frac{R}{\psi_i}\right) \right]. \quad (5.2.7)$$

If  $\psi_i \in \Psi_i^{(d)} = \{i^{-p} : p \in (-\infty, 1]\}$ , then  $\psi_i = i^{-p}$  for some  $p \in (-\infty, 1]$ . Suppose first that  $p \in (0, 1]$ . In that case,

$$1 - \exp\left(-\frac{R}{\psi_i}\right) \left(1 + \frac{R}{\psi_i}\right) \rightarrow 1, \text{ as } i \rightarrow \infty. \quad (5.2.8)$$

It follows from (5.2.8) that for any  $\varepsilon > 0$ , there exist  $i_0 \geq 1$  such that for  $i \geq i_0$ , the right hand side of (5.2.7) exceeds  $\psi_i(1 - \varepsilon)$ . Since  $\sum_{i=i_0}^{\infty} \psi_i(1 - \varepsilon) = \infty$  for  $\psi_i = i^{-p}$  where  $p \in (0, 1]$ , it follows that

$$\sum_{i=1}^{\infty} E(\theta_i \mathbb{I}_{\{\theta_i < R\}}) = \infty \text{ for } \psi_i = i^{-p}, \text{ with } p \in (0, 1], \text{ for any } R > 0.$$

By Kolmogorov's three series theorem it then follows that  $\sum_{i=1}^{\infty} \theta_i = \infty$ , almost surely.

Now let us consider the case where  $\psi_i = i^{-p}$ , with  $p \leq 0$ . If  $p = 0$ , then  $\theta_i$  are *iid*, so that trivially,  $\sum_{i=1}^{\infty} \theta_i = \infty$ , almost surely. So, let  $p < 0$ . Direct calculation shows that

$$P(\theta_i > R) = \exp\left(-\frac{R}{\psi_i}\right) = \exp(-Ri^p) \rightarrow 1, \text{ as } i \rightarrow \infty.$$

Hence,  $\sum_{i=1}^{\infty} P(\theta_i > R) = \infty$ , for any  $R > 0$ , so that by Kolmogorov's three series theorem,  $\sum_{i=1}^{\infty} \theta_i = \infty$ , almost surely.

Finally, consider the case  $\psi_i = q^{-i}$ ,  $q \in [0, 1]$ . If  $q = 1$ , then  $\theta_i$  are *iid*, so that  $\sum_{i=1}^{\infty} \theta_i = \infty$ , almost surely. So, let  $q \in [0, 1)$ . Then

$$P(\theta_i > R) = \exp(-Rq^i) \rightarrow 1, \text{ as } i \rightarrow \infty,$$

which leads to  $\sum_{i=1}^{\infty} \theta_i = \infty$ , almost surely. ■

Theorem 17 shows that in the case of independence,  $\sum_{i=1}^{\infty} \theta_i < \infty$  if and only if  $\psi_i \in \Psi_i^{(c)}$ , for  $i \geq 1$ . In the case of dependence, it can only be guaranteed that  $\sum_{i=1}^{\infty} \theta_i < \infty$  if  $\psi_i \in \Psi_i^{(c)}$ , for  $i \geq 1$ . It can not be asserted that  $\sum_{i=1}^{\infty} \theta_i = \infty$  if  $\psi_i \in \Psi_i^{(d)}$ , for  $i \geq 1$ . The implication is that, if  $X_i$  are also conditionally independent given  $\theta_i$ ,  $S_{j,n_j}^{\tilde{\theta}}$  of the form (5.2.6) corresponds in the hierarchical exponential setup to the maximal convergent series closest to divergence in the case of independence, but this need not be the case when  $\theta_i$  and/or  $X_i$  given  $\theta_i$  are dependent. This leads to the following theorem as a consequence of Theorem 17.

**Theorem 18** For  $i \geq 1$ , let  $\tilde{\theta}_i \sim G_{\tilde{\theta}_i}$ , and  $X_i \sim f_{\tilde{\theta}_i}(x_i) = \frac{1}{\tilde{\theta}_i} \exp\left(-\frac{x_i}{\tilde{\theta}_i}\right)$ . Then the partial sums  $\tilde{S}_{j,n_j}$  of the form (5.2.6) in the hierarchical exponential setup correspond to the maximal convergent series  $\sum_{i=1}^{\infty} X_i$  that is the closest to divergence, provided  $\theta_i$  are independent and conditionally on  $\theta_i$ ,  $X_i$  are also independent.

### 5.2.3 Construction of bounds for the partial sums in the general case

In the general situation where either  $X_i$  given  $\theta_i$  and  $\theta_i$  are not independent and/or  $X_i$  is supported on the real line, it is not possible to mathematically establish that  $S_{j,n_j}^{\tilde{\theta}}$  corresponds to the maximal convergent series closest to divergence.

In the general case we propose to construct bounds with arbitrary sequence of  $U_i$ 's, in the following way. First note that if  $\sum_{i=1}^{\infty} X_i < \infty$ , then, letting  $S_{j,n_j}$  denote the partial sum associated with the above series,  $|S_{j,n_j}| \rightarrow 0$  as  $j \rightarrow \infty$ , irrespective of the choice of the  $U_i$ 's. Theoretically, we need not have  $|S_{j,n_j}| \leq |S_{j,n_j}^{\tilde{\theta}}|$  even in the case of convergence, but we can expect that

$$|S_{j,n_j}| \leq |S_{j,n_j}^{\tilde{\theta}}| + \frac{a}{j} \quad (5.2.9)$$

holds in the case of convergence, where  $a (> 0)$  is some suitable constant. The idea is to slightly inflate  $|S_{j,n_j}^{\tilde{\theta}}|$  so that (5.2.9) holds. We propose (5.2.9) as an upper bound for the partial sums in the general setup.

#### Illustration with normal distribution

Assume that  $X_i \sim N(\mu_i, \sigma_i^2)$ , independently for  $i \geq 1$ . Assume also that for  $i \geq 1$ , independently,  $\mu_i \sim N(0, \phi_i^2)$  and  $\sigma_i^2 \sim \mathcal{E}(\vartheta_i)$ , that is, the exponential distribution with mean  $\vartheta_i$ . Let  $\phi_i^2 \in \Psi_i^{(c)} \cup \Psi_i^{(d)}$  and  $\vartheta_i \in \Psi_i^{(c)} \cup \Psi_i^{(d)}$ .

It is well-known (see, for example, Exercise 7.7.14 of Resnick (2014)) that  $\sum_{i=1}^{\infty} X_i < \infty$  almost surely if and only if  $\sum_{i=1}^{\infty} \mu_i < \infty$  and  $\sum_{i=1}^{\infty} \sigma_i^2 < \infty$  almost surely. This result,

along with its two different proofs can be found in page 319 of [Driver \(2010\)](#). Here, letting  $\Psi_i^{(c)} = \{i^{-p} : p \in [1 + \epsilon, M_1]\}$  or  $\Psi_i^{(c)} = \{q^{-i} : q \in [1 + \epsilon, M_2]\}$ , where  $M_1 > 1 + \epsilon$ ,  $M_2 > 1 + \epsilon$ , and  $\tilde{r}_i = \max\{i^{M_1}, M_2^i\}$ , we have

$$G_{\tilde{\mu}_i}(\mu) = \begin{cases} \Phi\left(\mu\sqrt{r_i(\epsilon)}\right) & \text{if } \mu \geq 0; \\ 1 - \Phi\left(-\mu\sqrt{\tilde{r}_i}\right) & \text{if } \mu < 0, \end{cases} \quad (5.2.10)$$

and

$$G_{\tilde{\sigma}_i^2}(\sigma^2) = 1 - \exp\left(-\sigma^2 r_i(\epsilon)\right), \quad (5.2.11)$$

where  $r_i(\epsilon) = \min\{i^{(1+\epsilon)}, (1+\epsilon)^i\}$ . In this case, due to independence, (5.2.10) and (5.2.11) do correspond to maximal convergent series for  $\sum_{i=1}^{\infty} \mu_i$  and  $\sum_{i=1}^{\infty} \sigma_i^2$ , and it holds that  $\mu_i \leq \tilde{\mu}_i$  and  $\sigma_i^2 \leq \tilde{\sigma}_i^2$ , but since  $X_i$  is supported on the entire real line, these do not guarantee that even  $X_i \leq X_{\tilde{\theta}_i}$  holds, where  $\tilde{\theta}_i = (\tilde{\mu}_i, \tilde{\sigma}_i^2)$ . For further clarity, note that  $X_i = \mu_i + \sigma_i Z_i$ , where  $Z_i \stackrel{iid}{\sim} N(0, 1)$ , for  $i \geq 1$ . Even though it is possible to theoretically ensure  $\mu_i \leq \tilde{\mu}_i$  and  $\sigma_i^2 \leq \tilde{\sigma}_i^2$ ,  $Z_i$  takes values on the entire real line, and hence  $X_i \leq X_{\tilde{\theta}_i}$  can not be guaranteed. Moreover, it is not possible to simulate from  $G_{\tilde{\mu}_i}$  by inverting the distribution function. However, we can still expect (5.2.9) to hold, for appropriate choice of  $a$  ( $> 0$ ).

An important point to observe is that as  $i \rightarrow \infty$ ,  $1 - \Phi\left(-\mu\sqrt{\tilde{r}_i}\right) \rightarrow 0$ , so that under (5.2.10) the distribution of  $\tilde{\mu}_i$  supports only non-negative values, as  $i \rightarrow \infty$ . This results in too large an upper bound, which makes it hard to detect divergences. Replacing this distribution of  $\tilde{\mu}_i$  with  $\tilde{\mu}_i \sim N\left(0, \sigma_{\tilde{\mu}_i}^2\right)$ , with  $\sigma_{\tilde{\mu}_i}^2 = 1/r_i(\epsilon)$ , resulted in more useful bounds for the partial sums in our simulation examples.

Note that in the case of independence, study of convergence of  $S_{1,\infty}(\omega)$  for only one  $\omega \in \mathfrak{S}$  is necessary, since  $S_{1,\infty}(\omega)$  either converges for almost all  $\omega \in \mathfrak{S}$  or diverges for almost all  $\omega \in \mathfrak{S}$ . The rest of the theory remains the same as that of Chapter 3.

## 5.3 Simulation experiments with parametric upper bound

### 5.3.1 Example 1: Hierarchical exponential distribution

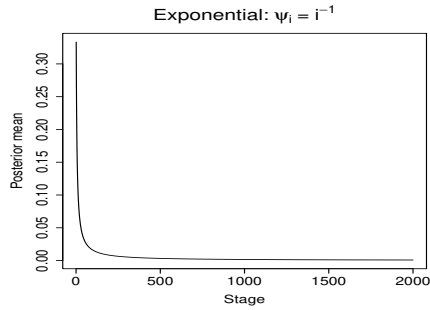
We first consider the setup  $X_i \sim \mathcal{E}(\theta_i)$  and  $\theta_i \sim \mathcal{E}(\psi_i)$ ;  $i \geq 1$ . Thus  $X_i$  has a two-stage hierarchical exponential distribution. Following the bound construction method detailed in Section 5.2.2, setting  $\epsilon = 0.001$  we considered the upper bound given by  $c_j = S_{j,n_j}^{\tilde{\theta}}$ , where  $n_j = 1000$ , for  $j = 1, \dots, K$ , with  $K = 2000$ .

We implement our recursive Bayesian procedure on an ordinary dual core laptop, splitting the sum of 1000 terms at each step of 2000 stages into the two processors using the Message Passing Interface (MPI) protocol in our C programming environment. In our implementation, the Bayesian recursive algorithm takes less than a second to yield result.

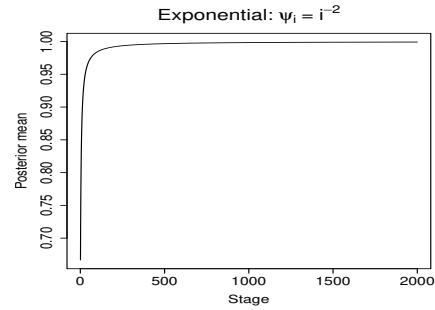
The results of our convergence analyses of this setup are depicted in Figure 5.3.1, which shows that the convergence behaviour of the random series are always correctly determined by our recursive Bayesian procedure with the aforementioned upper bound. That the method performs so well in spite of such small sample size, seems to very encouraging.

### 5.3.2 Example 2: Hierarchical normal distribution

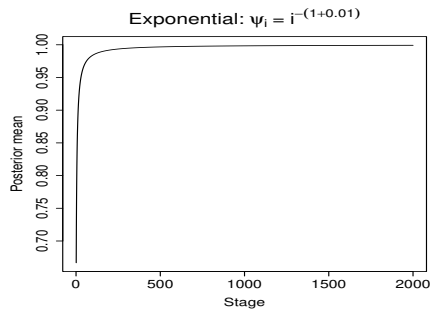
Now let  $X_i \sim N(\mu_i, \sigma_i^2)$ ,  $\mu_i \sim N(0, \phi_i^2)$  and  $\sigma_i^2 \sim \mathcal{E}(\vartheta_i)$ ;  $i \geq 1$ . This specifies a two-stage hierarchical normal distribution for  $X_i$ . For this setup, our results of convergence analyses are provided in Figure 5.3.2. Following the later discussion in Section 5.2.3 we construct  $S_{j,n_j}^{\tilde{\theta}}$  using  $\tilde{\mu}_i \sim N(0, \sigma_{\tilde{\mu}_i}^2)$ , with  $\sigma_{\tilde{\mu}_i}^2 = 1/r_i(\epsilon)$ . Consequently, setting  $\epsilon = 0.001$ , we consider the upper bound given by  $c_j = \left| S_{j,n_j}^{\tilde{\theta}} \right| + \frac{0.1}{j}$ , with  $n_j = 10^6$ ;  $j = 1, \dots, K$ , with  $K = 10^6$ . This many times longer run compared to the exponential simulation study setup detailed in Section 5.3.1 is required since mathematically valid parametric upper bound for the partial sums does not seem to be available in this case



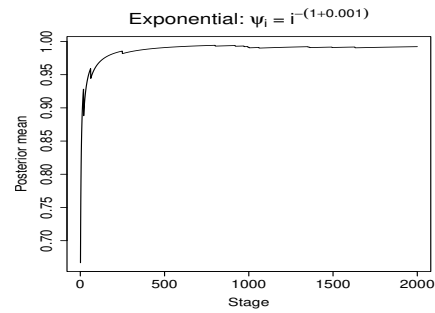
(a) Divergence.



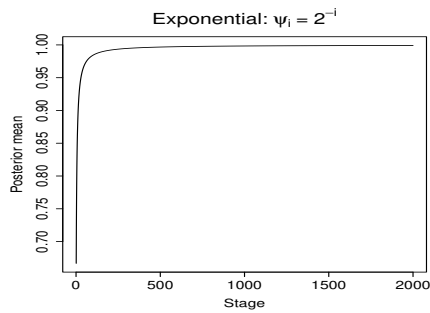
(b) Convergence.



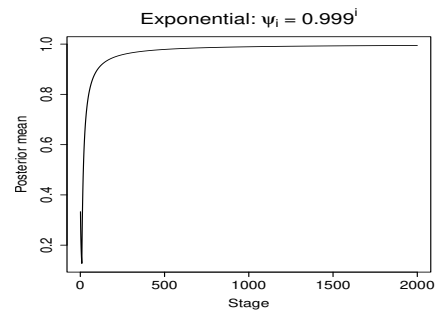
(c) Convergence.



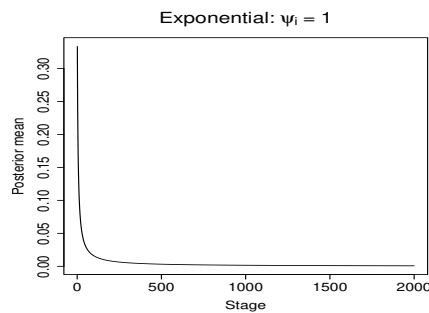
(d) Convergence.



(e) Convergence.



(f) Convergence.



(g) Divergence.

Figure 5.3.1: Example 1: Convergence and divergence for exponential series.

of normality. Indeed, as we shall see, even such enormously long runs turn out to be less than adequate in most cases.

Recall that in the case of exponential distribution,  $n_j = 1000$  for  $j = 1, \dots, K$ , with  $K = 2000$ . Thanks to such small sample, it has been possible to obtain the results in less than a second, even on an ordinary dual core laptop. However, in the current normality scenario, such pleasant computational perspective is unimaginable. Fortunately, we have access to a parallel computing architecture associated with a VMWare consisting of 100 64-bit cores, running at 2.80 GHz speed, and having 1 TB memory. Implementation of our parallelized C codes on the available 100 cores takes about 52 minutes.

The convergence behaviour of the random series are correctly determined, but panels (f) and (g) of Figures 5.3.2 indicate very slow divergence. Indeed, these figures depict the posterior means in the last  $5 \times 10^5$  iterations of the total  $K = 10^6$  iterations. We found that slow divergence is generally the case when one of  $\sum_{i=1}^{\infty} \mu_i$  or  $\sum_{i=1}^{\infty} \sigma_i^2$  is a divergent series of the form  $\sum_{i=1}^{\infty} i^{-p}$ , with  $1 - \zeta \leq p \leq 1$ , where  $\zeta (> 0)$  is small.

### 5.3.3 Example 3: Dependent hierarchical normal distribution

So far we have considered examples of random series where the terms are independent. The actual convergence properties of these random series are known by Kolmogorov's three series theorem, and knowledge of the convergence properties helped validate our Bayesian idea in these cases.

Since theoretically our Bayesian method characterizes all random series irrespective of their dependence structure, we now turn to empirical validation of our Bayesian method even in dependent situations. Note that Kolmogorov's three series theorem no longer holds for dependent situations, and we need to create examples where the actual convergence properties are known, in spite of dependence.

A simple example is as follows. We consider  $[X_i|\xi] \sim N(\mu_i, \xi\sigma_i^2)$ , independently, for  $i \geq 1$ , where  $\xi \sim U(0, 1)$ . Thus,  $X_i$  are conditionally independent given  $\xi$ , but



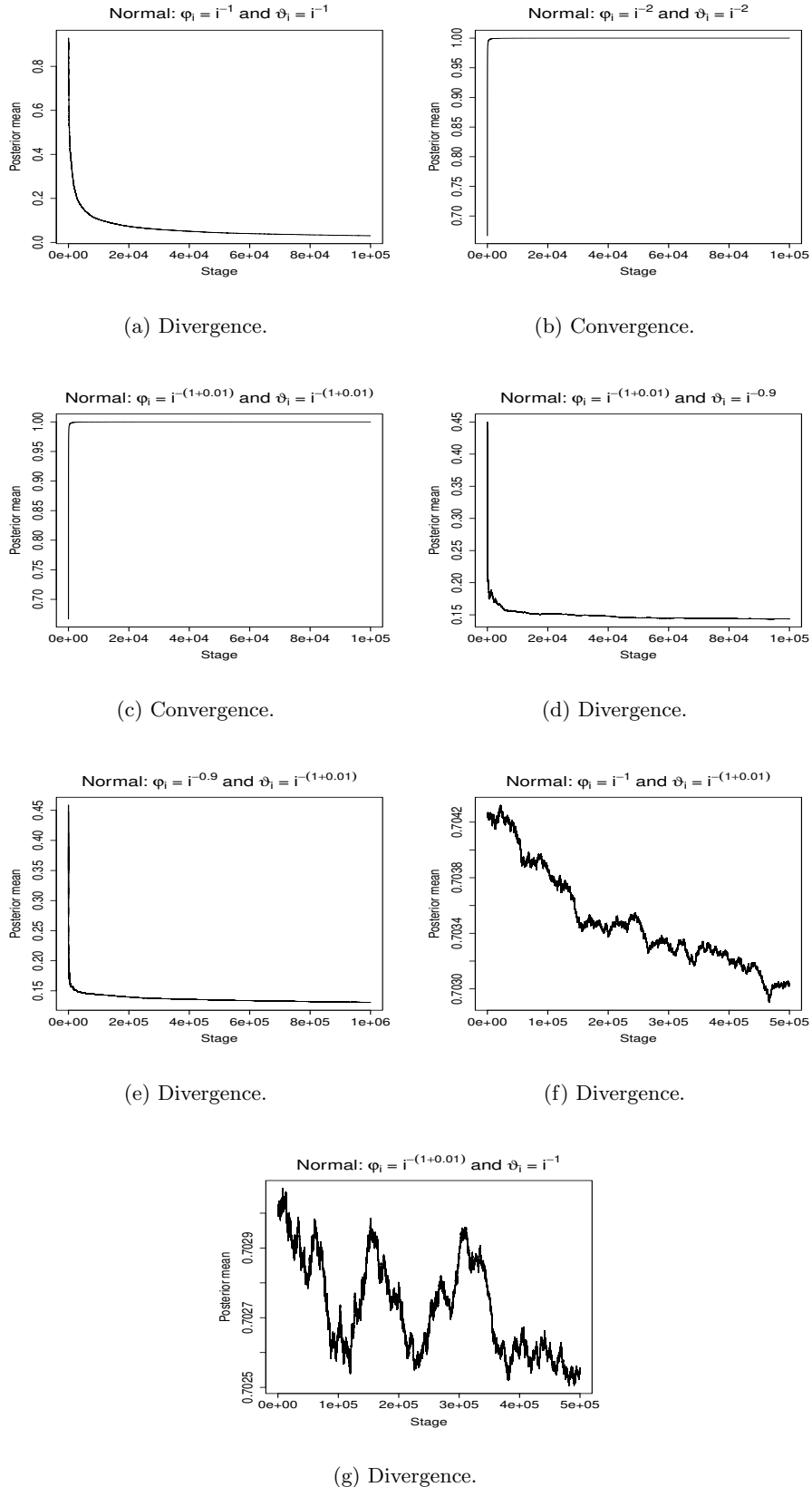


Figure 5.3.2: Example 2: Convergence and divergence for normal series.

unconditionally, they are dependent. As in the case of the independent normal example, we assume that  $\mu_i \sim N(0, \phi_i^2)$  and  $\sigma_i^2 \sim \mathcal{E}(\vartheta_i)$ . Hence, we now deal with a dependent, hierarchical normal setup for the  $X_i$ . Since given  $\xi$ , Kolmogorov's three series theorem is applicable and the series is either convergent or divergent almost surely, integrating over the finite random variable  $\xi$  does not alter the convergence properties, in spite of dependence. To see this, note that if almost surely  $\sum_{i=1}^{\infty} X_i < \infty$  given  $\xi$ , then letting  $P$  stand for the probability of events corresponding to  $X_i$  as well as the probability measure associated with  $\xi$ , the following hold:

$$\begin{aligned} P\left(\sum_{i=1}^{\infty} X_i < \infty\right) &= \int P\left(\sum_{i=1}^{\infty} X_i < \infty \middle| \xi\right) dP(\xi) \\ &= \int 1 \times dP(\xi) \\ &= 1. \end{aligned}$$

Similarly, if  $\sum_{i=1}^{\infty} X_i = \infty$  almost surely, given  $\xi$ , then

$$\begin{aligned} P\left(\sum_{i=1}^{\infty} X_i = \infty\right) &= \int P\left(\sum_{i=1}^{\infty} X_i = \infty \middle| \xi\right) dP(\xi) \\ &= \int 1 \times dP(\xi) \\ &= 1. \end{aligned}$$

Setting  $\epsilon = 0.001$ , as in the independent normal case we considered the upper bound  $c_j = \left|S_{j,n_j}^{\hat{\theta}}\right| + \frac{0.1}{j}$ , with  $n_j = 10^6$  for  $j = 1, \dots, K$ , where  $K = 10^6$ . VMWare implementation of our parallel codes again takes about 52 minutes with 100 cores. Convergence analyses for our dependent normal distribution are provided in Figure 5.3.3. Again, convergence behaviour of the random series are correctly determined, but as is evident from the figures, the rates of convergence and divergence turned out to be very slow in general. All these figures depict the posterior means in the last  $5 \times 10^5$  iterations

of a total  $10^6$  iterations.

#### 5.3.4 Example 4: Dependent state-space random series

We now consider the following random series:

$$\sum_{i=1}^{\infty} X_i \theta_i, \quad (5.3.1)$$

where for  $i \geq 1$ ,  $\theta_i \sim \mathcal{E}(\psi_i)$  independently, and  $X_i$  admits the following state-space representation:

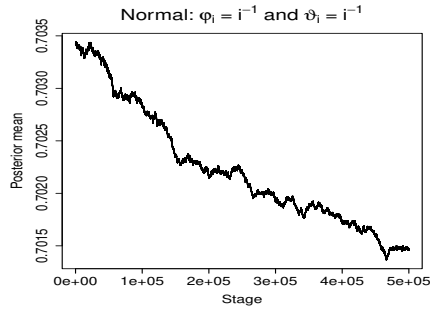
$$X_i = \alpha + \beta Z_i + \epsilon_i; \quad (5.3.2)$$

$$Z_i = \rho Z_{i-1} + \eta_i, \quad (5.3.3)$$

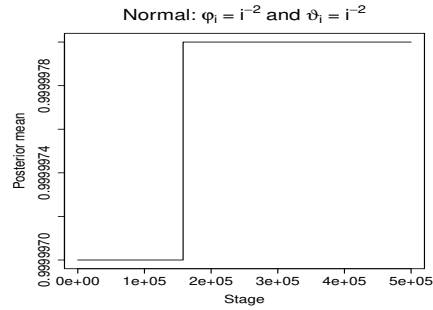
where  $Z_0, \alpha, \beta, \rho \stackrel{iid}{\sim} U(a, b)$ ,  $a = \varepsilon$ ,  $b = \varepsilon + 1$ , with  $\varepsilon > 0$ , and  $\epsilon_i, \eta_i \stackrel{iid}{\sim} N(0, 1)\mathbb{I}_{[a, b]}$ , that is the standard normal distribution truncated on  $[a, b]$ . It follows from the above representation that  $X_i$  are dependent, positive, and bounded random variables. Thus, the terms  $X_i \theta_i$  in (5.3.1) are also dependent, positive, but unbounded random variables. Since  $X_i$  are both upper and lower bounded, the convergence properties of (5.3.1) are dictated by the  $\theta_i$ 's.

In our simulation experiment, we generate  $\theta_i$  and  $X_i$  following the above model specifications, setting  $\varepsilon = 0.001$ . Thus, data  $Y_i = X_i \theta_i$ , for  $i \geq 1$ , are available for convergence analysis of (5.3.1).

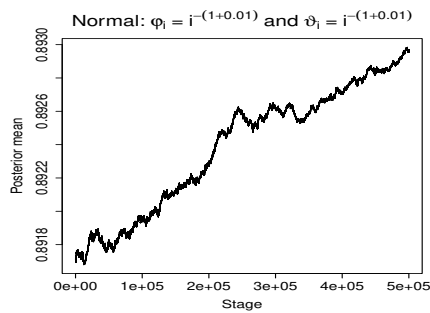
Since the exponential distribution dominates the convergence properties in this case, mathematically valid bound construction for the partial sums is possible in this case. Here we provide the details of our bound construction procedure. We first generate  $X_i^*$  following (5.3.2) and (5.3.3) and set  $Y_i = X_i^* \theta_i$ , with  $\theta_i = -\psi_i \log U_i$ . Combining these yields  $\log U_i = -Y_i / (\psi_i X_i^*)$ . We then set  $\tilde{Y}_i = X_i^* \tilde{\theta}_i$ , where  $\tilde{\theta}_i = -r_i^{-1}(\epsilon) \log U_i = (r_i^{-1}(\epsilon) Y_i) / (\psi_i X_i^*)$ ; as before, set  $\epsilon = 0.001$ . Letting  $S_{j, n_j}^{\tilde{\theta}}$  be the partial sums



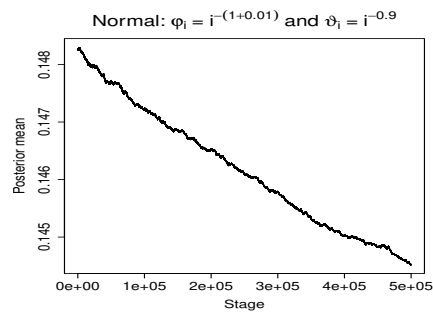
(a) Divergence.



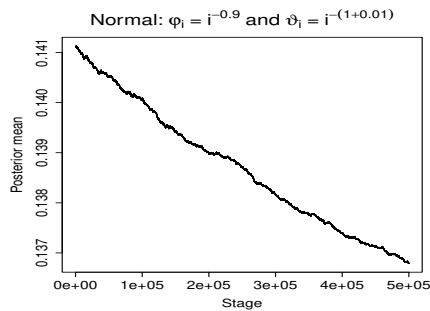
(b) Convergence.



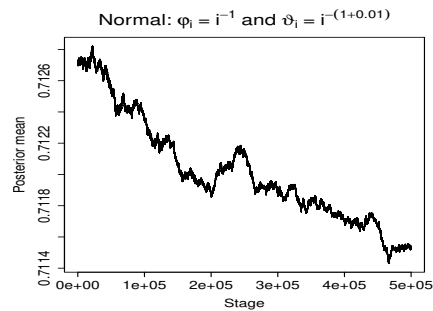
(c) Convergence.



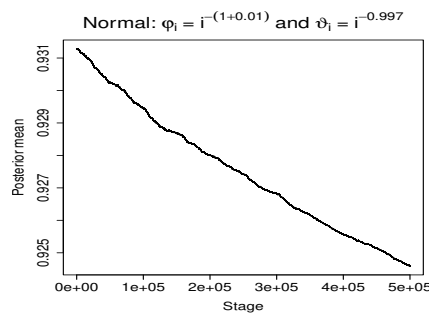
(d) Divergence.



(e) Divergence.



(f) Divergence.



(g) Divergence.

Figure 5.3.3: Example 3: Convergence and divergence for dependent normal series.

associated with  $\{\tilde{Y}_i\}_{i=1}^{\infty}$ , we set  $c_j = S_{j,n_j}^{\tilde{\theta}}$  as the upper bound for the partial sums associated with  $\{Y_i\}_{i=1}^{\infty}$ .

In this setup, as in Section 5.3.1 for the hierarchical exponential series, we set  $n_j = 1000$  for  $j = 1, \dots, K$ , where  $K = 2000$ . As before, with such small sample size, parallel implementation of this setup on our dual-core laptop takes less than a second to yield the results.

Figure 5.3.4 shows that the convergence behaviour of the random series are correctly and convincingly determined in all the cases despite the small sample sizes.

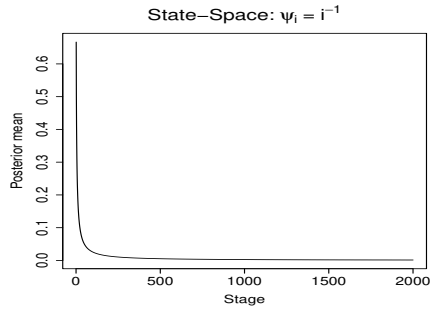
### 5.3.5 Example 5: Dependent state-space random series with hierarchical exponential distribution

In the state-space setup of Section 5.3.4 we considered  $\theta_i \sim \mathcal{E}(\psi_i)$ . Now we add an extra hierarchy to the exponential distribution by specifying, as in Section 5.3.1, that  $\theta_i \sim \mathcal{E}(\vartheta_i)$  and  $\vartheta_i \sim \mathcal{E}(\psi_i)$ . Thus, this state-space model is dominated by the hierarchical exponential distribution.

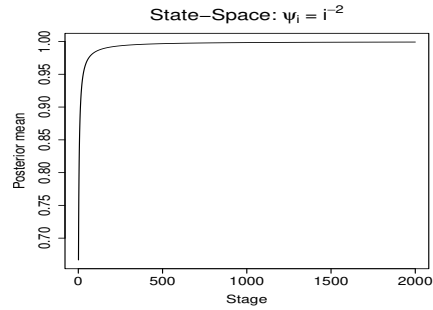
As before, let  $Y_i = X_i\theta_i$  be available. In our simulation experiment, we generate  $\theta_i$  and  $X_i$  following the hierarchical exponential driven state-space model specifications, setting  $\varepsilon = 0.001$ .

To obtain the bound  $c_j$  for the partial sums, we employ the following strategy. We first generate  $X_i^*$  following (5.3.2) and (5.3.3) and set  $Y_i = X_i^*\theta_i$ , with  $\theta_i = -\vartheta_i \log U_i$ . Combining these yields  $\log U_i = -Y_i/(\vartheta_i X_i^*)$ , where  $\vartheta_i = -\psi_i \log U_i^*$ . Here  $U_i$  and  $U_i^*$  are mutually independent *iid*  $U(0, 1)$  random variables for  $i \geq 1$ . We then set  $\tilde{Y}_i = X_i^*\tilde{\theta}_i$ , where  $\tilde{\theta}_i = -\tilde{\vartheta}_i \log U_i$ , and  $\tilde{\vartheta}_i = -r_i^{-1}(\varepsilon) \log U_i^*$ ; as before, we set  $\varepsilon = 0.001$ . Combining, we obtain  $\tilde{Y}_i = Y_i \left( \tilde{\vartheta}_i / \vartheta_i \right)$ . Letting  $S_{j,n_j}^{\tilde{\theta}}$  be the partial sums associated with  $\{\tilde{Y}_i\}_{i=1}^{\infty}$ , we set  $c_j = S_{j,n_j}^{\tilde{\theta}}$  as the upper bounds for the partial sums associated with  $\{Y_i\}_{i=1}^{\infty}$ .

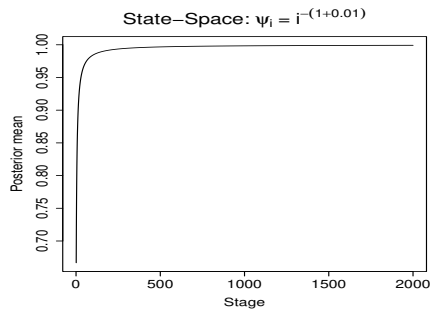
As before, we set  $n_j = 1000$ , for  $j = 1, \dots, K$ , where  $K = 2000$ , and our parallel computing procedure implemented in our laptop takes less than a second to complete



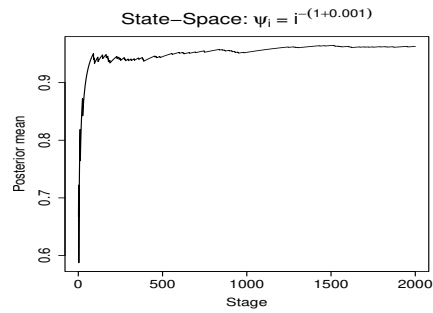
(a) Divergence.



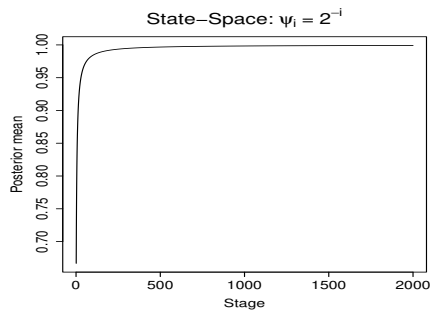
(b) Convergence.



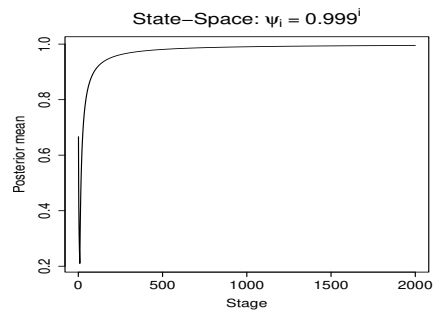
(c) Convergence.



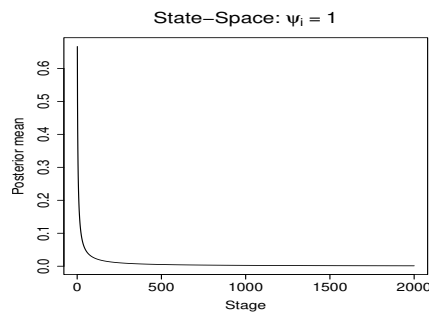
(d) Convergence.



(e) Convergence.



(f) Convergence.



(g) Divergence.

Figure 5.3.4: Example 4: Convergence and divergence for state-space series.

each exercise.

Figure 5.3.5 shows that in all the cases, our Bayesian procedure correctly detects convergence and divergence of the underlying series, even with such small sample size.

### 5.3.6 Example 6: Random Dirichlet series

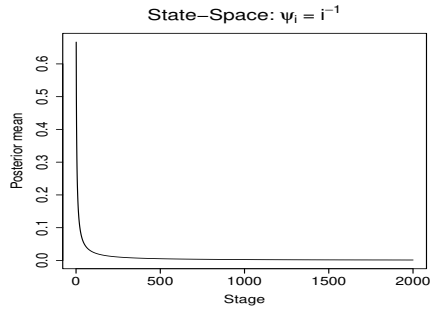
Consider the random Dirichlet series (RDS) given by

$$\sum_{i=1}^{\infty} \frac{X_i}{i^p}, \quad (5.3.4)$$

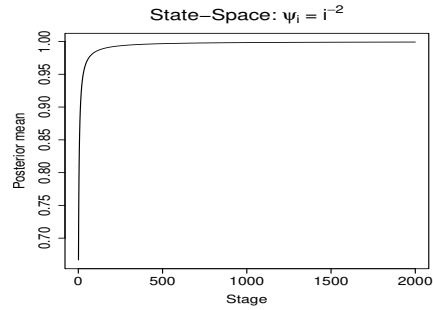
where  $X_i$  are *iid* random variables taking values  $-1$  and  $1$  with probabilities  $1/2$ , and  $p$  is a real number. Since  $|X_i| = 1$  almost surely, it follows that for any  $R > 0$ , there exists  $i_0$ , such that for  $i \geq i_0$ ,  $\frac{X_i}{i^p} < R$ , provided  $p > 0$ . Hence, for  $p > 0$ ,  $\mathbb{I}_{\left\{\frac{|X_i|}{i^p} < R\right\}} = 1$  almost surely, for  $i \geq i_0$ . With this, it follows by a simple application of Kolmogorov's three series theorem that the random series converges almost surely for  $p > 1/2$  and diverges almost surely for  $0 < p \leq 1/2$ . If  $p = 0$ , then the summands of (5.3.4) are *iid* and hence (5.3.4) diverges. Now, if  $p \in (-\infty, 0)$ , then for any  $R > 0$ , there exists  $i_0 \geq 1$  such that  $P\left(\frac{|X_i|}{i^p} > R\right) = 1$ , for  $i \geq i_0$ . Hence,  $\sum_{i=1}^{\infty} P\left(\frac{|X_i|}{i^p} > R\right) = \infty$ , for any  $R > 0$ . Consequently, by Kolmogorov's three series theorem, (5.3.4) diverges for  $p \in (-\infty, 0)$ . Combining the above arguments it follows that (5.3.4) converges almost surely for  $p > 1/2$  and diverges almost surely for  $p \leq 1/2$ .

Since  $X_i$  takes both positive and negative values with positive probabilities, application of the mathematically valid parametric upper bound is infeasible. Hence, we consider application of (5.2.9) where  $\tilde{\theta}$  in  $S_{j,n_j}^{\tilde{\theta}}$  corresponds to  $p = 1 + \epsilon$  in this case. Here we set  $\epsilon = 0.001$  as before. We experimented with various choices of the tuning parameter  $a$  on the right hand side of (5.2.9) and all of them yielded the same inference. Hence, we report our results with respect to  $a = 1$ .

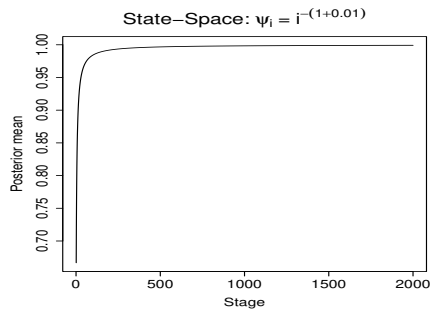
Figure 5.3.6 shows the results of our Bayesian application to this problem for various



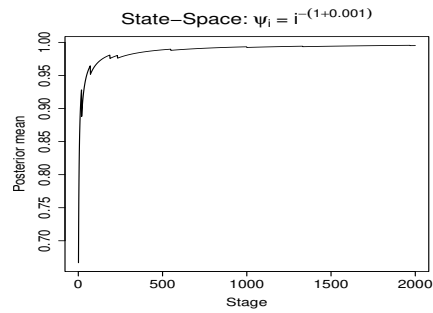
(a) Divergence.



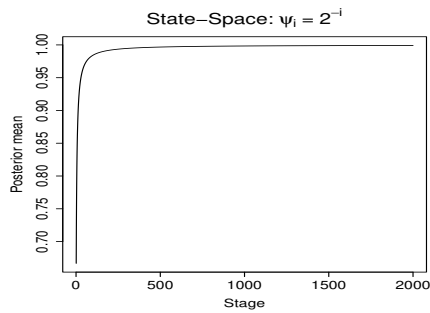
(b) Convergence.



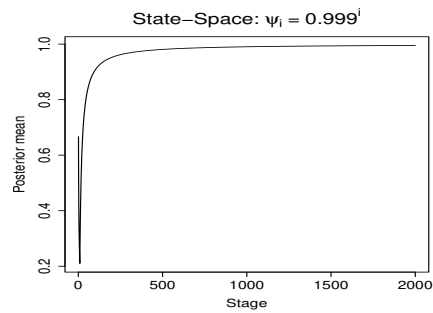
(c) Convergence.



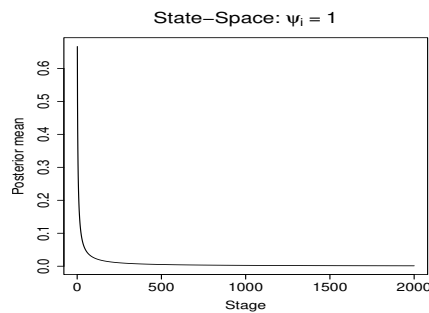
(d) Convergence.



(e) Convergence.



(f) Convergence.



(g) Divergence.

Figure 5.3.5: Example 5: Convergence and divergence for state-space series with hierarchical exponential distribution.



values of  $p$ , for  $n_j = 1000$ ;  $j = 1, \dots, K$ , with  $K = 2000$ . Note that for  $p = 0.501$  (panel (e) of Figure 5.3.6), we obtain the wrong result of divergence, whereas convergence is the correct result. This is a subtle situation as it may be difficult to distinguish divergence for  $p = 0.5$  and convergence for  $p = 0.501$ , but wrong results are obtained in many cases for  $p \in (0.5, 0.79)$ . Thus, effectiveness of the general upper bound (5.2.9) is again challenged in this example.

## 5.4 Nonparametric bounds for the partial sums and simulation experiments

The parametric upper bounds for the partial sums are quite restrictive in the sense of requiring non-negative supports. The general upper bound (5.2.9) is not theoretically sound and although it works well for exponential series and state-space series driven by exponential distributions (results not shown for the sake of brevity), we have shown that its performance for series driven by normal distributions is far from satisfactory, as very large number of iterations, with very large number of summands for the partial sums are required. Even then, the independent and dependent normal setups do not exhibit convergence of our Bayesian procedure adequately close to 1 and 0 for convergent and divergent random series, in many cases. Also in the RDS setup, incorrect results are obtained in a lot of cases with (5.2.9). Thus, the general bound is not expected to work well for distributions supported on the real line. Moreover, the bound construction methods require specific knowledge of the form of the underlying distribution  $f_{\theta_i}$  of the  $i$ -th element  $X_i$  of the random series. In reality, such information can not be expected to be available.

Hence, effective bounds, which are independent of supports of the summands and the underlying distributional assumptions, are desirable. To this end, we propose a nonparametric bound that, as we shall see later, also plays very important role in the

5.4. NONPARAMETRIC BOUNDS FOR THE PARTIAL SUMS AND SIMULATION EXPERIMENTS

103

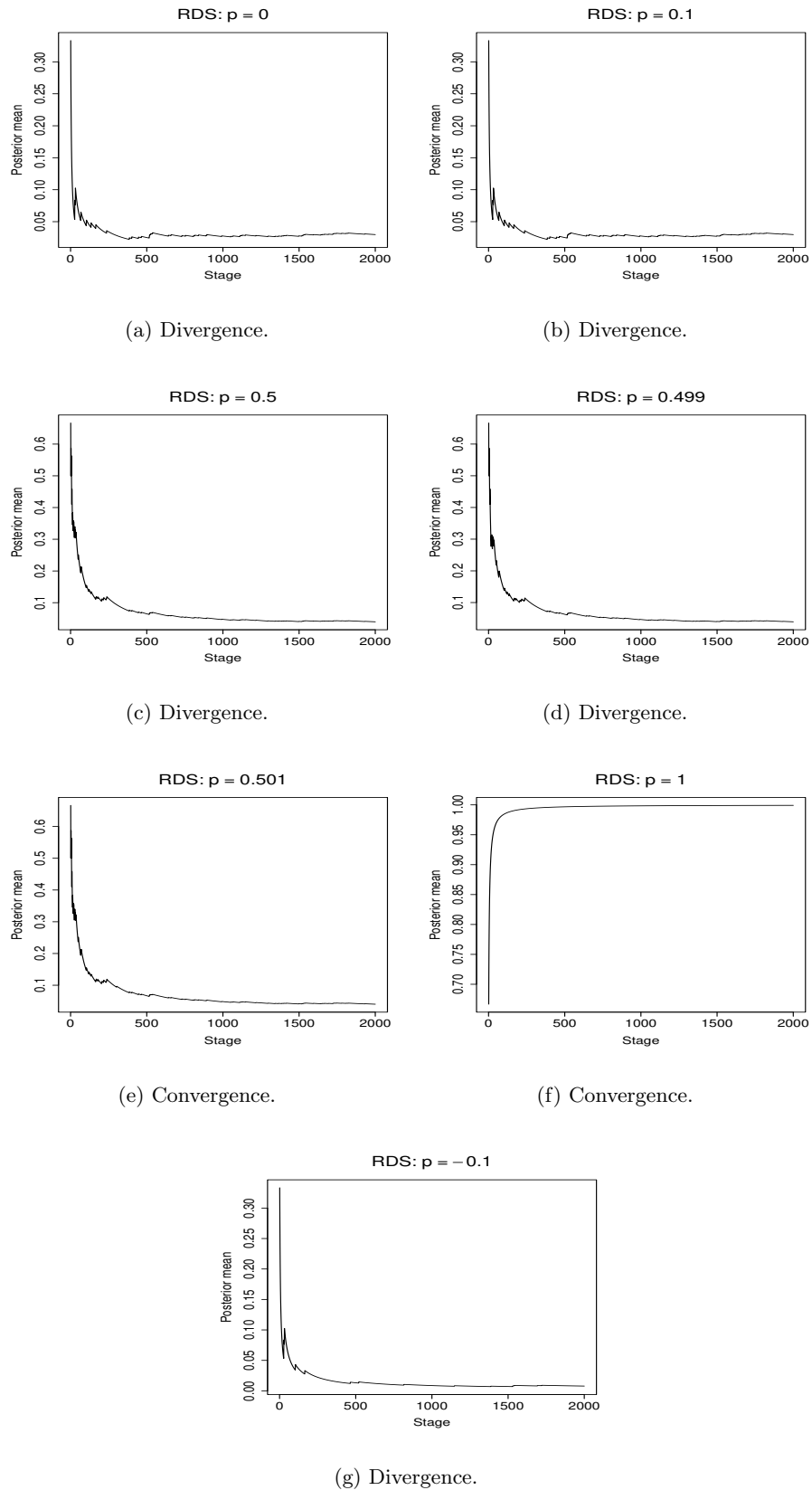


Figure 5.3.6: Example 6: Convergence and divergence for RDS.

context of Bayesian characterization of stochastic process properties. Specifically, we set

$$c_j = \hat{C}_j / \log(j + 1), \tag{5.4.1}$$

where  $\hat{C}_1$  is a chosen constant, and for  $j > 1$ ,  $\hat{C}_j = \hat{C}_{j-1} + 0.05$  if  $y_{j-1} = 1$  and  $\hat{C}_j = \hat{C}_{j-1} - 0.05$  if  $y_{j-1} = 0$ .

Thus, we favour convergence at the next,  $(j + 1)$ -th stage, if at the current stage convergence is supported ( $y_j = 1$ ), and favour divergence otherwise. The  $\log(j + 1)$  scale ensures that the rate of convergence of  $c_j$  to zero as  $j \rightarrow \infty$ , is neither too fast, nor too slow.

The choice of the initial value  $\hat{C}_1$  is an important issue and if chosen without utmost care, can yield wrong results regarding series convergence properties. The choice is also expected to be problem specific in general. However, in our examples involving normal and exponential based models, we find  $\hat{C}_1 = 0.71$  and  $0.725$ , respectively, to be quite appropriate. As we shall see later, this is somewhat in keeping with our results Chapter 7 in the time series context where  $\hat{C}_1 = 1$  turned to be adequate in most cases, in spite of the wide variety of examples. In the case of RDS we exploit the corresponding deterministic Dirichlet series to obtain an appropriate value of  $\hat{C}_1$ .

#### 5.4.1 Simulation experiments with the nonparametric bound form

We now conduct simulation experiments with this new, nonparametric bound form (5.4.1) applied to the setups considered in Section 5.3. For all the cases, we now consider  $n_j = 1000$  for  $j = 1, \dots, K$ , with  $K = 2000$ . Thus, even for the series driven by normal and dependent normal distributions we now consider situations where the number of summands in each partial sum, as well as the number of stages (iterations) for our Bayesian procedure are significantly smaller compared to those in Sections 5.3.2 and 5.3.3. Needless to mention, the time taken for the implementations of the Bayesian procedure with the nonparametric bound are less than a second. As we shall see, in

almost all the cases, the bound form (5.4.1) yields the correct answer, even for the normal driven series, in spite of many times smaller sample size as used in Sections 5.3.2 and 5.3.3. Importantly, in all the cases, the Bayesian method gets sufficiently close to 1 and 0 for convergent and divergent series, respectively. Recall that this was not the case for independent and dependent normal setups, even with extremely large sample sizes, and incorrect results were obtained for the RDS. Thus, the bound (5.4.1), in spite of having a nonparametric form, turns out to be far more effective and efficient than the previous general parametric bound (5.2.9). However, for the hierarchical exponential setup and the state-space hierarchical exponential setup, the nonparametric bound performs slightly worse in a very subtle situation compared to the mathematically valid parametric bound. On the other hand, the nonparametric bound slightly outperforms the mathematically sound parametric counterpart in a subtle situation of the state-space non-hierarchical exponential setup. Thus, the nonparametric bound seems to be very much comparable with the valid parametric bound when the latter is available, and emphatically outperforms the general parametric bound (5.2.9).

**Example 1 revisited: Hierarchical exponential distribution**

As in Section 5.3, we first consider the setup  $X_i \sim \mathcal{E}(\theta_i)$  and  $\theta_i \sim \mathcal{E}(\psi_i); i \geq 1$ . Here experimentation reveals that  $\hat{C}_1 = 0.725$  is an appropriate choice that can detect most convergent and divergent series driven by exponential distributions of the above form.

Figure 5.4.1 displays the results of our Bayesian analyses of different exponential series of the above form. Not only does the Bayesian procedure with the nonparametric bound captures the correct result even for such small sample sizes, it does so quite convincingly, as the method gets adequately close to 1 and 0 for convergent and divergent series, respectively. However, it is important to mention that for  $\psi_i = i^{-p}$ , for  $p \in (0.95, 1]$ , our method with the nonparametric bound failed to yield correct results. Thus, a little subtlety seems to have been sacrificed due to the small sample size. Indeed, increasing

$n_j$  led to increasing shrinkage of the offending interval  $(0.95, 1]$  towards 1.

**Example 2 revisited: Hierarchical normal distribution**

As in Section 5.3.2, we now let  $X_i \sim N(\mu_i, \sigma_i^2)$ ,  $\mu_i \sim N(0, \phi_i^2)$  and  $\sigma_i^2 \sim \mathcal{E}(\vartheta_i)$ ;  $i \geq 1$ . Here  $\hat{C}_1 = 0.71$  turned out to be appropriate. Notice its close similarity with  $\hat{C}_1 = 0.725$  for the exponential bound.

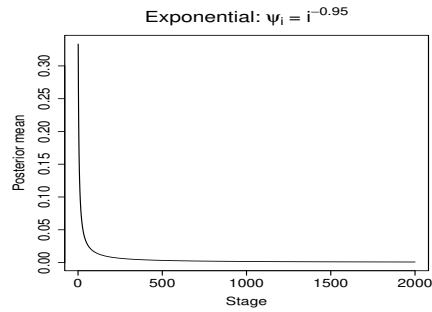
Figure 5.4.2 shows our results in this setup. In all the cases, correct results are convincingly obtained, even with such a small sample size. The results are convincing in the sense that the underlying Bayesian procedure gets sufficiently close to 1 and 0 for all the convergent and divergent series, respectively. Thus, compared to Figure 5.3.2 corresponding to the parametric bound, we have a huge gain in efficiency and effectiveness. However, it must be mentioned that for such small sample size, our method failed in the cases where  $\phi_i = \vartheta_i = i^{-(1+a)}$ , for  $a \in (0.0, 0.04)$ .

**Example 3 revisited: Dependent hierarchical normal distribution**

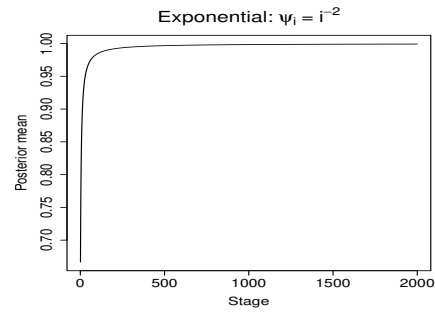
As in Section 5.3.3 we again consider  $[X_i|\xi] \sim N(\mu_i, \xi\sigma_i^2)$ , independently, for  $i \geq 1$ , where  $\xi \sim U(0, 1)$ ,  $\mu_i \sim N(0, \phi_i^2)$  and  $\sigma_i^2 \sim \mathcal{E}(\vartheta_i)$ , but now with the parametric bound for the partial sums replaced with the nonparametric form (5.4.1), with  $\hat{C}_1 = 0.71$ , the same initial constant used for the nonparametric bound for the normal setup in Section 5.4.1. Figure 5.4.3 shows the relevant results in this setup. The results are similar to the independent normal setup with nonparametric bound, and are very significant improvements to the results provided by the parametric bound displayed in Figure 5.3.3. Indeed, Figure 5.3.3 shows that none of the convergence and divergence results for the parametric bound is convincing, even for such huge samples, and even after such long run-times. In sharp contrast, the nonparametric bound results depicted by Figure 5.4.3 are highly persuasive, even with such small samples, requiring run-times of less than a second on our ordinary dual core laptop.

5.4. NONPARAMETRIC BOUNDS FOR THE PARTIAL SUMS AND SIMULATION EXPERIMENTS

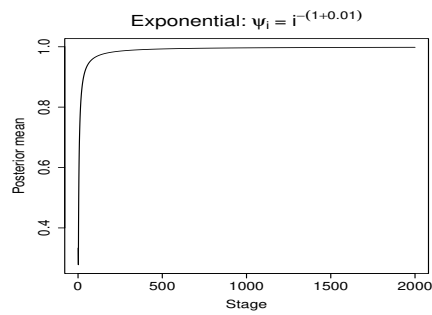
107



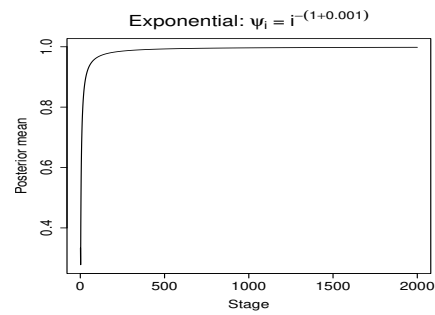
(a) Divergence.



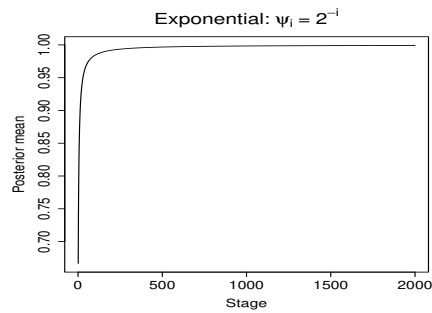
(b) Convergence.



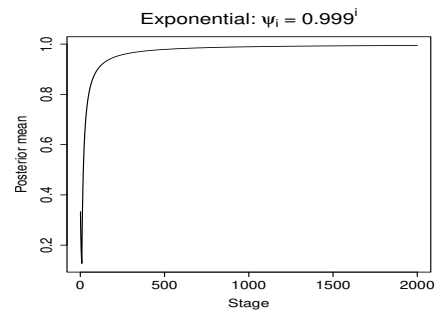
(c) Convergence.



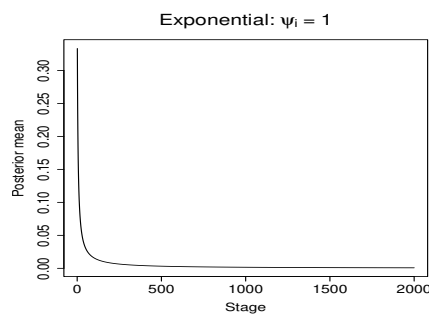
(d) Convergence.



(e) Convergence.



(f) Convergence.



(g) Divergence.

Figure 5.4.1: Example 1 revisited: Convergence and divergence for exponential series with nonparametric bound.

5.4. NONPARAMETRIC BOUNDS FOR THE PARTIAL SUMS AND SIMULATION EXPERIMENTS

108

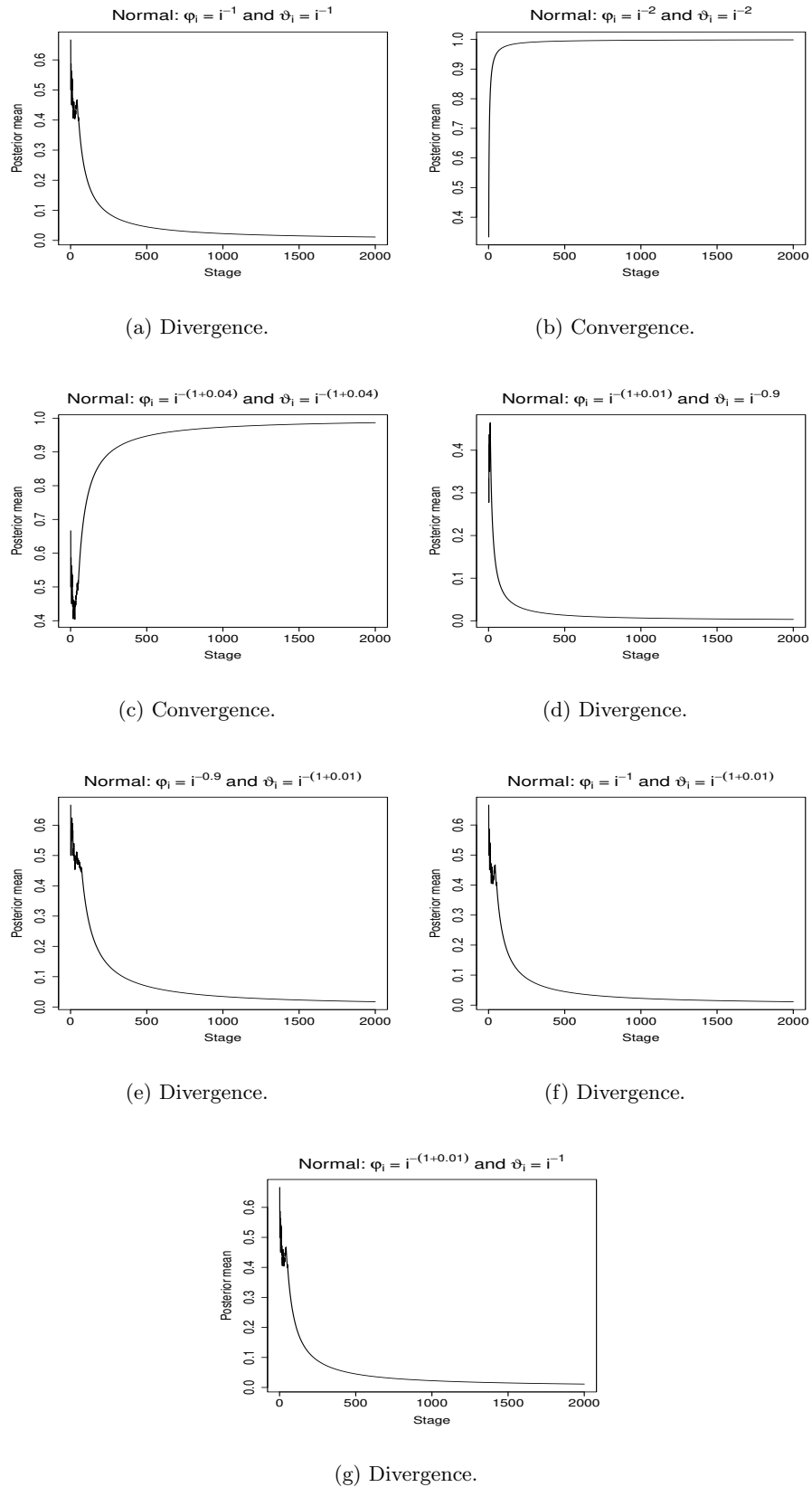


Figure 5.4.2: Example 2 revisited: Convergence and divergence for normal series with nonparametric bound.

5.4. NONPARAMETRIC BOUNDS FOR THE PARTIAL SUMS AND SIMULATION EXPERIMENTS

109

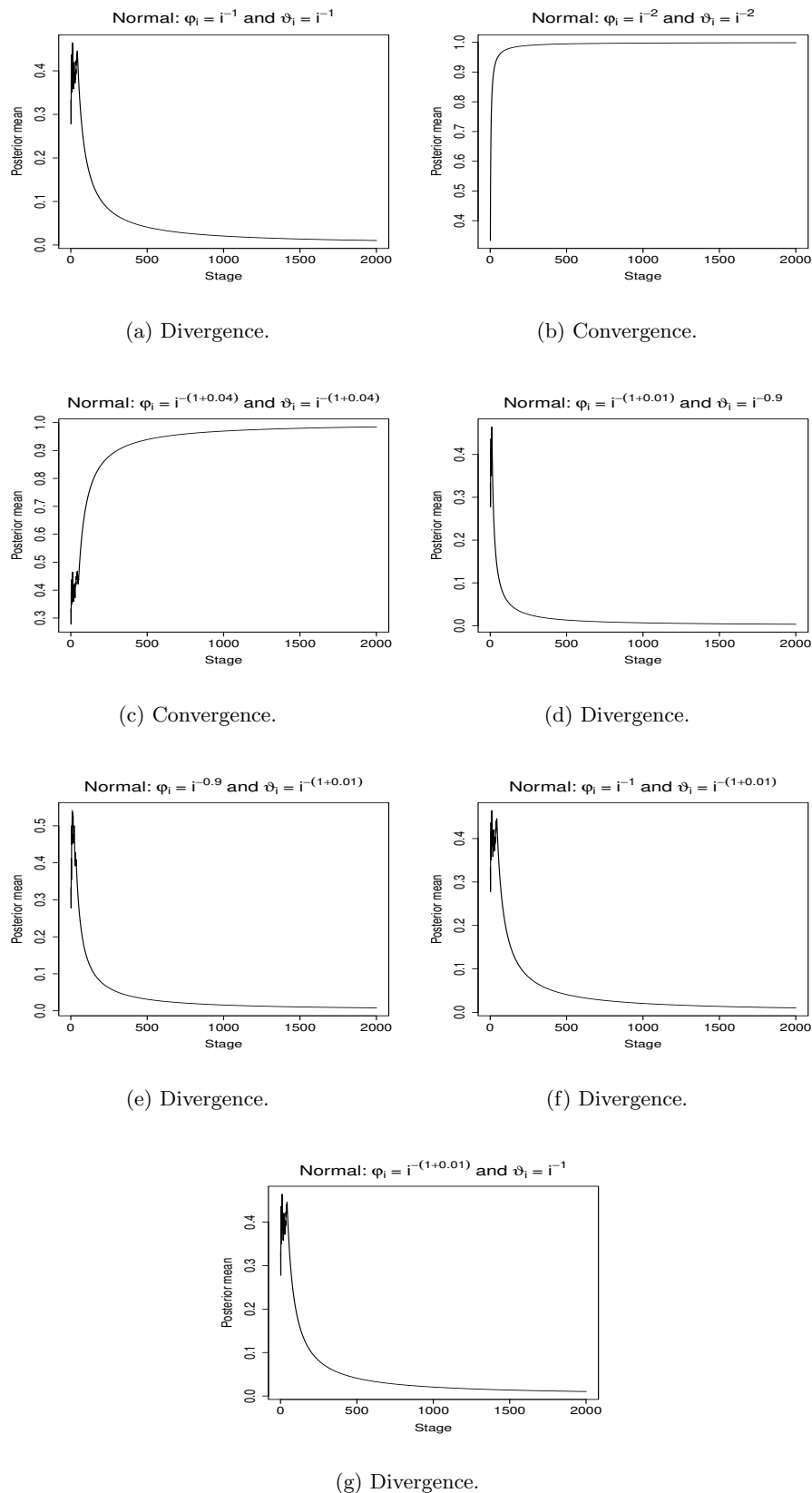


Figure 5.4.3: Example 3 revisited: Convergence and divergence for dependent normal series with nonparametric bound.



**Example 4 revisited: Dependent state-space random series**

Following Section 5.3.4 we consider random series of the form  $\sum_{i=1}^{\infty} X_i \theta_i$  where for  $i \geq 1$ ,  $\theta_i \sim \mathcal{E}(\psi_i)$  independently, and  $X_i$  has the state-space representation given by (5.3.2) and (5.3.3). The rest of the model details remain the same as in Section 5.3.4.

Application of our new nonparametric bound to the partial sums, with  $\hat{C}_1 = 0.725$ , which is the same as that of the exponential series with the nonparametric bound, we obtain correct results in all the cases, as displayed by Figure 5.4.4. In fact, the nonparametric bound not only matches the performance of the parametric bound method detailed in Section 5.3.4, it seems to outperform the latter for  $\psi = i^{-(1+0.001)}$  in terms of faster convergence.

**Example 5 revisited: Dependent state-space random series with hierarchical exponential distribution**

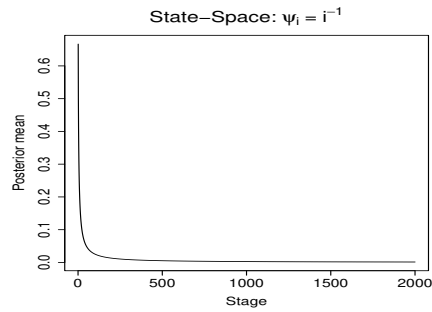
In the state-space model with hierarchical exponential distribution considered in Section 5.3.5, we now apply the nonparametric bound with  $\hat{C}_1 = 0.725$  to address convergence properties of  $\sum_{i=1}^{\infty} X_i \theta_i$  using our Bayesian methodology. The results displayed in Figure 5.4.5 again shows very accurate detection of convergence properties of the underlying infinite series even with small samples sizes. However, it is to be noted that because of the hierarchy in the exponential distribution, a little subtlety has been sacrificed by our method as it is unable to correctly diagnose divergence for  $\psi = i^{-p}$  when  $p \in (0.997, 1]$ .

**Example 6 revisited: Random Dirichlet series**

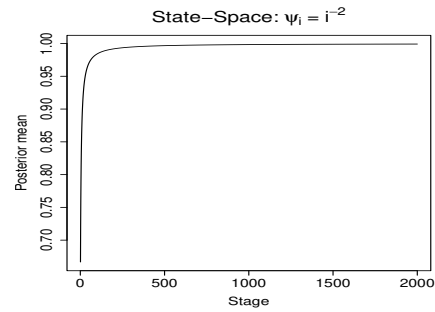
Again consider the RDS given by (5.3.4). Recall that this problem does not admit any theoretically valid upper bound since the summands take both positive and negative values with positive probabilities. Application of the general parametric upper bound (5.2.9) to this problem in Section 5.3.6 have led to wrong results in many cases of this

5.4. NONPARAMETRIC BOUNDS FOR THE PARTIAL SUMS AND SIMULATION EXPERIMENTS

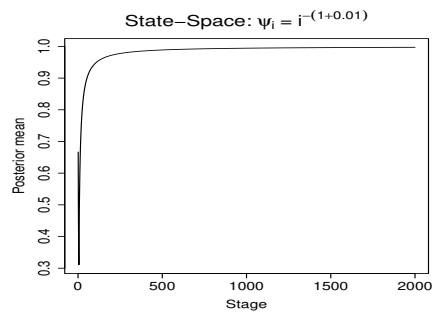
111



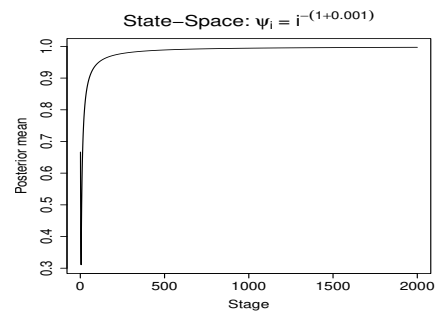
(a) Divergence.



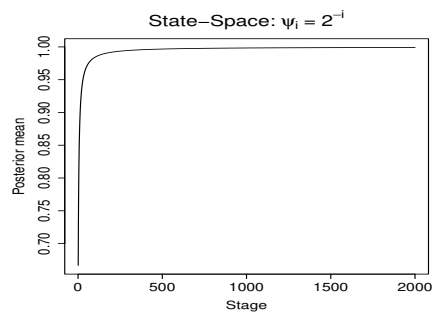
(b) Convergence.



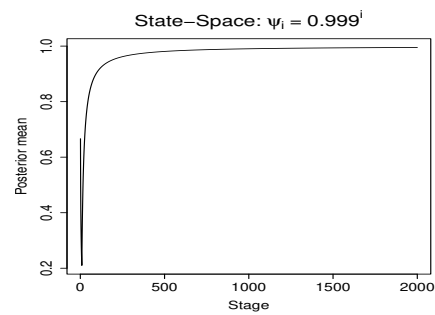
(c) Convergence.



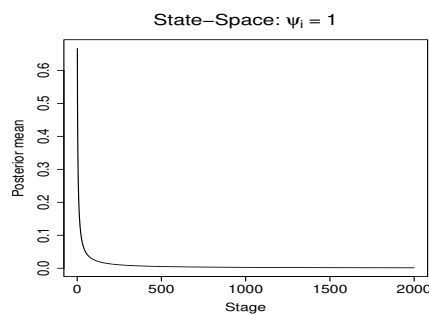
(d) Convergence.



(e) Convergence.



(f) Convergence.



(g) Divergence.

Figure 5.4.4: Example 4 revisited: Convergence and divergence for state-space series with nonparametric bound.

5.4. NONPARAMETRIC BOUNDS FOR THE PARTIAL SUMS AND SIMULATION EXPERIMENTS

112

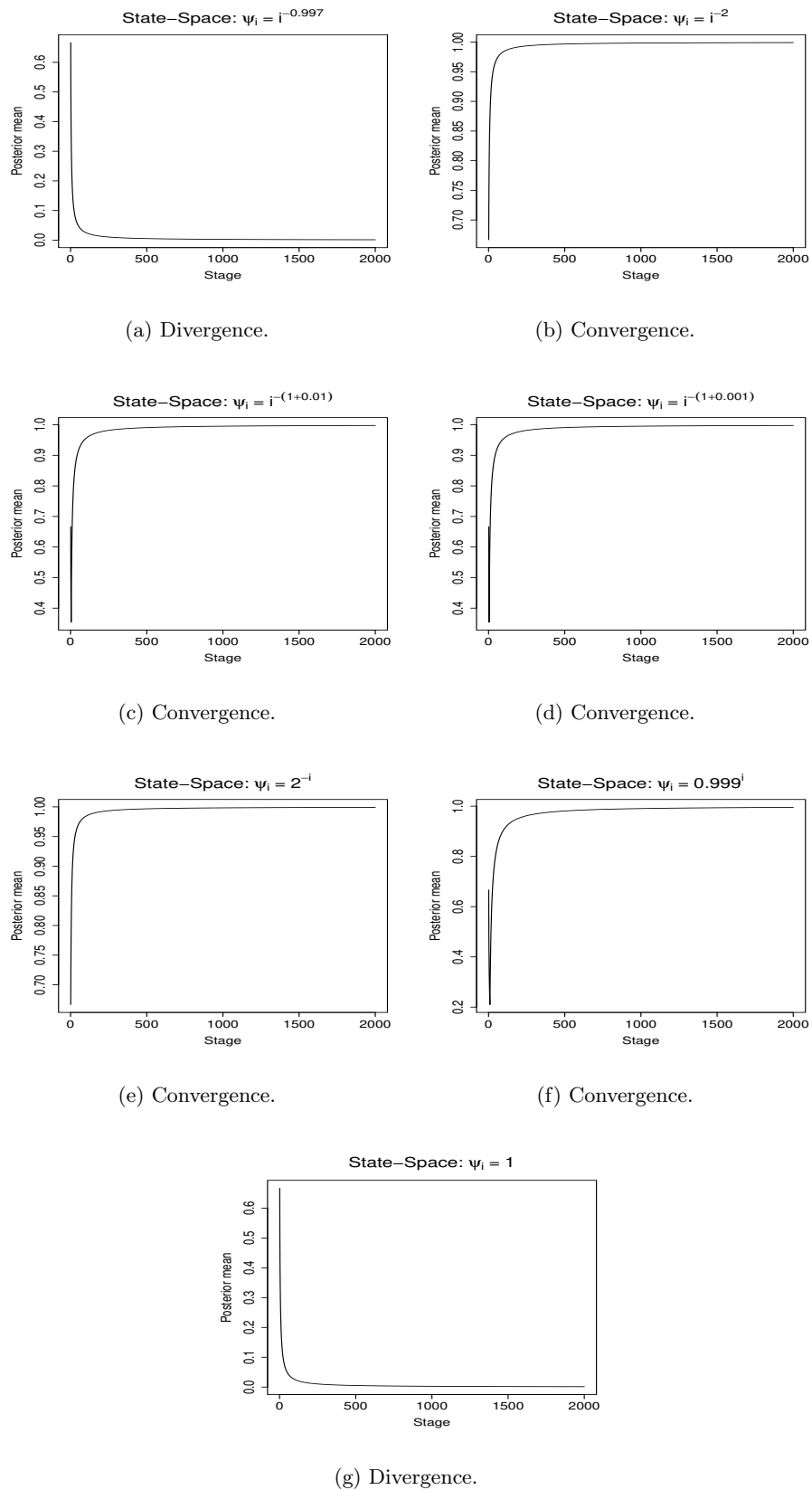


Figure 5.4.5: Example 5 revisited: Convergence and divergence for state-space series with hierarchical exponential distribution.

problem. Hence, we now employ our nonparametric bound to analyse convergence for the RDS.

As shown by Figure 5.4.6, application of our nonparametric bound to this problem for various values of  $p$  revealed correct convergence analysis by our Bayesian method in all the cases. To choose  $\hat{C}_1$  appropriately in this problem, we first considered the deterministic series  $\sum_{i=1}^{\infty} i^{-2p}$ , whose convergence properties are known. For this series we selected that value of  $\hat{C}_1$  which led to correct convergence diagnosis of our Bayesian procedure with the nonparametric bound, for all (in practice, most) values of  $p$ . This led to  $\hat{C}_1 = 0.44$ , and this value turned out to be an excellent choice even for the RDS given by (5.3.4).

In other words, the nonparametric bound in this problem soundly beats the parametric bound.

## 5.5 Application of random series convergence diagnostics to global climate change

### 5.5.1 Future global warming investigation

Global climate change, or gradual increase of the earth's average surface temperature, is arguably the most important issue plaguing the environmental scientists all over the world. Overwhelmingly strong evidence from various data sources have led the U.S. Global Change Research Program, the National Academy of Sciences, and the Intergovernmental Panel on Climate Change (IPCC) to declare that global warming in the recent decades is unquestionable.

Such a concern is supported by the HadCRUT4 observed near surface average global monthly temperature dataset during the years 1850 – 2020, available from the IPCC website; see <https://www.metoffice.gov.uk/hadobs/hadcrut4/data/current/download.html>. But since the year 2020 is still ongoing, data points for the last few years seem

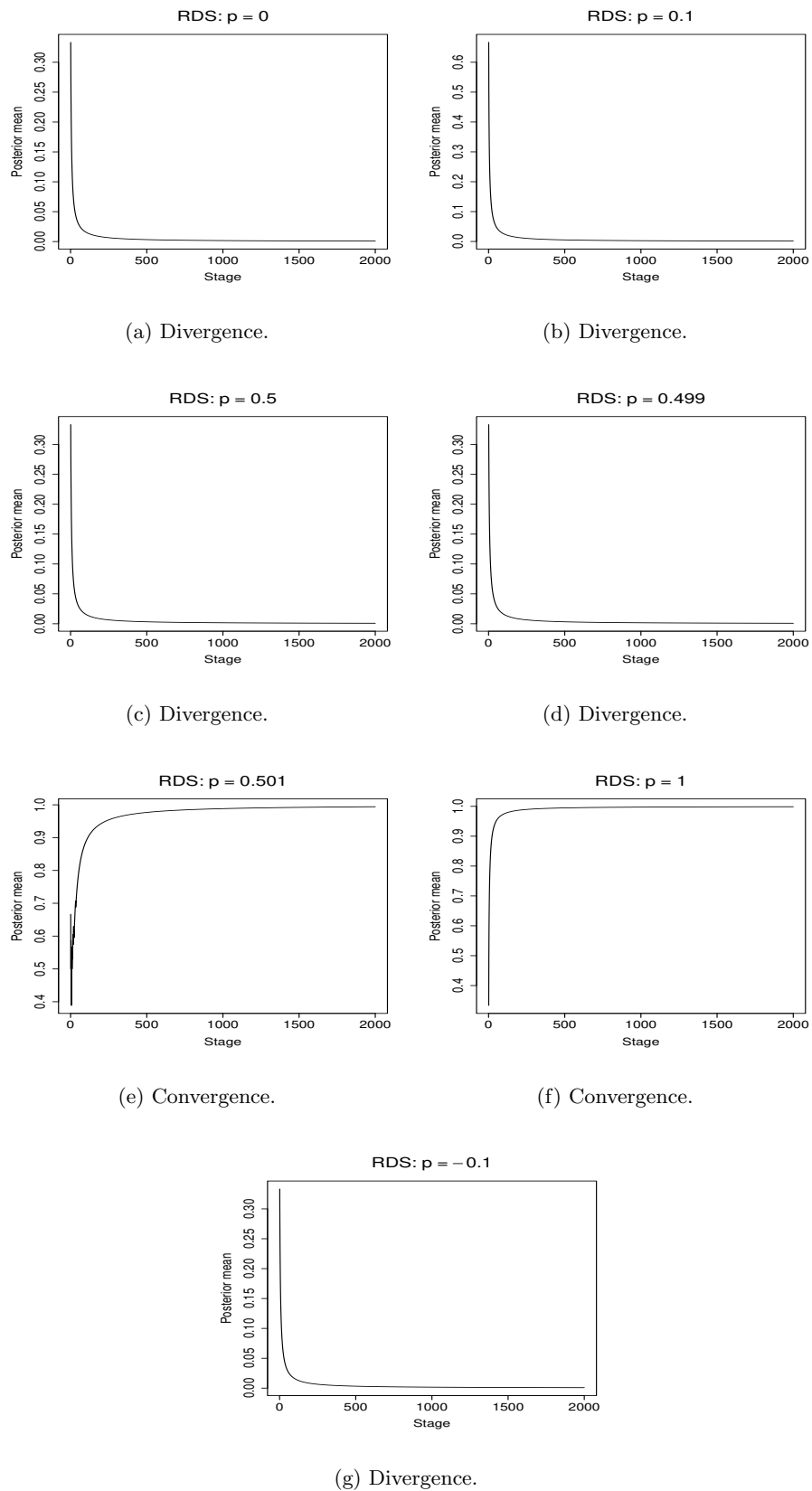
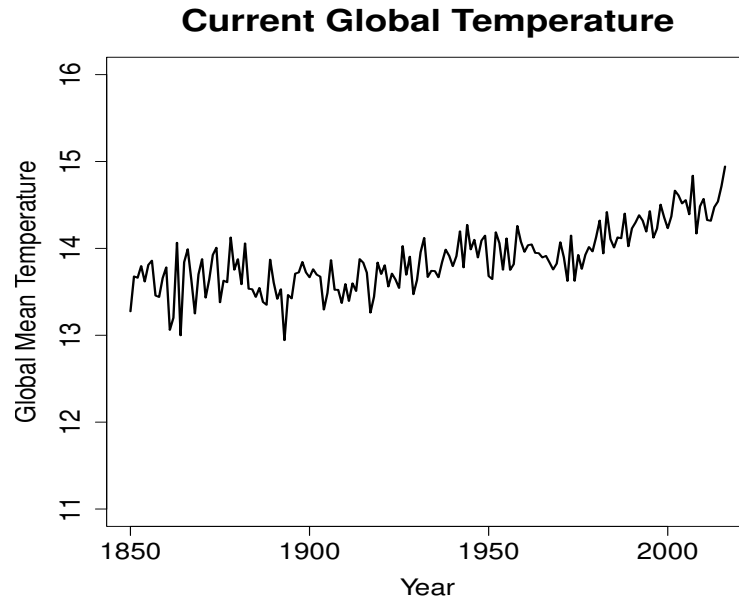


Figure 5.4.6: Example 6 revisited: Convergence and divergence for RDS.



**Figure 5.5.1:** Current, HadCRUT4 global mean temperature data.

somewhat doubtful to us, and hence we consider the monthly dataset in the range 1850 – 2016 (see also [Chatterjee and Bhattacharya \(2020\)](#) who analyzed the annual dataset). This dataset is only a record of temperature anomalies in degree celsius relative to the years 1961 – 1990, while we prefer the actual temperatures. As in [Chatterjee and Bhattacharya \(2020\)](#), we convert this anomaly data to (approximate) actual temperature data by adding  $14^{\circ}\text{C}$  to the anomalies, where  $14^{\circ}\text{C}$  is the most widely quoted value for the global average temperature for the 1961 – 1990 period (see [Jones \*et al.\* \(1999\)](#) for the detailed development). The IPCC website also provides 100 replications of the monthly HadCRUT4 data. Since these replications have very little variation we amalgamate these with the best estimate of the monthly global average temperature time series, to obtain a temperature time series for the 1850 – 2016 period consisting of  $167 \times 12 \times 100$  observations. A plot of the data is provided in Figure 5.5.1.

The dataset displayed in Figure 5.5.1 is not inconsistent with the IPCC records that compared to the pre-industrial baseline 1850 – 1900, the 2009 – 2015 time period was

warmer by about  $0.87^{\circ}\text{C}$ , and that each decade is getting warmer by about  $0.2^{\circ}\text{C}$ . Such an alarming rate of increase is (arguably) unprecedented, and continuation of such global warming may threaten life on earth in the future.

Thus, it is important to investigate if global warming will continue even in the future or if the temperature can be expected to “stabilize” in the near future around some value that does not threaten our existence on earth. Letting  $X_t$  denote global monthly average temperature at time point  $t$ , and  $\theta_0$  denote the temperature around which  $X_t$  is expected to concentrate for sufficiently large  $t$ , one may investigate convergence of the series  $\sum_{t=1}^{\infty} Y_{\theta_0,t}$ , where  $Y_{\theta_0,t} = X_t - \theta_0$ , or any other bijective transformation of  $X_t$ . Convergence of the series would imply that  $X_t \rightarrow \theta_0$ , as  $t \rightarrow \infty$ . In contrast, if the series diverges, then either global warming will continue or even if  $X_t \rightarrow \theta_0$ , as  $t \rightarrow \infty$ , the convergence would be much slower compared to the series convergence situation. Hence, in the case of divergence, stability can not be achieved in the near future.

Now, mean global temperature can not be assumed to be an unbounded quantity: even though Figure 5.5.1 shows a clearly increasing trend in the recent decades, it certainly must have an upper bound (say,  $U$ ), and a lower bound (say,  $L$ ) is even more obvious. Hence, if  $\sum_{t=1}^{\infty} Y_{\theta_0,t} = \infty$  for all  $\theta_0 \in [L, U]$  then  $X_t$  will not stabilize at any reasonable temperature value in the near future. This would also imply that global average temperature will randomly oscillate around various temperature values in the near future, ranging from hot to cold, and neither global warming nor global cooling can dominate the climate dynamics in the near future.

For the HadCRUT4 data shown in Figure 5.5.1, we set  $L = 11^{\circ}\text{C}$  and  $U = 16^{\circ}\text{C}$ , and consider the transformation  $Y_{\theta_0,t} = \log(\log(X_t)) - \log(\log(\theta_0))$ . Hence, for all  $\theta_0 \in [L, U]$ ,  $Y_{\theta_0,t} \in (-1, 1)$ . To implement our Bayesian procedure for random series convergence detection, we first note that there exists no standard model to represent the highly complex global climate dynamics. Thus the nonparametric method of bounding the partial sums using (5.4.1) is the only option. For  $\theta_0$ , we divide the interval  $[11, 16]$  into

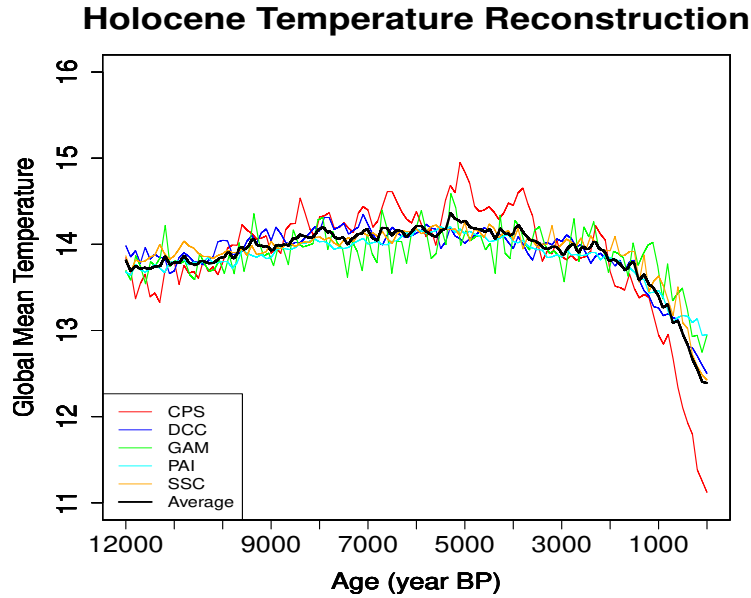
equidistant points with common gap 0.1 between any two consecutive points. Then, for each  $\theta_0$  in this grid of points, we apply our Bayesian procedure with  $n_j = 1200$  for  $j = 1, \dots, K = 167$ . In each case we obtain  $\sum_{t=1}^{\infty} Y_{\theta_0,t} = \infty$ , for  $\hat{C}_1 \in (0, 10)$ . Setting  $n_j$  and  $K$  to different values did not change the inference in any of the instances. Following the discussion in the previous paragraph, this helps us strongly conclude that in the near future the earth will not experience either global warming or global cooling. This conclusion is broadly consistent with the detailed future Bayesian nonparametric predictions of [Chatterjee and Bhattacharya \(2020\)](#).

### 5.5.2 Investigation of past climate stability

In Section 5.5.1 our Bayesian series convergence detection procedure helped us infer that future global warming or cooling is highly unlikely, and also that stability of the future climate can not be expected. We now investigate if stability, gradual warming or cooling can be expected of climate in the past. If neither is likely, then this would be consistent with our finding with the future climate dynamics, and would provide insight into general climate dynamics, both past and future.

To this end, we consider the Holocene global mean surface temperature reconstructions 12,000 years before present by [Kaufman \*et al.\* \(2020\)](#); here “present” refers to the year 1950. [Kaufman \*et al.\* \(2020\)](#) consider 5 methods of Holocene climate reconstruction, namely, Composite Plus Scale (CPS), Dynamic Calibrated Composite (DCC), General Additive Model (GAM), Pairwise Comparison (PAI) and Standard Calibrated Composite (SCC). We also consider the average of these 5 reconstructions, which we refer to as Average. The reconstructed Holocene temperatures by [Kaufman \*et al.\* \(2020\)](#) are available at <https://www.ncdc.noaa.gov/paleo-search/study/27330>. The reconstructions are provided at 100 years gap since 1950 to the past 12,000 years. We convert this to a monthly dataset by interpolation provided by the *R* software function “approx”. Our datasets thus consist of 144,000 Holocene temperature reconstruction values. The 5





**Figure 5.5.2:** Holocene global mean surface temperature reconstructions 12,000 years before present.

reconstructions, along with their average, are displayed in Figure 5.5.2.

To apply our Bayesian method for assessment of convergence in these past climate contexts, we first read the datasets in the reverse order, that is,  $\{X_1, X_2, \dots\}$  now stand for the temperatures during progressively past time points. Note that the reconstructions around the present (year 1950) are not quite consistent with the HadCRUT4 temperature around the same year (see Figure 5.5.1). Hence, such reconstructions are perhaps not unquestionable. However, for investigation of the respective series convergence these are unimportant since the first finite number of terms in the series do not influence convergence or divergence of the series.

As before, we set  $L = 11^\circ\text{C}$  and  $U = 16^\circ\text{C}$ , and consider the transformation  $Y_{\theta_0, t} = \log(\log(X_t)) - \log(\log(\theta_0))$ , where  $\theta_0$  takes values in the grid of points obtained by dividing the interval  $[11, 16]$  into equidistant points with common gap 0.1 between any two consecutive points. With  $n_j = 1000$  for  $j = 1, \dots, K = 144$ , and their variations, we obtained  $\sum_{t=1}^{\infty} Y_{\theta_0, t} = \infty$ , for  $\hat{C}_1 \in (0, 10)$ , with respect to each of the 6 time series

shown in Figure 5.5.2. Hence, again we strongly conclude that even Holocene global temperature did not exhibit either of stability, global warming or global cooling, at least in relatively recent past. This is in keeping with our inference regarding future climate change, and hence allows us to conclude that climate dynamics is subject to temporary variations, and long-term global warming or cooling is unlikely in the past as well as in the future.

## 5.6 Summary and conclusion

Fresh investigation of convergence properties of infinite series is an important undertaking in mathematical analysis, since the existing methods for detecting convergence and divergence fail for most infinite series. This, along with our willingness to challenge the ability of the Bayesian paradigm to address series convergence, stimulated us to develop Bayesian characterization of infinite series that indeed attempts to answer such questions of convergence. Our efforts further led to valuable insights regarding the celebrated Riemann Hypothesis. The details are presented in Chapters 3 and 4.

The key idea regarding the above is to embed the deterministic series within a random, stochastic process framework, and hence the Bayesian characterization of Chapter 3 is obviously and directly applicable to random infinite series. Interestingly, the Bayesian procedure is valid irrespective of any dependence structure among the random elements of the series. In this regard, note that the famous Kolmogorov's three series theorem requires independence among the elements.

In practice, success of our Bayesian procedure depends upon creation of efficient upper bounds for the partial sums. For deterministic infinite series the authors show how to achieve such bounds by judiciously exploiting the functional forms of the series elements. However, given any random infinite series, the functional forms of the series elements are of course unknown. For theoretical sake, the marginal distributions of the elements may be assumed known. If the series elements are independent, then Kolmogorov's three

series theorem is applicable in principle to directly assess convergence, but not in the case of dependence. Our Bayesian characterization holds in either case, but practical implementation requires bound construction for the partial sums. As we demonstrated in this chapter, even for known and simple standard distributions, construction of efficient parametric bounds is a highly non-trivial task. Although we could develop mathematically sound parametric upper bounds with non-negative distributional supports of the summands which also performed very well in our simulation experiments, the method of construction of valid parametric upper bounds in general setups still eluded us. The proposed general upper bound (5.2.9) can not be guaranteed to be a theoretically valid upper bound for arbitrary values of the tuning parameter  $a$ . Our properly tuned applications of (5.2.9) to the normal and dependent normal setups indicate correct results on convergence assessment in most cases, but with enormous sample sizes. Another concern is that in the normal based cases, even though the Bayesian algorithm shows eventual upward and downward trends for convergence and divergence respectively, it does not tend close enough to 1 and 0 even with such large sample sizes and run-times to persuasively demonstrate convergence and divergence with (5.2.9). Moreover, for the RDS, wrong convergence results are obtained with the general parametric upper bound in many cases. A further criticism of the parametric upper bound construction methods is that, the forms of  $\Psi_i^{(c)}$  and  $\Psi_i^{(d)}$  employed are too restrictive.

The aforementioned discussion points towards the requirement for constructing more effective and efficient bounds, reminding that parametric bounds can not be constructed in the first place if the underlying distributions are unknown. Indeed, given just the numerical values of the elements of the random series, formation of parametric bounds for the partial sums seems to be infeasible. As such, we propose a nonparametric bound structure for partial sums of general random series, irrespective of known and unknown distributions. The performance of this nonparametric bound structure depends upon the choice of the initial value  $\hat{C}_1$  associated with the first iteration of the Bayesian

algorithm. Experimentation demonstrates that  $\hat{C}_1 = 0.71$  and  $0.725$  are effective starting values for a wide range of random infinite series. As we shall see in Chapter 7, these values are also not much different from those found effective in the time series contexts, where  $\hat{C}_1 = 1$  tuned out to be adequate in all the time series examples considered. It is important to point out that if not much subtlety is required in practice in determination of convergence properties (such as divergence for  $p = 1$  but convergence for  $p = 1 + 0.001$ , many more values of  $\hat{C}_1$  can also be good candidates for our random series setup, and therefore in practice the Bayesian procedure can exhibit considerable robustness with respect to choice of  $\hat{C}_1$ . To obtain  $\hat{C}_1$  in the RDS context, we have demonstrated how the deterministic Dirichlet series can be exploited for our purpose.

Our experiments in the random series context with the nonparametric bound structure persuasively demonstrate correct detection of convergence properties with small sample sizes in all the setups, even in quite subtle situations. Indeed, our experiments reveal that performance of the nonparametric bound is very much comparable with the valid parametric bounds, whenever the latter are available. In the normal and dependent normal setups the nonparametric bound very significantly outperforms the parametric bound in terms of many times smaller sample size, far greater accuracy and huge computational gains. In the RDS setup, the nonparametric bound gives correct and persuasive results for all the cases even for small samples, while the parametric bound yields incorrect answers in many cases. Hence, overall the nonparametric bound quite emphatically outperforms the parametric bounds.

Although infinite series, both deterministic and random, have been topics of interest since ages, their applications in real data contexts are unheard of. This may be due to the reason that real data are always finite while here the topic of discussion is infinite series. However, if assessment of convergence properties is possible even with finitely many series elements, then there is no reason to stay away from relevant real applications. This is what we attempt in this work. With our Bayesian procedure, which assesses

convergence of the underlying infinite series with only a finite number of series elements, we proceed to address past and future climate change, a topic of great relevance and importance in the context of the current global warming scenario and climate change debate. The key issue that makes random infinite series applicable to such analysis is that convergence makes the series elements tend to zero and at fast rate. Exploiting this concept and applying our Bayesian procedure with our nonparametric upper bound for the partial sums on the current global temperature records and Holocene palaeoclimate temperature reconstructions, we obtain results that help us make interesting inferences regarding general global climate dynamics. Specifically, there does not seem to have been instances of prolonged global warming or cooling in the past, and nor such adverse climatic conditions are likely to prevail in the future. Indeed, global climate dynamics is subject to temporary variations only, and the current global warming phenomenon is just an instance of such variation.

# 6

## Bayesian Characterizations of Properties of Stochastic Processes with Applications

### 6.1 Introduction

In various areas of statistics dealing with stochastic processes, ascertainment of stationarity or nonstationarity of the process behind the observed data, is the primary requirement before postulating a stochastic model. In statistics, empirical plots of the data for visualizing stationarity is quite popular, particularly in the time series context. However, it is desirable that rigorous ascertainment of stationarity be carried out via appropriate hypotheses testing procedures. In the parametric time series context, stationarity is usually characterized by specific parameters, and by devising suitable testing

methods, inference regarding stationarity can be obtained. Using the result of such a test, appropriate stationarity or nonstationary models can then be built for statistical analysis of the given data. Although many tests exist in the time series literature, both parametric and nonparametric, they are meant for specific types of time series. In the real data scenario, where the parametric form may itself be called in question, reliability of the tests for stationarity need not be taken for granted.

A very important time series example where studying stationarity property is of utmost importance, is the Markov time series generated by Markov Chain Monte Carlo (MCMC) methods, particularly in the Bayesian posterior context. Although in principle there exist many formal theories for addressing MCMC convergence, they are usually difficult to establish for realistic problems. As a result, plenty of empirical (mostly ad-hoc) methods emerged for diagnosis of convergence of the MCMC sample to the target posterior distribution, and many such methods are based on visualizing the graphical plots of the MCMC sample. The available empirical diagnostic tools have the ill reputation of creating false impressions about convergence or non-convergence in realistic situations.

Compared to the time series literature, tests for stationarity in the spatial and spatio-temporal statistics domains are much less developed, and confined to checking covariance stationarity only, under assumptions that are often difficult to check in practice.

In the point process literature, except some simple tests for complete spatial randomness, there does not seem to exist any formal method to test for Poisson versus non-Poisson point process, or stationarity versus nonstationarity.

Motivated by the aforementioned problems, we seek a general principle that can attempt to effectively address all such issues. Interestingly, the recursive Bayesian idea proposed in Chapter 3 to characterize infinite series, turned out to have fruitful extension to our current situations. Indeed, the recursive Bayesian concept enabled us to study convergence of infinite series whose convergence properties are hitherto unknown. One

such infinite series is also a characterization of the most difficult unsolved problem of mathematics, namely, the Riemann hypothesis. The most surprising result obtained in Chapter 3 is the failure of our methods to accept Riemann hypothesis. Now, since the idea presented in Chapter 3 is primarily about studying deterministic infinite series, one may be left wondering how this can be useful from the statistical perspective. However, the key concept there is to view the deterministic terms of the series as realizations from some general stochastic process, then to relate convergence of the series to a quantity that can be interpreted as probability of convergence of the series under the stochastic process, and finally to build a recursive Bayesian procedure such that the posterior distribution of the probability of convergence tends to one if and only if the series converges and to zero if and only if it diverges.

From the above summary it can be perceived that the deterministic terms of the infinite series can be easily replaced with random elements if necessary. For study of stationarity and nonstationarity, we again relate stationarity to a quantity that admits interpretation as probability that the process is stationary, and apply the same concept of recursive Bayesian method for characterizations of stationarity and nonstationarity. Application of our idea of Bayesian characterization of stationarity and nonstationarity to time series contexts, including MCMC convergence diagnostics, as well as in spatial and spatio-temporal setups, yielded very encouraging results, as reported in Chapters 7 and 8. So did our Bayesian characterizations in the point process scenarios detailed in Chapter 9, where we characterized complete spatial randomness, stationarity and nonstationarity, and the Poisson assumption, via characterization of mutual independence among a set of random variables, using the general principles developed in this chapter. Note that our Bayesian characterization of mutual independence among a set of random variables may also be of general interest.

As we further show in Chapter 10, our Bayesian principle developed in this chapter can be used in another seemingly unrelated setup, namely, determination of frequencies of



oscillations of oscillating stochastic processes. The strategy consists of first transforming the observed data such that the oscillations become as prominent as possible and then relating the proportions of oscillations contained in various sub-intervals as frequencies, which we characterize using the principles developed in this work.

The rest of this chapter is structured as follows. We begin our treatise in Section 6.2 with some necessary definitions and prove results associated with them. With these, we elucidate the key concept behind our proposed ideas in Section 6.3, and then in Section 6.4, we characterize stationarity and nonstationarity using a recursive Bayesian procedure of the same form detailed in Section 3.3. Some relevant computational techniques and their theoretical validation are provided in Section 6.5, and issues related to discretization associated with our method are discussed in Section 6.6. Characterization of second order stationarity, that is stationarity of covariance structure, is considered in Section 6.7. Discussion of the role of non-recursive Bayesian procedures for characterizations is provided in Section 6.8.

## 6.2 Requisite definitions and associated results – prelude to the key concept

Consider a stochastic process  $\mathbf{X} = \{X_s : s \in \mathcal{S}\}$ , where  $\mathcal{S}$  is some arbitrary index set. We assume that  $\mathcal{S} = \cup_{i=1}^{\infty} \mathcal{M}_i$  such that  $\mathcal{M}_i$  are disjoint, and  $\{X_s : s \in \mathcal{M}_i\}$  is stationary. In other words, we assume that  $\mathbf{X}$  is locally stationary. We show below that most stochastic processes are approximately locally stationary. For simplicity of exposition, we consider the case where  $s$  is one-dimensional; the higher-dimensional case is a simple generalization.

**Theorem 19** *For any  $(s_1, \dots, s_m)$ , for  $m \geq 1$ , let  $F_{s_1, \dots, s_m}$  denote the joint distribution function of  $(X_{s_1}, \dots, X_{s_m})$ . Assume that for any  $(x_1, \dots, x_m)$ ,  $F_{s_1, \dots, s_m}$  is differentiable in sufficiently small neighborhoods of  $(x_1, \dots, x_m)$ , and that for  $i = 1, \dots, m$ ,*

$X_{s_i+h} = X_{s_i} + O_P(h)$ , as  $h \rightarrow 0$ . Then for any  $(x_1, \dots, x_m)$ ,  $F_{s_1+h, \dots, s_m+h}(x_1, \dots, x_m) = F_{s_1, \dots, s_m}(x_1, \dots, x_m) + O_P(h)$ , as  $h \rightarrow 0$ .

**Proof.** Let us first assume that  $X_s$  are deterministic variables satisfying  $X_{s_i+h} = X_{s_i} + O(h)$ , as  $h \rightarrow 0$ ,  $i = 1, \dots, m$ . Then by Taylor's series expansion up to the first order, using the above condition, reveals that  $F_{s_1+h, \dots, s_m+h}(x_1, \dots, x_m) = F_{s_1, \dots, s_m}(x_1, \dots, x_m) + O(h)$ . Hence, the result follows by an application of Theorem 7.15 of [Schervish \(1995\)](#). ■

**Remark 20** *The condition  $X_{s+h} = X_s + O_P(h)$ , as  $h \rightarrow 0$  is satisfied by stochastic processes  $X_s$  with almost surely differentiable paths, for example, Gaussian processes, with sufficiently smooth covariance structure (see, for example, [Adler \(1981\)](#), [Adler and Taylor \(2007\)](#)). Also, non-smooth processes that are mean square continuous, in the sense that  $E(X_{s+h} - X_s)^2 \rightarrow 0$ , as  $h \rightarrow 0$ , for any  $s$ , also satisfy the property. Furthermore, discrete processes such as Poisson processes satisfy the above property. Also note that the differentiability condition of  $F_{s_1, \dots, s_m}$  is satisfied by most distribution functions, including the step functions corresponding to discrete distributions.*

Note that local stationarity does not imply that the entire process is even asymptotically stationary. However, as we show below, global stationarity is also possible under our setup. Our goal is to distinguish between global (asymptotic) stationarity and nonstationarity.

For all practical purposes, we shall consider realizations of  $\mathbf{X}$  at discrete index points, that is, points on the set  $\tilde{\mathcal{S}} = \cup_{i=1}^{\infty} \mathcal{N}_i$ , where  $\mathcal{N}_i$  is a discretization of  $\mathcal{M}_i$  and  $\{X_s : s \in \mathcal{N}_i, |\mathcal{N}_i| = n_i\}$ , where  $|\mathcal{N}_i|$  is the cardinality of  $\mathcal{N}_i$ , is stationary. We assume that  $|\mathcal{N}_i| \rightarrow \infty$ , for each  $i$ . In particular, if  $s$  is one-dimensional, then  $\mathcal{N}_i = \{s_r : \sum_{k=1}^{i-1} n_k \leq r \leq \sum_{k=1}^i n_k\}$ , and  $|\mathcal{N}_i| = n_i \rightarrow \infty$  for each  $i$ ; we set  $n_0 = 0$ .

In practice, one can not observe the entire stochastic process  $\mathbf{X}$ , even on the discrete set  $\tilde{\mathcal{S}}$ . Hence, let us assume that only  $\mathbf{X}_K = \{X_s : s \in \cup_{i=1}^K \mathcal{N}_i\}$  has been observed, for sufficiently large  $K$ .

For any Borel set  $C$ , consider

$$\hat{P}_i(C) = n_i^{-1} \sum_{s \in \mathcal{N}_i} I(X_s \in C). \quad (6.2.1)$$

Now let

$$\begin{aligned} \tilde{P}_K(C) &= \frac{\sum_{s \in \cup_{i=1}^K \mathcal{N}_i} I(X_s \in C)}{\sum_{i=1}^K n_i} \\ &= \frac{\sum_{i=1}^K n_i \hat{P}_i(C)}{\sum_{i=1}^K n_i} = \sum_{i=1}^K \hat{p}_{iK} \hat{P}_i(C), \end{aligned} \quad (6.2.2)$$

where  $\hat{p}_{ik} = n_i / \sum_{j=1}^K n_j$ . By the Glivenko-Cantelli theorem for stationary random variables (see [Stute and Schumann \(1980\)](#))

$$\sup_C \left| \hat{P}_i(C) - P_i(C) \right| \xrightarrow{a.s.} 0, \text{ as } n_i \rightarrow \infty, \quad (6.2.3)$$

where  $P_i(C)$  is the probability that any random variable in  $\mathcal{N}_i$  belongs to  $C$ . Note that  $P_i(C)$  may itself be a random variable unless  $\{X_s : s \in \mathcal{N}_i, |\mathcal{N}_i| = n_i\}$  is also ergodic. Randomness of  $P_i(C)$  is not a cause for concern, however, for the methodology that we propose.

Let us now assume that

$$\hat{p}_{iK} = \frac{n_i}{\sum_{j=1}^K n_j} \rightarrow p_{iK} = \frac{p_i}{\sum_{j=1}^K p_j}, \quad (6.2.4)$$

as  $n_j \rightarrow \infty$ , for  $j = 1, \dots, K$ . Here  $0 \leq p_i \leq 1$ , such that  $\sum_{i=1}^K p_i = 1$ .

Let  $P_\infty(C) = \sum_{i=1}^K p_i P_i(C)$ . Then we have the following theorem.

**Theorem 21**

$$\lim_{K \rightarrow \infty} \lim_{n_i \rightarrow \infty, i=1, \dots, K} \sup_C \left| \tilde{P}_K(C) - P_\infty(C) \right| = 0, \text{ almost surely.} \quad (6.2.5)$$

**Proof.**

$$\begin{aligned}
 & \sup_C \left| \tilde{P}_K(C) - P_\infty(C) \right| \\
 &= \sup_C \left| \sum_{i=1}^K \hat{p}_{iK} \hat{P}_i(C) - \sum_{i=1}^K p_i P_i(C) - \sum_{i=K+1}^{\infty} p_i P_i(C) \right| \\
 &\leq \sup_C \left| \sum_{i=1}^K \hat{p}_{iK} \hat{P}_i(C) - \sum_{i=1}^K p_i P_i(C) \right| + \sup_C \left| \sum_{i=K+1}^{\infty} p_i P_i(C) \right| \\
 &\leq \sum_{i=1}^K p_i \left[ \sup_C \left| \hat{P}_i(C) - P_i(C) \right| \right] + \sum_{i=1}^K \left[ \sup_C \hat{P}_i(C) \right] |\hat{p}_{iK} - p_i| + \sum_{i=K+1}^{\infty} p_i \left[ \sup_C P_i(C) \right].
 \end{aligned} \tag{6.2.6}$$

Now, due to (6.2.3), given  $K$ ,

$$\sum_{i=1}^K p_i \left[ \sup_C \left| \hat{P}_i(C) - P_i(C) \right| \right] \rightarrow 0, \text{ almost surely as } n_i \rightarrow \infty, i = 1, \dots, K.$$

Hence,

$$\lim_{K \rightarrow \infty} \lim_{n_i \rightarrow \infty, i=1, \dots, K} \sum_{i=1}^K p_i \left[ \sup_C \left| \hat{P}_i(C) - P_i(C) \right| \right] = 0, \text{ almost surely.} \tag{6.2.7}$$

As  $n_i \rightarrow \infty$  for  $j = 1, \dots, K$  and  $K \rightarrow \infty$ , the second term of (6.2.6) can be shown to converge to zero in the following way:

$$\begin{aligned}
 & \lim_{n_i \rightarrow \infty, i=1, \dots, K} \sum_{i=1}^K \left[ \sup_C \hat{P}_i(C) \right] |\hat{p}_{iK} - p_i| \\
 &\leq \lim_{n_i \rightarrow \infty, i=1, \dots, K} \sum_{i=1}^K |\hat{p}_{iK} - p_i| = \sum_{i=1}^K |p_{iK} - p_i| = \sum_{i=K+1}^{\infty} p_i \rightarrow 0, \text{ as } K \rightarrow \infty.
 \end{aligned} \tag{6.2.8}$$

For the third term of (6.2.6), note that

$$\sum_{i=K+1}^{\infty} p_i \left[ \sup_C P_i(C) \right] \leq \sum_{i=K+1}^{\infty} p_i \rightarrow 0, \text{ as } K \rightarrow \infty. \quad (6.2.9)$$

The result follows by combining (6.2.6), (6.2.7), (6.2.8) and (6.2.9). ■

Note that stationarity of the process  $\mathbf{X}$  is characterized by  $P_i = P$  for  $i = 1, 2, \dots$ , in which case  $P_{\infty} = P$ . Observe that if  $P_i = P_{\infty}$  for  $i = 1, \dots, \infty$ , it then follows that  $P_{\infty} = P$ . Asymptotic stationarity is characterized by  $P_i = P$  for  $i \geq i_0$ , for some  $i_0 > 1$ . In this case, if  $P_j = P_{i_0, \infty} = \frac{\sum_{i=i_0+1}^{\infty} p_i P_i}{\sum_{i=i_0+1}^{\infty} p_i}$ , for  $j > i_0$ , then  $P_i = P$  for  $i > i_0$ . On the other hand, if  $\mathbf{X}$  is nonstationary and not even asymptotically stationary, then  $P_i \neq P_j$  for infinitely many  $j \neq i$ . The latter condition also implies that there does not exist  $i_0 > 1$  such that  $P_j = P_{i_0, \infty}$  for  $j > i_0$ . Hence, there exists no  $i_0 > 1$  such that  $P_i = P$  for  $i > i_0$ .

**Theorem 22**  $\mathbf{X}$  is stationary if and only if for  $i \geq 1$ ,  $\sup_C \left| \hat{P}_i(C) - \tilde{P}_K(C) \right| \rightarrow 0$  almost surely, as  $n_i \rightarrow \infty$  satisfying (6.2.4),  $i = 1, \dots, K$ ,  $K \rightarrow \infty$ .

**Proof.** Note that  $\sup_C \left| \hat{P}_i(C) - \tilde{P}_K(C) \right| \leq \sup_C \left| \hat{P}_i(C) - P_{\infty}(C) \right| + \sup_C \left| \tilde{P}_K(C) - P_{\infty}(C) \right|$ . The first part of the right hand side tends to zero almost surely as  $n_i \rightarrow \infty$  satisfying (6.2.4),  $i = 1, \dots, K$ ,  $K \rightarrow \infty$ , if and only if  $\mathbf{X}$  is stationary, and the second part tends to zero almost surely by Theorem 21. ■

**Theorem 23**  $\mathbf{X}$  is nonstationary if and only if  $\sup_C \left| \hat{P}_i(C) - \tilde{P}_K(C) \right| > 0$  almost surely, as  $n_i \rightarrow \infty$  satisfying (6.2.4),  $i = 1, \dots, K$ ,  $K \rightarrow \infty$ .

**Proof.** Note that

$$\left| \hat{P}_i(C) - \tilde{P}_K(C) \right| \geq \left| \left| \hat{P}_i(C) - P_{\infty}(C) \right| - \left| \tilde{P}_K(C) - P_{\infty}(C) \right| \right|. \quad (6.2.10)$$

By Theorem 21, for any  $\epsilon_1 > 0$ ,

$$\left| \tilde{P}_K(C) - P_\infty(C) \right| < \epsilon_1, \quad (6.2.11)$$

for all  $C$ , for sufficiently large  $n_i$  satisfying (6.2.4) and sufficiently large  $K$ . Also,

$$\left| \hat{P}_i(C) - P_\infty(C) \right| \geq \left| P_i(C) - P_\infty(C) \right| - \left| \hat{P}_i(C) - P_i(C) \right|. \quad (6.2.12)$$

By (6.2.3), for any  $\epsilon_2 > 0$ ,  $\left| \hat{P}_i(C) - P_i(C) \right| < \epsilon_2$ , for all  $C$ , as  $n_i \rightarrow \infty$ . But  $\left| P_i(C) - P_\infty(C) \right| > 0$ , at least for some  $C$ , since  $P_i \neq P_j$  for infinitely many  $j \neq i$ . Since  $\epsilon_2 (> 0)$  is arbitrary, it follows from these arguments and (6.2.12), that

$$\left| \hat{P}_i(C) - P_\infty(C) \right| > 0, \text{ for some } C, \text{ for sufficiently large } n_i. \quad (6.2.13)$$

Since  $\epsilon_1 (> 0)$  in (6.2.11) is also arbitrary, combining (6.2.13), (6.2.11) and (6.2.10) it is evident that the right hand side of (6.2.10) is positive for some  $C$  for sufficiently large  $n_i$  satisfying (6.2.4) and sufficiently large  $K$ . Hence,

$$\sup_C \left| \hat{P}_i(C) - \tilde{P}_K(C) \right| > 0$$

almost surely, as  $n_i \rightarrow \infty$  satisfying (6.2.4),  $i = 1, \dots, K$ ,  $K \rightarrow \infty$ . ■

### 6.3 The key concept

Let  $p_{j,n_j} = P \left( \sup_C \left| \hat{P}_j(C) - \tilde{P}_K(C) \right| \leq c_j \right)$ . As will be seen later, this can be interpreted as the probability that the underlying process is stationary when the observed data is  $\mathbb{I} \left\{ \sup_C \left| \hat{P}_j(C) - \tilde{P}_K(C) \right| \leq c_j \right\}$ . Note that, for stationarity, due to Theorem 22, for  $j = 1, \dots, K$ , as  $n_j \rightarrow \infty$ ,  $K \rightarrow \infty$ , the latter converges to one almost surely. Since  $p_{j,n_j} = E \left[ \mathbb{I} \left\{ \sup_C \left| \hat{P}_j(C) - \tilde{P}_K(C) \right| \leq c_j \right\} \right]$ , uniform integrability leads one to expect

that for  $j \geq 1$ , for any choice of the non-negative monotonically decreasing sequence  $\{c_j\}_{j=1}^{\infty}$ ,

$$\begin{aligned} & \lim_{K \rightarrow \infty} \lim_{n_j \rightarrow \infty, j=1, \dots, K} p_{j, n_j} \\ &= \lim_{K \rightarrow \infty} \lim_{n_j \rightarrow \infty, j=1, \dots, K} P \left( \sup_C \left| \hat{P}_j(C) - \tilde{P}_K(C) \right| \leq c_j \right) \\ &= \lim_{K \rightarrow \infty} \lim_{n_j \rightarrow \infty, j=1, \dots, K} E \left[ \mathbb{I} \left\{ \sup_C \left| \hat{P}_j(C) - \tilde{P}_K(C) \right| \leq c_j \right\} \right] \\ &= 1. \end{aligned}$$

Similarly, for nonstationarity, we expect, using Theorem 23 that for  $j \geq j_0 \geq 1$ ,

$$\lim_{K \rightarrow \infty} \lim_{n_j \rightarrow \infty, j=1, \dots, K} p_{j, n_j} = 0$$

almost surely, for any choice of the non-negative monotonically decreasing sequence  $\{c_j\}_{j=1}^{\infty}$ .

In reality it is not known if  $p_{j, n_j}$  converges to zero or one, since it is not known if  $\mathbf{X}$  is stationary or nonstationary. Thus, we consider learning about  $p_{j, n_j}$  from the data  $\mathbf{X}_K$  and some appropriate prior on  $p_{j, n_j}$  in the form of the posterior  $\pi(p_{j, n_j} | \mathbf{X}_K)$ . As we will show,

$$\lim_{K \rightarrow \infty} \lim_{n_j \rightarrow \infty, j=1, \dots, K} \pi(p_{j, n_j} | \mathbf{X}_K) = 1, \text{ almost surely}$$

for  $j \geq 1$  and any choice of the non-negative monotonically decreasing sequence  $\{c_j\}_{j=1}^{\infty}$ , characterizes stationarity of  $\mathbf{X}$  and

$$\lim_{K \rightarrow \infty} \lim_{n_j \rightarrow \infty, j=1, \dots, K} \pi(p_{j, n_j} | \mathbf{X}_K) = 0, \text{ almost surely}$$

for  $j \geq j_0 \geq 1$ , for any choice of the non-negative monotonically decreasing sequence  $\{c_j\}_{j=1}^{\infty}$ , characterizes nonstationarity of  $\mathbf{X}$ .

## 6.4 Characterization of stationarity properties of the underlying process

Let  $\{c_j\}_{j=1}^{\infty}$  be a non-negative decreasing sequence and

$$Y_{j,n_j} = \mathbb{I}\left\{\sup_C \left| \hat{P}_j(C) - \tilde{P}_K(C) \right| \leq c_j\right\}. \quad (6.4.1)$$

Let, for  $j \geq 1$ ,

$$P(Y_{j,n_j} = 1) = p_{j,n_j}. \quad (6.4.2)$$

Hence, the likelihood of  $p_{j,n_j}$ , given  $y_{j,n_j}$ , is given by

$$L(p_{j,n_j}) = p_{j,n_j}^{y_{j,n_j}} (1 - p_{j,n_j})^{1-y_{j,n_j}} \quad (6.4.3)$$

It is important to relate  $p_{j,n_j}$  to stationarity of the underlying series. Note that  $p_{j,n_j}$  is the probability that  $\sup_C \left| \hat{P}_j(C) - \tilde{P}_K(C) \right|$  falls below  $c_j$ . Thus,  $p_{j,n_j}$  can be interpreted as the probability that the process  $\mathbf{X}$  is stationary when the data observed is  $Y_{j,n_j}$ . If  $\mathbf{X}$  is stationary, then due to Theorem 22 it is to be expected *a posteriori*, that for  $j \geq 1$ , for any non-negative decreasing sequence  $\{c_j\}_{j=1}^{\infty}$ ,

$$p_{j,n_j} \rightarrow 1 \quad \text{as } n_j \rightarrow \infty, \text{ satisfying (6.2.4)}. \quad (6.4.4)$$

Indeed, as we will formally show, condition (6.4.4) is both necessary and sufficient for stationarity of  $\mathbf{X}$ .

On the other hand, if  $\mathbf{X}$  is nonstationary, then there exists  $j_0 \geq 1$  such that for every  $j > j_0$ , as  $n_j \rightarrow \infty$  satisfying (6.2.4),  $\sup_C \left| \hat{P}_j(C) - \tilde{P}_K(C) \right| > c_j$ , for any non-negative decreasing sequence  $\{c_j\}_{j=1}^{\infty}$ , due to Theorem 23. Here we expect, *a posteriori*, that

$$p_{j,n_j} \rightarrow 0 \quad \text{as } n_j \rightarrow \infty, \text{ satisfying (6.2.4)}, \quad (6.4.5)$$



for  $j \geq j_0 \geq 1$ . Again, we will prove formally that the above condition is both necessary and sufficient for divergence.

We assume that  $\{y_{j,n_j}; j = 1, 2, \dots\}$  is observed successively at stages indexed by  $j$ . That is, we first observe  $y_{1,n_1}$ , and based on our prior belief regarding the first stage probability,  $p_{1,n_1}$ , compute the posterior distribution of  $p_{1,n_1}$  given  $y_{1,n_1}$ , which we denote by  $\pi(p_{1,n_1}|y_{1,n_1})$ . Based on this posterior we construct a prior for the second stage, and compute the posterior  $\pi(p_{2,n_2}|y_{2,n_2})$ . We continue this procedure for as many stages as we desire. The details remain the same as Section 3.3.2.

Based on our recursive Bayesian theory we have the following theorem that characterizes stationarity of  $\mathbf{X}$  in terms of the limit of the posterior probability of  $p_{k,n_k}$ , as  $n_k \rightarrow \infty$  satisfying (6.2.4) and  $K \rightarrow \infty$ . We also assume, for the sake of generality, that for any  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ , where  $\mathfrak{N} (\subset \mathfrak{S})$  has zero probability measure, the non-negative monotonically decreasing sequence  $\{c_j\}_{j=1}^\infty$  depends upon  $\omega$ , so that we shall denote the sequence by  $\{c_j(\omega)\}_{j=1}^\infty$ . In other words, we allow  $\{c_j(\omega)\}_{j=1}^\infty$  to depend upon the corresponding data  $X(\omega)$ . Since  $\sup_C \left| \hat{P}_j(C) - \tilde{P}_K(C) \right| \leq 1$  and tends to zero in the case of stationarity, there exists a monotonically decreasing sequence  $\{c_j(\omega)\}_{j=1}^\infty$  such that for  $n_j; j = 1, \dots, K$  sufficiently large satisfying (6.2.4),

$$\sup_C \left| \hat{P}_j(C)(\omega) - \tilde{P}_K(C)(\omega) \right| \leq c_j(\omega), \text{ for } j \geq 1. \quad (6.4.6)$$

**Theorem 24** For all  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ , where  $\mathfrak{N}$  is some null set having probability measure zero,  $\mathbf{X}$  is stationary if and only if for any monotonically decreasing sequence  $\{c_j(\omega)\}_{j=1}^\infty$ ,

$$\pi(\mathcal{N}_1|y_{k,n_k}(\omega)) \rightarrow 1, \quad (6.4.7)$$

as  $k \rightarrow \infty$  and  $n_j \rightarrow \infty$  for  $j = 1, \dots, K$  satisfying (6.2.4) and  $K \rightarrow \infty$ , where  $\mathcal{N}_1$  is any neighborhood of 1 (one).

**Proof.** Let, for  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ , where  $\mathfrak{N}$  is some null set having probability measure

zero,  $\mathbf{X}$  be stationary. Then, by (6.4.6),  $\sup_C \left| \hat{P}_j(C)(\omega) - \tilde{P}_K(C)(\omega) \right| \leq c_j(\omega)$  for  $n_j$  sufficiently large satisfying (6.2.4), given any choice of the monotonically decreasing sequence  $\{c_j(\omega)\}_{j=1}^\infty$ . Hence,  $y_{j,n_j}(\omega) = 1$  for sufficiently large  $n_j$ , satisfying (6.2.4), for  $j \geq 1$ . Hence, in this case,  $\sum_{j=1}^k y_{j,n_j}(\omega) = k$ , Also,  $\sum_{j=1}^k \frac{1}{j^2} \rightarrow \frac{\pi^2}{6}$ , as  $k \rightarrow \infty$ . Consequently, it is easy to see that

$$\mu_k = E(p_{k,n_k} | y_{k,n_k}(\omega)) \sim \frac{\frac{\pi^2}{6} + k}{k + \frac{\pi^2}{3}} \rightarrow 1, \text{ as } k \rightarrow \infty, \text{ and,} \quad (6.4.8)$$

$$\sigma_k^2 = Var(p_{k,n_k} | y_{k,n_k}(\omega)) \sim \frac{(\frac{\pi^2}{6} + k)(\frac{\pi^2}{6})}{(k + \frac{\pi^2}{3})^2(1 + k + \frac{\pi^2}{3})} \rightarrow 0 \text{ as } k \rightarrow \infty. \quad (6.4.9)$$

In the above, for any two sequences  $\{a_k\}_{k=1}^\infty$  and  $\{b_k\}_{k=1}^\infty$ ,  $a_k \sim b_k$  indicates  $\frac{a_k}{b_k} \rightarrow 1$ , as  $k \rightarrow \infty$ . Now let  $\mathcal{N}_1$  denote any neighborhood of 1, and let  $\epsilon > 0$  be sufficiently small such that  $\mathcal{N}_1 \supseteq \{1 - p_{k,n_k} < \epsilon\}$ . Combining (6.4.8) and (6.4.9) with Chebychev's inequality ensures that (6.4.7) holds.

Now assume that (6.4.7) holds. Then for any given  $\epsilon > 0$ ,

$$\pi(p_{k,n_k} > 1 - \epsilon | y_{k,n_k}(\omega)) \rightarrow 1, \text{ as } k \rightarrow \infty. \quad (6.4.10)$$

Hence,

$$E(p_{k,n_k} | y_{k,n_k}(\omega)) \rightarrow 1; \quad (6.4.11)$$

$$Var(p_{k,n_k} | y_{k,n_k}(\omega)) \rightarrow 0, \quad (6.4.12)$$

as  $k \rightarrow \infty$ . If  $\mathbf{X}$  is nonstationary, then there exists  $j_0(\omega)$  such that for each  $j \geq j_0(\omega)$ , for sufficiently large  $n_j$  satisfying  $\sup_C \left| \hat{P}_j(C)(\omega) - \tilde{P}_K(C)(\omega) \right| > c_j(\omega)$ , for  $j \geq j_0(\omega)$ , for any choice of non-negative sequence  $\{c_j(\omega)\}_{j=1}^\infty$  monotonically converging to zero. Hence, in this situation,  $0 \leq \sum_{j=1}^k y_{j,n_j}(\omega) \leq j_0(\omega)$ . Substituting this in (3.3.14) and

(3.3.15), it is easy to see that, as  $k \rightarrow \infty$ ,

$$E(p_{k,n_k}|y_{k,n_k}(\omega)) \rightarrow 0; \quad (6.4.13)$$

$$\text{Var}(p_{k,n_k}|y_{k,n_k}(\omega)) \rightarrow 0, \quad (6.4.14)$$

so that (6.4.11) is contradicted.

■

We now prove the following theorem that provides necessary and sufficient conditions for nonstationarity of  $\mathbf{X}$  in terms of the limit of the posterior probability of  $p_{k,n_k}(\omega)$ , as  $n_k \rightarrow \infty$  satisfying (6.2.4).

**Theorem 25**  *$\mathbf{X}$  is nonstationary if and only if for any  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$  where  $\mathfrak{N}$  is some null set having probability measure zero, for any choice of the non-negative, monotonically decreasing sequence  $\{c_j(\omega)\}_{j=1}^\infty$ ,*

$$\pi(\mathcal{N}_0|y_{k,n_k}(\omega)) \rightarrow 1, \quad (6.4.15)$$

as  $k \rightarrow \infty$  and  $n_j \rightarrow \infty$ ,  $j = 1, \dots, K$  satisfying (6.2.4), and  $K \rightarrow \infty$ , where  $\mathcal{N}_0$  is any neighborhood of 0 (zero).

**Proof.** Assume that  $\mathbf{X}$  is nonstationary. Then there exists  $j_0(\omega) \geq 1$  such that for every  $j \geq j_0(\omega)$ ,  $\sup_C |\hat{P}_j(C)(\omega) - \tilde{P}_K(C)(\omega)| > c_j(\omega)$ , for sufficiently large  $n_j$ , for any choice of non-negative sequence  $\{c_j(\omega)\}_{j=1}^\infty$  monotonically converging to zero. From the proof of the sufficient condition of Theorem 24 it follows that (6.4.13) and (6.4.14) hold. Let  $\epsilon > 0$  be small enough so that  $\mathcal{N}_0 \supseteq \{p_{k,n_k} < \epsilon\}$ . Then combining Chebychev's inequality with (6.4.13) and (6.4.14) it is easy to see that (6.4.15) holds.

Now assume that (6.4.15) holds. Then for any given  $\epsilon > 0$ ,

$$\pi(p_{k,n_k} < \epsilon|y_{k,n_k}(\omega)) \rightarrow 1, \text{ as } k \rightarrow \infty. \quad (6.4.16)$$

It follows that

$$E(p_{k,n_k}|y_{k,n_k}(\omega)) \rightarrow 0; \quad (6.4.17)$$

$$Var(p_{k,n_k}|y_{k,n_k}(\omega)) \rightarrow 0, \quad (6.4.18)$$

as  $k \rightarrow \infty$ .

If  $\mathbf{X}$  is stationary, then by Theorem 24,  $\pi(\mathcal{N}_1|y_{k,n_k}(\omega)) \rightarrow 1$  as  $k \rightarrow \infty$ , for all sequences  $\{n_j\}_{j=1}^{\infty}$ , so that  $E(p_{k,n_k}|y_{k,n_k}(\omega)) \rightarrow 1$ , which is a contradiction to (6.4.17).

■

## 6.5 Computation of the sup norm between empirical distribution functions associated with $\hat{P}_j$ and $\tilde{P}_K$

In all practical applications that involves identifying stationarity or nonstationarity by our method, it is needed to compute the sup norms  $\sup_C |\hat{P}_j(C) - \tilde{P}_K(C)|$ ;  $j \geq 1$ . For this purpose, it is sufficient to compute  $\sup_{-\infty < x < \infty} |\hat{F}_j(x) - \tilde{F}_K(x)|$ , where  $\hat{F}_j(x)$  and  $\tilde{F}_K(x)$  stand for the empirical distribution functions corresponding to  $\hat{P}_j$  and  $\tilde{P}_K$ . Lemma 26 provides the formula for the desired sup norm.

**Lemma 26** *Let  $\hat{F}_j(x)$  and  $\tilde{F}_K(x)$  denote the empirical distribution functions corresponding to empirical probability distributions  $\hat{P}_j$  and  $\tilde{P}_K$ , respectively. Then it holds that*

$$\sup_{-\infty < x < \infty} |\hat{F}_j(x) - \tilde{F}_K(x)| = 1 - \tilde{F}_K(\hat{x}_j), \quad (6.5.1)$$

where  $\hat{x}_j = \max \mathcal{N}_j$ , provided that  $\hat{x}_j \neq \max \{\cup_{k=1}^K \mathcal{N}_k\}$ .

**Proof.** Since both  $\hat{F}_j(x)$  and  $\tilde{F}_K(x)$  are empirical distribution functions, their jumps occur at the order statistics associated with the sample data. Now, by inspection it can

be seen that, if  $\hat{x}_j \neq \max \{\cup_{k=1}^K \mathcal{N}_k\}$ , then

$$|\hat{F}_j(\hat{x}_j) - \tilde{F}_K(\hat{x}_j)| = 1 - \tilde{F}_K(\hat{x}_j). \quad (6.5.2)$$

For the  $r$ -th order statistic value  $x_{(t)}$ ,  $t \geq 1$  such that  $x_{(t)} \neq \hat{x}_j$ ,  $|\hat{F}_j(\hat{x}_j) - \tilde{F}_K(\hat{x}_j)|$  is of the form  $\left| \frac{\ell}{n_j} - \frac{r}{\sum_{k=1}^K n_k} \right|$ , where  $1 < \ell < n_j$ ,  $1 < r < \sum_{k=1}^K n_k$ . But, for  $1 \leq m \leq \sum_{k=1}^K n_k$ ,

$$1 - \frac{m}{\sum_{k=1}^K n_k} \geq \left| \frac{\ell}{n_j} - \frac{r}{\sum_{k=1}^K n_k} \right|. \quad (6.5.3)$$

Since  $1 - \tilde{F}_K(\hat{x}_j)$  in (6.5.2) is of the form  $1 - \frac{m}{\sum_{k=1}^K n_k}$ , it follows from (6.5.3) that (6.5.1) holds. ■

**Remark 27** Lemma 26 gives the formula for the sup norm when  $\hat{x}_j \neq \max \{\cup_{k=1}^K \mathcal{N}_k\}$ . In fact, (6.5.1) is no longer valid when  $\hat{x}_j = \max \{\cup_{k=1}^K \mathcal{N}_k\}$ . Note that there exists exactly one  $k \geq 1$  such that  $\hat{x}_{j^*} = \max \{\cup_{k=1}^K \mathcal{N}_k\}$ . For that  $j^*$ , there is no direct formula for the sup norm, and it is desirable to compute the sup norm by evaluating the differences between the empirical distribution functions at all the sample order statistics. However, just for a single  $k$ , such elaborate computation is not worthwhile. Instead it makes sense to construct  $\hat{F}_{j^*}$  based on all the observations in  $\mathcal{N}_{j^*}$  except  $\hat{x}_{j^*}$ . Hence, if  $\tilde{x}_{j^*}$  is the maximum of  $\mathcal{N}_{j^*} \setminus \{\hat{x}_{j^*}\}$ , then in that case,  $\sup_{-\infty < x < \infty} |\hat{F}_{j^*}(x) - \tilde{F}_K(x)| = 1 - \tilde{F}_K(\tilde{x}_{j^*})$ , which is what we shall use in our practical applications.

## 6.6 Choice of the cardinality of $\mathcal{N}_i$

An important ingredient of our method, particularly tied to practical implementation, is the choice of the number of random variables in the sets  $\mathcal{N}_i$ . Recall that  $\mathcal{N}_i$  is discretization of an index set  $\mathcal{M}_i$ , on which  $s$  varies continuously, such that  $\{X_s : s \in \mathcal{M}_i\}$  is stationary. Let the closure of  $\mathcal{M}_i$ , denoted by  $\overline{\mathcal{M}_i}$ , be compact.

Let the index  $s \in \mathbb{R}^p$ , for  $p \geq 1$ . For  $j = 1, 2, \dots$ , consider  $p$ -dimensional balls  $B_p(c_j, r)$  with centers  $c_j$  and radius  $r > 0$  such that for any  $s \in \overline{\mathcal{M}_i}$ , there exists  $j \geq 1$  such that  $s \in B_p(c_j, \epsilon)$ . Then the set  $\{B_p(c_j, \epsilon) : j \geq 1\}$  constitutes an open cover for  $\overline{\mathcal{M}_i}$ . By compactness, there exists a set  $\{B_p(c_{j_k}, \epsilon) : k = 1, \dots, n_i\}$ , for finite  $n_i \geq 1$  such that  $\overline{\mathcal{M}_i} \subseteq \cup_{k=1}^{n_i} B_p(c_{j_k}, \epsilon)$ . It follows that

$$\text{Vol}(\overline{\mathcal{M}_i}) \leq \sum_{k=1}^{n_i} \text{Vol}(B_p(c_{j_k}, \epsilon)), \quad (6.6.1)$$

where for any set  $S$ ,  $\text{Vol}(S)$  denotes the volume of  $S$ . Since  $\text{Vol}(B_p(c_{j_k}, \epsilon)) = \text{Vol}(B_p(\mathbf{0}, \epsilon))$ , the  $p$ -dimensional ball with center  $\mathbf{0}$ , and since  $\text{Vol}(B_p(\mathbf{0}, \epsilon)) = \frac{\pi^{p/2}}{\Gamma(p/2+1)} \epsilon^p$ , it follows from (6.6.1) that

$$n_i \geq \left( \frac{\text{Vol}(\overline{\mathcal{M}_i})}{\epsilon^p} \right) \left( \frac{\Gamma(p/2+1)}{\pi^{p/2}} \right). \quad (6.6.2)$$

For example, if  $\mathcal{M}_i$  is a  $p$ -dimensional hypercube with  $c_i (> 0)$  being the length of each edge, then it follows from (6.6.2) that  $n_i \geq \left(\frac{c_i}{\epsilon}\right)^p \left(\frac{\Gamma(p/2+1)}{\pi^{p/2}}\right)$ . For example, if  $p = 1$  and  $c = 3\epsilon$ , then  $n \geq 1.5$ ; if  $p = 2$  and  $c = 3\epsilon$ , then  $n \geq 2.865$ ;  $p = 3$  and  $c = 3\epsilon$ , implies  $n \geq 6.446$ , etc. Similar idea has been considered in Section 1.2.1 of Giraud (2015), in the context of large  $p$ . In our illustrations, the total number of observations are allocated to a substantially large number of cubes of dimensions one, two and three. Consequently,  $c/\epsilon$  is not expected to be significantly larger than one. As such, we take care such that the cube containing the minimum number of observations has at least three observations.

## 6.7 Stationarity of covariance structure

Let  $Y_{(s_1, s_2)} = X_{s_1} X_{s_2}$ ,  $\mathcal{N}_{ih} = \{(s_1, s_2) \in \mathcal{N}_i : \|s_1 - s_2\| = h\}$ , and  $n_{ih} = |\mathcal{N}_{ih}|$ .

$$\widehat{Cov}_{ih} = \frac{\sum_{(s_1, s_2) \in \mathcal{N}_{ih}} Y_{(s_1, s_2)}}{2n_{ih}} - \left( \frac{\sum_{s_1 \in \mathcal{N}_{ih}} X_{s_1}}{n_{ih}} \right) \left( \frac{\sum_{s_2 \in \mathcal{N}_{ih}} X_{s_2}}{n_{ih}} \right). \quad (6.7.1)$$

Noting that  $Y_{(s_1, s_2)}$ , where  $(s_1, s_2) \in \mathcal{N}_i$ , is stationary, it follows by the ergodic theorem that

$$\widehat{Cov}_{ih} \xrightarrow{a.s.} Cov_{ih} = Cov(X_{s_1}, X_{s_2}) \text{ where } \|s_1 - s_2\| = h. \quad (6.7.2)$$

Let

$$\widetilde{Cov}_{Kh} = \sum_{i=1}^K \tilde{p}_{iKh} \widehat{Cov}_{ih}, \quad (6.7.3)$$

where  $\tilde{p}_{iKh} = n_{ih} / \sum_{j=1}^K n_{jh}$ , with  $\sum_{i=1}^K p_{ih} = 1$ , and

$$Cov_{\infty, h} = \sum_{i=1}^{\infty} \tilde{p}_{ih} Cov_{ih}, \quad (6.7.4)$$

We assume that

$$\tilde{p}_{iKh} \rightarrow p_{iKh} = \frac{p_{ih}}{\sum_{j=1}^K p_{jh}}, \text{ as } n_{ih} \rightarrow \infty; i = 1, \dots, K. \quad (6.7.5)$$

**Theorem 28** *Let*

$$\sum_{i=1}^{\infty} p_{ih} |Cov_{ih}| < \infty. \quad (6.7.6)$$

*Then*

$$\lim_{K \rightarrow \infty} \lim_{n_{ih} \rightarrow \infty; i=1, \dots, K} \left| \widetilde{Cov}_{Kh} - Cov_{\infty, h} \right| = 0. \quad (6.7.7)$$

**Proof.**

$$\left| \widetilde{Cov}_{Kh} - Cov_{\infty, h} \right| \leq \sum_{i=1}^K \left| \widehat{Cov}_{ih} \right| \left| \tilde{p}_{iKh} - p_{ih} \right| + \sum_{i=1}^K p_{ih} \left| \widehat{Cov}_{ih} - Cov_{ih} \right| + \sum_{K+1}^{\infty} p_i |Cov_{ih}|. \quad (6.7.8)$$

Due to (6.7.5),  $\sum_{i=1}^K \left| \widehat{Cov}_{ih} \right| \left| \tilde{p}_{iKh} - p_{ih} \right| \rightarrow \sum_{i=1}^K |Cov_{ih}| |p_{iKh} - p_{ih}|$  as  $n_{ih} \rightarrow \infty$ ;  $i = 1, \dots, K$ . Due to (6.7.6),  $|Cov_{ih}| < L$ , for some  $L > 0$ , for all  $i \geq 1$ . Hence, the first term on the right hand side of (6.7.8) is bounded above by  $L \sum_{j=K+1}^{\infty} p_j$ , which tends to zero, as  $K \rightarrow \infty$ , since  $\sum_{i=1}^{\infty} p_i = 1$ .

Using (6.7.2), it is seen that the second term of the right hand side of (6.7.8) also tends to zero as  $n_{ih} \rightarrow \infty$ ;  $i = 1, \dots, K$ , satisfying (6.7.5) and as  $K \rightarrow \infty$ .

The last term on the right hand side of (6.7.8) tends to zero as  $K \rightarrow \infty$  due to (6.7.6).

■

Note that the covariance structure of  $\mathbf{X}$  is stationary if and only if  $Cov_{ih} = Cov_{\infty,h}$  for all  $i \geq 1$  and all  $h > 0$ , and is nonstationary if and only if  $Cov_{ih} \neq Cov_{\infty,h}$  for all  $i \geq 1$  for some  $h > 0$ .

**Theorem 29** *The covariance structure of  $\mathbf{X}$  is stationary if and only if for  $i \geq 1$ , for all  $h > 0$ ,*

$$\lim_{K \rightarrow \infty} \lim_{n_{jh} \rightarrow \infty; j=1, \dots, K} \left| \widehat{Cov}_{ih} - \widetilde{Cov}_{Kh} \right| = 0.$$

**Proof.** Using Theorem 28, the proof follows in the same way as the proof of Theorem 22, with the probabilities replaced with the respective covariances.

■

**Theorem 30** *The covariance structure of  $\mathbf{X}$  is nonstationary if and only if for  $i \geq 1$ , for some  $h > 0$ ,*

$$\lim_{K \rightarrow \infty} \lim_{n_{jh} \rightarrow \infty; j=1, \dots, K} \left| \widehat{Cov}_{ih} - \widetilde{Cov}_{Kh} \right| > 0.$$

**Proof.** Using Theorem 28, the proof follows in the same way as the proof of Theorem 23, with the probabilities replaced with the respective covariances. ■

Now define  $Y_{j,n_{jh}} = \mathbb{I} \left\{ \left| \widehat{Cov}_{ih} - \widetilde{Cov}_{Kh} \right| < c_{jh} \right\}$ . Then the following characterization theorems hold, the proofs of which are the similar to those of Theorems 24 and 25.

**Theorem 31** *For all  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ , where  $\mathfrak{N}$  is some null set having probability measure zero,  $\mathbf{X}$  is stationary if and only if for any  $h > 0$ , there exists a monotonically decreasing sequence  $\{c_{jh}(\omega)\}_{j=1}^{\infty}$  such that*

$$\pi(\mathcal{N}_1 | y_{k,n_{kh}}(\omega)) \rightarrow 1, \tag{6.7.9}$$



as  $k \rightarrow \infty$  and  $n_{jh} \rightarrow \infty$  for  $j = 1, \dots, K$  satisfying (6.2.4) and  $K \rightarrow \infty$ , where  $\mathcal{N}_1$  is any neighborhood of 1 (one).

**Theorem 32**  $\mathbf{X}$  is nonstationary if and only if for some  $h > 0$ , and for any  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$  where  $\mathfrak{N}$  is some null set having probability measure zero, for any choice of the non-negative, monotonically decreasing sequence  $\{c_{jh}(\omega)\}_{j=1}^\infty$ ,

$$\pi(\mathcal{N}_0 | y_{k, n_{kh}}(\omega)) \rightarrow 1, \quad (6.7.10)$$

as  $k \rightarrow \infty$  and  $n_{jh} \rightarrow \infty$ ,  $j = 1, \dots, K$  satisfying (6.2.4), and  $K \rightarrow \infty$ , where  $\mathcal{N}_0$  is any neighborhood of 0 (zero).

## 6.8 Characterization of stationarity and nonstationarity using non-recursive Bayesian posteriors

Observe that it is not strictly necessary for the prior at any stage to depend upon the previous stage. Indeed, we may simply assume that  $\pi(p_{j, n_j}) \equiv \text{Beta}(\alpha_j, \beta_j)$ , for  $j = 1, 2, \dots$ . In this case, the posterior of  $p_{k, n_k}$  given  $y_{k, n_k}$  is simply  $\text{Beta}(\alpha_k + y_{k, n_k}, 1 + \beta_k - y_{k, n_k})$ . The posterior mean and variance are then given by

$$E(p_{k, n_k} | y_{k, n_k}(\omega)) = \frac{\alpha_k + y_{k, n_k}(\omega)}{1 + \alpha_k + \beta_k}; \quad (6.8.1)$$

$$\text{Var}(p_{k, n_k} | y_{k, n_k}(\omega)) = \frac{(\alpha_k + y_{k, n_k}(\omega))(1 + \beta_k - y_{k, n_k}(\omega))}{(1 + \alpha_k + \beta_k)^2(2 + \alpha_k + \beta_k)}. \quad (6.8.2)$$

Since  $y_{k, n_k}(\omega)$  (or  $y_{k, n_{kh}}(\omega)$ ) converges to 1 or 0 as  $n_k \rightarrow \infty$ , accordingly as  $\mathbf{X}$  is stationary or nonstationary (or the covariance structure of  $\mathbf{X}$  is stationary or nonstationary), it is easily seen, provided that  $\alpha_k \rightarrow 0$  and  $\beta_k \rightarrow 0$  as  $k \rightarrow \infty$ , that (6.8.1) converges to 1 (respectively, 0) if and only if  $\mathbf{X}$  is (covariance) stationary (respectively, (covariance) nonstationary). Importantly, if we choose  $\alpha_k = \beta_k = 0$  for all  $k \geq 1$ , then  $k \rightarrow \infty$  is no

longer needed, and the results continue to hold if  $n_k \rightarrow \infty$ .

Thus, characterization of stationarity or nonstationarity of  $\mathbf{X}$  is possible even with the non-recursive approach. Indeed, note that the prior parameters  $\alpha_k$  and  $\beta_k$  are more flexible compared to those associated with the recursive approach. This is because, in the non-recursive approach we only require  $\alpha_k \rightarrow 0$  and  $\beta_k \rightarrow 0$  as  $k \rightarrow \infty$ , so that convergence of the series  $\sum_{j=1}^{\infty} \alpha_j$  and  $\sum_{j=1}^{\infty} \beta_j$  are not necessary, unlike the recursive approach. However, choosing  $\alpha_k$  and  $\beta_k$  to be of sufficiently small order ensures much faster convergence of the posterior mean and variance as compared to the recursive approach.

Unfortunately, an important drawback of the non-recursive approach is that it does not admit extension to the case of general oscillatory stochastic processes. On the other hand, as we show subsequently, the principles of our recursive theory can be easily adopted to develop a Bayesian theory for determining (multiple) frequencies of oscillating stochastic processes. In other words, the recursive approach seems to be more powerful from the perspective of development of a general Bayesian principle for learning about the basic characteristics of the underlying stochastic process. Moreover, as our examples demonstrate, the recursive posteriors converge sufficiently fast to the correct degenerate distributions, obviating the need to consider the non-recursive approach. Consequently, we do not further pursue the non-recursive approach, as before.

## 6.9 Summary and conclusion

The main purpose of this chapter is to propose and develop a key idea based on the principles of our original Bayesian characterization concept of infinite series (Chapters 3 and 5) that promises to unify various areas of statistics that deal with properties of stochastic processes. Our motivation for this work is derived from the dearth of statistical tests for stationarity and nonstationarity in important areas of statistics such as time series, spatial and spatio-temporal processes and point processes. As we

elucidated, the point process area requires development with respect to characterizations of complete spatial randomness and the Poisson assumption, apart from stationarity and nonstationarity. The general developments presented in this chapter serve as prelude to our specific developments with respect to the above areas in the later chapters, namely, Chapters 7, 8 and 9. Furthermore, adopting these ideas, we develop a novel Bayesian characterization theory in the context of frequency determination in oscillatory stochastic processes with considerable applications in periodic time series.

An important goal of our research is to render our characterization theories amenable to practical applications. To this end, in the later chapters, we shall provide ample illustrations of our methods and implementations with simulated and real data sets, in each of the aforementioned areas of statistics. Most of our codes are written in C, parallelised using MPI (Message Passing Interface), and implemented in parallel architectures. Some parallelized R codes are also used in conjunction with our parallel C codes. Very fast and efficient computation is the result of our efforts.

# 7

## Application of Bayesian Characterization of Stationarity and Nonstationarity to Time Series and Markov Chain Monte Carlo

### 7.1 Introduction

In statistics, the importance of time series analysis is undeniable, from both theoretical and application perspectives. Applications of time series stretches to almost all branches of science; economics, medicine, biology, physics, environment, ecology, engineering, to name few. Comprehensive theoretical treatises on time series analysis can be found in [Hamilton \(1994\)](#), [Brockwell and Davis \(2002\)](#) and [Brockwell and Davis \(2009\)](#), while [Montgomery \*et al.\* \(2016\)](#), [Hyndman and Athanasopoulos \(2018\)](#), [Chatfield and Xing](#)

(2002) focus on applications. A balanced blend of theory and applications is offered by Shumway and Stoffer (2006).

For fitting any stochastic process to the given time series data, the important task of ascertaining stationarity or nonstationarity of the underlying data-generating time series, must be undertaken. The appropriate stationary or nonstationary process can then be selected for the purpose at hand, given the observed data. Although in some Bayesian hierarchical model scenarios priors can be assigned to the parameters controlling stationarity (for example, the process parameter of the well-known AR(1) model, the first order auto-regressive model) such that *a priori* both stationarity and nonstationarity are considered plausible, in the classical time series modeling it is required to ascertain stationarity or nonstationarity before postulating a model.

The above discussion points towards the importance of the existence of appropriate tests for stationarity in the time series literature. However, somewhat surprisingly, except for some specific instances of parametric and nonparametric model setups (see, for example, Dickey and Fuller (1979), Kwiatkowski *et al.* (1992), Philips and Perron (1988), Breitung (2002), Basu *et al.* (2009), Cardinali and Nason (2018), van Delft *et al.* (2018)), such tests are non-existent. Moreover, even when such tests are available, they are usually unable to distinguish between subtle cases distinguishing stationarity and nonstationarity. The subtle situations arise when the underlying stochastic process lie on the verge of stationarity and nonstationarity. For instance, the AR(1) process is stationary if and only if the absolute value of the process parameters is less than one. However, if the true absolute value of the parameter is quite close to one, then the existing tests often yields the wrong conclusion. Also, we are not aware of any sound test for stationarity in the nonparametric setup.

Thus, it is of great importance to devise adequate tests for stationarity in time series scenarios that can distinguish between stationarity and nonstationarity even in nonparametric and subtle situations, and even if the same size is not desirably large.

In this chapter, we demonstrate the practical utility of our Bayesian characterization of stationarity and nonstationarity of general stochastic processes developed in Chapter 6, in response to the aforementioned concerns with respect to time series. In particular, we apply our Bayesian approach to the popular time series models AR(1), autoregressive models of order 2 (AR(2)), first order autoregressive conditional heteroscedastic model (ARCH(1)) and generalized ARCH of order one (GARCH(1,1)), under many simulated settings, with large as well as relatively small sample sizes. Our results seem promising enough to address all the concerns mentioned above, and convincingly outperforms the existing tests, whenever they are applicable. A special mention must be made of the ability of our Bayesian characterization approach to correctly distinguish between stationarity and nonstationarity, even in very subtle situations. Since our approach does not require any modeling assumptions, the results vouch for the effectiveness of our approach in general situations. We also demonstrate that effectiveness of our approach can be substantially enhanced when the underlying model structure is correctly assumed.

As already pointed out in Section 6.1, there is also an extremely important special case of the problem of ascertaining stationarity and nonstationarity of general time series, namely, convergence diagnostics of MCMC algorithms. As is well-known, the key idea of MCMC is to generate a Markov chain in such a way that it converges to the desired distribution. For instance, in the Bayesian paradigm, MCMC is used to simulate from complex posterior distributions. Indeed, the extreme popularity of MCMC can be attributed to its utility in the Bayesian paradigm. For MCMC theory, techniques and existing convergence diagnostics, see, for example, [Meyn and Tweedie \(1993\)](#), [Gilks and Roberts \(1996\)](#), [Liu \(2001\)](#), [Robert and Casella \(2004\)](#), [Brooks \*et al.\* \(2011\)](#).

However, although MCMC can be designed such that it theoretically converges to the target distribution, in reality, it is a challenging task to ascertain if convergence has taken place, since only a finite number of iterations of the MCMC algorithm can be implemented in practice. Although there exists a plethora of methods for MCMC

convergence diagnostics (see, for example, [Gelman and Rubin \(1992\)](#), [Geweke \(1992\)](#), [Raftery and Lewis \(1992\)](#), [Robert \(1995\)](#), [Gilks and Roberts \(1996\)](#), [Cowles and Carlin \(1996\)](#), [Brooks and Gelman \(1998\)](#), [Brooks and Roberts \(1998\)](#), [Brooks \*et al.\* \(2011\)](#), [Robert and Casella \(2004\)](#), [Roy \(2019\)](#)), almost all of them are heuristic in nature, often involving subjective judgment, while the theoretical establishment of rates of MCMC convergence is too difficult in general situations. Needless to mention, the popular MCMC convergence diagnostics can often mislead the practitioner about the actual convergence scenario. Moreover, in reality, the target posteriors can often be multimodal, and in such cases, the performances of such diagnostic tools can be even poorer.

In response to the above concerns on MCMC convergence diagnostics, in this chapter we also demonstrate the usefulness of our Bayesian characterization approach in correctly assessing convergence of MCMC algorithms. We specifically apply our methods to Transformation based Markov Chain Monte Carlo (TMCMC) introduced by [Dutta and Bhattacharya \(2014\)](#). The major feature of TMCMC is its effective dimension-reduction property achieved by mapping deterministic transformations of some low-dimensional (often, one-dimensional) random variable to the actual high-dimensional random variable associated with the target distribution. Naturally, this conceptualization permits drastic improvements of computational speed and mixing properties. However, although in general, good convergence properties of TMCMC can be expected, not all TMCMC algorithms possess good mixing and convergence properties. The theoretical developments regarding these issues are provided in [Dey and Bhattacharya \(2016\)](#), [Dey and Bhattacharya \(2017\)](#), [Dey and Bhattacharya \(2019\)](#). In this chapter, we evaluate convergence of various implementations of TMCMC, ranging from inefficient to efficient, using our Bayesian characterization theory, and demonstrate that our results are very much consistent with the theoretical underpinnings developed in the aforementioned works.

In what follows, in Section 7.2 we begin with the applications of our Bayesian character-

ization theory with the relatively simple AR(1) model, along with the comparisons with the existing methods of stationarity detection. We deal with the relatively more complex time series models AR(2), ARCH(1) and GARCH(1,1) in Section 7.3. Assessment of TMCMC convergence diagnostics with our Bayesian approach is detailed in Section 7.4.

## 7.2 First illustration: AR(1) model

Let us consider the following AR(1) model:  $X_t = \rho X_{t-1} + \epsilon_t$ ;  $t \geq 1$ , where  $\epsilon_t \stackrel{iid}{\sim} N(0, 1)$ , and  $X_0 \sim U(-1, 1)$ , the uniform distribution on  $(-1, 1)$ . It is well-known that  $\{X_t : t \geq 1\}$  is (asymptotically) stationary if and only if  $|\rho| < 1$ . We illustrate the performance of our methodology after generating the data from the above AR(1) model for various values of  $\rho$ , which we pretend to be unknown for illustration. In particular, we consider three different setups in this regard. In the first setup, we consider samples of sizes  $2 \times 10^8$  from from the AR(1) model, and assume that the form of the true model is known, and that only  $\rho$  is unknown. In the second setup, we generate samples of sizes 2500 from from the AR(1) model, and assume as before that only  $\rho$  is unknown. In the last setup, we draw samples of sizes 2500 from from the AR(1) model, and assume that the entire data-generating model is unknown.

### 7.2.1 Case 1: Large sample size, form of the model known

#### Sample size

We draw samples of sizes  $2 \times 10^8$  from the AR(1) model for various values of  $\rho$  and evaluate the performance of our Bayesian methodology, setting  $n = 10^4$  and  $K = 2 \times 10^4$ .

#### Construction of bound

An important ingredient of our proposed method is the construction of the bounds  $c_j(\omega)$ . In this case, we construct the bounds as follows. We first draw a sample of size  $2 \times 10^8$



from the AR(1) model with  $\rho = 0.99999$ . With this sample, for  $j = 1, \dots, K$ , we form the sup norms  $\tilde{c}_j = \sup_{-\infty < x < \infty} |\hat{F}_j(x) - \tilde{F}_K(x)|$  according to Lemma 26 and Remark 27. We then set  $c_j$  as

$$c_j = \tilde{c}_j + 10^6 \times (0.99999 - |\hat{\rho}|) / \log(\log(j + 1)), \quad (7.2.1)$$

where  $\hat{\rho}$  is the maximum likelihood estimator (MLE) of  $\rho$  based on the observed sample. If the MLE of  $\rho$  does not exist, we set  $\hat{\rho} \equiv 1$ .

To explain the strategy behind (7.2.1), note that for  $\rho = 0.99999$ , the AR(1) process, although stationary, is very close to nonstationarity. So, for any value of  $\rho$  such that  $|\rho| < 0.99999$ ,  $\tilde{c}_j$  is expected to be larger than  $c_j$ . Hence, in such cases, stationarity is to be expected. On the other hand, if  $|\rho| \geq 1$ ,  $\tilde{c}_j$  is expected to be smaller than  $c_j$ , so that nonstationarity is implied. For simplicity we assume that values of  $\rho$  such that  $0.99999 < |\rho| < 1$  are not of interest.

To further improve the bound, we add the quantity  $10^6 \times (0.99999 - |\hat{\rho}|) / \log(\log(j+1))$  to  $\tilde{c}_j$ . The significance of this addition is as follows. If  $|\hat{\rho}| < 0.99999$ , this quantity is positive but tends to zero at a slow rate. This enhances the conclusion of stationarity. Similarly, if  $|\hat{\rho}| > 0.99999$ , the quantity is negative and tends to zero slowly, favouring nonstationarity. Multiplication with  $10^6$  inflates the quantity for more prominence.

### Implementation

Note that at each stage  $j$ , we need to compute the sup norm given by Lemma 26 (also, Remark 27). This requires evaluation of  $\tilde{F}_K$  at  $\hat{x}_j$  (or  $\tilde{x}_{j^*}$ ). We carry out this evaluations by splitting the summations of the indicator functions associated with  $\tilde{F}_K$  on 104 parallel cores on a VMWare, and obtaining the final result on a single node, which also carries out the iterative procedure. The entire exercise takes about 6 minutes in the case of stationarity and about 3 minutes in the case of nonstationarity.

## Results

We implement our method when the data is generated from the AR(1) model with  $\rho$  randomly selected from  $U(-1, 1)$ , and with  $\rho$  taking the values 0.99, 0.995, 0.999, 0.9999, 1, 1.00005, 1.05 and 2. Figure 7.2.1 shows that in all the cases, our method correctly detects stationarity and nonstationarity. That even with such subtle differences among the true values of  $\rho$  our method performs so well, is quite encouraging.

### 7.2.2 Case 2: Relatively small sample size, form of the model known

#### Sample size

We draw samples of sizes 2500 from the AR(1) model for those values of  $\rho$  as in Section 7.2.1 and evaluate the performance of our Bayesian methodology, setting  $n = 50$  and  $K = 50$ .

#### Construction of bound

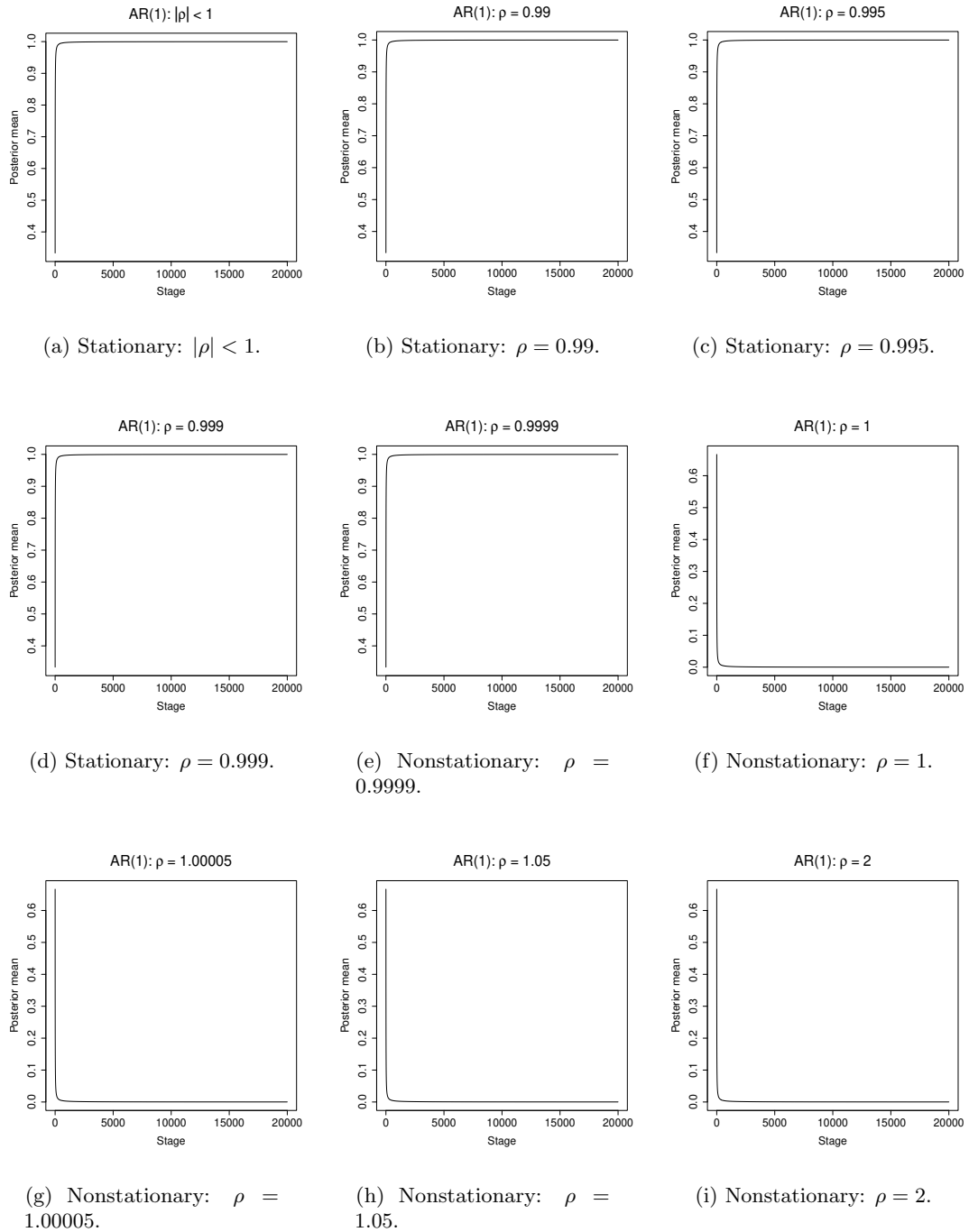
In this case, we choose the basic form of the bounds in a similar manner as in Section 7.2.1, but make it adaptive with the iterations to suit the small sample situation.

As before, we first draw a sample of size  $2 \times 10^8$  from the AR(1) model with  $\rho = 0.99999$ . With this sample, for  $j = 1, \dots, K$ , we form the sup norms  $\tilde{c}_j = \sup_{-\infty < x < \infty} |\hat{F}_j(x) - \tilde{F}_K(x)|$  according to Lemma 26 and Remark 27. We then set  $c_j$  as

$$c_j = \tilde{c}_j + \hat{C}_j \times (0.99999 - |\hat{\rho}| + \hat{\epsilon}_j) / \log(\log(j + 1)), \quad (7.2.2)$$

where  $\hat{C}_1 = 1$ ,  $\hat{\epsilon}_1 = 0$ , and for  $j > 1$ , we adaptively modify these values as follows:

- If  $|\hat{\rho}| > 0.9985$ ,
  1. If  $y_j = 1$ , then  $\hat{\epsilon}_{j+1} = \hat{\epsilon}_j + 0.001$  and  $\hat{C}_{j+1} = \hat{C}_j + 1$ .
  2. If  $y_j = 0$ , then  $\hat{\epsilon}_{j+1} = \hat{\epsilon}_j - 0.001$  and  $\hat{C}_{j+1} = \hat{C}_j + 1$ .



**Figure 7.2.1:** Parametric AR(1) example with  $K = 20000$  and  $n = 10000$ .

- If  $0.9955 < |\hat{\rho}| \leq 0.9985$ ,
  1.  $y_j = 1$ , then  $\hat{\epsilon}_{j+1} = \hat{\epsilon}_j + 0.01$  and  $\hat{C}_{j+1} = \hat{C}_j + 1$ .
  2.  $y_j = 0$ , then  $\hat{\epsilon}_{j+1} = \hat{\epsilon}_j - 0.01$  and  $\hat{C}_{j+1} = \hat{C}_j + 1$ .
- If  $0 < |\hat{\rho}| \leq 0.9955$ ,
  1. If  $y_j = 1$ , then  $\hat{\epsilon}_{j+1} = \hat{\epsilon}_j + 0.05$  and  $\hat{C}_{j+1} = \hat{C}_j + 1$ .
  2. If  $y_j = 0$ , then  $\hat{\epsilon}_{j+1} = \hat{\epsilon}_j - 0.05$  and  $\hat{C}_{j+1} = \hat{C}_j + 1$ .

To appreciate the above strategy, first note that for small samples, the MLE of  $\rho$  need not be adequately close to the true value of  $\rho$ , and hence we need to add a quantity  $\hat{\epsilon}_j$  to make up for the inadequacy. We select  $\hat{\epsilon}_j$  adaptively, increasing its value for the next iteration if  $y_j = 1$ , so that in the next iteration stationarity is preferred, given the current value of  $y_j$ . If  $y_j = 0$  in the current iteration, we decrease the current value of  $\hat{\epsilon}_j$ , so that nonstationarity is favoured in the next iteration. We also increase the value of  $\hat{C}_j$  by one, at every iteration, rather than keeping it constant over the iterations. Thus, the prominence of the quantity  $\hat{C}_j \times (0.99999 - |\hat{\rho}| + \hat{\epsilon}_j) / \log(\log(j + 1))$  increases with the iterations.

The increment and decrement of  $\hat{\epsilon}_j$  depends upon the magnitude of  $\hat{\rho}$ . If  $|\hat{\rho}| > 0.9985$ , that is, when the model is close to nonstationarity, we increase/decrease  $\hat{\epsilon}_j$  by 0.001 only, since larger quantities, if added, can wrongly indicate stationarity.

When  $0.9955 < |\hat{\rho}| \leq 0.9985$ , we consider adding/subtracting 0.01 to  $\hat{\epsilon}_j$ ; this larger quantity is expected to make up for the uncertainty associated with stationarity and nonstationarity when  $0.9955 < |\hat{\rho}| \leq 0.9985$ .

On the other hand, when  $0 < |\hat{\rho}| \leq 0.9955$ , we add/subtract 0.05 to  $\hat{\epsilon}_j$ , since we expect our algorithm to favour stationarity in this situation. The choice 0.05, which is larger than the quantities in the previous cases, is expected to facilitate diagnosis of stationarity.

### Implementation

The implementation remains the same as before. For this small sample, even with 2 cores, the results are delivered almost instantly.

### Results

As before, we implement our method when the data is generated from the AR(1) model with  $\rho$  randomly selected from  $U(-1, 1)$ , and with  $\rho$  taking the values 0.99, 0.995, 0.999, 0.9999, 1, 1.00005, 1.05 and 2. Figure 7.2.2 shows that, except in the case where the true value of  $\rho$  is 0.9999, our method correctly detects stationarity and nonstationarity. That even with such small sample, and with such subtle differences among the true values of  $\rho$ , our method performs well, is quite encouraging, despite its fallibility at  $\rho = 0.9999$ . Indeed, with such small sample, correct detection of stationarity in the case of so subtle difference with nonstationarity is perhaps not to be expected.

### 7.2.3 Case 3: Relatively small sample size, form of the model unknown

#### Sample size

We draw samples of sizes 2500 from the AR(1) model for those values of  $\rho$  as in Sections 7.2.1 and 7.2.2 and evaluate the performance of our Bayesian methodology, setting  $n = 50$  and  $K = 50$ , assuming that the model itself is unknown.

#### Construction of bound

Since we assume now that the model itself is unknown, there is no provision of obtaining the MLE of  $\rho$  and constructing bounds on its basis. We also can not compute  $\tilde{c}_j$ , since it requires knowledge of the underlying model. Hence, in the absence of such information, we set

$$c_j = \hat{C}_j / \log(j + 1), \quad (7.2.3)$$

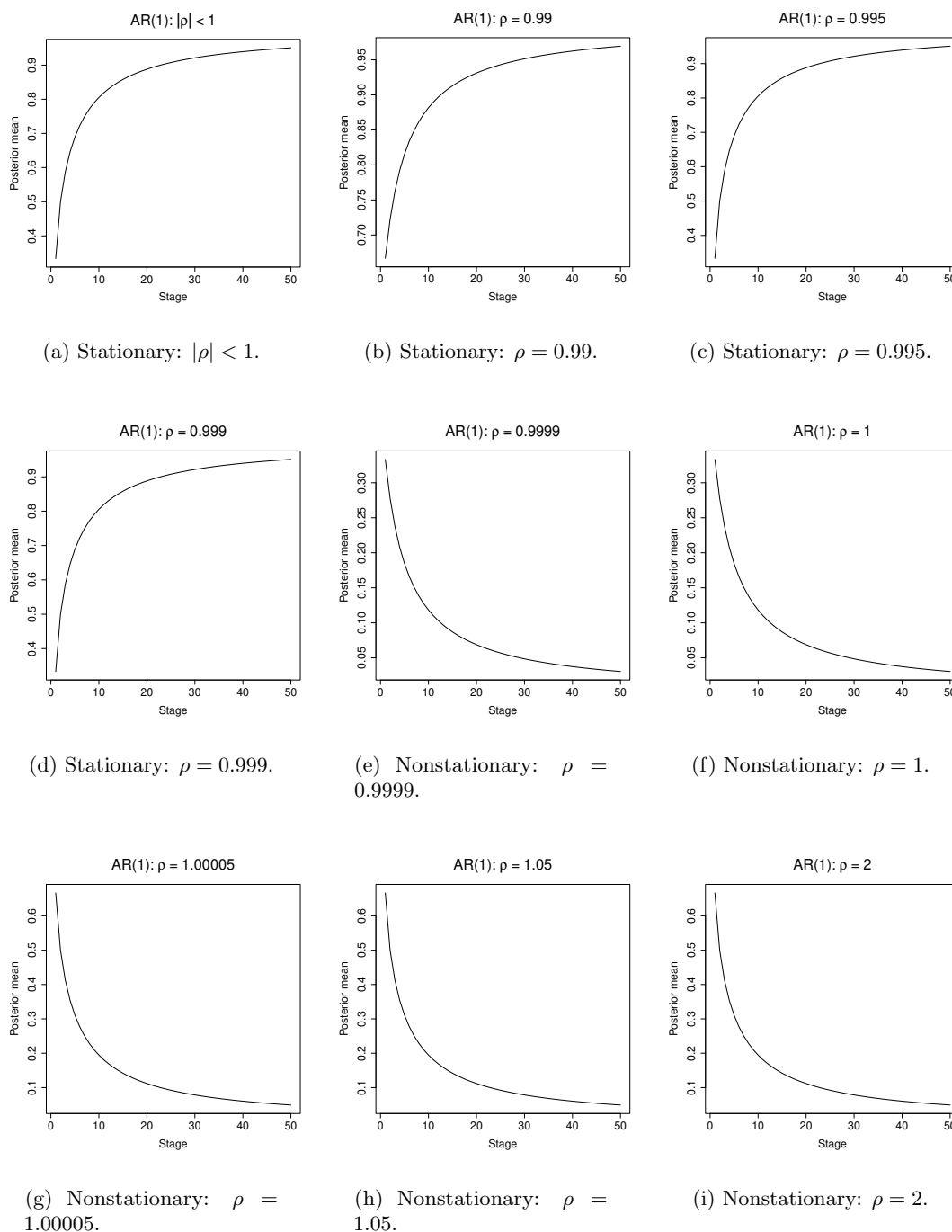


Figure 7.2.2: Parametric AR(1) example with  $K = 50$  and  $n = 50$ .

which is of the same form as (5.4.1). Here we set  $\hat{C}_1 = 1$ , and for  $j > 1$ ,  $\hat{C}_j = \hat{C}_{j-1} + 0.05$  if  $y_{j-1} = 1$  and  $\hat{C}_j = \hat{C}_{j-1} - 0.05$  if  $y_{j-1} = 0$ .

Thus, as before, we favour stationarity at the next stage if at the current stage stationarity is favoured ( $y_j = 1$ ) and nonstationarity otherwise. Note that unlike the previous cases, we have considered  $\log(j + 1)$  instead of  $\log(\log(j + 1))$ . This faster rate turned out to be more appropriate in this situation of very less information about the true model.

### Implementation

The implementation remains the same as before, only that here it is much simpler because of the simple structure of the bound. Again, for this small sample, even with 2 cores, the results are delivered almost instantaneously.

### Results

As before, we implement our method when the data is generated from the AR(1) model with  $\rho$  randomly selected from  $U(-1, 1)$ , and with  $\rho$  taking the values 0.99, 0.995, 0.999, 0.9999, 1, 1.00005, 1.05 and 2. Figure 7.2.3 shows that, again except in the case where the true value of  $\rho$  is 0.9999, our method correctly detects stationarity and nonstationarity, albeit in a less precise manner as in Figure 7.2.2. That even with such small sample, with no assumption about the true model, and with such subtle differences among the true values of  $\rho$ , our method performs well, is quite encouraging, again, despite its fallibility at  $\rho = 0.9999$ , which is perhaps not expected to be detected correctly in this situation of so less information.

### Comparison with classical tests of nonstationarity

To test stationarity of AR(1) model, there are well-known classical hypotheses tests, namely, the augmented Dickey-Fuller (ADF) test (Dickey and Fuller (1979)), the Philips-

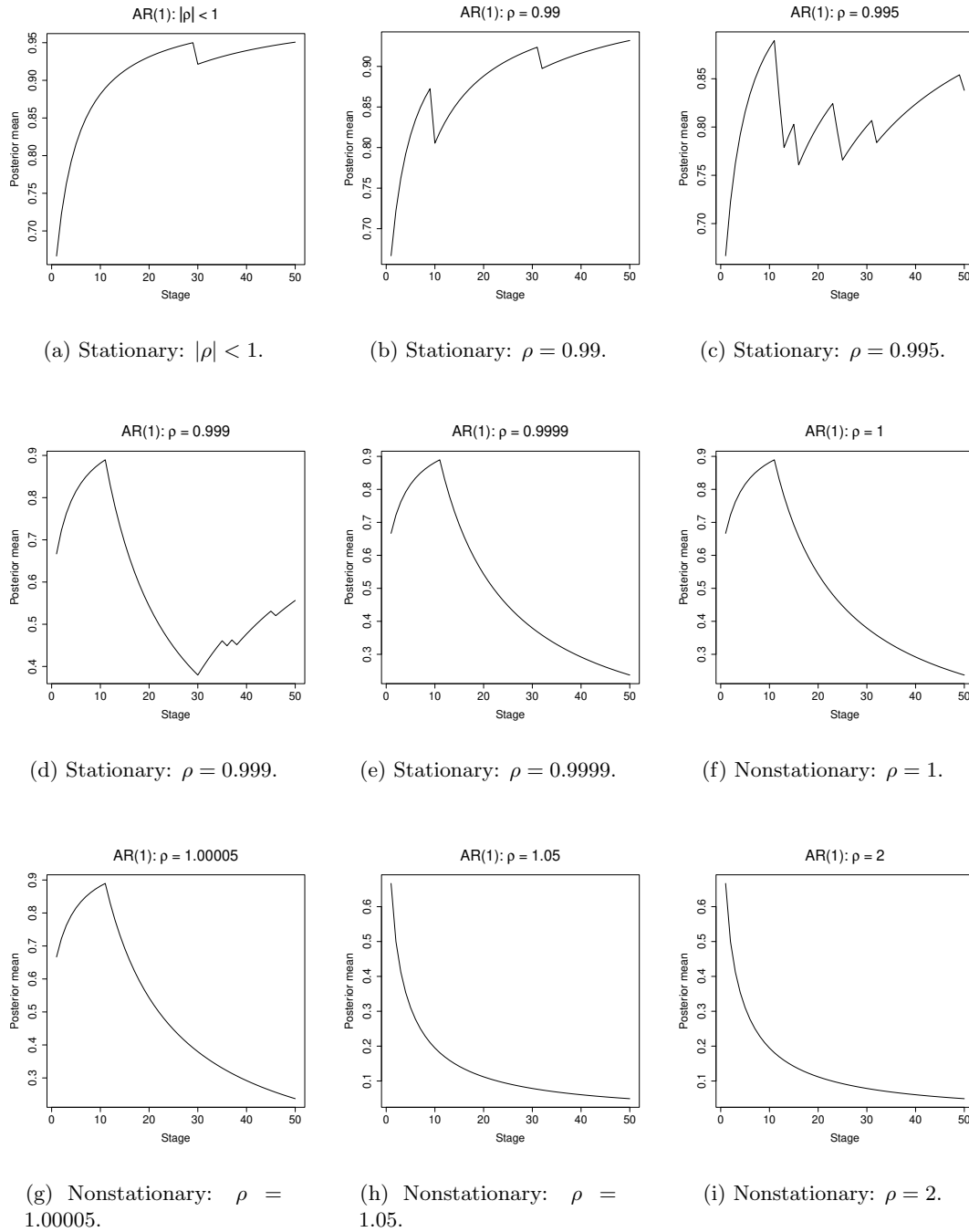


Figure 7.2.3: Nonparametric AR(1) example with  $K = 50$  and  $n = 50$ .



Perron (PP) test (Philips and Perron (1988)), and the Kwiatkowski, Phillips, Schmidt, Shin (KPSS) test (Kwiatkowski *et al.* (1992)).

Researchers have noticed that the first two tests, PP and ADF, are not very efficient in distinguishing between stationarity and nonstationarity when the process is stationary, but at the verge of stationarity and nonstationarity. Indeed, when we apply these tests on our datasets with sample size 2500, we find that these two tests correctly determines stationarity/nonstationarity of the process when  $\rho$  is randomly chosen between  $(-1, 1)$ ,  $\rho = 0.99$  and  $\rho = 0.995$ , at the 5% level of significance, but fails when  $\rho = 0.999$ ,  $0.9999$  and  $1.05$ . However, both these tests correct conclude nonstationarity when  $\rho = 1$  and  $1.00005$ . For  $\rho = 2$ , both the tests turn out to be inapplicable.

On the other hand, at the 5% level of significance, the KPSS test provides correct answers whenever  $|\rho| < 1$ , but fails when  $\rho \geq 1$ .

Thus, our proposed method outperforms all the three existing popular methods of testing stationarity in AR(1) models. Here we emphasize that the testing methods ADF, PP and KPSS are particularly designed to detect stationarity in autoregressive models, while ours is a completely general method. That our method still managed to outperform the existing specialized testing methods, is very encouraging.

### 7.3 Second illustration: AR(2), ARCH(1) and GARCH(1,1) models

We now test our ideas on relatively more complex time series models. In particular, we consider autoregressive models of order 2 (AR(2)), first order autoregressive conditional heteroscedastic model (ARCH(1)) and generalized ARCH of order one (GARCH(1,1)). We consider samples of size 2500 for our investigation, since the relatively small sample size, as we observed in the context of AR(1), can pose beneficial challenge to our Bayesian method.

### 7.3.1 Application to AR(2)

The AR(2) model is given by

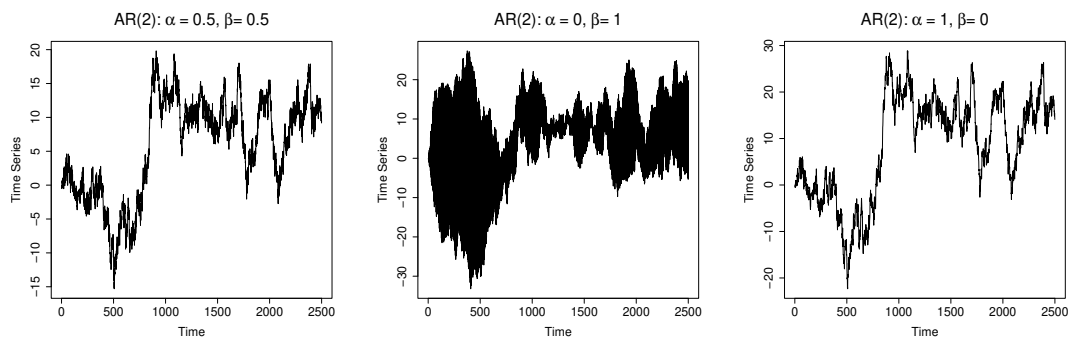
$$x_t = \alpha x_{t-1} + \beta x_{t-2} + \epsilon_t; \quad t = 1, 2, \dots, \quad (7.3.1)$$

where we set  $x_1 = x_2 = 0$  and  $\epsilon_t \stackrel{iid}{\sim} N(0, 1)$ , for  $t = 1, 2, \dots$ . The necessary and sufficient conditions for stationarity of the AR(2) model (7.3.1) are given by (see, for example, [Shumway and Stoffer \(2006\)](#))

$$\begin{aligned} \alpha + \beta &< 1; \\ \beta - \alpha &< 1; \\ \beta &> -1. \end{aligned} \quad (7.3.2)$$

We simulate samples of size 2500 from (7.3.1) with various fixed values of  $\alpha$  and  $\beta$  that satisfy and do not satisfy (7.3.2), and apply our Bayesian procedure to ascertain stationarity and nonstationarity, with the bound of the form (7.2.3), starting with  $\hat{C}_1 = 1$ . We initially consider ( $n = 50, K = 50$ ) but in a few nonstationary cases ( $(\alpha = 1, \beta = 0)$ ,  $(\alpha = 0, \beta = 1)$  and  $(\alpha = 0.5, \beta = 0.5)$ ) this failed to work satisfactorily, since a relatively large value of  $n$  in the context of relatively small sample size has the tendency to create overlaps among neighboring regions of local stationarity, in effect, destroying local stationarity which is at the heart of our Bayesian procedure. This happens when the underlying time series diverges slowly, as in the aforementioned values of  $(\alpha, \beta)$ . Figure 7.3.1 captures such behaviours of such slowly diverging nonstationary processes in comparison to fast diverging nonstationary processes.

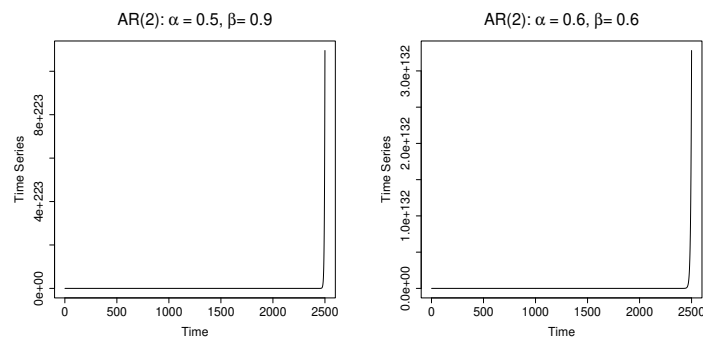
On the other hand, the choice ( $n = 5, K = 500$ ) turned out to work very well in all the cases that we considered. Figure 7.3.2, depicting the results of our Bayesian method for various values of  $\alpha$  and  $\beta$  for ( $n = 5, K = 500$ ), shows that all the stationarity and



(a) Slow divergence:  $\alpha = 0.5, \beta = 0.5$ .

(b) Slow divergence:  $\alpha = 0, \beta = 1$ .

(c) Slow divergence:  $\alpha = 1, \beta = 0$ .



(d) Fast divergence:  $\alpha = 0.5, \beta = 0.9$ .

(e) Fast divergence:  $\alpha = 0.6, \beta = 0.6$ .

**Figure 7.3.1:** Slow and fast divergence tendencies of AR(2) model for several values of  $\alpha$  and  $\beta$ .

nonstationarity situations are correctly identified.

### 7.3.2 Application to ARCH(1)

The ARCH models introduced by Engle (1982) attempts to take into account the heteroscedasticity of financial time series, which is often ignored by other popular financial models such as Black-Scholes (Black and Scholes (1973)) and the Ornstein-Uhlenbeck process (Ornstein and Uhlenbeck (1930)). In the ARCH( $p$ ) model, the conditional variance is modeled as an autoregressive process of order  $p$ . For details on ARCH models, see Bera and Higgins (1993), Giraitis *et al.* (2005), Straumann (2005).

The ARCH(1) model is of the following form: for  $t = 1, 2, \dots$ ,

$$\begin{aligned}x_t &= \epsilon_t \sigma_t \\ \sigma_t^2 &= \omega + \alpha x_{t-1}^2,\end{aligned}\tag{7.3.3}$$

where  $\omega > 0$ ,  $\alpha \geq 0$  and  $\epsilon_t \stackrel{iid}{\sim} N(0, 1)$ , for  $t = 1, 2, \dots$ . The necessary and sufficient condition for stationarity of (7.3.3) is  $0 < \alpha < 1$ . We set  $\omega = 1$  and  $x_1 = 0$  for our purpose.

As in the AR(2) situations, here we considered  $n = 5$ ,  $K = 500$ , and the bound (7.2.3) with  $\hat{C}_1 = 1$ . With these, Figure 7.3.3 provides the results of our Bayesian analyses of the realizations of (7.3.3) for  $\omega = 1$  and various values of  $\alpha$ . Although for  $0 < \alpha < 1$ , our method correctly identifies stationarity in all the cases, for  $\alpha = 1, 1.5, 2$ , our procedure falsely declares nonstationarity as stationarity.

To understand the reason for this, it is necessary to recall some of the properties of the ARCH(1) model. Note that  $E(x_t) = 0$  for  $t \geq 1$  and for any  $t \geq 1$ ,  $Var(x_t) = \frac{\omega}{1-\alpha}$ , provided  $0 < \alpha < 1$ . For  $\alpha \geq 1$ ,  $Var(x_t)$  increases with  $t$ . Moreover,  $Cov(x_t, x_{t+j}) = 0$  for  $j \geq 1$ . The last fact shows that the ARCH(1) model is serially uncorrelated. Thus, even though for  $\alpha \geq 1$ ,  $Var(x_t)$  increases with  $t$ , the realizations will be centered around

162 7.3. SECOND ILLUSTRATION: AR(2), ARCH(1) AND GARCH(1,1) MODELS

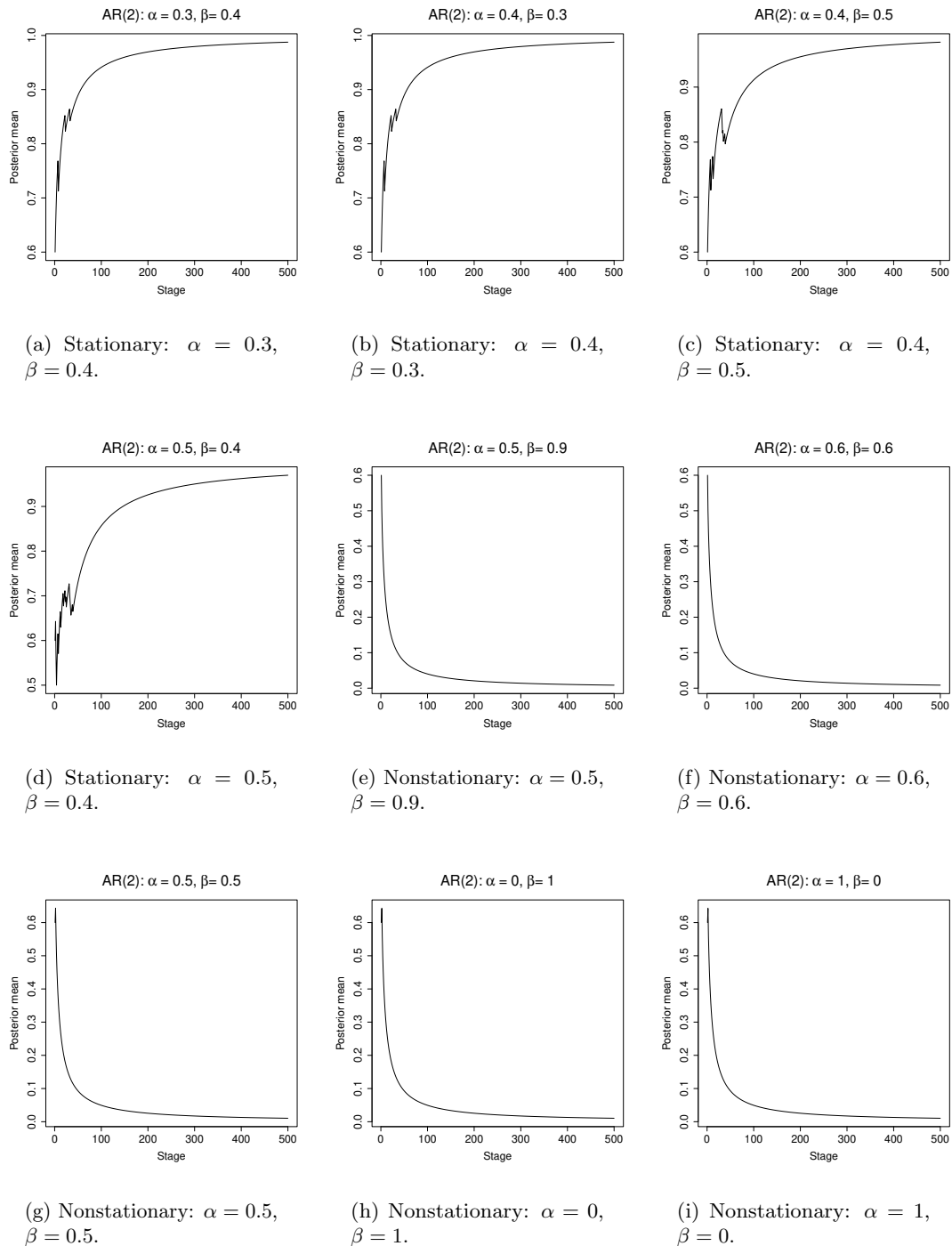


Figure 7.3.2: Nonparametric AR(2) example with  $K = 500$  and  $n = 5$ .

163 7.3. SECOND ILLUSTRATION: AR(2), ARCH(1) AND GARCH(1,1) MODELS

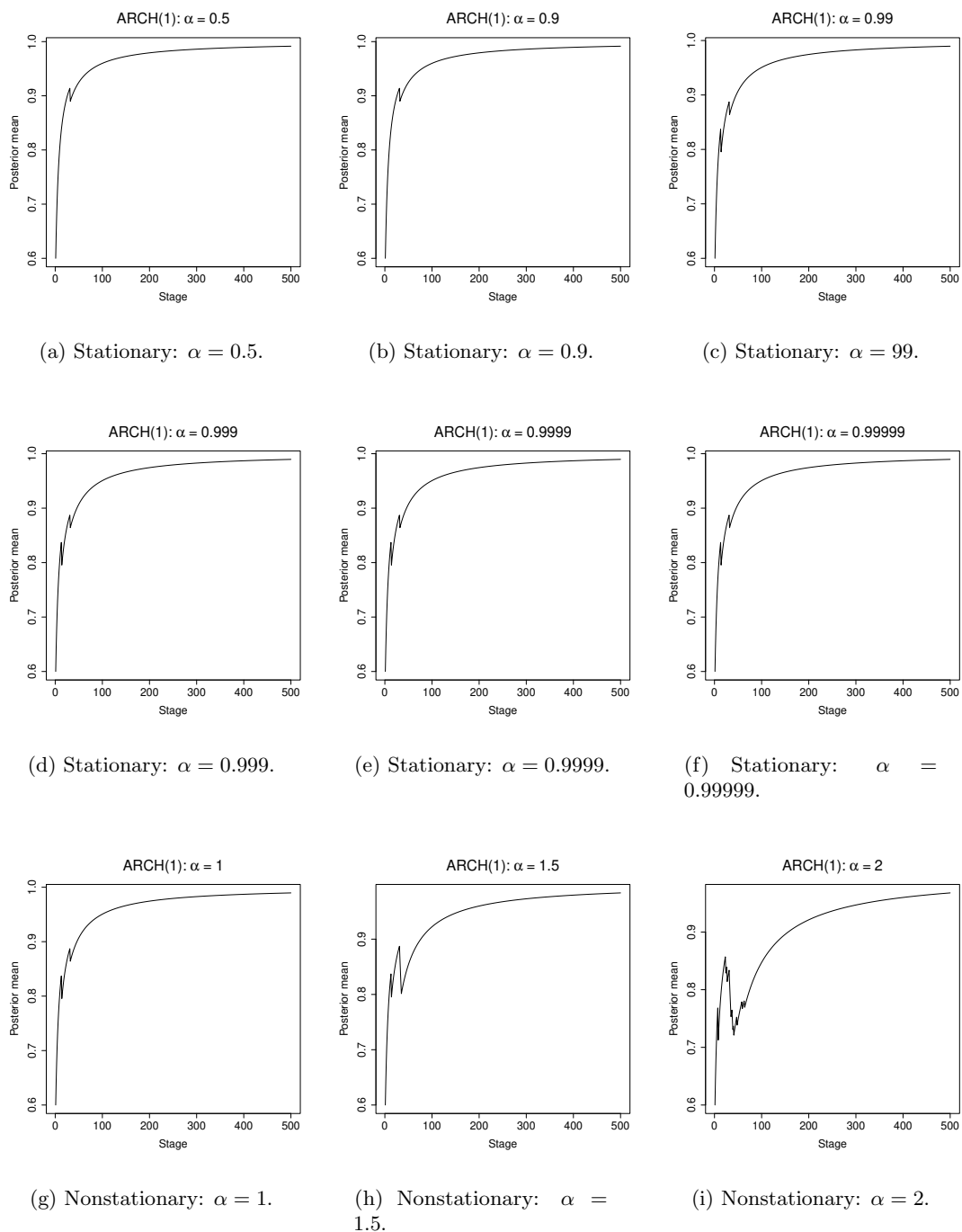


Figure 7.3.3: Nonparametric ARCH(1) example with  $K = 500$  and  $n = 5$ .

zero and will be serially uncorrelated, and these are instrumental in rendering the pattern of the realizations seem like stationary time series. Although the variances are increasing in such cases, the realizations need not have an increasing range pattern due to absence of serial correlation. Figure 7.3.4 shows ARCH(1) realizations for  $\alpha = 0.9, 1, 1.5$  and  $2$ . Note that none of the realizations exhibit any trend of increasing range, even though only  $\alpha = 0.9$  corresponds to stationarity. Moreover, the pattern of the nonstationary realization for  $\alpha = 1$  is quite similar to that of the stationary realization  $\alpha = 0.9$ . Indeed, all the four realizations shown in Figure 7.3.4 have similar patterns; they essentially differ only at a few time points, where the realizations have different ranges.

In other words, the realizations for  $\alpha = 1, 1.5$  and  $2$  shown in Figure 7.3.4 do not seem to have enough information to distinguish them from stationarity. Hence, it is not surprising that our Bayesian method declared these realizations as stationary.

### 7.3.3 Application to GARCH(1,1)

The ARCH model has been generalized by [Bollerslev \(1986\)](#) and [Taylor \(1986\)](#) independently to let  $\sigma_t^2$  to have an autoregressive structure as well. This generalized ARCH, or GARCH model, is arguably the most widely used model in financial time series, particularly, for modeling stochastic volatility. For details on GARCH, see [Bougerol and Picard \(1992\)](#), [Giraitis et al. \(2005\)](#), [Berkes et al. \(2003\)](#) and [Straumann \(2005\)](#).

The GARCH(1,1) model, which generalizes ARCH(1), is of the following form: for  $t = 1, 2, \dots$ ,

$$\begin{aligned} x_t &= \epsilon_t \sigma_t \\ \sigma_t^2 &= \omega + \alpha x_{t-1}^2 + \beta \sigma_{t-1}^2, \end{aligned} \tag{7.3.4}$$

where  $\omega > 0$ ,  $\alpha \geq 0$ ,  $\beta \geq 0$  and  $\epsilon_t \stackrel{iid}{\sim} N(0, 1)$ , for  $t = 1, 2, \dots$ . The necessary and sufficient condition for stationarity of (7.3.4) is  $0 < \alpha + \beta < 1$ . We set  $\omega = 1$  and  $x_1 = 0$

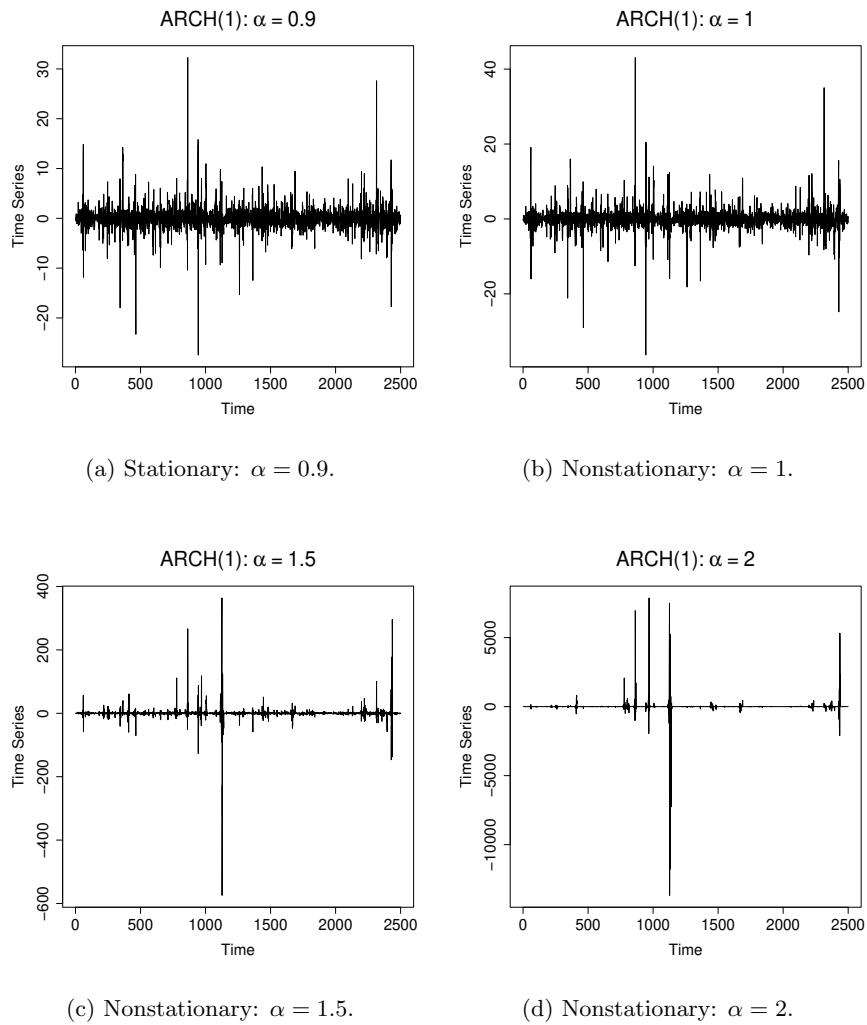


Figure 7.3.4: Comparison of ARCH(1) samples for several values of  $\alpha$  where our Bayesian method failed.



and  $\sigma_1 = 0$  for our purpose.

Again we set  $n = 5$  and  $K = 500$  and consider the nonparametric bound (7.2.3) for applying our Bayesian idea to model (7.3.4) for different values of  $\alpha$  and  $\beta$  leading to stationarity and nonstationarity. Figure 7.3.5, summarizing the results of our Bayesian experiments, show that all the cases have been correctly identified, except the cases of  $(\alpha = 1, \beta = 0)$  and  $(\alpha = 0.5, \beta = 0.5)$ . Note that the first case is the same as ARCH(1) with  $\alpha = 1$ , and the reason for failure of our Bayesian method for this case has already been explained in Section 7.3.2.

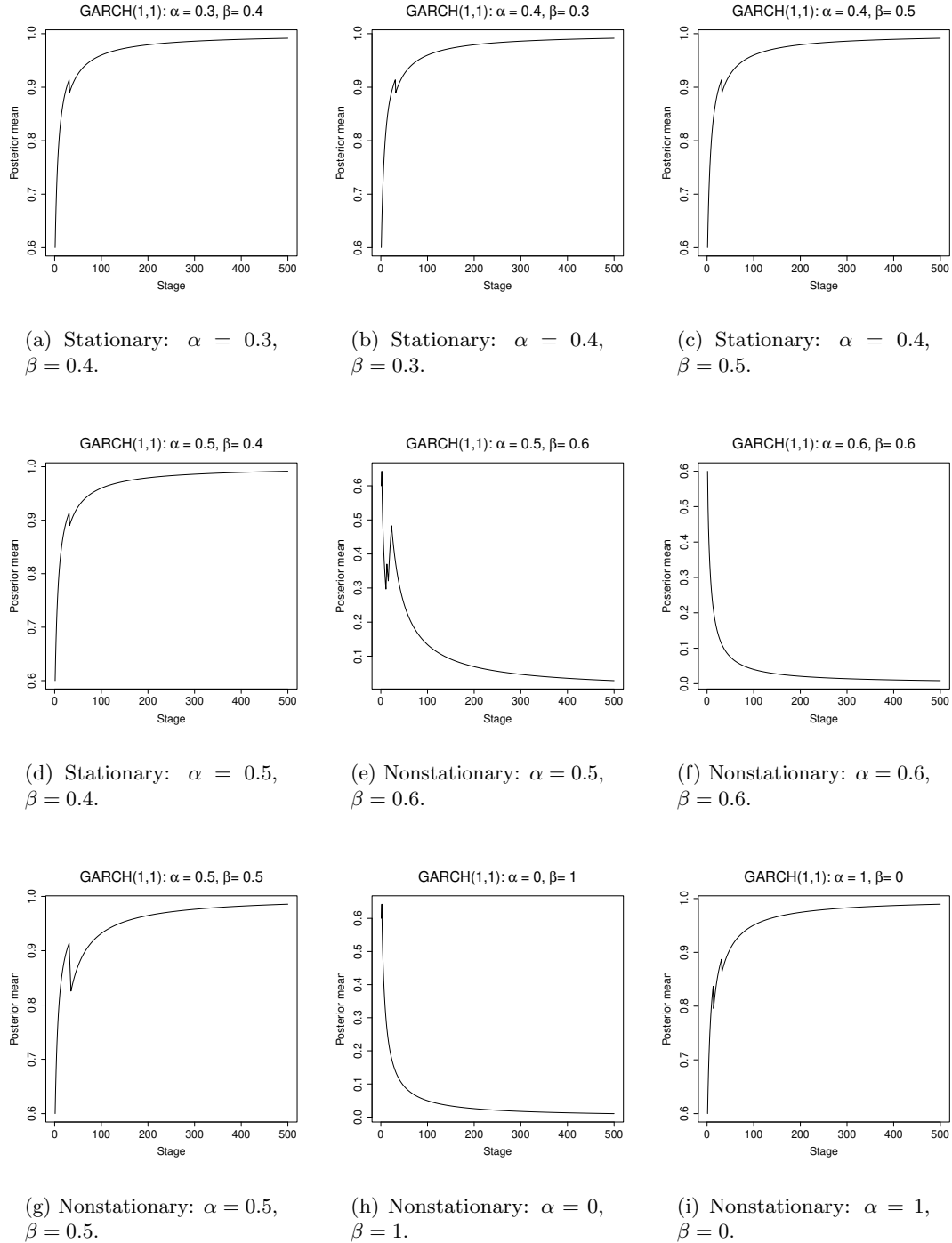
The diagram for the case of  $(\alpha = 0.5, \beta = 0.5)$  is provided in Figure 7.3.6. Note that this realization is essentially of the same pattern as panels (a) and (b) of Figure 7.3.6 associated with ARCH(1) models with  $\alpha = 0.9$  and 1, respectively, which do not seem to show any evidence of nonstationarity. Hence, again, quite unsurprisingly, our Bayesian method declared this case as stationary.

## 7.4 Third illustration: MCMC convergence diagnostics

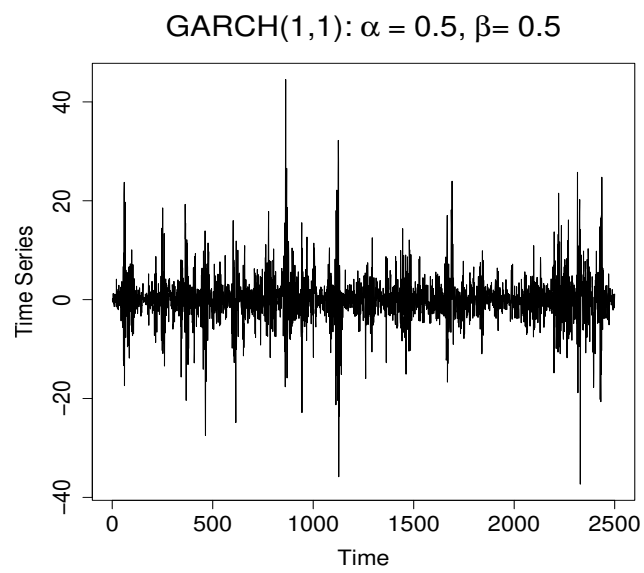
We now test our Bayesian method on the very relevant problem of MCMC convergence diagnosis. As already mentioned, for our purpose, we consider application of our characterization approach to TMCMC introduced by [Dutta and Bhattacharya \(2014\)](#). We consider three examples: in the first example, we assume that the target distribution is a product of 100 standard normal densities, and consider seven instances of additive TMCMC. Here we make use of the optimal scaling theory for additive TMCMC. In the next two examples, we consider mixtures of two normal densities. In all the cases, we evaluate convergence of TMCMC using our proposed Bayesian method.

### 7.4.1 A brief overview of TMCMC

TMCMC enables updating an entire block of parameters using deterministic bijective transformations of some arbitrary low-dimensional random variable. Thus very high-



**Figure 7.3.5:** Nonparametric GARCH(1,1) example with  $K = 500$  and  $n = 5$ .



(a) Nonstationary:  $\alpha = 0.5, \beta = 0.5$ .

**Figure 7.3.6:** GARCH(1,1) sample for  $\alpha = 0.5$  and  $\beta = 0.5$  where our Bayesian method failed.

dimensional parameter spaces can be explored using simple transformations of very low-dimensional random variables. In fact, transformations of some one-dimensional random variable always suffices, which we shall adopt in our examples. The underlying idea also greatly improves computational speed and acceptance rate compared to block Metropolis-Hastings methods. Interestingly, the TMCMC acceptance ratio is independent of the proposal distribution chosen for the arbitrary low- dimensional random variable. For implementation in our cases, we shall consider the additive transformation, since it is shown in [Dutta and Bhattacharya \(2014\)](#) that many fewer number of “move types” are required by this transformation compared to non-additive transformations. To elaborate the additive TMCMC mechanism, assume that a block of parameters  $\mathbf{x} = (x_1, \dots, x_r)$  is to be updated simultaneously using additive TMCMC, where  $r (\geq 2)$  is some positive integer. At the  $t$ -th iteration ( $t \geq 1$ ) we shall then simulate  $\theta \sim g(x)I_{\{x>0\}}$ , where  $g(\cdot)$  is some arbitrary distribution and  $I_{\{x>0\}}$  is the indicator function of the set  $\{x > 0\}$ .

We then propose, for  $j = 1, \dots, r$ ,  $x_j^{(t)} = x_j^{(t-1)} \pm a_j \eta$ , with equal probability (although equal probability is a convenience, not a necessity), where  $(a_1, \dots, a_r)$  are appropriate scaling constants. Thus, using additive transformations of a single, one-dimensional  $x$ , we update the entire block  $\mathbf{x}$  at once.

#### 7.4.2 Optimal scaling of TMCMC

In our examples we shall choose  $r = d$ , where  $d$  is the total number of parameters to be updated. In other words, we shall update all the parameters simultaneously, in a single block. We shall consider  $a_i = 1$ , for  $i = 1, \dots, d$  and  $g(\cdot)$  to be the  $N(0, \frac{\ell^2}{d})$  density, so that  $\eta$  is simulated from a truncated normal distribution, with mean zero and variance  $\ell^2/d$ . The optimum choice of  $\ell$  is directly related to the optimal scaling problem (see [Dey and Bhattacharya \(2017\)](#) and [Dey and Bhattacharya \(2019\)](#)). Under appropriate regularity conditions it turns out that the optimal value of  $\ell$  corresponds to the optimal additive TMCMC acceptance rate 0.439. When the target distribution  $\pi(x_1, \dots, x_d)$  is

a product of  $d$  iid standard normal densities, as we consider, then it turns out that the optimum choice of  $\ell$  is 2.4.

### 7.4.3 TMCMC example 1: product of 100 standard normal densities

We apply additive TMCMC to generate  $10^6$  realizations from  $\pi(x_1, \dots, x_d)$  being a product of  $d$  standard normal densities with  $d = 100$ . We consider seven values of  $\ell$ , and hence seven different TMCMC chains, each corresponding to a value of  $\ell$ . In particular, we set  $\ell = 0.001, 0.01, 0.1, 2.4, 10, 100$  and  $1000$ . Of these,  $\ell = 2.4$  is the optimum value that maximizes the “diffusion speed” associated with the TMCMC chain. The values relatively closer to 2.4, although not optimal, can still generate TMCMC chains with reasonable convergence properties. Significantly small values of  $\ell$  generates TMCMC chains with very high acceptance rates but with very slow convergence rates, as at each iteration, the chain is allowed to take only small steps for movement. On the other hand, for significantly large values of  $\ell$ , large steps are generally proposed, which are often rejected. Thus, the chain again has slow convergence, with poor acceptance rate.

It transpires from the above discussion that for values of  $\ell$  equal to, or relatively close to 2.4, good convergence properties of the TMCMC chains can be expected, and it is desirable that our Bayesian method indicates convergence to stationarity for such cases. For other values of  $\ell$ , since the convergence properties of the chains are expected to be poor, our Bayesian method must reflect so.

Generation of  $10^6$  TMCMC realizations from  $\pi(x_1, \dots, x_d)$  with  $d = 100$  takes less than 0.05 seconds on an ordinary 64 bit laptop. For implementation of our Bayesian idea, we need the bounds  $c_j$ . The general-purpose nonparametric bound (7.2.3) turned out to be quite appropriate in all the TMCMC examples that we consider. Indeed, in general there is no provision for parametric bounds in MCMC situations, as such bounds would require direct generation from  $\pi$  or some distribution close to  $\pi$ , but if such direct generation were at all possible, MCMC would not be needed in the first place.

For  $K = 1000$  and  $n = 1000$ , Figures 7.4.1, 7.4.2 and 7.4.3 display the trace plots (presented after thinning the original chain of length  $10^6$  by 100, to reduce the file sizes) and the corresponding Bayesian posterior means associated with our Bayesian stationarity detection idea, for different values of  $\ell$ , for the first co-ordinate  $x_1$  of  $(x_1, \dots, x_{100})$ . It takes a few seconds even on a 64-bit dual core laptop for parallel implementation of our Bayesian idea in these cases.

The results are very much in keeping with our prior expectation that for significantly small and large values of  $\ell$  convergence to stationarity for the given sample size is not expected, while for  $\ell = 2.4$  and values relatively close to 2.4, stationarity is expected. Specifically, the figures for Bayesian stationarity detection strongly indicate convergence for  $\ell = 0.1, 2.4$  and 10, but strongly indicate that the chains corresponding to  $\ell = 0.001, 0.01, 100$  and 1000, are yet to achieve stationarity. Note that these results are also in accordance with the visual information obtained from the corresponding trace plots.

#### 7.4.4 TMCMC example 2: mixture normal densities

We now consider two mixtures of normal densities. The first mixture is of the form

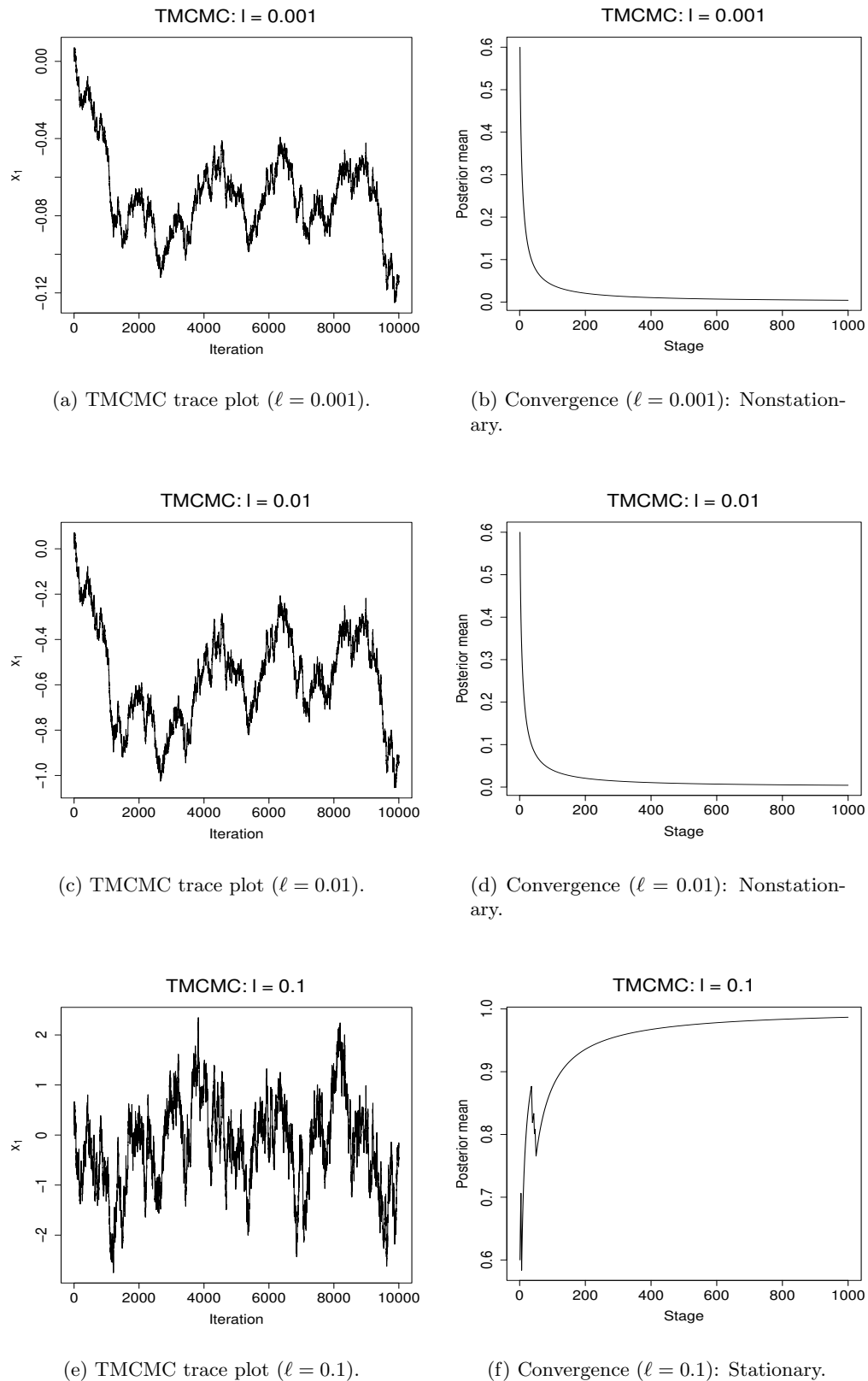
$$\pi(x) = \frac{1}{2}N(x : 0, 1) + \frac{1}{2}N(x, 10, 1), \quad (7.4.1)$$

where  $N(x, \mu, \sigma^2)$  denotes the normal density with mean  $\mu$  and variance  $\sigma^2$ , evaluated at  $x$ . The second mixture is of the form

$$\pi(x) = \frac{1}{2}N(x : 0, 1) + \frac{1}{2}N(x, 15, 1), \quad (7.4.2)$$

The mixtures differ slightly only in the means of the second mixture, but with TMCMC implementation, they reveal significant difference.

With the same implementation as before, with  $\ell = 2.4$ , and with the same bound  $c_j$ , we obtain Figure 7.4.4. The TMCMC trace plot and the Bayesian idea of stationarity



**Figure 7.4.1:** Additive TCMCMC convergence example, with  $K = 1000$  and  $n = 1000$ .

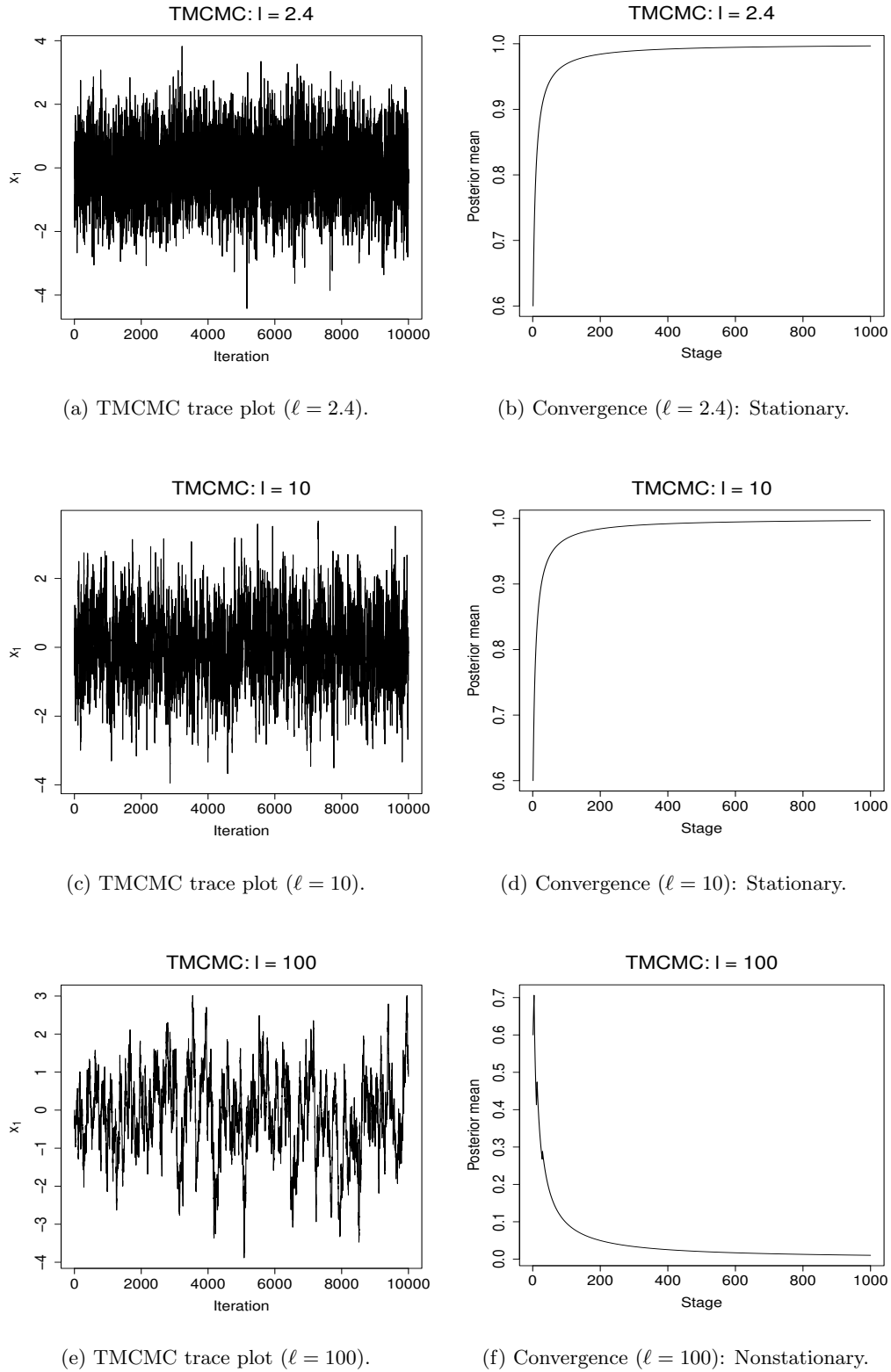
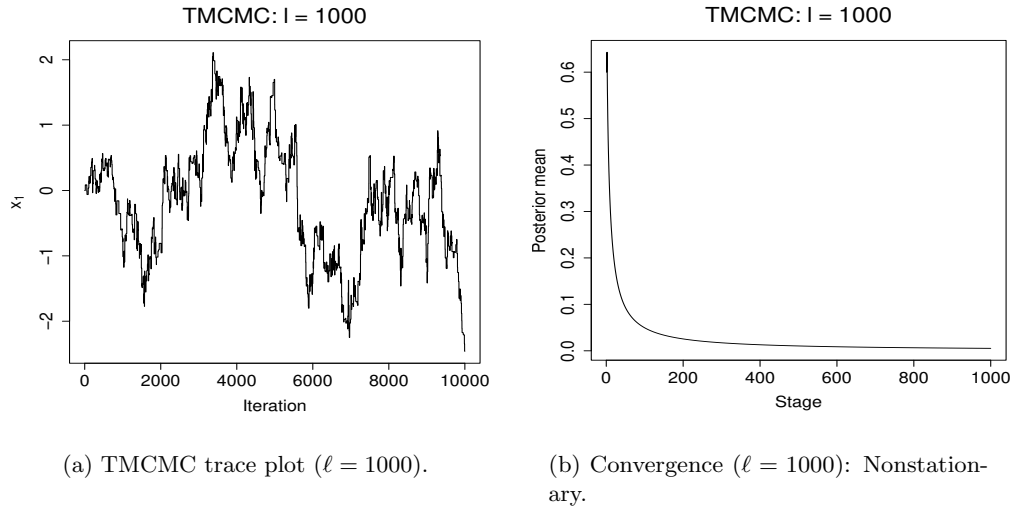


Figure 7.4.2: Additive TCMC convergence example, with  $K = 1000$  and  $n = 1000$ .

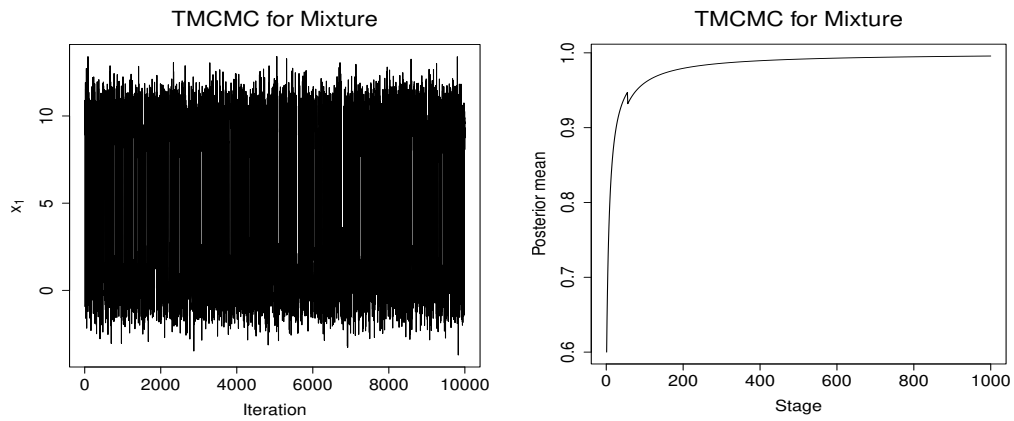




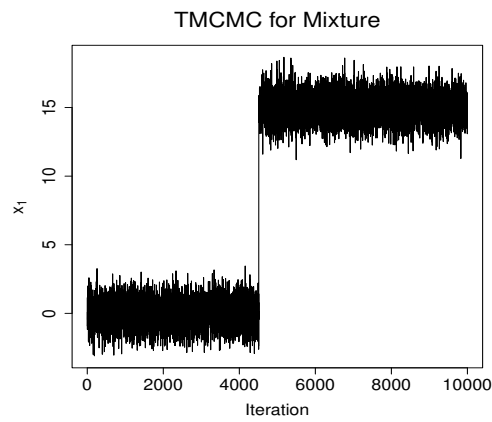
**Figure 7.4.3:** Additive TCMC convergence example, with  $K = 1000$  and  $n = 1000$ .

detection reveals that for (7.4.1) stationarity is clearly reached. That this is achieved even though the chain concentrates around two values 0 and 10, is quite encouraging.

The trace plot for (7.4.2), with the same implementation as before displays two instances of very distinct and significant local stationarity. Consequently, for stationarity detection for this case,  $K = 1000$  and  $n = 1000$  is no longer appropriate. Rather,  $K = 2$  and  $n = 500000$ , seems to be natural and appropriate. With this we obtain the posterior means for the two iterations (corresponding to  $K = 2$ ) to be 0.6 and 0.5, respectively, with the associated posterior variances 0.04 and 0.03125. This is an indication that the chain did not yet reach stationarity, which is also evident from the trace plot. Indeed, for just two instances of significant local stationarities, global stationarity can not be ensured.



(a) TCMC trace plot for first mixture.

(b) Convergence: Stationary ( $K = 1000$ ,  $n = 1000$ ).(c) TCMC trace plot for second mixture.  
Convergence: Nonstationary ( $K = 2$ ,  $n = 500000$ )**Figure 7.4.4:** Additive TCMC convergence example for mixture densities.

## 7.5 Summary and conclusion

This chapter brings out the power of the Bayesian characterization of stationarity and nonstationarity in the time series contexts. As our simulation experiments demonstrate, the strategy is quite effective even when the true model is assumed to be of unknown form, and even if the sample size is not adequately large. A major role in the effectiveness of our strategy is played by the nonparametric bound form (5.4.1), which turns out to be appropriate for all the time series model setups considered. As we also clarified with the AR(1) model, the accuracy of the bound can be significantly improved if the form of the underlying true time series model is assumed to be known. Interestingly,  $\hat{C}_1 = 1$  turns out to be the appropriate starting value of the bound for all the time series models that we considered.

Of particular interest is the reliability of our Bayesian approach when the underlying time series is at the verge of stationarity and nonstationarity. That even in such crucially challenging setups our approach has been able to come off with flying colours in most cases, is quite satisfying. Indeed, only with inadequately small sample sizes our Bayesian approach fails to detect the subtleties.

An important contribution of this chapter, relevant to the current era of Bayesian inference and MCMC methods, is the successful application of our Bayesian characterization approach to MCMC convergence diagnostics. Focussing particularly on TMCMC convergence, we demonstrated the effectiveness of our Bayesian approach to convergence diagnosis with the nonparametric bound form (5.4.1). It is important to appreciate at this point that there is no provision of application of any parametric bound form for MCMC. That even in our MCMC experiments the bound form (5.4.1) with  $\hat{C}_1 = 1$  turned out to be appropriate, is highly encouraging. Thus, our experiments suggest that  $\hat{C}_1 = 1$  is probably a robust choice, at least in time series setups.

# 8

## Applications of Bayesian Characterization of Stochastic Processes to Spatial and Spatio-Temporal Data

### 8.1 Introduction

Spatial and spatio-temporal statistics have received wide attention in the literature due to their important roles in scientific disciplines as varied as agricultural field trials, environmental and ecological sciences, neurosciences, engineering, epidemiology, biosciences, genetics, forestry, geosciences, etc. Glimpses of various applications along with theoretical and methodological developments can be found in [Cressie \(1993b\)](#), [Cressie \(1993a\)](#), [Waller and Gotway \(2004\)](#), [Bivand \*et al.\* \(2008\)](#), [Lawson \(2009\)](#), [Gelfand \*et al.\* \(2011\)](#), [Cressie and Wikle \(2011\)](#).

In spatial statistics, the task is to model the data observed at various locations, and then to make predictions at locations of interest, given the model and the methods employed. Spatio-temporal statistics involves data observed at various locations at different times; often time series data at the locations are of interest. The goal in the spatio-temporal setup is to model the data observed in space and time and to make predictions at locations and times of interest. Usually, future forecasts at locations of interest, given the model, is the purpose of the underlying scientific investigation.

It is clear from the above discussion that the models for spatial and spatio-temporal data must take account of the dependence of the data in space and in space-time, respectively, for reliable inference. The phenomena giving rise to the observed data are most realistically modeled by stochastic processes, and appropriate spatial and spatio-temporal dependence structures emulating the reality can, in principle, be constructed to be inherited by the stochastic process. However, the task is not as simple as it sounds, since the data usually arises from complex real phenomena with complicated dependence structures, and hence postulating stochastic processes that match the complexities, is a highly non-trivial problem. Even if a suitably complex stochastic process is considered, the statistical inferential procedure gets even more complicated, to the dismay of the ordinary statistician, who seeks simple ways to avoid technical difficulties.

As such, the simple-minded Gaussian process heavily dominates the spatial and spatio-temporal literature. Moreover, the Gaussian process structure is most commonly assumed to be stationary and isotropic, the latter meaning that the covariance between responses at any two spatial indices or any two space-time indices depend only upon the distances between the indices. Such drastic simplifying assumptions facilitate methods that are simple to comprehend and are amenable to relatively cheap computations. However, as is easily anticipated, such simplicity does not reflect the reality in general. Indeed, usually in practice, neither Gaussianity, nor stationarity, can be expected, let alone isotropy. Discussions in these directions can be found in [Das \(2018\)](#) and [Guha \(2020\)](#).

Although [Das and Bhattacharya \(2020\)](#) and [Guha \(2020\)](#) proposed reasonable non-Gaussian, nonparametric, non-stationary spatial and spatio-temporal stochastic processes with the desirable property that the covariance tends to zero as the distance between the indices tends to infinity, the important issue of detecting stationarity and nonstationarity of the real spatial and spatio-temporal processes yielding the observed data, is hitherto almost unexplored. Even the weaker condition of detecting covariance stationarity and nonstationarity has received very little attention in the literature, with some of the few known works in this regard being [Ephraty \*et al.\* \(2001\)](#), [Fuentes \(2002\)](#), [Guan \*et al.\* \(2004\)](#), [Fuentes \(2005\)](#), [Li \*et al.\* \(2008\)](#), [Jun and Genton \(2012\)](#), [Bandopadhyay and Rao \(2017\)](#), [Bandopadhyay \*et al.\* \(2017\)](#); the last two works perhaps being the most general and effective among them. Although usually nonstationarity is expected in reality, it is of course not possible to rule out stationarity in practice. Indeed, there are very many real datasets for which stationarity is conjectured. We shall provide such a real example in this chapter. It is thus of great importance to come up with formal methods for detecting stationarity and nonstationarity of the underlying spatial and spatio-temporal processes generating the data.

In this chapter, we show that the developments in Chapter 6 for characterizing stationarity and nonstationarity of general stochastic processes, which yielded very encouraging results for distinguishing between stationarity and nonstationarity in time series processes, including MCMC convergence diagnostics, is also as useful in our current scenario of spatial and spatio-temporal statistics. Specifically, we extend our methods proposed in Chapter 6 for detection of both strong and weak (covariance) spatial and spatio-temporal stationarity of the underlying true but unknown stochastic process that generates the observed data. Our simulation experiments with many examples yield quite encouraging results. Moreover, comparisons of our results with those of [Bandopadhyay and Rao \(2017\)](#) and [Bandopadhyay \*et al.\* \(2017\)](#) whenever applicable, demonstrate superiority of our methods. Noting further that development

of our theoretical results require far less assumptions than those of [Bandopadhyay and Rao \(2017\)](#) and [Bandopadhyay et al. \(2017\)](#), winning convincingly in the simulation experiments seem to render our efforts all the more fruitful.

The rest of this chapter is structured as follows. In Section 8.2 we illustrate detection of strict and covariance stationarity and nonstationarity in spatial setups, along with comparisons with existing tests for covariance stationarity. Section 8.3 is about application of our ideas in spatio-temporal contexts, with comparisons with existing tests for covariance stationarity. Applications to real spatial and spatio-temporal data sets are considered in Section 8.4.

## 8.2 Detection of stationarity and nonstationarity in spatial data

In this illustration, we shall consider detecting both strict and weak stationarity of the spatial processes that gave rise to the observed data.

### 8.2.1 Data generation

We now conduct simulation experiments with our theory for detecting stationarity and nonstationarity in spatial data. To conduct the experiment, we simulate two datasets from stationary and nonstationary zero-mean Gaussian processes (GPs) with covariance functions

$$\text{Cov}(X_{s_1}, X_{s_2}) = \exp(-5\|s_1 - s_2\|^2) \quad (8.2.1)$$

and

$$\text{Cov}(X_{s_1}, X_{s_2}) = C_1(\|s_1 - s_2\|) = \exp(-5\|\sqrt{s_1} - \sqrt{s_2}\|^2), \quad (8.2.2)$$

for all spatial locations  $s_1, s_2 \in \mathbb{R}^2$ . For our simulation studies, we restrict the spatial locations to  $[0, 1]^2$ . We simulate partial realizations of length 10000 from the two GPs.

We begin by simulating first, for  $i = 1, \dots, 10000$ ,  $\tilde{s}_i \sim U([0, 1]^2)$ , and then setting  $s_i = \sqrt{\tilde{s}_i}$ . Here for any  $s = (u, v)^T \in [0, 1]^2$ ,  $\sqrt{s} = (\sqrt{u}, \sqrt{v})^T$ . The strategy of taking square roots of the components of  $\tilde{s}_i$  ensured numerical stability of the corresponding covariance matrices. We then simulate from 10000 zero-mean multivariate normals with covariance matrices defined by the above stationary and nonstationary covariance functions. Generating from the multivariate normal distributions by parallelising the required Cholesky decomposition of the covariance matrix and subsequent multiplication of the Cholesky factor with the vector of standard normal random variables using ScaLAPACK (Scalable Linear Algebra Package) takes less than 40 seconds in our C code implementation on our 64 bit laptop (8 GB RAM and 2.3 GHz CPU speed), with just 4 cores.

### 8.2.2 Implementation of our method to detect strict stationarity

For our purpose, we first need to form  $\mathcal{N}_i$ ;  $i = 1, \dots, K$ . In the spatial setting, the  $K$ -means clustering of the locations  $s_i$ ;  $i = 1, \dots, 10000$ , seems to be very appropriate. The nearby locations based on the distances from the centroid, will be classified within the same cluster, which is desirable from the spatial perspective. Thus, once we select  $K$ , the  $K$ -means clustering yields the  $K$  clusters, which are  $\mathcal{N}_i$ ;  $i = 1, \dots, K$  in our notation. In our example, we select  $K = 250$ , so that there are about 40 observations per cluster on the average. We choose the clusterings such that there are at least 15 observations per cluster. As before, we consider the general purpose nonparametric bound  $c_j$  given by (7.2.3) for implementation of our method.



### Choice of $\hat{C}_1$

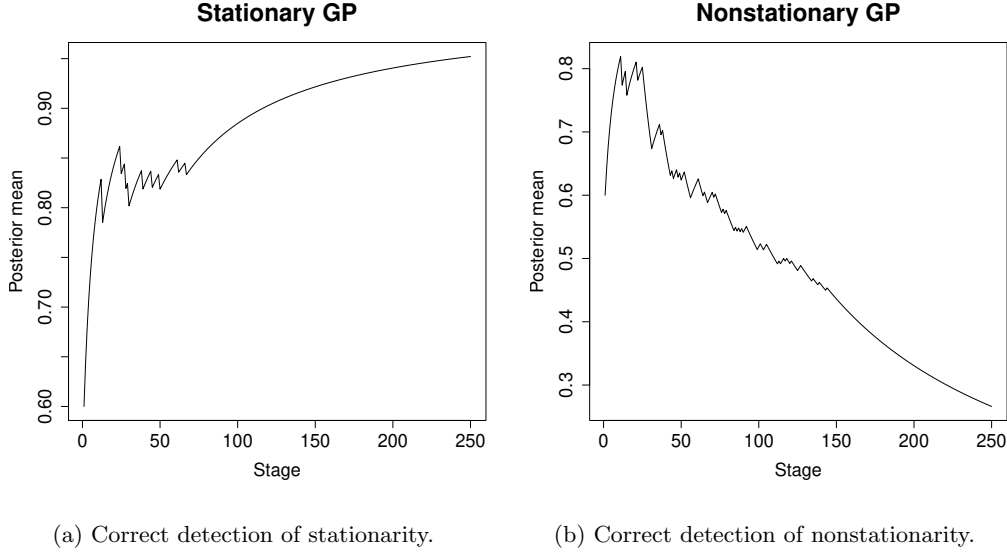
For the choice of  $\hat{C}_1$ , we first generate a sample of size 10000 from a zero mean GP with the Whittle covariance function of the form

$$Cov(X_{s_1}, X_{s_2}) = (\|s_1 - s_2\|/\psi)\mathcal{K}_1(\|s_1 - s_2\|/\psi), \quad (8.2.3)$$

where  $\mathcal{K}_1$  is the second kind modified Bessel function of order 1. For the same value of  $\psi$ , this covariance function has thicker tails than exponential correlation functions of the forms  $\exp(-\|s_1 - s_2\|^2/\psi)$  and  $\exp(-\|s_1 - s_2\|/\psi)$ . We set  $\psi = 0.8$  to achieve reasonable thickness of the tail of (8.2.3). With this covariance function, we then use the bound (7.2.3) and set  $\hat{C}_1$  to be the minimum positive value such that convergence to 1 is achieved. This  $\hat{C}_1$  can be interpreted as providing a reasonable bound for spatial processes with covariance functions with reasonably thick tails, but thinner than that of (8.2.3) with  $\psi = 0.8$ . With this method, we obtain  $\hat{C}_1 = 0.89$ . This value, being close to 1, suggests that the default choice  $\hat{C}_1 = 1$  still makes sense. Indeed, both the choices yielded the same results regarding the decision on stationarity or nonstationarity of the underlying process.

### 8.2.3 Results

Figure 8.2.1 shows the results of implementation of our theory to detect strong stationarity and nonstationarity of the data obtained from the two GPs. The bounds (7.2.3) correspond to  $\hat{C}_1 = 0.89$  obtained using the strategic procedure using (8.2.3). Panel (a) correctly asserts strict stationarity when the covariance is of the form (8.2.1) and correctly detects strict nonstationarity when the covariance is of the form (8.2.2). The entire methodology takes less than a second for parallel implementation on our 64 bit laptop using 4 cores.



**Figure 8.2.1:** Detection of strong stationarity and nonstationarity in spatial data drawn from GPs.

#### 8.2.4 Implementation of our method to detect covariance stationarity

As we demonstrated, our proposed method does an excellent job in capturing strict stationarity and nonstationarity of the underlying spatial stochastic process. In routine spatial modeling, however, strict stationarity and nonstationarity plays little role compared to covariance stationarity and covariance nonstationarity. Thus, it is more important to detect if the covariance in question is stationary or not. Although in our example it directly follows from our tests of strict stationarity that the covariances for the two GPs must be stationary and nonstationary, we directly check covariance stationarity using our Bayesian method formalized in Theorems 31 and 32.

For practical implementation, we convert the covariances  $\widehat{Cov}_{ih}$  given by (6.7.1) into correlations by dividing them by the relevant standard errors and initially set  $\mathcal{N}_{i,h_j,h_{j+1}} = \{(s_1, s_2) \in \mathcal{N}_i : h_j \leq \|s_1 - s_2\| < h_{j+1}\}; j = 1, \dots, 10$ , where  $h_1 = 0$  and  $h_j = h_{j-1} + 0.1$ , for  $j = 2, \dots, 10$ . We consider the nonparametric bound  $c_j$  given by (7.2.3) for all  $j = 1, \dots, 10$ , for both the GPs. But we found that these  $\mathcal{N}_{i,h_j,h_{j+1}}$  are too

large to be useful, as  $0 < \|s_1 - s_2\| < 0.04$ , for all  $(s_1, s_2)$  in most of the  $K$ -means clusters that we obtained. Indeed, only three neighborhoods defined by  $h_1 = 0$ ,  $h_2 = 0.02$ ,  $h_3 = 0.03$  and  $h_4 = 0.04$ , turned out to be appropriate.

We again fix  $K = 250$  clusters such that each cluster contains at least 15 observations.

### Choice of $\hat{C}_1$

To obtain appropriate choice of  $\hat{C}_1$  for detecting covariance stationarity, we consider three strategies. Our first method in this regard corresponds to using  $\hat{C}_1$  for strict stationarity. Thus, the first strategy yields  $\hat{C}_1 = 0.89$ .

For the second strategy, we utilize the GP realization with covariance function (8.2.3). Here we choose the minimum value of  $\hat{C}_1$  such that (7.2.3) yielded convergence to 1 for all  $\mathcal{N}_{i,h_j,h_{j+1}}$ ;  $j = 1, 2, 3$ . This gave  $\hat{C}_1 = 0.412$ .

In the third strategy, we chose the minimum value of  $\hat{C}_1$  that yielded convergence to 1 for all  $\mathcal{N}_{i,h_j,h_{j+1}}$ ;  $j = 1, 2, 3$  for one dataset and convergence to 0 for the other dataset. In our case, this strategy again gave  $\hat{C}_1 = 0.412$ .

The strategic choice  $\hat{C}_1 = 0.412$  successfully detected covariance stationarity and nonstationarity. However, the choice  $\hat{C}_1 = 0.89$  turned out to be too large to detect covariance nonstationarity. This is in keeping with the issue that detection of strict stationarity requires a bound that must also ensure covariance stationarity, and hence such a bound must be larger than that for covariance stationarity.

Again, our parallel implementation takes less than a second on our laptop, for each  $\mathcal{N}_{i,h_j,h_{j+1}}$ . This quick computation ensures that choice of  $\hat{C}_1$  is not a computationally demanding exercise.

Figure 8.2.2 shows the results associated with  $\mathcal{N}_{i,h_1,h_2}$ ,  $\mathcal{N}_{i,h_2,h_3}$  and  $\mathcal{N}_{i,h_3,h_4}$ , for  $i = 1, \dots, K$ , where  $K = 250$  as before, and  $\hat{C}_1 = 0.89$ . The figure shows that whenever the data arises from the GP with covariance of the form (8.2.1), our Bayesian method correctly identifies covariance stationarity for every  $j$ . Indeed, for all  $j = 1, 2, 3$ ,

covariance stationarity is clearly indicated. On the other hand, when the data arises from the GP with the nonstationary covariance (8.2.2), convergence to 0 is indicated with  $\mathcal{N}_{i,h_3,h_4}$ . As per Theorem 32, this shows nonstationarity of the covariance structure.

### 8.2.5 Detection of strict nonstationarity in mixtures of stationary and nonstationary covariances

We now consider realizations from zero-mean GPs with covariances of the form

$$\text{Cov}(X_{s_1}, X_{s_2}) = p \exp(-5\|s_1 - s_2\|^2) + (1 - p) \exp(-5\|\sqrt{s_1} - \sqrt{s_2}\|^2), \quad (8.2.4)$$

where  $0 < p < 1$ . In particular, using our Bayesian theory, we attempt to detect strict and weak nonstationarity of the process when  $p = 0.9, 0.99, 0.999, 0.9999, 0.99999$ . Note that in these cases, although most of the weight concentrates on the stationary part of (8.2.4), the little mass on the nonstationary part makes the covariance nonstationary, and it is important to detect such subtle difference between stationarity and nonstationarity. As before, we set  $K = 250$  clusters with each cluster containing at least 15 observations.

We consider the same way of data generation from GP as before, and the same way of implementation. We again use the same form of the bound  $c_j$  as (7.2.3), with  $\hat{C}_1 = 0.89$  and  $\hat{C}_1 = 1$  for detection of strict nonstationarity, as before. These choices put up excellent performances and are in agreement with each other, in spite of the subtlety involved in this exercise. Figure 8.2.3, corresponding to  $\hat{C}_1 = 0.89$ , shows that our Bayesian method correctly identifies nonstationarity in all the cases.

### 8.2.6 Detection of covariance nonstationarity in mixtures of stationary and nonstationary covariances

The same strategies discussed in Section 8.2.4, adapted in this situation, yielded effective bounds of the form (7.2.3) with  $\hat{C}_1 = 0.412$ , as before. We briefly discuss the second

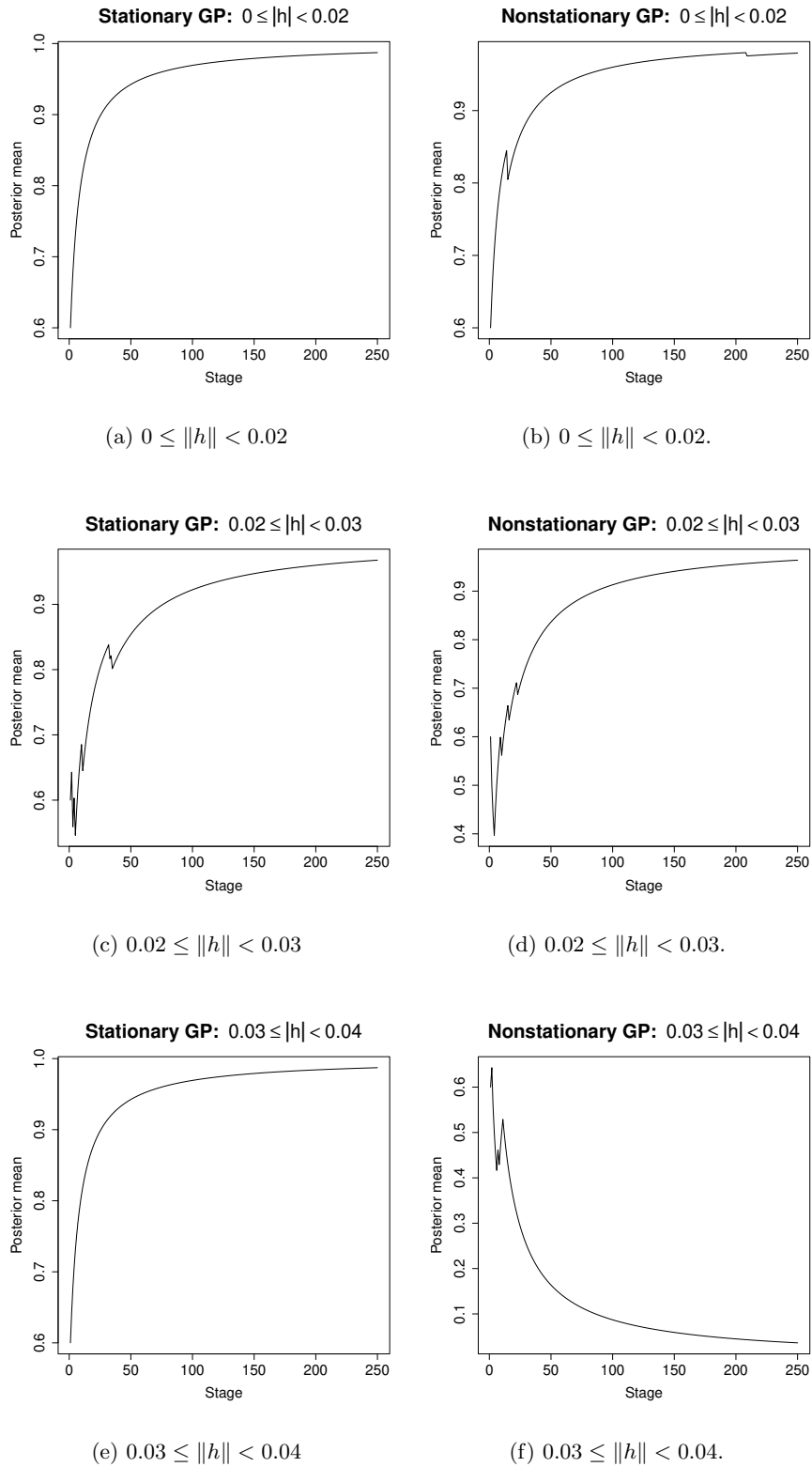
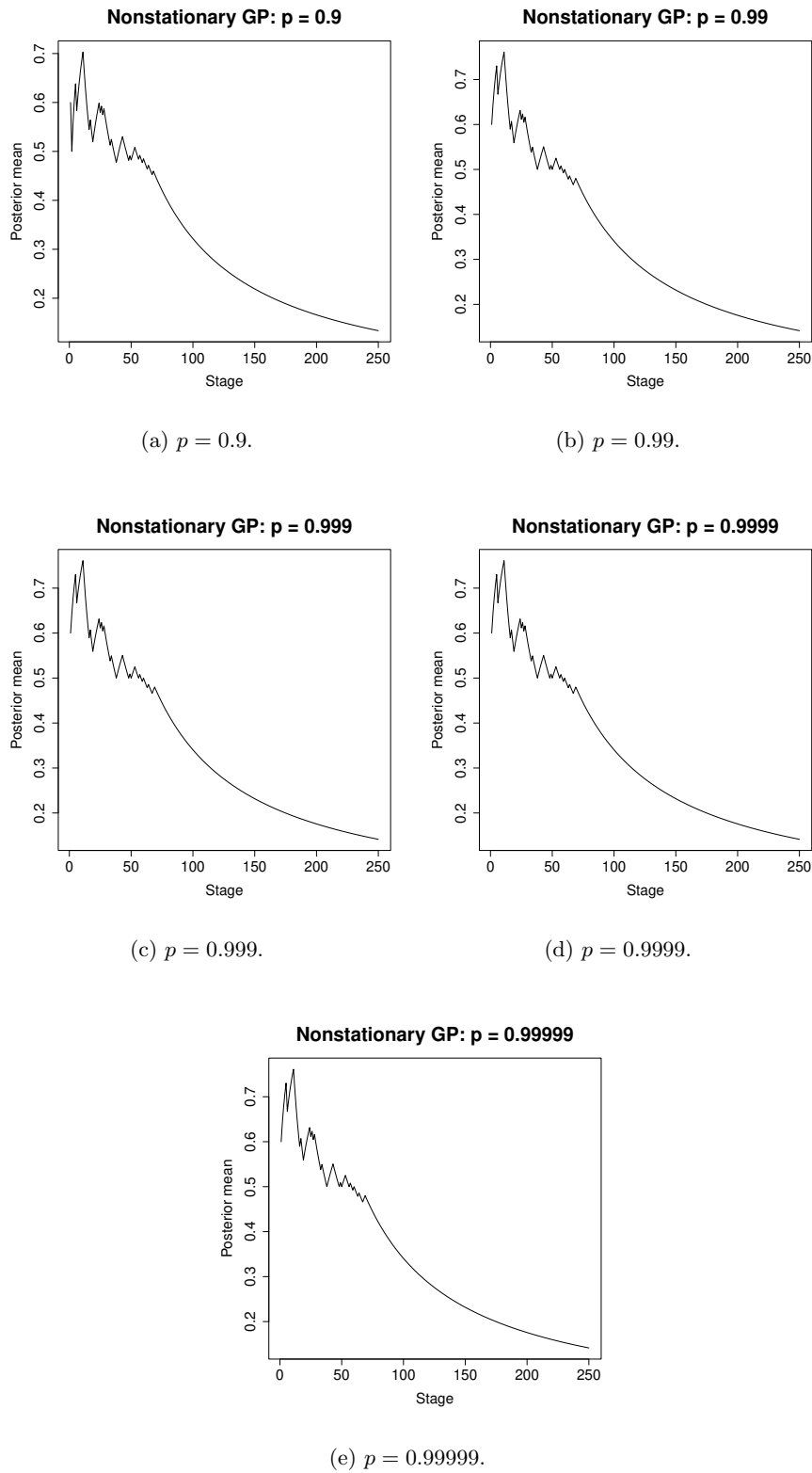


Figure 8.2.2: Detection of covariance stationarity and nonstationarity in spatial data drawn from GPs.



**Figure 8.2.3:** Detection of strong nonstationarity in spatial data drawn from GP with covariance structure (8.2.4) with  $p = 0.99999$ .

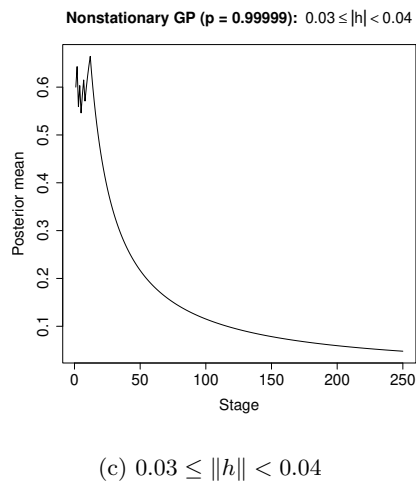
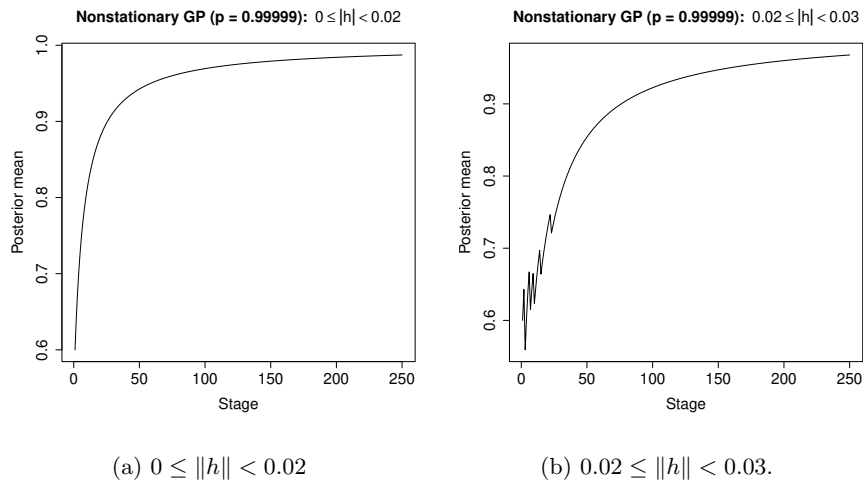
procedure of adapting the strategy to the current scenario. Note that the first procedure does not need any change at all.

To implement our second strategy in this case, we need a benchmark dataset for which covariance stationarity has been established. We thus consider the GP data with covariance of the form (8.2.1), whose covariance stationarity is established. For any new dataset for which covariance stationarity needs to be checked, in this case, any dataset with covariance structure of the form (8.2.4), we consider the same bound starting with  $\hat{C}_1 = 0.89$ . We then gradually decrease  $\hat{C}_1$  for both the datasets until we arrive at a point that discriminates covariance stationarity and nonstationarity, in the same way as discussed in Section 8.2.4. With this method, we obtain  $\hat{C}_1 = 0.412$ , which shows covariance stationarity for (8.2.1) but covariance nonstationarity for (8.2.4). Recall that  $\hat{C}_1 = 0.412$  also resulted with respect to the GP realization for the Whittle covariance function (8.2.3).

Again we set  $K = 250$ , with each cluster consisting of a minimum of 15 observations. Figure 8.2.4, corresponding to  $\hat{C}_1 = 0.412$  and  $p = 0.99999$  in the covariance structure (8.2.4), shows that this procedure does an excellent job in detecting covariance nonstationarity even in such a subtle situation. Indeed, the same  $\hat{C}_1 = 0.412$  very successfully captured covariance nonstationarity for all other values of  $p$ , namely,  $p = 0.9, 0.99, 0.999, 0.9999$  (figures omitted for brevity).

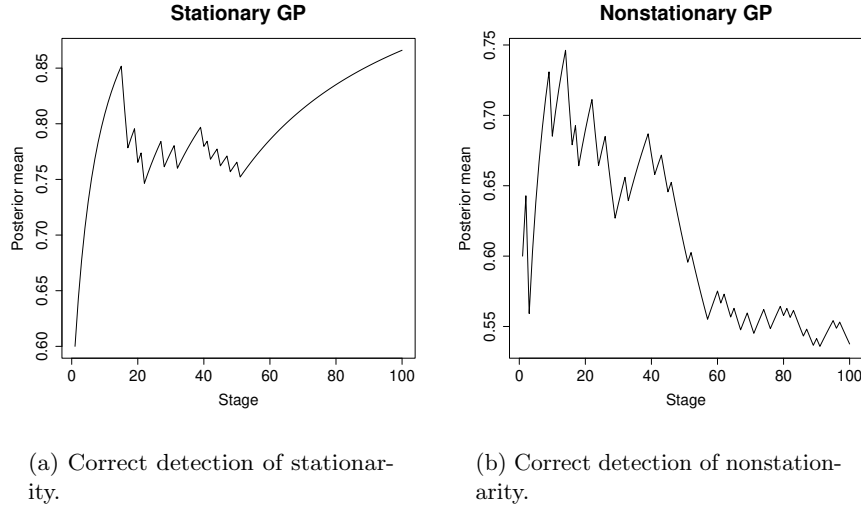
### 8.2.7 Spatial experiments with smaller data sets

We now repeat all the above experiments with datasets of sizes 1000. We consider  $K = 100$  clusters with average cluster size 10. For checking strict stationarity, our first strategy of fixing  $\hat{C}_1$ , using the Whittle covariance function (8.2.3) yielded  $\hat{C}_1 = 0.02$ , which produced too small bounds to be useful. On the other hand, the second procedure gave  $\hat{C}_1 = 1.24$ , which yielded reliable results, even for these small data sets. Figures 8.2.5 and 8.2.6 depict the results for  $\hat{C}_1 = 1.24$ . For covariance stationarity, these



**Figure 8.2.4:** Detection of covariance nonstationarity in spatial data drawn from GP with covariance structure (8.2.4) with  $p = 0.99999$ .



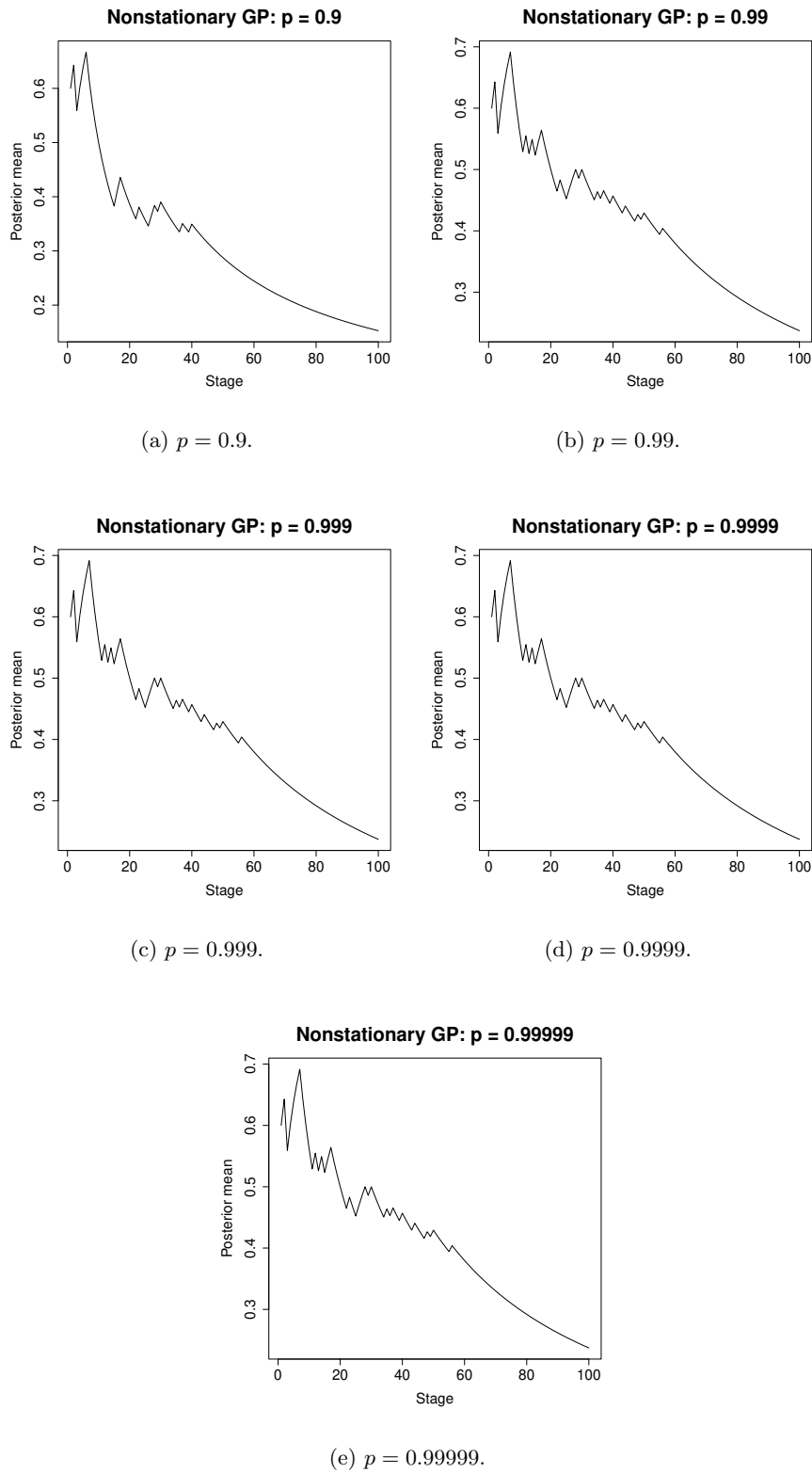


**Figure 8.2.5:** Detection of strong stationarity and nonstationarity in spatial data of size 1000 drawn from GPs.

small data sets were able to produce a single valid region  $\mathcal{N}_{i,h_1,h_2}$ , defined by  $h_1 = 0$  and  $h_2 = 0.1$ , and hence, with only this region, verification of covariance stationarity or nonstationarity is not possible. But since the underlying model is GP, covariance stationarity is equivalent to strict stationarity, and even for non-Gaussian processes, strict stationarity would imply covariance stationarity (although strict nonstationarity need not imply covariance nonstationarity).

### 8.2.8 Comparison with existing methods

In spatial statistics, formal methods of testing stationarity or nonstationarity are rare, and mostly exploratory data analysis is used to informally check stationarity. However, [Bandopadhyay and Rao \(2017\)](#) have introduced some tests for checking covariance stationarity, under a variety of assumptions. These methods seem to be more general compared to the existing ones. An *R*-code for implementing their method is available at the webpage of the first author. Given a dataset, the code calculates two test statistics,



**Figure 8.2.6:** Detection of strong nonstationarity in spatial data of size 1000 drawn from GP with covariance structure (8.2.4) with  $p = 0.99999$ .

denoted by  $T$  and  $V$ , along with the corresponding  $P$ -values under the null hypothesis of stationarity. The statistic  $V$  has been proposed in [Bandopadhyay et al. \(2017\)](#).

We apply their methods to our simulated spatial datasets in order to compare with our results. However, with data size 10000, it turned out that obtaining a result within reasonable time limits with the aforementioned  $R$  code is almost infeasible. Instead, we applied their methods to data sets of sizes 1000, 3000 and 5000. The run times for the  $R$  code for these data sizes are about 28 seconds, 5 minutes and 12 minutes, respectively.

Table 8.2.1 presents the results of the tests applied to our simulated datasets. In all the cases, the  $T$  statistic failed to reject the null hypothesis of stationarity, even though there is only one case of true null stationarity. On the other hand, the  $V$ -statistic performs much better, with its performance consistently improving with increasing sample size, as vindicated by the corresponding  $P$ -values. But observe that for sample size 1000, even the  $V$ -statistic fails to reject the null hypothesis of stationarity at the 5% level for most cases where the actual model is nonstationary. Moreover, at the 5% level, this statistic rejects the true null stationary model for sample sizes 3000 and 5000.

Thus, compared to our Bayesian idea, the overall performance of both the statistics  $T$  and  $V$  does not seem to be satisfactory for the models that we considered.

Moreover, from the methodological perspective, the tests of [Bandopadhyay and Rao \(2017\)](#) check covariance stationarity only, not strict stationarity. Various assumptions, which may be difficult to verify in practice, are also required. In contrast, our Bayesian method requires the only assumption of local stationarity that is expected to hold in practice, and allows for identification of both weak and strict stationarity.

### 8.3 Detection of stationarity and nonstationarity in spatio-temporal data

We now apply our techniques in ascertaining stationarity and nonstationarity in spatio-temporal data, where both spatial and temporal components play important roles. For our simulation studies, we consider covariance functions of the following forms:

$$\text{Cov}(X_{(s_1, t_1)}, X_{(s_2, t_2)}) = \exp(-5\|s_1 - s_2\|^2) \times \frac{\rho^{|t_1 - t_2|}}{1 - \rho^2}, \quad (8.3.1)$$

$$\text{Cov}(X_{(s_1, t_1)}, X_{(s_2, t_2)}) = \exp(-5\|\sqrt{s_1} - \sqrt{s_2}\|^2) \times \frac{\rho^{|t_1 - t_2|}}{1 - \rho^2}, \quad (8.3.2)$$

and

$$\text{Cov}(X_{(s_1, t_1)}, X_{(s_2, t_2)}) = (p \exp(-5\|s_1 - s_2\|^2) + (1 - p) \exp(-5\|\sqrt{s_1} - \sqrt{s_2}\|^2)) \times \frac{\rho^{|t_1 - t_2|}}{1 - \rho^2}, \quad (8.3.3)$$

for all  $s_1, s_2 \in \mathbb{R}^2$ ,  $t_1, t_2 \in \mathbb{R}^+$  and  $\rho \in \mathbb{R}$ . Note that  $\frac{\rho^{|t_1 - t_2|}}{1 - \rho^2}$  is the covariance function associated with an  $AR(1)$  model with parameter  $\rho$ . The forms of the covariance functions (8.3.1), (8.3.2) and (8.3.3) show that the covariance parts associated with spatial and temporal components are separated from each other, thanks to the product forms. Covariance functions with such a property are known as separable covariance functions. In (8.3.3),  $p \in [0, 1]$ , as before. If  $p = 0$ , then (8.3.3) reduces to (8.3.2) and to (8.3.1) if  $p = 1$ .

Note that if  $|\rho| < 1$ , then (8.3.1) is a stationary covariance function, and nonstationary otherwise. On the other hand, (8.3.2) and (8.3.3) are both nonstationary covariance functions, irrespective of the value of  $\rho$ .

For our simulation experiments, we consider zero-mean GPs  $X_{(s, t)}$  with the above covariance functions, restricting the spatial locations on  $[0, 1]^2$  and setting the time

points  $t_i = i$ , for  $i \geq 1$ . We simulate, for  $i = 1, \dots, 100$ ,  $\tilde{s}_i \sim U([0, 1]^2)$  and set  $s_i = \sqrt{\tilde{s}_i}$ . We set  $t_i = i$ , for  $i = 1, \dots, 100$ . This defines covariance matrices for 10000-dimensional multivariate normal associated with the underlying GPs. Note that such covariance matrices are Kronecker products of the spatial and temporal covariance matrices, thanks to separability.

Observe that the above separable covariance matrices correspond to separable spatio-temporal processes of the form

$$X_{(s,t)} = X_{(s,t-1)} + \epsilon_{(s,t)}, \quad (8.3.4)$$

for  $t = 1, 2, \dots$ , where  $X_{(s,0)} = \mathbf{0}$  (null vector), and  $\epsilon_{(s,t)}$  are zero-mean GPs independent in time, but with spatial covariance with forms same as the spatial parts in (8.3.1), (8.3.2) and (8.3.3). With the above representation, generation of 10000 realization takes about a second, even in  $R$ .

To construct  $\mathcal{N}_i$ ,  $i = 1, \dots, K$ , we consider  $K$ -means clustering of the points

$$\{(s_i, t_j); i = 1, \dots, 100; j = 1, \dots, 100\},$$

into  $K = 250$  clusters.

### 8.3.1 Choice of the bound $c_j$ in the spatio-temporal case

We consider the bound of the form (7.2.3) as before. As regards,  $\hat{C}_1$ , we found that  $\hat{C}_1 = 0.5$  performed adequately for the entire suite of our simulation experiments in the spatio-temporal scenario. However, we also consider a strategy for obtaining  $\hat{C}_1$  using ideas similar to the spatial setup, detailed below.

We first generate a sample of size 10000 from a zero mean GP with the covariance

function of the following form:

$$\text{Cov}(X_{(s_1,t_1)}, X_{(s_2,t_2)}) = (\|s_1 - s_2\|/\psi)\mathcal{K}_1(\|s_1 - s_2\|/\psi) \times \frac{\xi^{|t_1-t_2|}}{1 - \xi^2}, \quad (8.3.5)$$

with  $\psi = 0.8$  and  $\xi = 0.999999$ . Note that this covariance function corresponds to a model of the form (8.3.4) with  $X_{(s,0)} = \mathbf{0}$  and zero-mean GPs  $\epsilon_{(s,t)}$  independent in time, with spatial covariance given by the spatial form in (8.3.5). The parameter values  $\psi = 0.8$  and  $\xi = 0.999999$  are chosen to make the underlying spatio-temporal process reasonably close to nonstationarity with respect to space and time.

We then choose that minimum value of  $\hat{C}_1$  such that the spatio-temporal process remains stationary. This minimum value, for checking strict stationarity, is given by  $\hat{C}_1 = 0.37$ , which is reasonably close to  $\hat{C}_1 = 0.5$  that worked well for our experiments. Again, we obtained same results for both the values of  $\hat{C}_1$ , and we report results for  $\hat{C}_1 = 0.37$ .

However, for weak stationarity, we again failed to obtain multiple valid intervals for realizations of size 10000 from the zero-mean GP with covariance (8.3.5). Indeed, we could obtain only a single interval  $[0, 0.15]$ . Hence, in that case we consider  $\hat{C}_1 = 0.5$ .

Below we discuss the experimental designs for our various simulation experiments.

### 8.3.2 Spatial and temporal stationarity

We generate partial realizations of length 10000 from the zero mean GP with covariance function (8.3.1) using the formulation (8.3.4), with  $\rho = 0.8$  and also with  $\rho = 0.999999$ . Thus, the spatio-temporal GPs are strictly stationary, and our Bayesian method is expected to reflect this. The latter situation is quite subtle, as the difference with temporal nonstationarity is negligible.

Apart from strict stationarity, we also investigate weak stationarity, focussing on the subtle situation where  $\rho = 0.999999$ .

### 8.3.3 Spatio-temporal nonstationarity

Recall that spatio-temporal nonstationarity occurs in our cases when  $|\rho| \geq 1$  in (8.3.1) and when covariances (8.3.2) or (8.3.3) are chosen. We experiment with (8.3.1) with  $\rho = 1$ , (8.3.2) with  $\rho = 0.8$  and  $\rho = 1$ , (8.3.3) with  $p = 0.99999$  and  $\rho = 0.8$ . The latter is a subtle situation where nonstationarity is quite difficult to ascertain. Note that if nonstationarity can be captured by our Bayesian method in this situation, then so is possible for larger values of  $\rho$  taking the temporal part closer to nonstationarity. With the last, subtle situation, we also investigate covariance nonstationarity.

### 8.3.4 Results

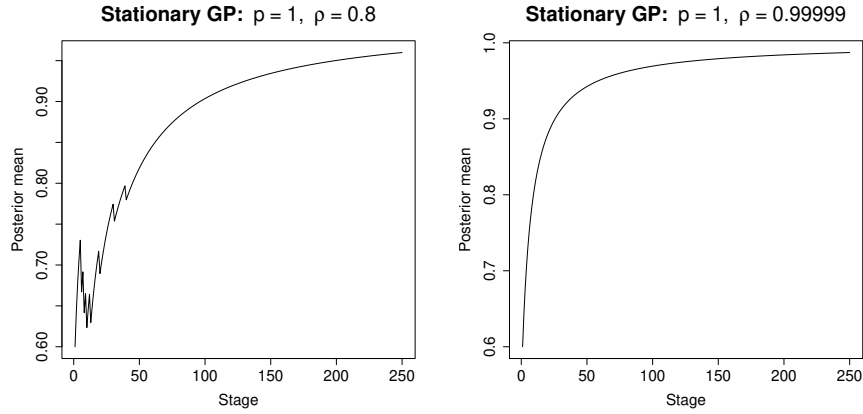
Figure 8.3.1, diagrammatically representing our Bayesian procedure, vindicates that the stochastic processes associated with covariance function (8.3.1) with  $\rho = 0.8$  and  $\rho = 0.99999$ , are indeed strictly stationary. On the other hand, the processes corresponding to (8.3.1) with  $\rho = 1$ , (8.3.2) with  $\rho = 0.8$  and  $\rho = 1$ , (8.3.3) with  $p = 0.99999$  and  $\rho = 0.99999$ , are all correctly detected by our Bayesian method as strictly nonstationary.

Figure 8.3.2 depicts the results of investigation of weak stationarity for the covariance (8.3.1) with  $\rho = 0.99999$ . For the covariance (8.3.3) with  $p = 0.99999$  and  $\rho = 0.8$ , Figure 8.3.3 presents the results of our Bayesian technique. In both the cases, success of our Bayesian proposal is clearly borne out.

### 8.3.5 Investigation of spatio-temporal stationarity with smaller sample size

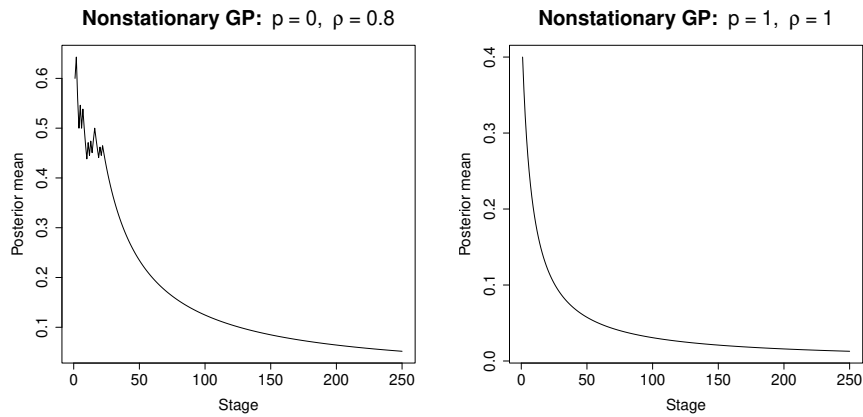
We now investigate stationarity of the above spatio-temporal models using much smaller sample sizes. In particular, we consider 50 locations and 20 time points only, and  $K = 100$  clusters. We ensured at least 3 data points in each cluster. Our strategy for choosing  $\hat{C}_1$ , detailed in Section 8.3.1, gave  $\hat{C}_1 = 0.87$  for investigating strict stationarity. Again,  $\hat{C}_1 = 0.5$  yielded the same conclusions. Figure 8.3.4, depicting the results of our analysis

8.3. DETECTION OF STATIONARITY AND NONSTATIONARITY IN SPATIO-TEMPORAL DATA



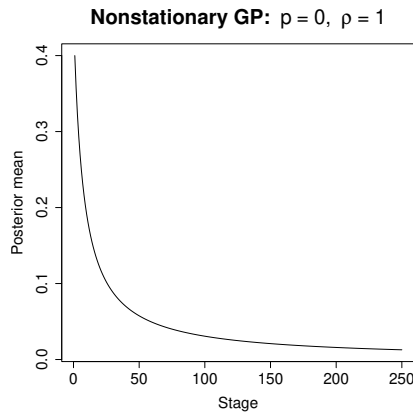
(a) Correct detection of stationarity.

(b) Correct detection of stationarity.



(c) Correct detection of nonstationarity.

(d) Correct detection of nonstationarity.

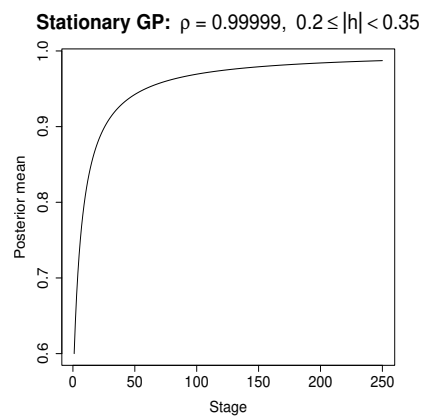
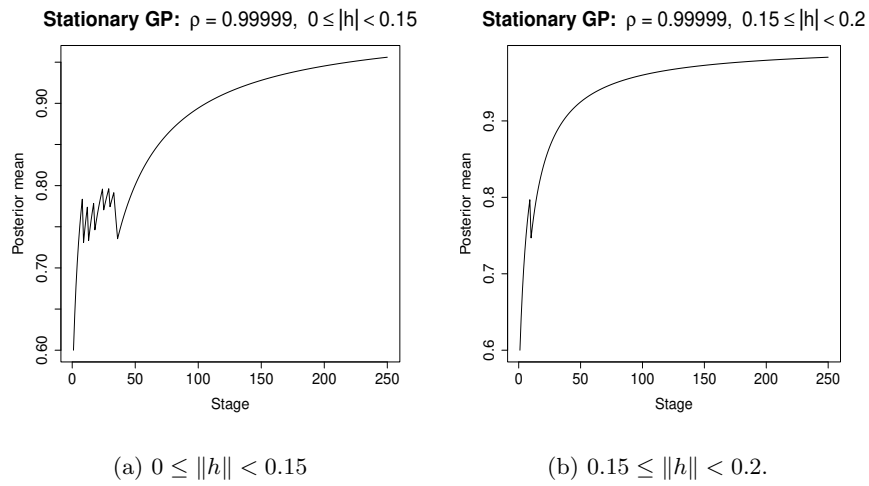


(e) Correct detection of nonstationarity.

Figure 8.3.1: Detection of strong stationarity and nonstationarity in spatio-temporal data drawn from GPs.

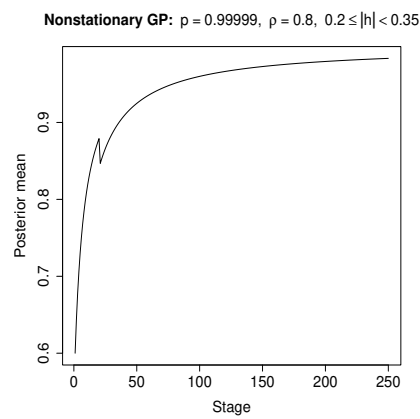
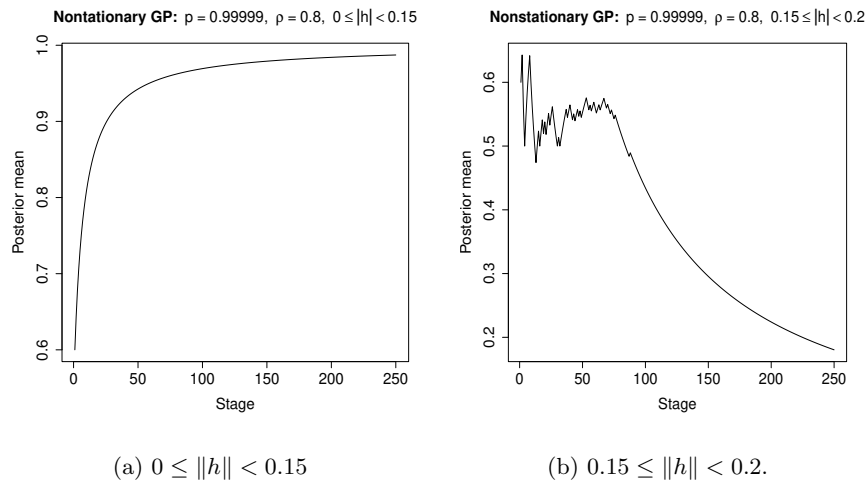


8.3. DETECTION OF STATIONARITY AND NONSTATIONARITY IN SPATIO-TEMPORAL DATA



**Figure 8.3.2:** Detection of covariance stationarity in spatio-temporal data drawn from GP with covariance structure (8.3.1) with  $\rho = 0.99999$ .

### 8.3. DETECTION OF STATIONARITY AND NONSTATIONARITY IN SPATIO-TEMPORAL DATA



**Figure 8.3.3:** Detection of covariance nonstationarity in spatio-temporal data drawn from GP with covariance structure (8.3.3) with  $p = 0.99999$  and  $\rho = 0.8$ .

for  $\hat{C}_1 = 0.87$ , indicates correct decisions on strict stationarity and nonstationarity in all the cases, even for such small data size.

However, validating covariance stationarity could not be achieved for such small samples, as we again ended up with the single interval  $\mathcal{N}_{i,h_1,h_2}$  with  $h_1 = 0$  and  $h_2 = 0.2$ .

### 8.3.6 Comparison with existing methods

As in the spatial case, for the spatio-temporal setup, formal methods of testing stationarity are very rare in the literature. Recently, some methods in this direction are proposed in [Bandopadhyay \*et al.\* \(2017\)](#). Indeed, the authors propose as many as 10 test statistics to detect covariance stationarity, under a variety of assumptions. The main ideas are similar to the testing ideas in the spatial setup proposed in [Bandopadhyay and Rao \(2017\)](#). A relevant *R* code is provided in the webpage of the first author, but it failed to work for our simulated spatio-temporal datasets, possibly because the methods are heavily dependent on choices of the underlying parameters involved in their methods. Instead, we apply our Bayesian methodology on the spatio-temporal models and simulation designs to which [Bandopadhyay \*et al.\* \(2017\)](#) applied their testing methods.

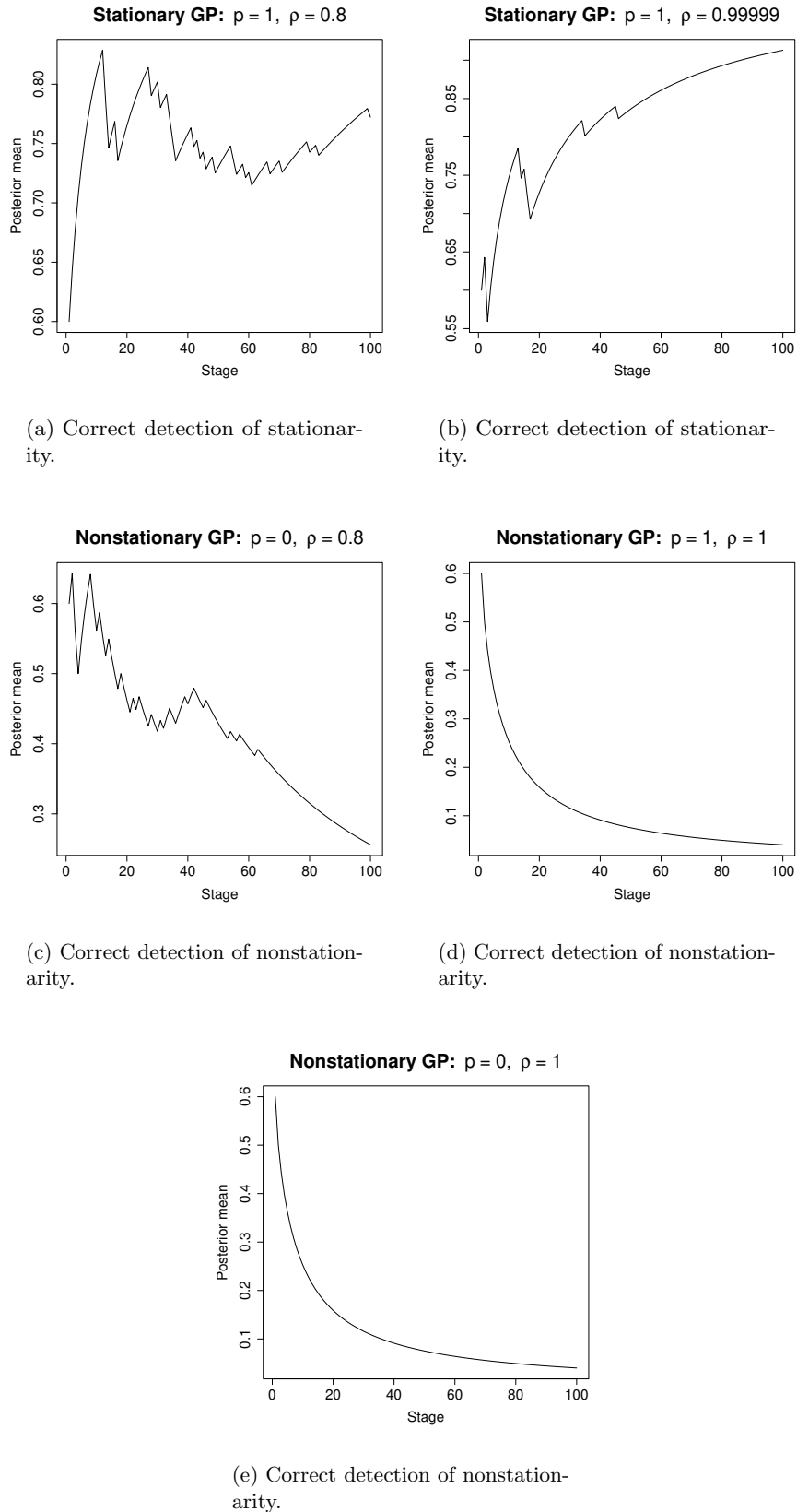
Following [Bandopadhyay \*et al.\* \(2017\)](#), we consider zero mean spatio-temporal processes, with  $T = 200$  time points and  $m = 100$  or  $500$  locations drawn uniformly from  $[-\frac{\lambda}{2}, \frac{\lambda}{2}]$ . We then apply our Bayesian procedure to the 5 spatio-temporal models considered by [Bandopadhyay \*et al.\* \(2017\)](#), under the same setups, described below.

#### Simulations under stationarity with exponential spatial covariance function

We generate data from the following stationary models:

$$(S1) \quad X_{(s,t)} = 0.5X_{(s,t-1)} + \epsilon_{(s,t)}, \text{ where } X_{s,0} = \mathbf{0} \text{ and } \epsilon_{(s,t)} \text{ are zero mean GPs independent}$$

### 8.3. DETECTION OF STATIONARITY AND NONSTATIONARITY IN SPATIO-TEMPORAL DATA



**Figure 8.3.4:** Detection of strong stationarity and nonstationarity in spatio-temporal data drawn from GPs with 50 locations and 20 time points.

over time with spatial covariance structure

$$\text{Cov}(\epsilon_{(s_1,t)}, \epsilon_{(s_2,t)}) = \exp(-\|s_1 - s_2\|/\psi). \quad (8.3.6)$$

The above model defines a spatially and temporally stationary Gaussian random field.

(S2)  $X_{(s,t)} = 0.5X_{(s,t-1)} + 0.4X_{(s,t-1)}\epsilon_{(s,t-1)} + \epsilon_{(s,t)}$ , where  $X_{s,0} = \mathbf{0}$  and  $\epsilon_{(s,t)}$  are zero mean GPs independent over time with spatial covariance (8.3.6). This model is a spatially and temporally non-Gaussian random field.

For both the above models, we set  $\lambda = 5$  for simulating the locations, and fix  $\psi = 0.5$  and 1 for two sets of data simulations for each of  $(m = 100, T = 200)$  and  $(m = 500, T = 200)$  sample sizes.

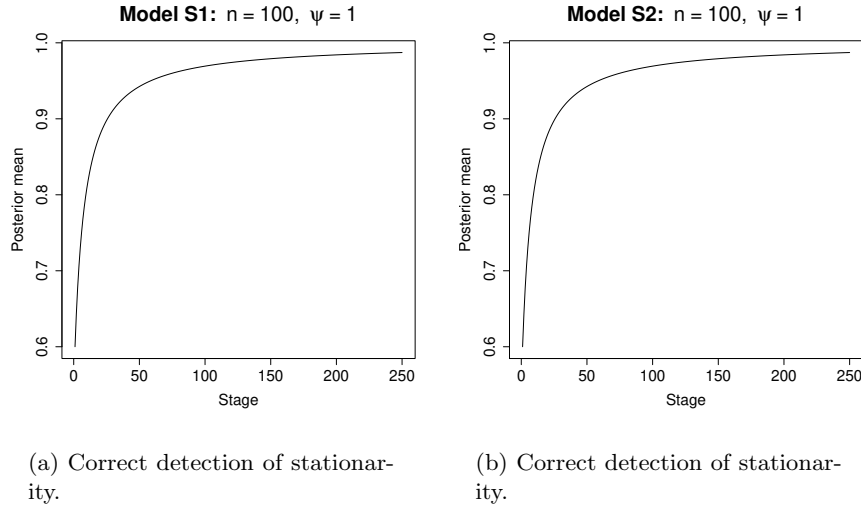
For checking strict stationarity, for sample size  $(m = 100, T = 200)$ , our strategy for choosing  $\hat{C}_1$ , detailed in Section 8.3.1, gave  $\hat{C}_1 = 0.042$ , and for  $(m = 500, T = 200)$ , we obtained  $\hat{C}_1 = 0.045$ . As before, we consider  $K = 250$  clusters in both the cases.

For covariance stationarity, we obtained  $\hat{C}_1 = 0.4$  for both  $(m = 100, T = 200)$  and  $(m = 500, T = 200)$ . For the first sample size, we obtained  $\mathcal{N}_{i,h_j,h_{j+1}}$  defined by  $h_1 = 0$ ,  $h_2 = 0.4$ ,  $h_3 = 0.7$ ,  $h_4 = 0.9$ ,  $h_5 = 2$ ,  $h_6 = 3$ . For the second sample size, we also obtained  $h_7 = 4$  for model  $S1$  when  $\psi = 5$  and for model  $S2$  when  $\psi = 1$  and  $\psi = 5$ .

For brevity we show the strict and weak stationarity convergence results only for  $(m = 100, T = 200)$ , with  $\psi = 1$ , depicted as Figures 8.3.5, 8.3.6 and 8.3.7.

### Simulations under stationarity with Whittle spatial covariance function

Following [Bandopadhyay et al. \(2017\)](#) we now repeat the above experiments with the same models  $S1$  and  $S2$  but with the exponential covariance functions replaced with the Whittle covariance function (8.2.3), with  $\psi = 0.37$  and  $0.72$ . Note that the values of



**Figure 8.3.5:** Detection of strong stationarity in spatio-temporal data drawn from models  $S1$  and  $S2$  with sample size 100 locations and 200 time points, with  $\psi = 1$  and  $\lambda = 5$ .

$\hat{C}_1$  remain the same as before; however, the minimum values of  $\hat{C}_1$  for which covariance stationarities were achieved, varied between 0.15, 0.2 and 0.3.

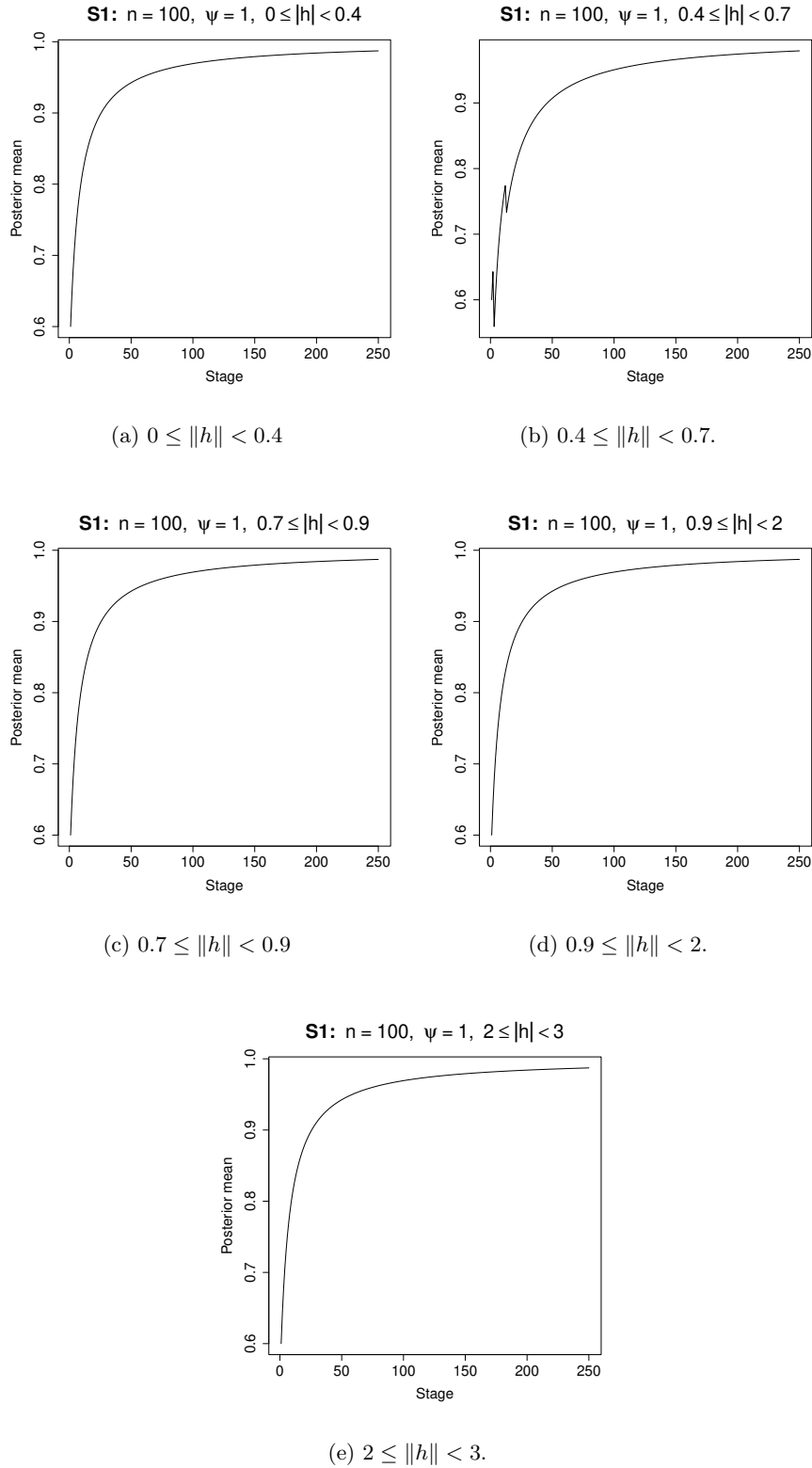
As expected, we obtained excellent results in all the cases, but present the results corresponding to  $(m = 100, T = 200)$  and  $\psi = 0.72$  for brevity. Figures 8.3.8, 8.3.9 and 8.3.10 depict our Bayesian results regarding strict and weak stationarities of the models  $S1$  and  $S2$ .

### Simulations under nonstationarity

We now apply our Bayesian methodology to the three nonstationary models and setups considered by [Bandopadhyay et al. \(2017\)](#).

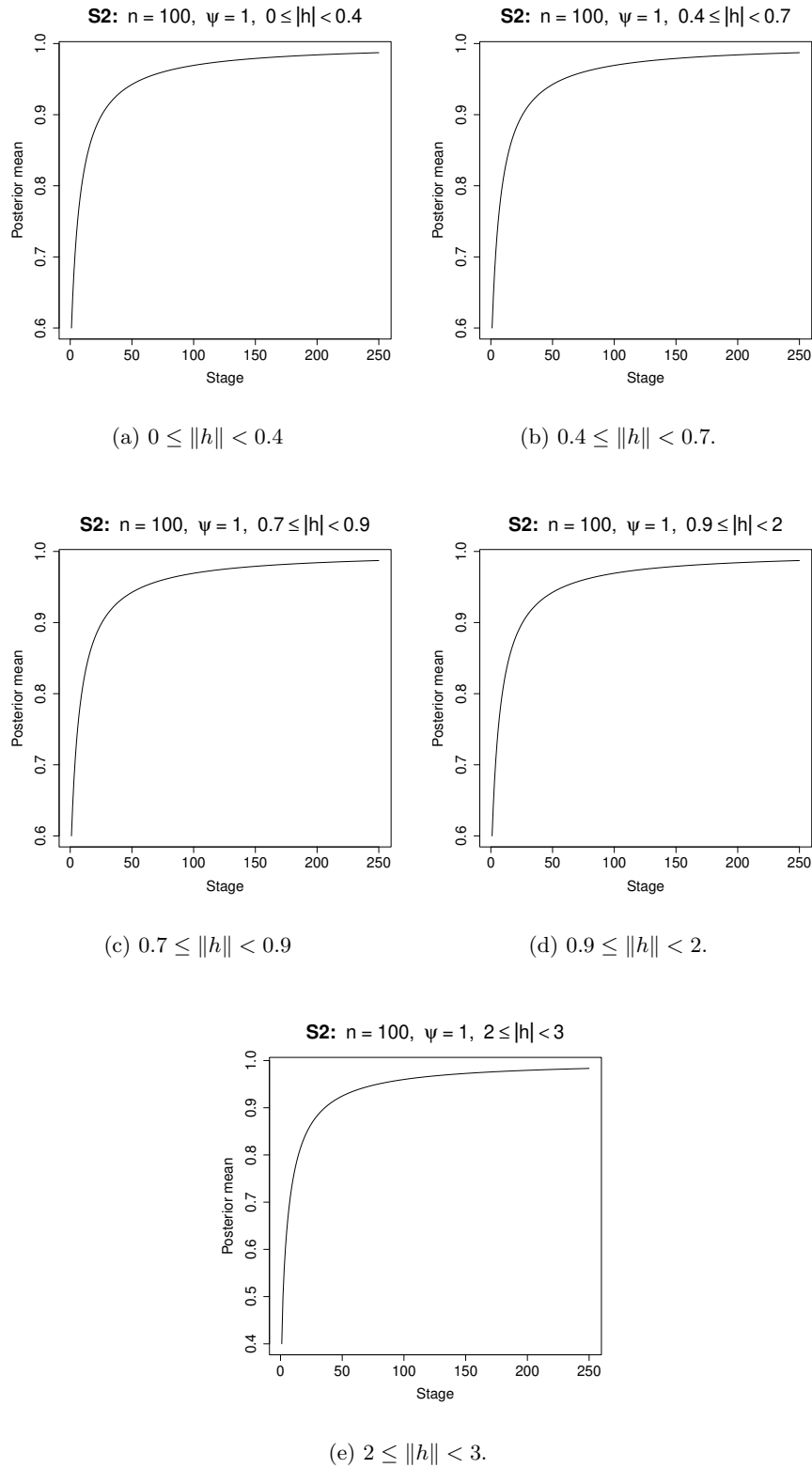
(NS1)  $X_{(s,t)} = 0.5X_{(s,t-1)} + \left(1.3 + \sin\left(\frac{2\pi t}{400}\right)\right)\epsilon_{(s,t)}$ , where  $X_{s,0} = \mathbf{0}$  and  $\epsilon_{(s,t)}$  are zero mean GPs independent over time with spatial covariance structure (8.3.6). Note that this is a temporally nonstationary but spatially stationary Gaussian random field. We consider  $\psi = 0.5$  and 1, and  $\lambda = 5$  for the simulations.

8.3. DETECTION OF STATIONARITY AND NONSTATIONARITY IN SPATIO-TEMPORAL DATA



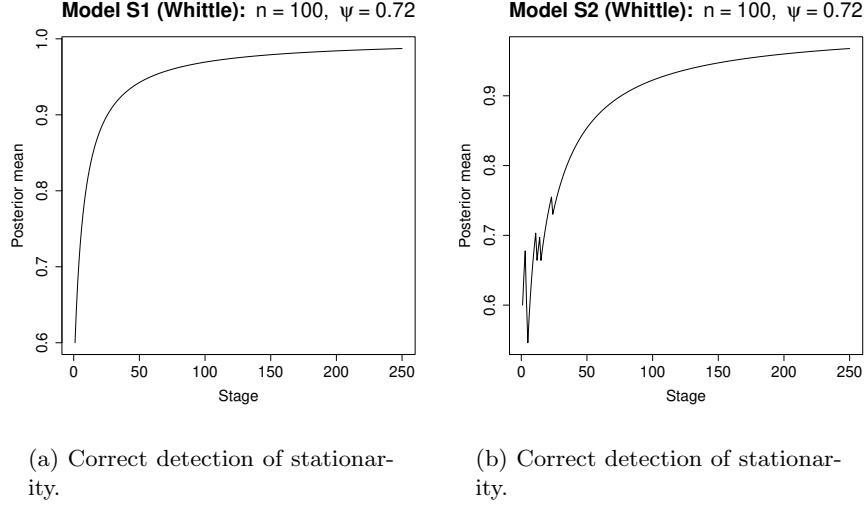
**Figure 8.3.6:** Detection of covariance stationarity in spatio-temporal data drawn from model  $S1$  with sample size 100 locations and 200 time points, with  $\psi = 1$  and  $\lambda = 5$ .

8.3. DETECTION OF STATIONARITY AND NONSTATIONARITY IN SPATIO-TEMPORAL DATA



**Figure 8.3.7:** Detection of covariance stationarity in spatio-temporal data drawn from model  $S2$  with sample size 100 locations and 200 time points, with  $\psi = 1$  and  $\lambda = 5$ .





**Figure 8.3.8:** Detection of strong stationarity in spatio-temporal data drawn from models  $S1$  and  $S2$  with sample size 100 locations and 200 time points, corresponding to Whittle spatial covariance with  $\psi = 0.72$  and  $\lambda = 5$ .

(NS2)  $X_{(s,t)} = 0.5X_{(s,t-1)} + 0.4X_{(s,t-1)}\epsilon_{(s,t-1)} + \eta_{(s,t)}$ , where  $X_{s,0} = \mathbf{0}$  and  $\eta_{(s,t)}$  are zero mean GPs independent over time with nonstationary spatial covariance given as follows.

$$Cov(\eta_{(s_1,t)}, \eta_{(s_2,t)}) = \left| \Sigma\left(\frac{s_1}{\lambda}\right) \right|^{\frac{1}{4}} \left| \Sigma\left(\frac{s_2}{\lambda}\right) \right|^{\frac{1}{4}} \left| \frac{\Sigma\left(\frac{s_1}{\lambda}\right) + \Sigma\left(\frac{s_2}{\lambda}\right)}{2} \right|^{-\frac{1}{2}} \exp\left[-\sqrt{Q_\lambda(s_1, s_2)}\right], \quad (8.3.7)$$

where  $Q_\lambda(s_1, s_2) = 2(s_1 - s_2)^T [\Sigma(\frac{s_1}{\lambda}) + \Sigma(\frac{s_2}{\lambda})]^{-1} (s_1 - s_2)$  and  $\Sigma(\frac{s}{\lambda}) = \Gamma(\frac{s}{\lambda}) \Lambda \Gamma(\frac{s}{\lambda})^T$ .

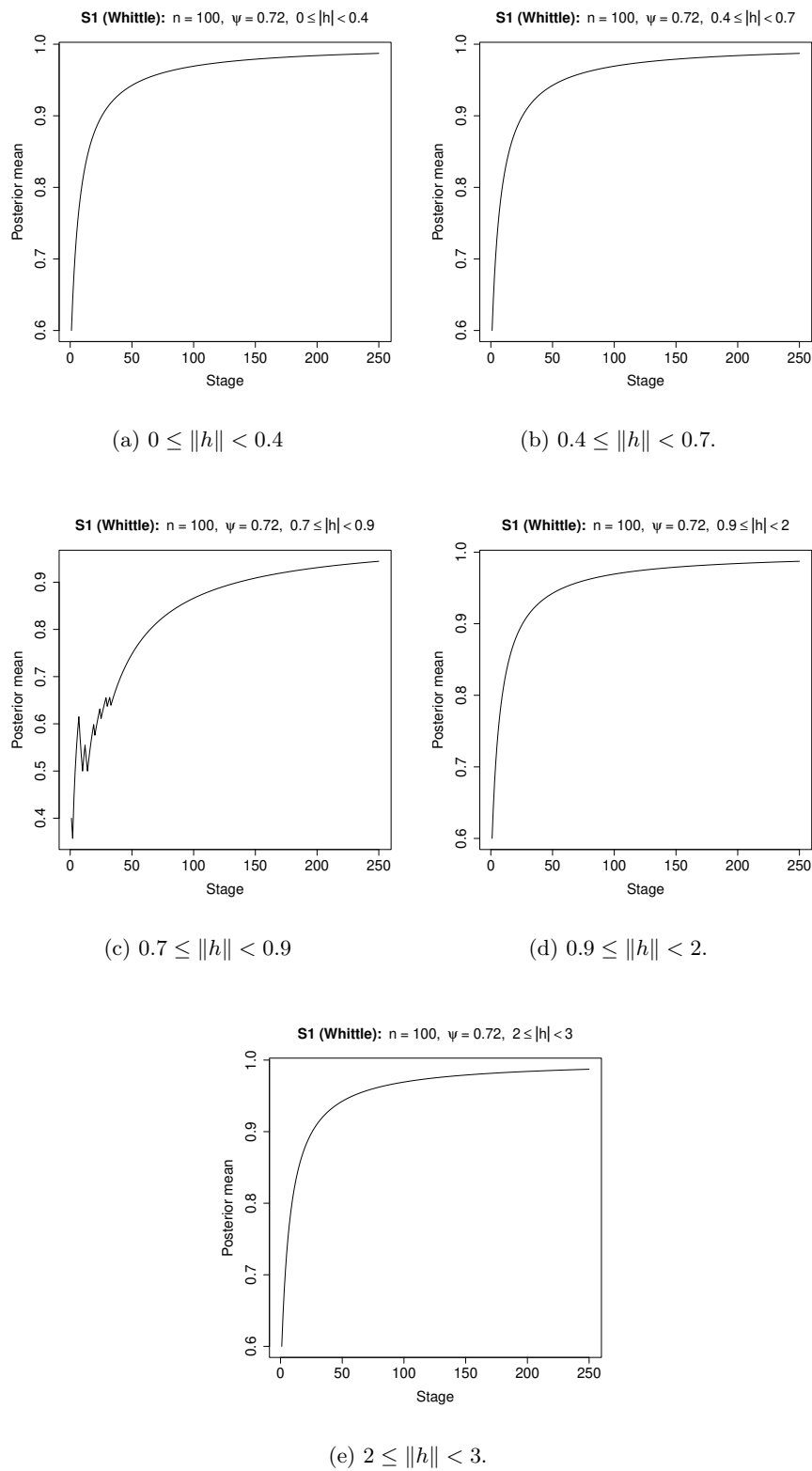
In the above,

$$\Gamma\left(\frac{s}{\lambda}\right) = \begin{pmatrix} \gamma_1\left(\frac{s}{\lambda}\right) & -\gamma_2\left(\frac{s}{\lambda}\right) \\ \gamma_2\left(\frac{s}{\lambda}\right) & \gamma_1\left(\frac{s}{\lambda}\right) \end{pmatrix}; \quad \Lambda = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix},$$

where  $\gamma_1\left(\frac{s}{\lambda}\right) = \log(u/\lambda + 0.75)$ ,  $\gamma_2\left(\frac{s}{\lambda}\right) = (u/\lambda)^2 + (v/\lambda)^2$ , and  $s = (u, v)^T$ .

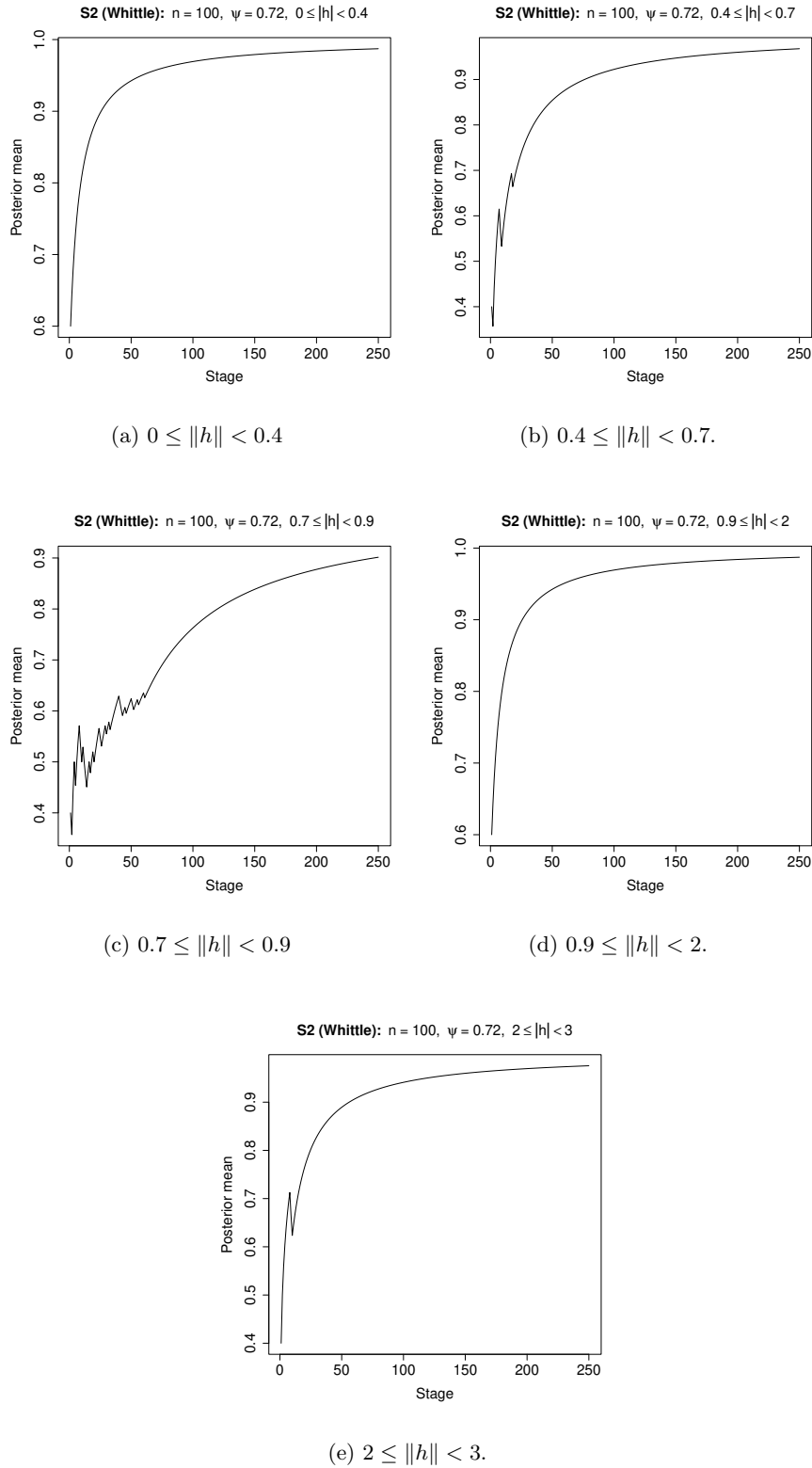
With this, the model is a temporally stationary and spatially nonstationary Gaus-

### 8.3. DETECTION OF STATIONARITY AND NONSTATIONARITY IN SPATIO-TEMPORAL DATA



**Figure 8.3.9:** Detection of covariance stationarity in spatio-temporal data drawn from model  $S1$  with sample size 100 locations and 200 time points, corresponding to Whittle spatial covariance with  $\psi = 0.72$  and  $\lambda = 5$ .

### 8.3. DETECTION OF STATIONARITY AND NONSTATIONARITY IN SPATIO-TEMPORAL DATA



**Figure 8.3.10:** Detection of covariance stationarity in spatio-temporal data drawn from model  $S2$  with sample size 100 locations and 200 time points, corresponding to Whittle spatial covariance with  $\psi = 0.72$  and  $\lambda = 5$ .

sian random field. For simulations, we consider  $\lambda = 20$ , following [Bandopadhyay et al. \(2017\)](#).

(NS3)  $X_{(s,t)} = 0.5X_{(s,t-1)} + (1.3 + \sin(\frac{2\pi t}{400}))\eta_{(s,t)}$ , where  $X_{s,0} = \mathbf{0}$  and  $\eta_{(s,t)}$  are zero mean GPs independent over time with nonstationary spatial covariance given by (8.3.7). This defines a temporally and spatially nonstationary Gaussian random field. Again, we set  $\lambda = 20$  for simulations, following [Bandopadhyay et al. \(2017\)](#).

We obtained the right results in all the cases of nonstationarity, but present the results corresponding to  $(m = 100, T = 200)$  and  $\psi = 1$  for brevity. Figure 8.3.11 provides the results on strong stationarity and the result on covariance stationarity of *NS1* is depicted in Figure 8.3.12. For detection of strict nonstationarity,  $\hat{C}_1$  varied between 0.04 and 0.05. The same values also yielded respective covariance nonstationarities in these examples. However, the maximum values of  $\hat{C}_1$  for detecting covariance nonstationarities varied between 0.05, 0.15, 0.2 and 0.3.

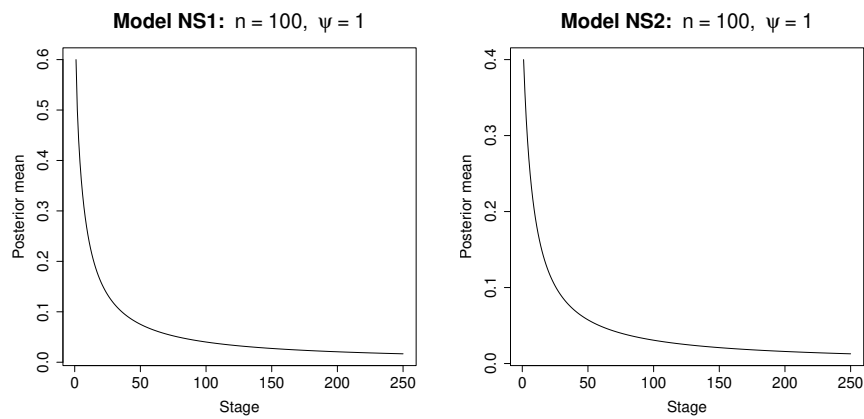
#### Overall comparison of our results with those of [Bandopadhyay et al. \(2017\)](#)

First, our Bayesian procedure is designed to identify both weak and strict stationarity of the underlying spatio-temporal process, while the methods of [Bandopadhyay et al. \(2017\)](#) are meant for detection of weak stationarity only, and not for strict stationarity.

Second, our method requires the only assumption of local stationarity, which is expected to hold in general. In contrast, the methods of [Bandopadhyay et al. \(2017\)](#) require a variety of assumptions, which may be difficult to verify in practice.

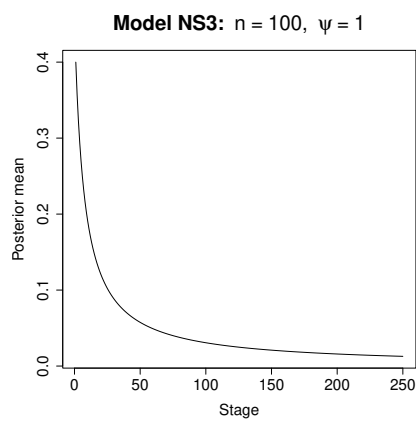
Overall, our Bayesian procedure worked adequately for all the strict stationarity and nonstationarity cases that we considered. The method also performed satisfactorily whenever there existed well-defined regions  $\mathcal{N}_{i,h_j,h_{j+1}}$  in the data set. On the other hand, the methods of [Bandopadhyay et al. \(2017\)](#) did not yield satisfactory results particularly when the underlying process is non-Gaussian.

## 8.3. DETECTION OF STATIONARITY AND NONSTATIONARITY IN SPATIO-TEMPORAL DATA



(a) Correct detection of nonstationarity.

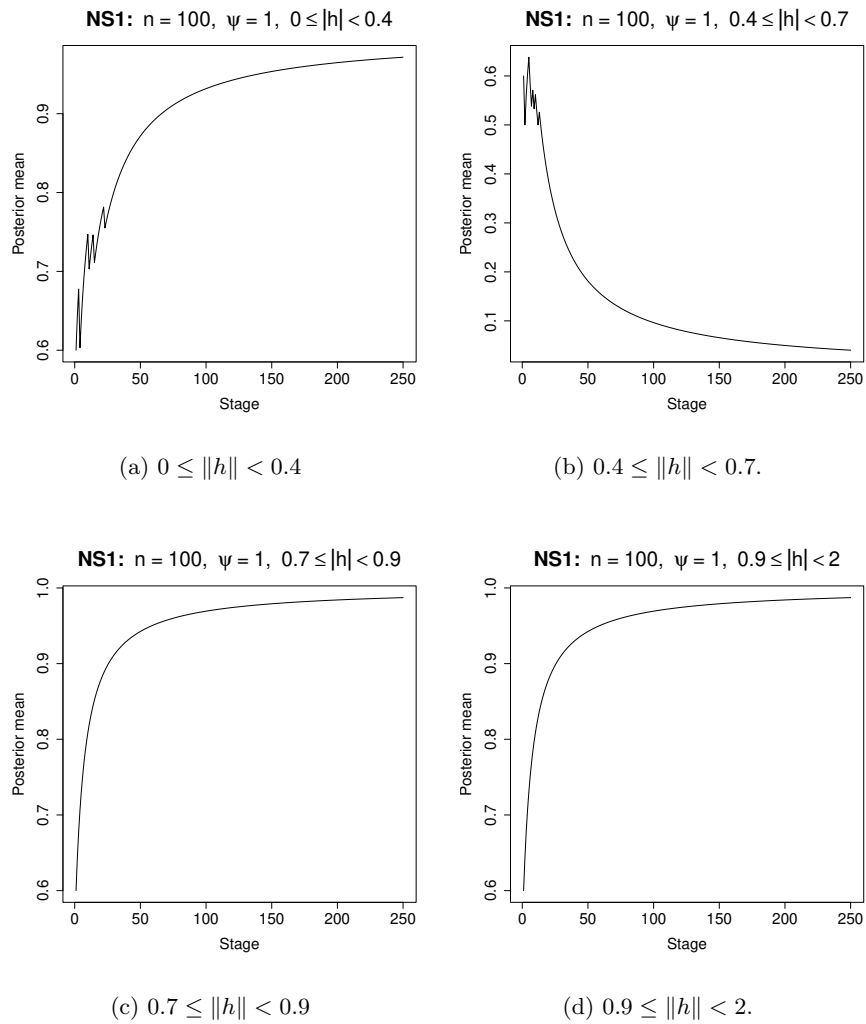
(b) Correct detection of nonstationarity.



(c) Correct detection of nonstationarity.

**Figure 8.3.11:** Detection of strong nonstationarity in spatio-temporal data drawn from models  $NS1$ ,  $NS2$  and  $NS3$  with sample size 100 locations and 200 time points.

8.3. DETECTION OF STATIONARITY AND NONSTATIONARITY IN SPATIO-TEMPORAL DATA



**Figure 8.3.12:** Detection of covariance nonstationarity in spatio-temporal data drawn from model *NS1* with sample size 100 locations and 200 time points.

## 8.4 Real data analyses for spatial and spatio-temporal data

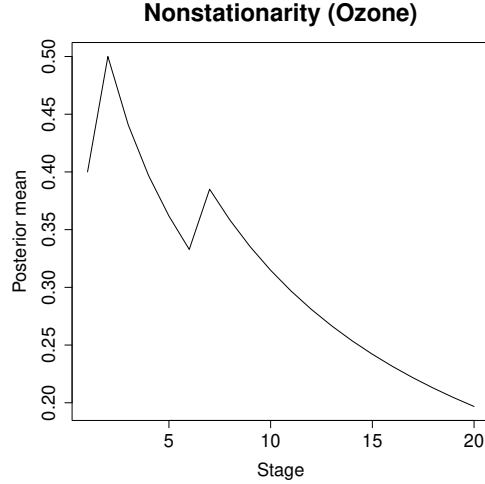
Das and Bhattacharya (2020) considered three real spatial and spatio-temporal data sets on pollutants for illustration of their new general nonparametric spatial and spatio-temporal model and methods. One is an ozone data set, which is a spatial data. Initially, Das and Bhattacharya (2020) fitted a stationary model, a special case of their general model, to the ozone data, but obtained unsatisfactory fit. This prompted them to fit the nonstationary instance of their model, which yielded adequate results. Thus, nonstationarity of the ozone data seems to be more plausible than stationarity. Here we establish with our Bayesian method that this is indeed the case.

The other two data sets are spatio-temporal data sets on particulate matters (PM), which are mixtures of solid particles and liquid droplets found in the air. The data sets correspond to measurements of air concentrations of two different size ranges – PM 10 and PM 2.5. The first one, PM 10, is suspected to be nonstationary, while PM 2.5 is suspected to be stationary in the literature (see, for example, Paciorek *et al.* (2009)). With our Bayesian method for characterizing stationarity and nonstationarity, we establish that such intuitions are correct.

For details regarding the three data sets, see Das and Bhattacharya (2020). There are also covariates associated with the three data sets, which have been utilized by Das and Bhattacharya (2020) for their modeling purpose. However, for checking stationarity and nonstationarity, only the responses are necessary. Hence, for our current purpose, the covariates are unnecessary. We evaluate all the final responses in their log scales.

### 8.4.1 Spatial ozone data

After appropriate data transformations (see Das and Bhattacharya (2020)), we obtain 76 observations, evaluated in the log scale. To obtain  $\hat{C}_1$ , we first generate 76 observations from a GP with the Whittle covariance function given by (8.2.3), with  $\psi = 0.8$ , and with the same set of locations as the ozone data. We set  $K = 20$  for this small data set, and



(a) Nonstationarity (ozone data).

**Figure 8.4.1:** Detection of nonstationarity of the ozone data with our Bayesian method.

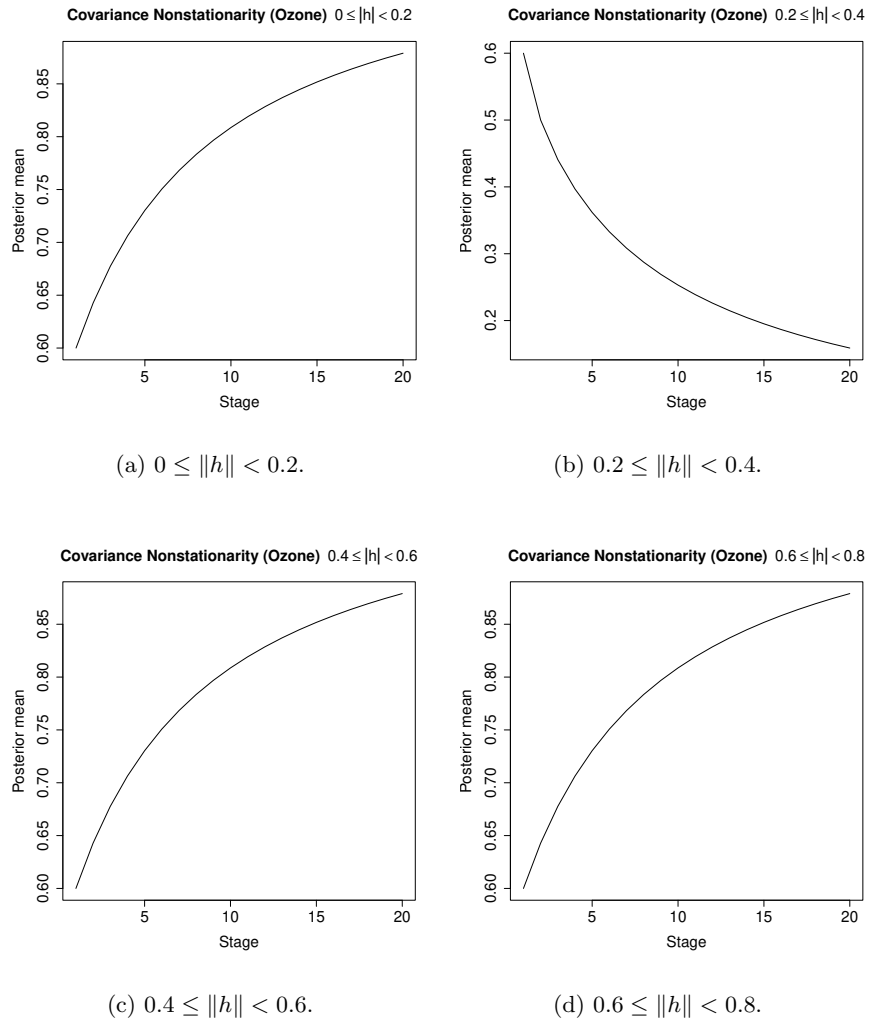
obtain the minimum value of  $\hat{C}_1$  that ensured stationarity for this GP data with our Bayesian method, to be 0.38. With this value of  $\hat{C}_1$  and larger (even with  $\hat{C}_1 = 0.43$ ), we obtained clear evidence of nonstationarity for the ozone data, as depicted in Figure 8.4.1.

To check covariance stationarity, we obtain four neighborhoods  $\mathcal{N}_{i,h_j,h_{j+1}}$ , for  $j = 1, 2, 3, 4$ , where  $h_1 = 0.0$ ,  $h_2 = 0.2$ ,  $h_3 = 0.4$ ,  $h_4 = 0.6$  and  $h_5 = 0.8$ . With  $K = 20$  and the same Whittle covariance based GP data for strict stationarity, the same value  $\hat{C}_1 = 0.38$  turned out to be the minimum value ensuring covariance stationarity for the GP data. Figure 8.4.2 shows covariance nonstationarity for the ozone data with  $\hat{C}_1 = 0.38$ . Indeed, convergence to zero is indicated with  $\mathcal{N}_{i,h_2,h_3}$ .

### 8.4.2 Spatio-temporal PM 10 data

This data set consists of 70572 observations, a part of which has been used by [Das and Bhattacharya \(2020\)](#) for model fitting. However, here we use all 70572 log-response





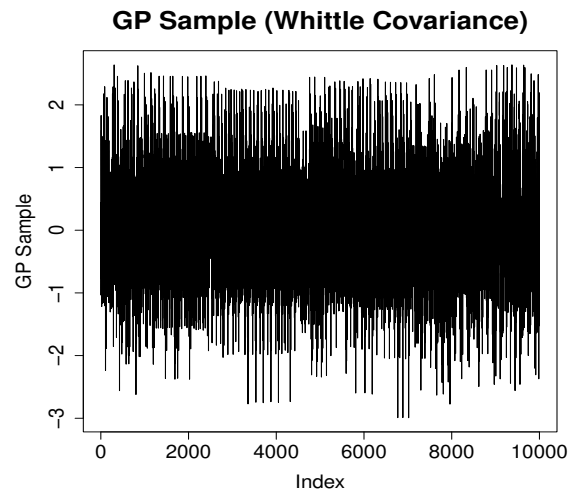
**Figure 8.4.2:** Detection of covariance nonstationarity of the ozone data.

values to check strict and covariance stationarity. To obtain  $\hat{C}_1$ , we need to generate GP samples of size 70572 with the Whittle covariance function and the locations, time points corresponding to the real PM 10 data set. However, generation of such a large GP sample turned out to be prohibitive with our current infrastructure. But more of concern is the issue that the stability of the covariance matrix turned out to steadily deteriorate for dimensions larger than 100000. Figure 8.4.3 shows two GP samples of sizes 10000 and 20000 generated using the *R*-package “mvnfast”, using 80 parallel cores. Although the sample of size 10000 is stable, the other shows increasing variability from index 10000 onwards. Hence, to obtain  $\hat{C}_1$  we consider the GP sample of size 10000. Setting  $K = 250$  as in the simulation studies, we obtain  $\hat{C}_1 = 0.16$  for checking strict stationarity. For the real PM 10 data of size 70572, we then set  $\hat{C}_1 = 0.16$  and  $K = 1764$ . The latter is chosen such that the number of observations per cluster is on the average 40, to match the average number of observations per cluster in the simulated GP data. Figure 8.4.4 clearly indicates strict nonstationarity of the PM 10 data.

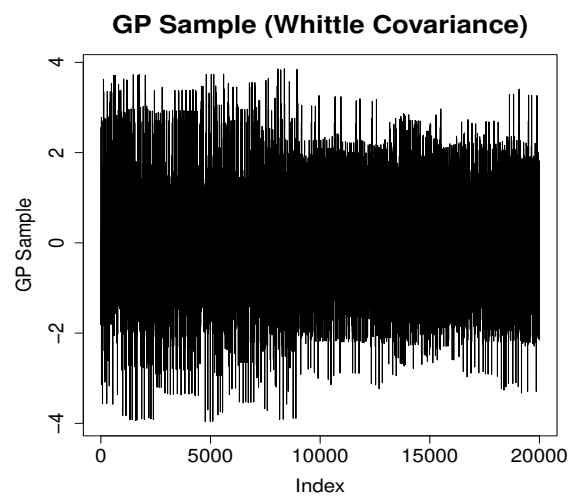
For checking covariance stationarity, our method with Whittle covariance failed to yield a valid  $\hat{C}_1$  since we could obtain only a single neighborhood  $\mathcal{N}_{i,h_1,h_2}$ , with  $h_1 = 0.0$  and  $h_2 = 0.15$ . Hence, we set  $\hat{C}_1 = 0.16$ , the same value obtained for checking strict stationarity. Again, for obtaining valid intervals, we needed to decrease the number of clusters and increase the number of observations per cluster. In this regard, setting  $K = 500$  let us obtain four valid neighborhoods  $\mathcal{N}_{i,h_j,h_{j+1}}$ ;  $j = 1, 2, 3, 4$ , with  $h_1 = 0.0$ ,  $h_2 = 0.1$ ,  $h_3 = 0.2$ ,  $h_4 = 0.3$  and  $h_5 = 0.4$ . Figure 8.4.5 shows covariance nonstationarity for the PM 10 data, as convergence to zero is indicated with  $\mathcal{N}_{i,h_1,h_2}$  and  $\mathcal{N}_{i,h_2,h_3}$ .

### 8.4.3 Spatio-temporal PM 2.5 data

The PM 2.5 data set consists of 17496 observations. For checking strict stationarity, we generated a GP sample of size 17496 with the Whittle covariance function with  $\psi = 0.8$ , with the same locations and time points as the real PM 2.5 data. Unlike the PM 10 case,

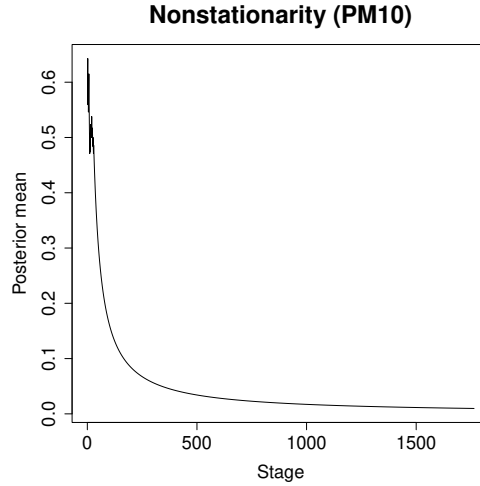


(a) GP sample size 10000.



(b) GP sample size 20000.

**Figure 8.4.3:** GP samples of sizes 10000 and 20000 for Whittle covariance with  $\psi = 0.8$  for PM 10 data.



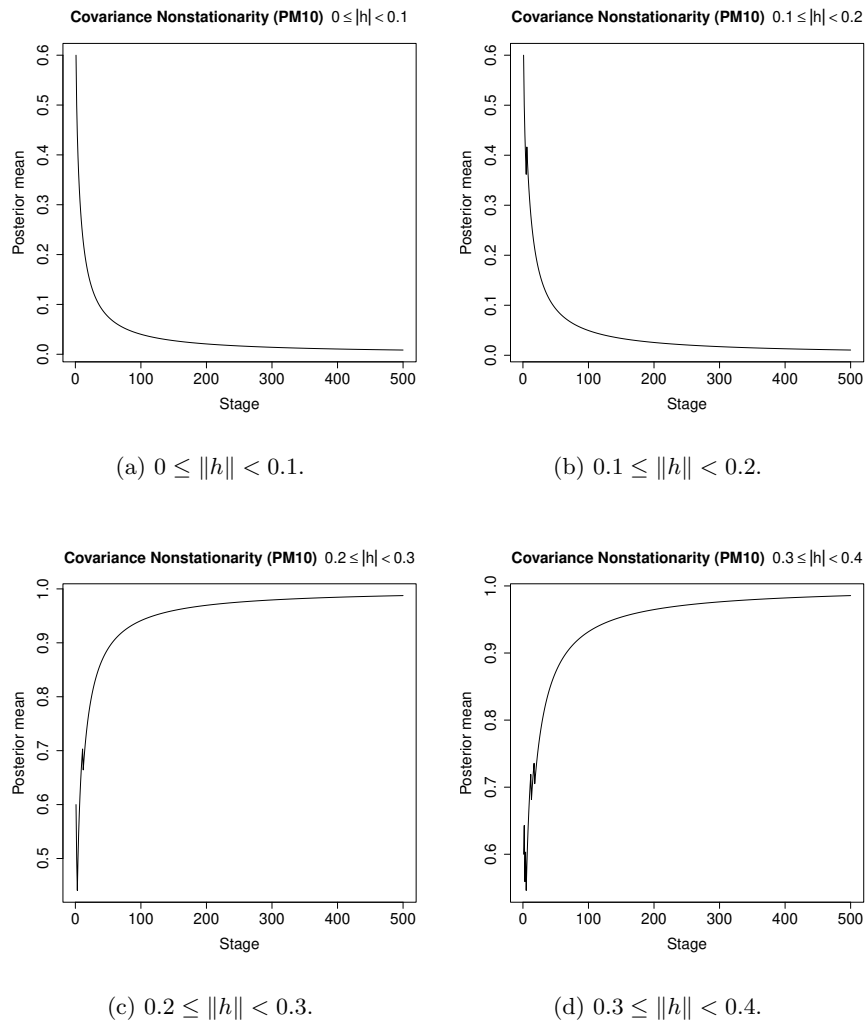
(a) Nonstationarity (PM 10 data).

**Figure 8.4.4:** Detection of nonstationarity of the PM 10 data with our Bayesian method.

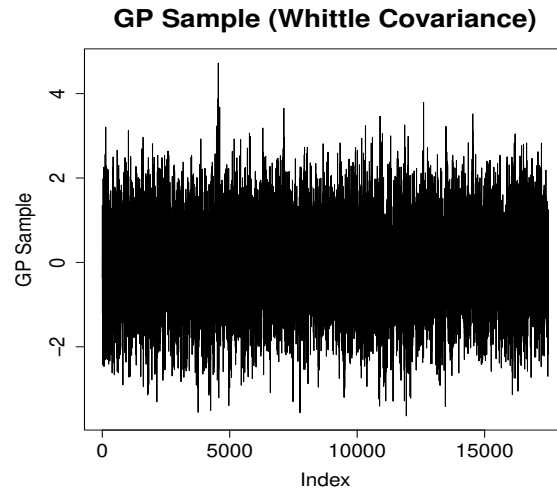
here the GP sample turned out to be stable, as shown in Figure 8.4.6. Setting  $K = 437$ , so that there are 40 observations on the average in each cluster, we obtained  $\hat{C}_1 = 0.02$  with the Whittle based GP sample. Figure 8.4.7 shows that the PM 2.5 data is strongly stationary. Hence, it is not necessary to check covariance stationarity of this data.

## 8.5 Summary and conclusion

In this chapter, we have illustrated our Bayesian characterizations of strong and weak notions of stationarity and nonstationarity of spatial and spatio-temporal processes with considerable number of simulation experiments and three real data examples. Correct results are obtained by our approach in almost all the simulation study cases, and even the real data analyses reveal results that are in accordance with the research endeavors of [Das and Bhattacharya \(2020\)](#) and the opinions of some other investigators based on their preliminary intuitions (see [Das and Bhattacharya \(2020\)](#) and the references

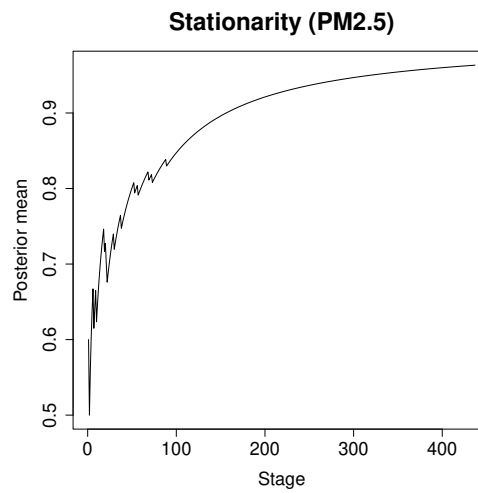


**Figure 8.4.5:** Detection of covariance nonstationarity of the PM 10 data.



(a) GP sample size 17496.

**Figure 8.4.6:** GP sample of size 17496 for Whittle covariance with  $\psi = 0.8$  for PM 2.5 data.



(a) Stationarity (PM 2.5 data).

**Figure 8.4.7:** Detection of stationarity of the PM 2.5 data with our Bayesian method.

therein). In other words, our theoretical results are aptly supported by the results of our simulated and real applications in the spatial and spatio-temporal scenarios.

It is important to point towards the fact that although the nonparametric bound of the form (5.4.1) turned out to be appropriate in our examples in this chapter, the initial value  $\hat{C}_1 = 1$  is no longer adequate in general, unlike in the time series and TCMC setups investigated in Chapter 7. However, we have presented and recommended methods of obtaining suitable values of  $\hat{C}_1$  and have demonstrated effectiveness of the methods with our simulation experiments.

Since [Bandopadhyay and Rao \(2017\)](#) and [Bandopadhyay \*et al.\* \(2017\)](#) present perhaps the most formal and effective tests of covariance stationarity and nonstationarity for spatial and spatio-temporal processes among those existing in the literature, we compare our results with theirs, with respect to their experimental setups and ours. It is interesting to observe that our ideas exhibit superior performances, in spite of far less assumptions compared to theirs. More importantly, our approach is valid for detection of both strong and weak stationarities, whereas the approaches of [Bandopadhyay and Rao \(2017\)](#) and [Bandopadhyay \*et al.\* \(2017\)](#) are meant for testing weak stationarity only.

**Table 8.2.1**  
The performance evaluation of the test statistics  $T$  and  $V$  of Bandopadhyay and Rao (2017) and Bandopadhyay et al. (2017) applied to our simulated spatial datasets.

	Model		Nonstationary	$p = 0.9$	$p = 0.99$	$p = 0.999$	$p = 0.9999$	$p = 0.99999$
	Test	Stationary						
1000	$T$	7.692	2.444	2.290	4.717	9.151	8.105	9.254
	$P$ -value ( $T$ )	0.158	0.751	0.776	0.387	0.103	0.141	0.099
	$V$	11.405	4.643	4.429	9.999	12.411	11.766	12.603
3000	$P$ -value ( $V$ )	0.056	0.398	0.424	0.080	0.043	0.051	0.041
	$T$	4.921	11.466	6.743	5.286	5.162	4.964	5.106
	$P$ -value ( $T$ )	0.361	0.055	0.206	0.322	0.335	0.356	0.341
5000	$V$	13.483	16.527	16.631	14.073	13.432	13.508	13.432
	$P$ -value ( $V$ )	0.031	0.014	0.014	0.023	0.031	0.031	0.031
	$T$	3.307	4.234	3.196	3.313	3.385	3.342	3.385
5000	$P$ -value ( $T$ )	0.595	0.451	0.615	0.595	0.583	0.589	0.583
	$V$	18.160	20.233	16.787	17.843	18.160	18.160	18.238
	$P$ -value ( $V$ )	0.010	0.006	0.014	0.011	0.010	0.010	0.010



# 9

## Bayesian Characterization of Point Processes

### 9.1 Introduction

Point pattern analysis is the study involving analysis of the spatial distribution of the observed events and to infer about the underlying data-generating process. The importance of such study is easy to perceive in diverse scientific fields such as particle physics, chemistry, archaeology, biology, ecology, environment, astronomy, to name a few, where investigations involving atoms, molecules, cells, animals, plants, trees, particles, pores, stars, galaxies are indispensable tasks. As can be anticipated, the field of spatial point pattern analysis is very much distinct from the area of traditional statistical analysis, and the methods used in classical statistics are often not appropriate for point process setups. For details regarding spatial point process theories, methods and various

applications, see Daley and Vere-Jones (2003), Møller and Waagepetersen (2004), Illian *et al.* (2008), Bivand *et al.* (2008).

Before embarking on a model for the underlying spatial point process, the first pertinent question to ask is whether or not interactions exist between the events (points). Hence, a relevant test that is often used in point pattern analysis is the test of complete spatial randomness (CSR), that is, if the points are independently and uniformly distributed over the study area (see, for example, O'Sullivan and Unwin (2003), Waller and Gotway (2004) and Schabenberger and Gotway (2005), for some simple existing tests).

Theoretically, homogeneous Poisson point process (HPP) corresponds to CSR, and thus tests for CSR can be devised on such basis, assuming the Poisson process framework for independent disjoint sets of events. However, rejecting CSR only rejects the HPP assumption and does not facilitate conclusion of stationarity or nonstationarity, Poisson or non-Poisson process. Bayesian characterization of stationarity and nonstationarity can be achieved with the same principles as before with relatively minor adjustments, while Bayesian characterization of CSR and Poisson assumption require additional innovative work. To characterize the Poisson assumption we exploit mutual independence of disjoint sets of events, under the assumption of orderliness and almost sure boundedly finite property of the process without fixed atoms. For the purpose, we provide a novel Bayesian characterization of mutual independence among random variables using Dirichlet processes. We believe that such a characterization is also of independent interest.

The rest of this chapter is structured as follows. We begin with a brief overview of the traditional CSR test in Section 9.2, the basic concept of which will be exploited in our Bayesian characterization. The complete details of our Bayesian characterization of CSR is provided in Section 9.3 In Section 9.4 we provide the Bayesian characterization of stationarity and nonstationarity of point processes. We proceed towards Bayesian characterization of the Poisson assumption of point processes by first providing, in Section

9.5, our novel Bayesian characterization of mutual independence among random variables using Dirichlet processes. Exploiting Bayesian characterization of mutual independence among random variables, we then provide Bayesian characterization of Poisson point process in Section 9.6. Finally, in Section 9.7, we bring out the practical utilities and efficacies of our theoretical Bayesian characterizations of general point processes with ample simulation experiments, all of which yielded quite encouraging results.

## 9.2 A brief overview of the existing CSR test

Testing for CSR can be found in O'Sullivan and Unwin (2003), Waller and Gotway (2004) and Schabenberger and Gotway (2005). The key ingredient in such tests is the so-called  $G$  function that provides the distribution of the distance from any arbitrary event to its nearest event. Specifically, let  $d_{ij}$  denote the distance between the  $i$ -th and  $j$ -th events in a set of  $n$  events, and for  $s = 1, \dots, n$ , let  $d_s = \min \{d_{st} : t \neq s\}$ . Consider the empirical distribution function

$$\hat{G}(x) = \frac{\sum_{s=1}^n I(d_s \leq x)}{n}. \quad (9.2.1)$$

Under CSR, that is, under the assumption of homogeneous Poisson point process,  $\hat{G}(x)$  has expectation

$$G(x) = 1 - \exp(-\lambda \pi x^2), \quad (9.2.2)$$

the  $G$ -function. Here  $\lambda$  is the intensity, or the number of events per unit area, the maximum likelihood estimator of which is given by  $\tilde{\lambda} = n/|W|$ , where  $W$  is the bounded region where the points are observed, and  $|W|$  denotes the volume of  $W$ . Indeed, the entire point process  $\mathbf{X}$  defined on some region  $\mathbf{S} \subset \mathbb{R}^d$ , for some  $d \geq 1$  can not be observed, and hence a bounded region  $W \subset \mathbf{S}$  is considered where points are observed.

Let

$$\tilde{G}(x) = 1 - \exp(-\tilde{\lambda}\pi x^2). \quad (9.2.3)$$

If the plot of  $\hat{G}(\cdot)$  versus  $\tilde{G}(\cdot)$  is approximately a straight line, then the CSR assumption may be accepted. Note that some quantification of uncertainty can be made by the point-wise envelopes under CSR which can be obtained by Monte Carlo simulation of a CSR point process with intensity  $\tilde{\lambda}$  in the observation window. If the CSR assumption holds good then  $\hat{G}(\cdot)$  is expected to be contained inside the simulated envelope.

### 9.3 Bayesian characterization of CSR

Let us assume that  $\mathbf{X}_K = \{X_s : s \in \cup_{i=1}^K \mathcal{N}_i\}$  has been observed, for  $K > 1$ . Here  $\cup_{i=1}^K \mathcal{N}_i$  corresponds to the observation window  $W$ . For the purpose of asymptotics, we assume that  $|\mathbf{S}|$ , the volume of  $\mathbf{S}$ , tends to infinity, so that even though  $|W|$  remains finite,  $n$ , the number of points in  $W$  tends to infinity, almost surely.

For any  $x > 0$ , consider

$$\hat{G}_i(x) = n_i^{-1} \sum_{s \in \mathcal{N}_i} I(d_s \leq x), \quad (9.3.1)$$

where  $n_i = |\mathcal{N}_i|$ , as before. Note that  $n = \sum_{i=1}^K n_i$ .

Now let

$$\begin{aligned} \hat{G}_K(x) &= \frac{\sum_{s \in \cup_{i=1}^K \mathcal{N}_i} I(d_s \leq x)}{\sum_{i=1}^K n_i} \\ &= \frac{\sum_{i=1}^K n_i \hat{G}_i(x)}{\sum_{i=1}^K n_i} = \sum_{i=1}^K \hat{p}_{iK} \hat{G}_i(x), \end{aligned} \quad (9.3.2)$$

where  $\hat{p}_{ik} = n_i / \sum_{j=1}^K n_j$ , as before. Let us now assume (6.2.4), which we recall as

$$\hat{p}_{iK} = \frac{n_i}{\sum_{j=1}^K n_j} \rightarrow p_{iK} = \frac{p_i}{\sum_{j=1}^K p_j},$$

as  $n_j \rightarrow \infty$ , for  $j = 1, \dots, K$ . Here  $0 \leq p_i \leq 1$ , such that  $\sum_{i=1}^{\infty} p_i = 1$ .

Let  $W_d$  denote the space where the distances  $d_i$ ,  $i = 1, \dots, n$ , associated with the observation window  $W$ , lie upon. However, for the asymptotic theory, we must let the window  $W$  and corresponding  $W_d$  to grow, otherwise the number of points  $n$  can not tend to infinity. Indeed, for fixed  $W$ , even the MLE  $\tilde{\lambda} = n/|W|$  is not a consistent estimator for  $\lambda$  in the HPP case. Thus, in this regard, we consider the sequences  $W_r$ ,  $W_{dr}$ ,  $K = K_r$ ,  $n_{ir}$ ,  $n_r$ ,  $K_r$ ,  $\hat{p}_{iK_r}$  and  $\tilde{\lambda}_r$ , for  $r = 1, 2, \dots$ , where the suffix  $r$  is incorporated to our previous notation to signify sequences. Let  $|W_r| \rightarrow \infty$  as  $r \rightarrow \infty$ . Note that  $K_r$  may remain finite even as  $r \rightarrow \infty$ . Let us also denote by  $G_{true}$  the true point process generating the data. Note that for HPP,  $G_{true} = G$ . In reality, the true point process, and hence  $G_{true}$ , is unknown.

A problem associated with HPP is that it is hard to establish  $\sup_{x \in W_{dr}} |\hat{G}_K - \tilde{G}(x)| \rightarrow 0$ , in either weak or strong sense. To see this, note that

$$\begin{aligned} \sup_{x \in W_{dr}} |\tilde{G}(x) - G(x)| &= \sup_{x \in W_{dr}} \left| \exp(-\lambda \pi x^2) \left( 1 - \exp\left(-\pi x^2 (\tilde{\lambda}_r - \lambda)\right) \right) \right| \\ &\leq 1 - \inf_{x \in W_{dr}} \exp\left(-\pi x^2 |\tilde{\lambda}_r - \lambda|\right). \end{aligned}$$

Since  $\exp\left(-\pi x^2 |\tilde{\lambda}_r - \lambda|\right)$  is decreasing in  $x^2$  and  $W_{dr}$  is bounded, the infimum over  $W_{dr}$  is given by  $\exp\left(-\pi \xi_r^2 |\tilde{\lambda}_r - \lambda|\right)$ , where  $\xi_r$  is the maximum interpoint distance in  $W_{dr}$ . In other words,

$$\sup_{x \in W_d} |\tilde{G}(x) - G(x)| \leq 1 - \exp\left(-\pi \xi_r^2 |\tilde{\lambda}_r - \lambda|\right). \quad (9.3.3)$$

By Markov's inequality, for any  $\epsilon > 0$ ,

$$P\left(\xi_r^2 \left|\tilde{\lambda}_r - \lambda\right| > \epsilon\right) < \epsilon^{-2} \xi_r^4 E\left(\frac{n_r}{|W_r|} - \lambda\right)^2 = \epsilon^{-2} \lambda \frac{\xi_r^4}{|W_r|},$$

which tends to zero if  $\frac{\xi_r^4}{|W_r|} \rightarrow 0$  as  $r \rightarrow \infty$ . But as can be easily verified, this does not hold for regular window shapes such as squares, rectangles, circles, triangles, etc. Indeed, for these shapes,  $\frac{\xi_r^4}{|W_r|} \rightarrow \infty$  as  $r \rightarrow \infty$ .

Instead of  $\sup_{x \in W_{dr}} \left|\hat{G}_K(x) - \tilde{G}(x)\right|$  we shall thus deal with  $\int_{W_{dr}} \left|\hat{G}_K(x) - \tilde{G}(x)\right| dG_{true}(x)$  in the following theorem.

**Theorem 33** *Assume that  $\mathbf{X}$  follows homogeneous Poisson point process, and that the points are observed in the window  $W_r$ , where  $|W_r| \rightarrow \infty$  as  $r \rightarrow \infty$ . Let  $W_{dr}$  denote the space of the distances associated with  $W_r$ . Then, for all values of  $K_\infty = \lim_{r \rightarrow \infty} K_r$ ,*

$$\lim_{r \rightarrow \infty, n_{ir} \rightarrow \infty, i=1, \dots, K_r} \int_{W_{dr}} \left|\hat{G}_{K_r}(x) - \tilde{G}(x)\right| dG_{true}(x) = 0, \quad (9.3.4)$$

almost surely if  $\sum_{r=1}^{\infty} |W_r|^{-1} < \infty$ .

**Proof.** Observe that

$$\begin{aligned} & \int_{W_{dr}} \left|\hat{G}_{K_r}(x) - \tilde{G}(x)\right| dG_{true}(x) \\ & \leq \sup_{x \in W_{dr}} \left|\hat{G}_{K_r}(x) - G(x)\right| G_{true}(W_{dr}) + \int_{W_{dr}} \left|\tilde{G}(x) - G(x)\right| dG_{true}(x) \\ & \leq \sup_{x \in W_{dr}} \left|\hat{G}_{K_r}(x) - G(x)\right| + \int_{W_{dr}} \left|\tilde{G}(x) - G(x)\right| dG_{true}(x). \end{aligned} \quad (9.3.5)$$

Since

$$\sup_{x \in W_{dr}} \left|\hat{G}_{K_r}(x) - G(x)\right| = \sup_{x \in W_{dr}} \left|\sum_{i=1}^{K_r} \left(\hat{G}_i(x) - G(x)\right)\right| \leq \sum_{i=1}^{K_r} \hat{p}_{iK_r} \sup_{x \in W_{dr}} \left|\hat{G}_i(x) - G(x)\right|. \quad (9.3.6)$$

Now, as  $r \rightarrow \infty$ , the right hand side of (9.3.6) converges almost surely to

$$\sum_{i=1}^{K_\infty} p_{iK_\infty} \lim_{r \rightarrow \infty} \sup_{x \in W_{dr}} \left| \hat{G}_i(x) - G(x) \right|, \quad (9.3.7)$$

since  $\hat{p}_{iK_r} \rightarrow p_{iK_\infty}$  in the same way as (6.2.4). Also,  $\sup_{x \in W_{dr}} \left| \hat{G}_i(x) - G(x) \right| \xrightarrow{a.s.} 0$ , as  $r \rightarrow \infty$  and  $n_{ir} \rightarrow \infty$  by Glivenko-Cantelli theorem for stationary random variables (Stute and Schumann (1980)). That is, given any  $K_\infty$ , (9.3.7) converges to zero almost surely. Thus, (9.3.7) converges to zero almost surely, even as  $K_\infty \rightarrow \infty$ . Hence, it follows from these arguments and (9.3.6) that for all values of  $K_\infty$ ,

$$\sup_{x \in W_{dr}} \left| \hat{G}_K(x) - G(x) \right| \xrightarrow{a.s.} 0, \text{ as } n_{ir} \rightarrow \infty, i = 1, \dots, K_r, r \rightarrow \infty,$$

and hence

$$\int_{W_{dr}} \left| \hat{G}_K(x) - G(x) \right| dG_{true}(x) \xrightarrow{a.s.} 0, \text{ as } n_{ir} \rightarrow \infty, i = 1, \dots, K_r, r \rightarrow \infty. \quad (9.3.8)$$

Now note that, for  $\tilde{\lambda}_r = n_r/|W_r|$ ,

$$\begin{aligned} \int_{W_{dr}} \left| \tilde{G}(x) - G(x) \right| dG_{true}(x) &= \int_{W_{dr}} \left| \exp(-\lambda\pi x^2) \left( 1 - \exp\left(-\pi x^2 (\tilde{\lambda} - \lambda)\right) \right) \right| dG_{true}(x) \\ &\leq G_{true}(W_{dr}) - \int_{W_{dr}} \exp\left(-\pi x^2 |\tilde{\lambda} - \lambda|\right) dG_{true}(x). \end{aligned} \quad (9.3.9)$$

In (9.3.9),

$$G_{true}(W_{dr}) \rightarrow 1, \text{ as } r \rightarrow \infty. \quad (9.3.10)$$

Now, by Markov's inequality, for any  $\epsilon > 0$ ,

$$\begin{aligned} \sum_{r=1}^{\infty} P\left(\left|\tilde{\lambda}_r - \lambda\right| > \epsilon\right) &= \sum_{r=1}^{\infty} P\left(\left|\frac{n_r}{|W_r|} - \lambda\right| > \epsilon\right) \\ &< \epsilon^{-2} \sum_{r=1}^{\infty} E\left(\frac{n_r}{|W_r|} - \lambda\right)^2 = \epsilon^{-2} \lambda \sum_{r=1}^{\infty} \frac{1}{|W_r|} < \infty, \end{aligned}$$

where the last step is due to our assumption. Hence, by Borel-Cantelli lemma,  $\left|\tilde{\lambda}_r - \lambda\right| \xrightarrow{a.s.} 0$ , as  $r \rightarrow \infty$ . By dominated convergence theorem, it follows that

$$\int_{W_{dr}} \exp\left(-\pi x^2 \left|\tilde{\lambda} - \lambda\right|\right) dG_{true}(x) \xrightarrow{a.s.} 1, \text{ as } r \rightarrow \infty. \quad (9.3.11)$$

It follows from (9.3.9), (9.3.10) and (9.3.11) that

$$\int_{W_{dr}} \left|\tilde{G}(x) - G(x)\right| dG_{true}(x) \xrightarrow{a.s.} 0, \text{ as } r \rightarrow \infty. \quad (9.3.12)$$

The result follows by combining (9.3.5), (9.3.8) and (9.3.12). ■

**Remark 34** Note that unlike in the previous cases where we required  $K \rightarrow \infty$ , here we did not require the assumption  $K_{\infty} \rightarrow \infty$ . Theorem 33 explicitly mentions that the result holds for all values of  $K_{\infty}$ . This difference is due to the fact that in the asymptotics of point process we assumed that the observation window  $W_r$  is growing with  $r$ , and with such growing observation window, the entire point process can be ultimately captured. Hence increasing the number of clusters is not required. From a more mathematical perspective, note that  $\hat{G}_K$  uses all the observations in the observation window, and so the value of  $K$  is irrelevant mathematically.

**Remark 35** Note that by direct application of Glivenko-Cantelli theorem for stationary random variables we can obtain,

$$\sup_{x \in W_{dr}} \left|\hat{G}_{K_r}(x) - G(x)\right| \xrightarrow{a.s.} 0, \text{ as } r \rightarrow \infty. \quad (9.3.13)$$



This does not require breaking up the observation window  $W_r$  into sub-regions  $\mathcal{N}_1, \dots, \mathcal{N}_{K_r}$ , and the assumption  $n_{i_r} \rightarrow \infty$  for  $i = 1, \dots, K_r$ . However, it is important to detect which sub-regions of  $W_r$  are not representatives of CSR. From this perspective, it is important to consider the sub-regions  $\mathcal{N}_1, \dots, \mathcal{N}_{K_r}$ , and consideration of the form (9.3.2), which we formalize in our Bayesian characterization.

Let  $\{c_j\}_{j=1}^\infty$  be a non-negative decreasing sequence and

$$Y_{j,n_{j_r}} = \mathbb{I} \left\{ \int_{W_{d_r}} \left| \hat{G}_j(x) - \tilde{G}(x) \right| dG(x) \leq c_j \right\}. \quad (9.3.14)$$

In practice, we shall approximate  $\int_{W_{d_r}} \left| \hat{G}_j(x) - \tilde{G}(x) \right| dG(x)$  by  $\frac{1}{n_{j_r}} \sum_{i=1}^{n_{j_r}} \left| \hat{G}_j(d_i) - \tilde{G}(d_i) \right|$ , where the distances  $d_i$  are assumed to correspond to the true data-generating point process  $G_{true}$ .

As before, let, for  $j \geq 1$ ,

$$P(Y_{j,n_{j_r}} = 1) = p_{j,n_{j_r}}. \quad (9.3.15)$$

Hence, the likelihood of  $p_{j,n_{j_r}}$ , given  $y_{j,n_{j_r}}$ , is given by the form (6.4.3).

As before, we construct a recursive Bayesian methodology that formally characterizes homogeneous Poisson process and otherwise in terms of formal posterior convergence. The relevant theorems in this regard, the proofs of which are similar to stationarity and nonstationarity characterizations, are presented below as Theorems 36 and 37.

**Theorem 36** For all  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ , where  $\mathfrak{N}$  is some null set having probability measure zero,  $\mathbf{X} \cap W$  follows homogeneous Poisson process if and only if for any monotonically decreasing sequence  $\{c_j(\omega)\}_{j=1}^\infty$ ,

$$\pi(\mathcal{N}_1 | y_{k,n_{k_r}}(\omega)) \rightarrow 1, \quad (9.3.16)$$

as  $k \rightarrow \infty$  and  $n_{j_r} \rightarrow \infty$  for  $j = 1, \dots, K_r$  satisfying (6.2.4) and  $K_r \rightarrow \infty$  as  $r \rightarrow \infty$ , where  $\mathcal{N}_1$  is any neighborhood of 1 (one).

**Theorem 37**  $\mathbf{X} \cap W$  does not follow homogeneous Poisson process if and only if for any  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$  where  $\mathfrak{N}$  is some null set having probability measure zero, for any choice of the non-negative, monotonically decreasing sequence  $\{c_j(\omega)\}_{j=1}^\infty$ ,

$$\pi(\mathcal{N}_0 | y_{k, n_{k_r}(\omega)}(\omega)) \rightarrow 1, \quad (9.3.17)$$

as  $k \rightarrow \infty$  and  $n_{j_r} \rightarrow \infty$ ,  $j = 1, \dots, K_r$  satisfying (6.2.4), and  $K_r \rightarrow \infty$  as  $r \rightarrow \infty$ , where  $\mathcal{N}_0$  is any neighborhood of 0 (zero).

**Remark 38** Note that Theorems 36 and 37 require  $K_r \rightarrow \infty$  as  $r \rightarrow \infty$ , even though Theorem 33 does not have this requirement. But this arises entirely for convergence of the recursive Bayesian algorithm as the stage number  $k \rightarrow \infty$ .

### Discussion on edge correction

Since the data are observed in the bounded window  $W$ , the minimum distance  $d_i$  in the window may be larger than the true minimum distance had the complete point process  $\mathbf{X}$  been observed. In classical point process analysis, this may induce a bias in estimating the true distribution function, which is known as edge effect. Needless to mention, various corrections for such edge effect is available in the literature.

However, in the way we proceed with our Bayesian method, the edge effects do not influence our final results. The reason for this is the following. We partition the point pattern in the observation window  $W$  into  $K$  clusters using the K-means clustering algorithm. Thus, within each cluster in the interior of  $W$ , the edge effect is minimized. This is because the K-means clustering algorithm guarantees that within cluster variation is minimized and the between cluster variation is maximized, which entails that the minimum distance  $d_i$  of any point  $i$  within each cluster is often indeed the minimum when all the points are considered. Note that this is actually the case for ‘empty distances’, if the distances are measured from the centroid of each cluster. Our

experiments demonstrate the validity of our aforementioned arguments in this regard.

## 9.4 Bayesian characterization of stationarity and nonstationarity of point processes

The characterization of stationarity and nonstationarity in the point process setup remains essentially the same as in the general situation, with the conceptual difference being consideration of  $W_r$  in the point process setup, with  $|W_r| \rightarrow \infty$ . We present the main results regarding stationarity and nonstationarity in the point process setup, which are slight modifications of Theorems 21, 22, 23, 24 and 25.

**Theorem 39** *Let  $K_r \rightarrow \infty$  as  $r \rightarrow \infty$ . Then*

$$\lim_{r \rightarrow \infty} \lim_{n_{ir} \rightarrow \infty, i=1, \dots, K_r} \sup_C \left| \tilde{P}_{K_r}(C) - P_\infty(C) \right| = 0, \text{ almost surely.}$$

**Theorem 40** *The point process  $\mathbf{X}$  is stationary if and only if  $\sup_C \left| \hat{P}_j(C) - \tilde{P}_{K_r}(C) \right| \rightarrow 0$  almost surely, as  $n_{jr} \rightarrow \infty$  satisfying (6.2.4),  $j = 1, \dots, K_r$ ,  $K_r \rightarrow \infty$ , as  $r \rightarrow \infty$ .*

**Theorem 41**  *$\mathbf{X}$  is nonstationary if and only if  $\sup_C \left| \hat{P}_j(C) - \tilde{P}_{K_r}(C) \right| > 0$  almost surely, as  $n_{jr} \rightarrow \infty$  satisfying (6.2.4),  $j = 1, \dots, K_r$ ,  $K_r \rightarrow \infty$ , as  $r \rightarrow \infty$ .*

Let  $\{c_j\}_{j=1}^\infty$  be a non-negative decreasing sequence and

$$Y_{j,n_{jr}} = \mathbb{I} \left\{ \sup_C \left| \hat{P}_j(C) - \tilde{P}_{K_r}(C) \right| \leq c_j \right\}.$$

Let, for  $j \geq 1$ ,

$$P(Y_{j,n_{jr}} = 1) = p_{j,n_{jr}}.$$

**Theorem 42** For all  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ , where  $\mathfrak{N}$  is some null set having probability measure zero,  $\mathbf{X}$  is stationary if and only if for any monotonically decreasing sequence  $\{c_j(\omega)\}_{j=1}^\infty$ ,

$$\pi(\mathcal{N}_1 | y_{k, n_{k_r}}(\omega)) \rightarrow 1,$$

as  $k \rightarrow \infty$  and  $n_{j_r} \rightarrow \infty$  for  $j = 1, \dots, K_r$  satisfying (6.2.4) and  $K_r \rightarrow \infty$  as  $r \rightarrow \infty$ , where  $\mathcal{N}_1$  is any neighborhood of 1 (one).

**Theorem 43**  $\mathbf{X}$  is nonstationary if and only if for any  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$  where  $\mathfrak{N}$  is some null set having probability measure zero, for any choice of the non-negative, monotonically decreasing sequence  $\{c_j(\omega)\}_{j=1}^\infty$ ,

$$\pi(\mathcal{N}_0 | y_{k, n_{k_r}(\omega)}(\omega)) \rightarrow 1,$$

as  $k \rightarrow \infty$  and  $n_{j_r} \rightarrow \infty$ ,  $j = 1, \dots, K_r$  satisfying (6.2.4), and  $K_r \rightarrow \infty$  as  $r \rightarrow \infty$ , where  $\mathcal{N}_0$  is any neighborhood of 0 (zero).

## 9.5 Bayesian characterization of mutual independence among random variables

In this section we first characterize mutual independence among a general set of random variables  $\mathbf{X}_K = (X_1, \dots, X_K)$ , as  $K \rightarrow \infty$ , and then specialize the characterization in the point process setup. Indeed, although characterizations and tests for mutual independence among a set of random variables is available in the literature (see, for example, Puri and Sen (1971), Gieser and Randles (1997), Um and Randles (2001), Cl eroux *et al.* (1995), Bilodeau and L de Micheaux (2005), Hoeffding (1948), Blum *et al.* (1961), Ghoudi *et al.* (2001), Beran *et al.* (2007), Bilodeau and Nangu e (2017)), they are meant for a finite set of random variables. Moreover, such characterizations are often not computationally manageable. Here we attempt to provide a characterization for

number of random variables tending to infinity, with manageable computation. Also, unlike the previous approaches, we need only asymptotic stationarity of the realizations of the random variables, not even independence.

The key idea is to consider the differences

$$\zeta_i = \sup_{t_1, \dots, t_i \in \mathbb{R}} |P(X_i \leq t_i | X_1 \leq t_1, \dots, X_{i-1} \leq t_{i-1}) - P(X_i \leq t_i)|, \quad (9.5.1)$$

for  $i = 2, \dots, K$ , with  $\zeta_1 = 0$ . If all  $\zeta_i$ ;  $i = 2, \dots, K$ , are sufficiently small, then the random variables  $(X_1, \dots, X_K)$  are mutually independent. For practical purposes, we must replace

$$P(X_i \leq t_i | X_1 \leq t_1, \dots, X_{i-1} \leq t_{i-1})$$

and  $P(X_i \leq t_i)$  with their corresponding empirical probabilities. In other words, we write

$$P(X_i \leq t_i | X_1 \leq t_1, \dots, X_{i-1} \leq t_{i-1}) = \frac{P(X_1 \leq t_1, \dots, X_{i-1} \leq t_{i-1}, X_i \leq t_i)}{P(X_1 \leq t_1, \dots, X_{i-1} \leq t_{i-1})}, \quad (9.5.2)$$

and replace  $P(X_1 \leq t_1, \dots, X_{i-1} \leq t_{i-1}, X_i \leq t_i)$  and  $P(X_1 \leq t_1, \dots, X_{i-1} \leq t_{i-1})$  with their corresponding empirical distribution functions

$$F_{n,1:i}(X_1 \leq t_1, \dots, X_{i-1} \leq t_{i-1}, X_i \leq t_i)$$

and

$$F_{n,1:(i-1)}(X_1 \leq t_1, \dots, X_{i-1} \leq t_{i-1}),$$

respectively. We also replace  $P(X_i \leq t_i)$  with its empirical distribution function  $F_{n,i}(X_i \leq t_i)$ . We denote the differences of the empirical distribution functions corresponding to (9.5.1) by  $\hat{\zeta}_i$ ;  $i = 2, \dots, k$ , with  $\hat{\zeta}_1 = 0$ .

However, computation of the joint empirical distribution functions  $F_{n,1:i}$  often turn out

to be zero numerically, even if  $i$  is not too large. To address this, we resort to Bayesian nonparametrics, with Dirichlet process prior for the joint distribution of  $\mathbf{X}_K$ . In fact, more generally, we consider a stochastic process prior for the sequence of random variables  $\mathbf{X} = (X_1, X_2, X_3, \dots)$ . Let  $G_0$  denote the expected parametric stochastic process for  $\mathbf{X}$ . Specifically, we assume that  $\mathbf{X} \sim G$  and  $G \sim DP(\alpha G_0)$ , where  $DP(\alpha G_0)$  stands for Dirichlet process with base measure  $G_0$  and strength parameter  $\alpha > 0$ . More transparently, let  $\mathbf{X}_{i_1, i_2, \dots, i_K} = (X_{i_1}, X_{i_2}, \dots, X_{i_K})$ , for any set of indices  $i_1, \dots, i_K$ . Then  $\mathbf{X}_{i_1, i_2, \dots, i_K} \sim G_{i_1, i_2, \dots, i_K}$  and  $G_{i_1, i_2, \dots, i_K} \sim DP(\alpha G_{0, i_1, i_2, \dots, i_K})$ , where  $G_{i_1, i_2, \dots, i_K}$  and  $G_{0, i_1, i_2, \dots, i_K}$  are  $k$ -dimensional distributions associated with  $\mathbf{X}_{i_1, i_2, \dots, i_K}$ .

Now, if data  $\mathbf{X}_{i_1, i_2, \dots, i_K}^j; j = 1, 2, \dots$ , are available which are not necessarily *iid* or not even independent, we consider the following recursive strategy for sequentially updating the posterior distribution of the Dirichlet process. We assume that

$$\mathbf{X}_{i_1, i_2, \dots, i_K}^1 \sim G_1; G_1 \sim DP(\alpha G_{0, i_1, i_2, \dots, i_K}). \quad (9.5.3)$$

so that the posterior distribution of the random distribution given  $\mathbf{X}_{i_1, i_2, \dots, i_K}^1$  is given by

$$[G_1 | \mathbf{X}_{i_1, i_2, \dots, i_K}^1] \sim DP\left(\alpha G_{0, i_1, i_2, \dots, i_K} + \delta_{\mathbf{X}_{i_1, i_2, \dots, i_K}^1}\right). \quad (9.5.4)$$

Now, assuming  $[G_1 | \mathbf{X}_{i_1, i_2, \dots, i_K}^1]$  to be the prior for the distribution of  $\mathbf{X}_{i_1, i_2, \dots, i_K}^2$ , we have

$$[G_2 | \mathbf{X}_{i_1, i_2, \dots, i_K}^2] \sim DP\left(\alpha G_{0, i_1, i_2, \dots, i_K} + \delta_{\mathbf{X}_{i_1, i_2, \dots, i_K}^1} + \delta_{\mathbf{X}_{i_1, i_2, \dots, i_K}^2}\right). \quad (9.5.5)$$

Continuing as (9.5.3), (9.5.4) and (9.5.5), we obtain in general, for  $j \geq 1$ ,

$$[G_j | \mathbf{X}_{i_1, i_2, \dots, i_K}^j] \sim DP\left(\alpha G_{0, i_1, i_2, \dots, i_K} + \sum_{r=1}^j \delta_{\mathbf{X}_{i_1, i_2, \dots, i_K}^r}\right). \quad (9.5.6)$$

Note that the posterior in this case is of the same form as that of  $[G_j | \mathbf{X}_{i_1, i_2, \dots, i_K}^r; r =$

$1, \dots, j]$ , had  $\mathbf{X}_{i_1, i_2, \dots, i_K}^r; r = 1, \dots, j$  been *iid* with distribution  $G_{i_1, i_2, \dots, i_K}$  and  $G_{i_1, i_2, \dots, i_K} \sim DP(\alpha G_{0, i_1, i_2, \dots, i_K})$ .

In particular, for  $n$  data points  $\{\mathbf{X}_K^j; j = 1, 2, \dots, n\}$ , following (9.5.6) we obtain the posterior mean as

$$E[G_n | \mathbf{X}_K^n] = \frac{\alpha G_{0, 1:K} + \sum_{r=1}^n \delta_{\mathbf{X}_K^r}}{\alpha + n}, \quad (9.5.7)$$

which involves all the available data points  $\{\mathbf{X}_K^j; j = 1, 2, \dots, n\}$ . With (9.5.7), we deal with the following form of the conditional distribution function of  $[X_j | X_1, \dots, X_{j-1}]$  for  $j \geq 1$ :

$$\tilde{\zeta}_{jn}(t_1, \dots, t_j) = \frac{E[G_n(X_1 \leq t_1, \dots, X_j \leq t_j) | \mathbf{X}_j^n]}{E[G_n(X_1 \leq t_1, \dots, X_{j-1} \leq t_{j-1}) | \mathbf{X}_{j-1}^n]}. \quad (9.5.8)$$

The marginal distribution of  $X_j$  in this case that we shall consider is

$$\tilde{\zeta}_{jn}(t_j) = \frac{\alpha G_{0, j}(X_j \leq t_j) + \sum_{r=1}^n \delta_{X_j^r}(X_j^r \leq t_j)}{\alpha + n} \quad (9.5.9)$$

With these, we have the following result.

**Theorem 44** *For any  $K \geq 2$ , let  $\mathbf{X}_K^j; j \geq 1$ , be stationary. Then  $(X_1, \dots, X_K)$  are mutually independent if and only if, for  $j = 1, \dots, K$ ,*

$$\sup_{t_1, \dots, t_j \in \mathbb{R}} \left| \tilde{\zeta}_{jn}(t_1, \dots, t_j) - \tilde{\zeta}_{jn}(t_j) \right| \xrightarrow{a.s.} 0, \quad n \rightarrow \infty. \quad (9.5.10)$$

**Proof.** Let  $(X_1, \dots, X_K)$  be mutually independent. Then  $[X_j | X_1, \dots, X_{j-1}] = [X_j]$ , for  $j \geq 2$ . In other words, it holds that  $P(X_j \leq t_j | X_1 \leq t_1, \dots, X_{j-1} \leq t_{j-1}) = P(X_j \leq t_j)$ , for all  $t_1, \dots, t_j \in \mathbb{R}$ , and  $j \geq 2$ . Now,

$$\sup_{t_j \in \mathbb{R}} \left| \tilde{\zeta}_{jn}(t_1, \dots, t_j) - \tilde{\zeta}_{jn}(t_j) \right| \leq \sup_{t_j \in \mathbb{R}} \left| \tilde{\zeta}_{jn}(t_1, \dots, t_j) - P(X_j \leq t_j) \right| + \sup_{t_j \in \mathbb{R}} \left| P(X_j \leq t_j) - \tilde{\zeta}_{jn}(t_j) \right|. \quad (9.5.11)$$

Let us first focus on the first term of (9.5.11). For fixed  $\alpha$ , as  $n \rightarrow \infty$ , due to Glivenko-

Cantelli theorem for stationarity, it is easily seen that

$$E[G_n(X_1 \leq t_1, \dots, X_{j-1} \leq t_{j-1}) | \mathbf{X}_{j-1}^n] \xrightarrow{a.s.} P(X_1 \leq t_1, \dots, X_{j-1} \leq t_{j-1}), \quad (9.5.12)$$

for any  $t_1, \dots, t_{j-1} \in \mathbb{R}$ . Also, for any  $t_1, \dots, t_{j-1} \in \mathbb{R}$ , again due to Glivenko-Cantelli theorem for stationarity,

$$\sup_{t_j \in \mathbb{R}} |E[G_n(X_1 \leq t_1, \dots, X_j \leq t_j) | \mathbf{X}_j^n] - P(X_1 \leq t_1, \dots, X_j \leq t_j)| \xrightarrow{a.s.} 0, \text{ as } n \rightarrow \infty. \quad (9.5.13)$$

Combining (9.5.12) and (9.5.13) yields

$$\sup_{t_j \in \mathbb{R}} \left| \frac{E[G_n(X_1 \leq t_1, \dots, X_j \leq t_j) | \mathbf{X}_j^n]}{E[G_n(X_1 \leq t_1, \dots, X_{j-1} \leq t_{j-1}) | \mathbf{X}_{j-1}^n]} - \frac{P(X_1 \leq t_1, \dots, X_j \leq t_j)}{P(X_1 \leq t_1, \dots, X_{j-1} \leq t_{j-1})} \right| \xrightarrow{a.s.} 0, \text{ as } n \rightarrow \infty,$$

for all  $t_1, \dots, t_{j-1} \in \mathbb{R}$ . That is, for all  $t_1, \dots, t_{j-1} \in \mathbb{R}$ ,

$$\sup_{t_j \in \mathbb{R}} \left| \tilde{\zeta}_{jn}(t_1, \dots, t_j) - P(X_j \leq t_j | X_1 \leq t_1, \dots, X_{j-1} \leq t_{j-1}) \right| \xrightarrow{a.s.} 0,$$

and since under mutual independence,  $P(X_j \leq t_j | X_1 \leq t_1, \dots, X_{j-1} \leq t_{j-1}) = P(X_j \leq t_j)$ ,

$$\sup_{t_j \in \mathbb{R}} \left| \tilde{\zeta}_{jn}(t_1, \dots, t_j) - P(X_j \leq t_j) \right| \xrightarrow{a.s.} 0, \text{ as } n \rightarrow \infty,$$

for all  $t_1, \dots, t_{j-1} \in \mathbb{R}$ , under mutual independence. More transparently, since  $\tilde{\zeta}_{jn}(t_1, \dots, t_j)$  is asymptotically independent of  $t_1, \dots, t_{j-1}$ , for any  $\epsilon > 0$  under mutual independence, there exists  $n_0(\epsilon) \geq 1$  such that for  $n > n_0(\epsilon)$ ,

$$\sup_{t_j \in \mathbb{R}} \left| \tilde{\zeta}_{jn}(t_1, \dots, t_j) - P(X_j \leq t_j) \right| < \epsilon,$$



for all  $t_1, \dots, t_{j-1} \in \mathbb{R}$ . That is, (9.5.10)

$$\sup_{t_1, \dots, t_j \in \mathbb{R}} \left| \tilde{\zeta}_{jn}(t_1, \dots, t_j) - P(X_j \leq t_j) \right| \xrightarrow{a.s.} 0, \text{ as } n \rightarrow \infty. \quad (9.5.14)$$

For the second term of (9.5.11), note that

$$\sup_{t_j \in \mathbb{R}} \left| P(X_j \leq t_j) - \tilde{\zeta}_{jn}(t_j) \right| \xrightarrow{a.s.} 0, \quad (9.5.15)$$

as  $n \rightarrow \infty$ , due to Glivenko-Cantelli theorem for stationarity,

Combining (9.5.11), (9.5.14) and (9.5.15) yields (9.5.10) under mutual independence.

Now if (9.5.10) holds for  $j \geq 2$ , then this clearly implies mutual independence of the random variables. ■

**Remark 45** *Apart from being much more stable numerically compared to the approach of comparison between classical empirical conditional and marginal distributions, our DP-based approach also allows incorporation of the dependence structure, if any, through the base measure  $G_0$ . This can be achieved by empirically estimating the dependence structure from the data, and incorporating it in  $G_0$ . For example, if  $G_0$  corresponds to Gaussian process, then its mean and the covariance structure can be estimated from the data. This is expected to improve efficiency of inference regarding mutual independence. Note that such dependence structure can not be exploited in the approach of comparison between classical empirical conditional and marginal distributions.*

For our Bayesian characterization of mutual independence, let  $n_j$  denote the minimum number of observations associated with  $(X_1, \dots, X_j)$ , for  $j \geq 2$ . Now let  $\{c_j\}_{j=1}^\infty$  be a non-negative decreasing sequence and

$$Y_{j,n_j} = \mathbb{I} \left\{ \sup_{t_1, \dots, t_j \in \mathbb{R}} \left| \tilde{\zeta}_{jn_j}(t_1, \dots, t_j) - \tilde{\zeta}_{jn_j}(t_j) \right| \leq c_j \right\}.$$

Let, for  $j \geq 1$ ,

$$P(Y_{j,n_j} = 1) = p_{j,n_j}.$$

Let the rest of the recursive Bayesian procedure be the same as in Section 3.3. Then, using Theorem 44, the following theorem can be proved in almost the same way as Theorem 24.

**Theorem 46** *Let  $X^i; i = 1, 2, \dots$ , be stationary. Then  $(X_1, X_2, \dots)$  are mutually independent if and only if for all  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ , where  $\mathfrak{N}$  is some null set having probability measure zero, for any monotonically decreasing sequence  $\{c_j(\omega)\}_{j=1}^\infty$ ,*

$$\pi(\mathcal{N}_1 | y_{k,n_k}(\omega)) \rightarrow 1,$$

as  $k \rightarrow \infty$  and  $n_j \rightarrow \infty$  for  $k = 2, 3, \dots, K$  and  $K \rightarrow \infty$ , where  $\mathcal{N}_1$  is any neighborhood of 1 (one).

## 9.6 Bayesian characterization of Poisson point process

Recall that for a Poisson point process, if for any set of disjoint regions  $C_i; i = 1, \dots, K$ , where  $C_i \subset \mathcal{S}$ ,  $\mathbf{X}_{C_i}$ , denoting the set of points in  $C_i$ , are independent, for any  $K > 1$ . This is referred to as the complete independence property in Daley and Vere-Jones (2003). However, complete independence alone is not sufficient to characterize Poisson point process. In this regard, let us consider the following assumptions.

- (A1) Let  $N(A)$ , the number of points in the set  $A$ , be defined and finite for every bounded set  $A$  in the Borel sigma-field generated by the open spheres of  $\mathcal{S}$ . This can be simply expressed by saying that the trajectories of  $N(\cdot)$  are almost surely boundedly finite (Daley and Vere-Jones (2003)).
- (A2)  $P_r \{N(S_\epsilon(x)) > 1\} = o(P_r \{N(S_\epsilon(x)) > 1\})$ , as  $\epsilon \rightarrow 0$ . Here  $S_\epsilon(x)$  denotes the open sphere with radius  $\epsilon$  and center  $x$ . This property is called orderliness.

With these, the Poisson process can be characterized as follows.

**Theorem 47 (Daley and Vere-Jones (2003))** *Let  $N(\cdot)$  be almost surely boundedly finite and without fixed atoms. Then  $N(\cdot)$  is a Poisson process if and only if it is orderly and has the complete independence property.*

We also note the following lemma.

**Lemma 48 (Daley and Vere-Jones (2003))** *A point  $x_0$  is an atom of the parameter measure  $\Lambda$  if and only if it is a fixed atom of the process.*

**Corollary 49** *Theorem 47 and Lemma 48 together imply that if  $\Lambda$  corresponds to a continuous distribution, then (A1)-(A2) along with complete independence characterize Poisson process.*

We now characterize Poisson process in a recursive Bayesian framework using our Bayesian characterization of mutual independence assuming (A1)–(A2) and non-atomicity of the process. In all our examples, we consider  $\Lambda$  to be associated with continuous distributions, hence non-atomic; (A1)–(A2) also hold in all our simulation studies.

Assume that  $\mathbf{X}_{C_i}$  are locally stationary and let  $D_{C_i}$  denote the set of minimum inter-point distances associated with  $\mathbf{X}_{C_i}$ . As before, for  $r = 1, 2, \dots$ , let  $W_r$  and  $W_{dr}$  be the observation window and the space of inter-point distances corresponding to  $W_r$  at the  $r$ -th stage, where  $|W_r| \rightarrow \infty$  as  $r \rightarrow \infty$ . Let us also replace  $n_j$  and  $K$  with  $n_{jr}$  and  $K_r$ , respectively, as before.

Now let  $\{c_j\}_{j=1}^\infty$  be a non-negative decreasing sequence and

$$Y_{j,n_{jr}} = \mathbb{I} \left\{ \sup_{t_1, \dots, t_j \in \mathbb{R}} \left| \tilde{\zeta}_{jn_{jr}}(t_1, \dots, t_j) - \tilde{\zeta}_{jn_{jr}}(t_j) \right| \leq c_j \right\},$$

and, for  $j \geq 1$ ,

$$P(Y_{j,n_{jr}} = 1) = p_{j,n_{jr}}.$$

Then we have the following result for point processes corresponding to Theorem 46.

**Theorem 50** *Let  $\mathbf{X}$  be a point process in  $\mathcal{S}$ . Assume that for the disjoint regions  $C_i \subset \mathcal{S}; i = 1, \dots, K_r$ ,  $\mathbf{X}_{C_i}$  are locally stationary. Then  $(D_{C_1}, \dots, D_{C_{K_r}})$  are mutually independent if and only if for all  $\omega \in \mathfrak{S} \cap \mathfrak{N}^c$ , where  $\mathfrak{N}$  is some null set having probability measure zero, for any monotonically decreasing sequence  $\{c_j(\omega)\}_{j=1}^\infty$ , and any set of disjoint regions  $C_i; i = 1, \dots, K_r$ , where  $C_i \subset \mathcal{S}$ ,*

$$\pi(\mathcal{N}_1 | y_{k, n_{kr}}(\omega)) \rightarrow 1, \tag{9.6.1}$$

as  $k \rightarrow \infty$  and  $n_{kr} \rightarrow \infty$  for  $k = 2, 3, \dots, K_r$  and  $K_r \rightarrow \infty$  as  $r \rightarrow \infty$ , where  $\mathcal{N}_1$  is any neighborhood of 1 (one).

**Proof.** Using Theorem 44, the proof follows in almost the same way as that of Theorem 24. ■

**Theorem 51** *Consider any point process  $\mathbf{X} \in \mathcal{S}$ . Assume that the  $\sigma$ -algebra for  $\mathcal{S}$  is separable and generated by the mutually disjoint sets  $\{C_i; i \geq 1\}$ , and that  $\mathbf{X}_{C_i}$  are locally stationary. Then, provided that (A1)–(A2) hold and the process is non-atomic,  $\mathbf{X}$  is a Poisson point process if and only if (9.6.1) holds.*

**Proof.** By Theorem 50,  $(D_{C_1}, \dots, D_{C_{K_r}})$  are mutually independent if and only if (9.6.1) holds. Since the mutually disjoint sets  $\{C_i; i \geq 1\}$  generates the  $\sigma$ -field for  $\mathcal{S}$ , it follows that any set of mutually disjoint sets  $\{B_1, \dots, B_\ell\}$  in the  $\sigma$ -field for  $\mathcal{S}$ , for any  $\ell > 1$ ,  $(D_{B_1}, \dots, D_{B_\ell})$ , are mutually independent.

Also, it is easy to see that  $(D_{B_1}, \dots, D_{B_\ell})$  are mutually independent if and only if  $(\mathbf{X}_{B_1}, \dots, \mathbf{X}_{B_\ell})$  are mutually independent.

Hence, by the hypothesis of the theorem it follows that  $\mathbf{X}$  is a Poisson point process if and only if (9.6.1) holds. ■

### 9.6.1 Computational strategy for mutual independence assessment

Note that for relatively large  $j$ , it may not be feasible to directly compute  $\sup_{t_1, \dots, t_j \in \mathbb{R}} \left| \tilde{\zeta}_{jn_j}(t_1, \dots, t_j) - \tilde{\zeta}_{jn_j}(t_j) \right|$ . Hence we consider the following strategy. For  $j = 2$ , let  $\tilde{t}_1, \tilde{t}_2$  be the maximizers of  $\left| \tilde{\zeta}_{jn_j}(t_1, \dots, t_j) - \tilde{\zeta}_{jn_j}(t_j) \right|$ , and for  $j \geq 3$ , let

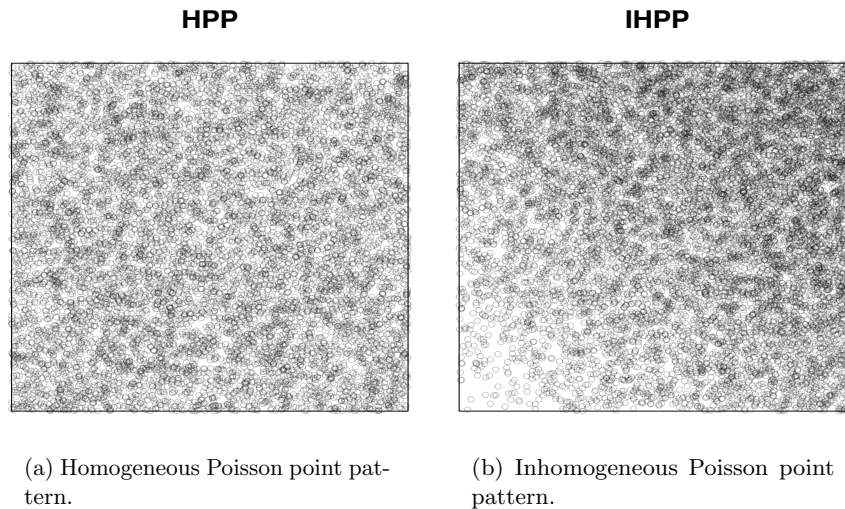
$$\begin{aligned} & \sup_{t_1, \dots, t_j \in \mathbb{R}} \left| \tilde{\zeta}_{jn_j}(t_1, \dots, t_j) - \tilde{\zeta}_{jn_j}(t_j) \right| \\ &= \sup_{t_j \in \mathbb{R}} \left| \frac{E[G_n(X_1 \leq \tilde{t}_1, \dots, X_{j-1} \leq \tilde{t}_{j-1}, X_j \leq t_j) | \mathbf{X}_j^n]}{E[G_n(X_1 \leq \tilde{t}_1, \dots, X_{j-1} \leq \tilde{t}_{j-1}) | \mathbf{X}_{j-1}^n]} - \frac{\alpha G_{0,j}(X_j \leq t_j) + \sum_{r=1}^n \delta_{X_j^r}(X_j^r \leq t_j)}{\alpha + n} \right|, \end{aligned} \quad (9.6.2)$$

where  $\tilde{t}_3, \dots, \tilde{t}_{j-1}$  are the maximizers of  $\left| \tilde{\zeta}_{j-1n_{j-1}}(t_1, \dots, t_{j-1}) - \tilde{\zeta}_{j-1n_{j-1}}(t_{j-1}) \right|$ , for  $j \geq 3$ .

## 9.7 Simulation experiments

### 9.7.1 Example 1: Detection of HPP and IHPP and their properties

We generate a HPP with intensity  $\lambda = 1$  on a window of the form  $[0, 100] \times [0, 100]$ , using the R package “spatstat” (Baddeley and Turner (2005)), and obtain 9949 points in this exercise. We also simulate an IHPP using the spatstat package with  $\lambda(x, y) = 100(x + y)$  on  $[0, 5] \times [0, 5]$ , generating 12447 observations. The plots of the point patterns are provided in Figure 9.7.1. Observe that while the HPP pattern in panel (a) is reasonably uniform on the observed window, the IHPP pattern in panel (b) shows sparsity in the bottom left corner and density in the top right corner of the observation window. Our goal is to identify the true point processes that generated the data, pretending that they are unknown and that only the data are observed.



**Figure 9.7.1:** Homogeneous and inhomogeneous Poisson point processes.

### Homogeneity detection

Let us first concentrate on the HPP data. With  $K = 1000$  clusters, we use bound (7.2.3) and obtain  $\hat{C}_1 = 0.25$  as the minimum value of  $\hat{C}_1$  that led to convergence of our recursive Bayesian algorithm to 1. The result is depicted in panel (a) of Figure 9.7.2. Panel (b) of Figure 9.7.2 is the simultaneous critical envelope associated with classical test of HPP, prepared using spatstat with 1000 simulations of CSR. Here  $r$  stands for the distance argument, and  $\hat{G}_{obs}(r)$ ,  $\hat{G}_{theo}(r)$ ,  $\hat{G}_{lo}(r)$  and  $\hat{G}_{hi}(r)$  stand for the observed empirical distribution function for the distances with Kaplan-Meier edge correction, the theoretical distribution function under CSR, the lower critical boundary and the upper critical boundary for the distribution functions under CSR, respectively. Here the significance level of simultaneous Monte Carlo test is given by 0.000999. Since the observed distribution function fall well within the lower and upper critical boundaries, the result is in agreement with our Bayesian result and indeed, the truth.

We now analyse the point pattern obtained from the IHPP. Panel (c) of Figure

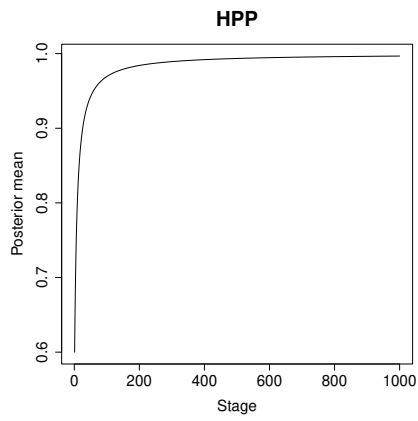
9.7.2 shows the result of our Bayesian analysis with  $K = 1000$  clusters and  $\hat{C}_1 = 0.25$ . Divergence to zero, that is, inhomogeneity is clearly indicated. However, this does not validate or invalidate Poisson process. To validate Poisson process, we need to create a characterization of mutual independence between the points contained in the  $K$  clusters. Panel (d) of Figure 9.7.2 is similar to panel (b) except that the observed distribution function in this case now corresponds to IHPP. Note that the observed distribution function  $\hat{G}_{obs}(r)$  falls almost entirely within the limits  $\hat{G}_{lo}(r)$  and  $\hat{G}_{hi}(r)$ , which makes it considerably difficult to distinguish this IHPP from HPP. The advantage of our Bayesian method depicted in panel (c) is clearly pronounced over this classical method in this regard.

### Stationarity detection

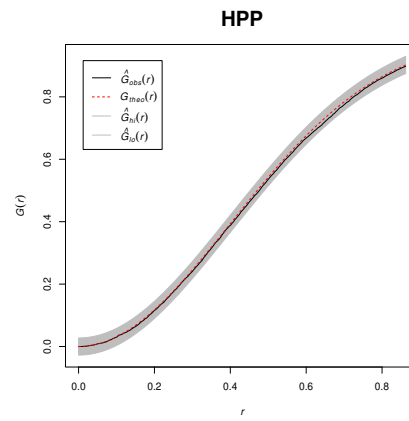
The traditional tests of CSR tests for HPP only. But inhomogeneity neither rejects the Poisson assumption, nor either of stationarity and nonstationarity. In this regard, we first address the question of stationarity and nonstationarity with our Bayesian method in our current examples of HPP and IHPP. Recall that for point processes, we regard the minimum distances  $d_i; i = 1, \dots, n$ , as the spatial data, along with their corresponding locations. Indeed, with this, we obtain the correct results with  $K = 1000$  clusters, bound (7.2.3) with  $\hat{C}_1 = 0.06$ , the minimum value for which convergence to 1 is obtained under the HPP example. The results presented in Figure 9.7.3, correctly identifies HPP and IHPP as stationary and nonstationary, respectively. Larger values of  $\hat{C}_1$ , such as  $\hat{C}_1 = 0.1$  led to the same result.

### Validation of Poisson assumption

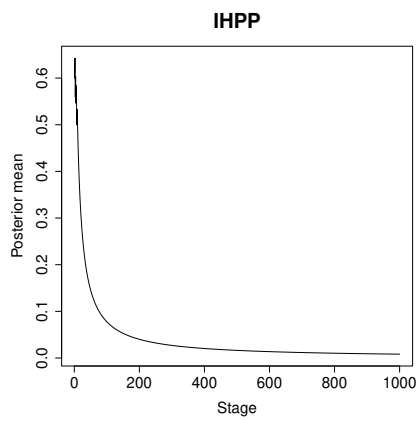
We finally examine, with our recursive Bayesian method for characterizing mutual independence, if the two point patterns that we generated can be safely assumed to be Poisson point patterns. Note that Poisson point process is equivalent to mutual



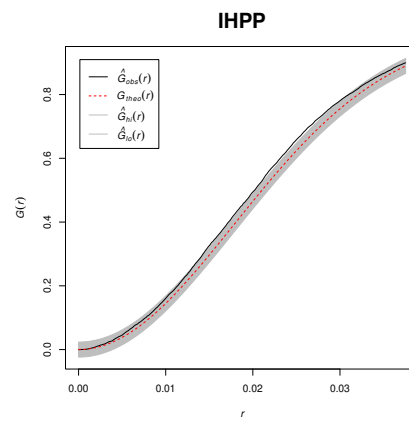
(a) HPP detection with Bayesian method.



(b) HPP detection with classical method.



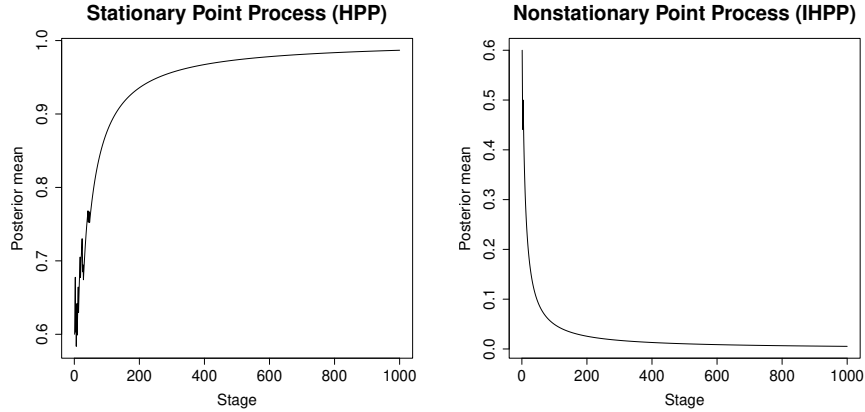
(c) IHPP detection with Bayesian method.



(d) IHPP detection with classical method.

**Figure 9.7.2:** Detection of CSR with our Bayesian method and traditional classical method.



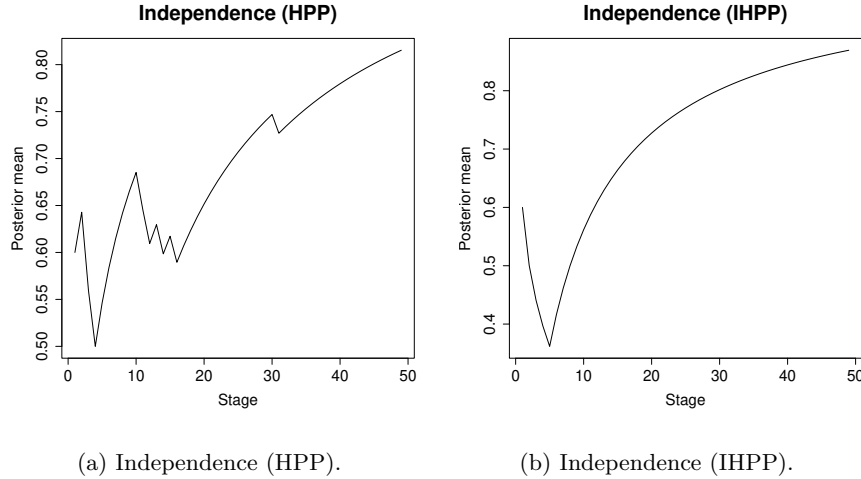


(a) Stationary point process (HPP).

(b) Nonstationary point process (IHPP).

**Figure 9.7.3:** Detection of stationarity and nonstationarity of point processes (here HPP and IHPP) with our Bayesian method.

independence of the points in disjoint subsets of  $W$ . In this regard, for  $i = 1, \dots, K$ , let  $\mathbf{X}_{C_i}$  denote the points in cluster  $C_i$ . If  $\mathbf{X}_{C_i}$  are mutually independent for all possible clusters  $C_i$  and  $K$ , then  $\mathbf{X}$  can be regarded as Poisson point process. For practical purposes, we restrict attention to a single set of clusters  $C_1, \dots, C_K$ . For numerical stability of the computations, we set  $K = 50$ , so that in most cases we investigate mutual independence among  $K = 50$  variables, where each variable is considered to take values in one and only one of the clusters. We set the strength parameter  $\alpha$  of the Dirichlet process to 1, which is quite standard, and use the ‘emcdf’ function of the ‘Emcdf’ package in R to parallelise the computations of the joint empirical distribution functions required for our Bayesian method. Here the joint distribution functions are those of the log-distances associated with the clusters. For the base distribution  $G_0$  of the Dirichlet process, we considered the multivariate normal distribution with mean vector and covariance matrices obtained empirically from the log-distances associated with  $\mathbf{X}_{C_i}$ ’s. Specifically, for  $K$  dimensions,  $G_0$  is a  $K$ -variate normal distribution with mean



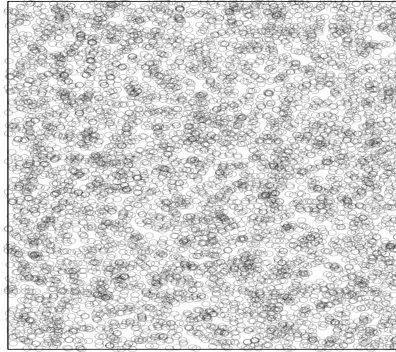
**Figure 9.7.4:** Detection of independence in point patterns (here HPP and IHPP) with our Bayesian method, suggesting that both the point processes are Poisson point processes.

vector being the  $K$ -component vector obtained by taking the means of the log-distances in  $\mathbf{X}_{C_i}$ ;  $i = 1, \dots, K$  and the covariance matrix being the empirical covariance obtained from the log-distances in the  $K$  clusters. The lower-dimensional distributions are then simply the marginalized versions of the higher-dimensional cases.

The entire exercise beginning from clustering the observed point pattern to yielding the maximum absolute differences between the conditional distribution functions and the marginal distribution functions, takes about 20 minutes in a 4-core laptop. The results of our Bayesian analyses with the bound (7.2.3) and  $\hat{C}_1 = 0.5$ , the minimum value for convergence in the HPP case, are provided in Figure 9.7.4. Indeed, both the panels indicate convergence, and hence independence. Hence, both the point processes can be safely assumed to be Poisson point processes.

### 9.7.2 Example 2: Homogeneous log-Gaussian Cox process

We now consider analyses of simulated data obtained from log-Gaussian Cox process.  $\mathbf{X}$  is a Cox process if conditional on a non-negative process  $\{\Lambda(u) : u \in \mathbf{S}\}$ ,  $\mathbf{X}$  is a Poisson

**Homogeneous LGCP****Figure 9.7.5:** Homogeneous LGCP.

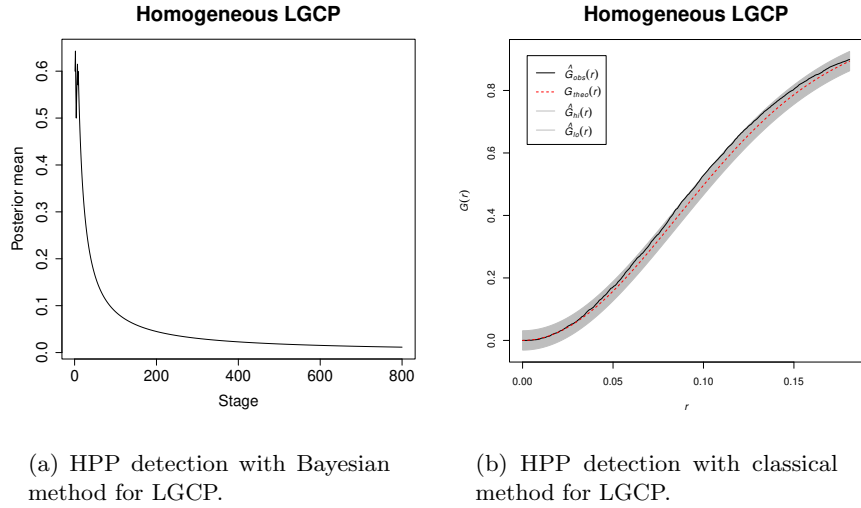
process with intensity function  $\Lambda$  (see, for example, Daley and Vere-Jones (2003)), and  $\mathbf{X}$  is a log-Gaussian Cox process if  $\log \Lambda$  is a Gaussian process. In this example, let us consider a log-Gaussian Cox process with mean function  $E[\log \Lambda(u)] = \mu(u) = 3$  for all  $u$ , and exponential covariance function given by  $Cov(\log \Lambda(u), \log \Lambda(v)) = \sigma^2 \times \exp(-a \|u - v\|)$ , where  $\|\cdot\|$  denotes Euclidean distance,  $\sigma^2 = 0.2$  and  $a = 10$ . This is a stationary non-Poisson point process, and homogeneous in the sense that the marginalized intensity  $E[\Lambda(u)]$ , is constant.

We choose  $W = [0, 15] \times [0, 20]$  and obtain 6553 observations from this point process using spatstat, which are displayed in Figure 9.7.5.

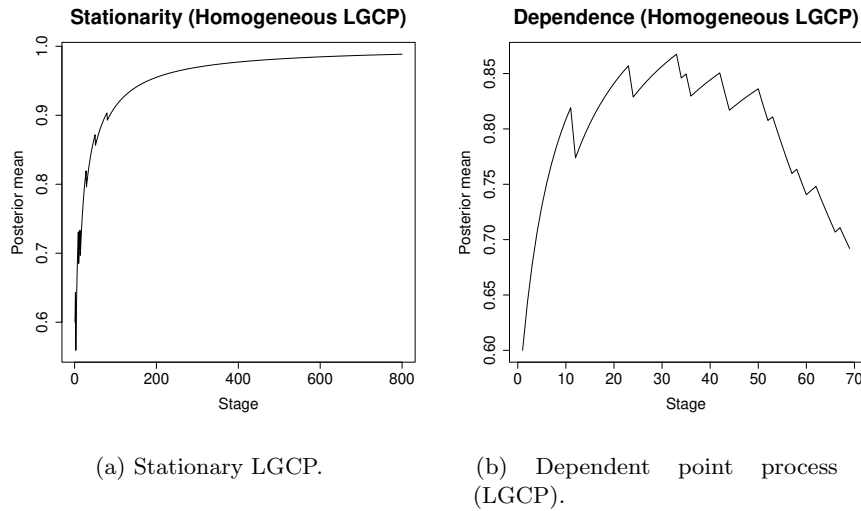
We consider  $K = 800$  and algorithm (7.2.3) with  $\hat{C}_1 = 0.24$  for our Bayesian method. Figure 9.7.6 compares our Bayesian method with the classical method regarding CSR detection. Observe that the Bayesian method correctly identifies that the point process is not CSR, while the classical method fails to correctly recognize the process.

For addressing stationarity, we set  $K = 800$  and  $\hat{C}_1 = 0.15$ . Panel (a) of Figure 9.7.7 shows that stationarity is clearly indicated by our Bayesian approach.

For testing if the underlying point process is Poisson process, we test independence as before, among  $K = 70$  random variables  $\mathbf{X}_{C_i}$ ;  $i = 1, \dots, K$ . With  $\hat{C}_1 = 0.5$ , panel (b) of Figure 9.7.7 indicates dependence, validating the non-Poisson assumption.



**Figure 9.7.6:** Detection of CSR with our Bayesian method and traditional classical method for LGCP. The Bayesian method correctly identifies that the underlying point process is not CSR, but the classical method falsely indicates CSR.



**Figure 9.7.7:** Detection of stationarity and dependence of homogeneous LGCP with our Bayesian method.

### Inhomogeneous LGCP

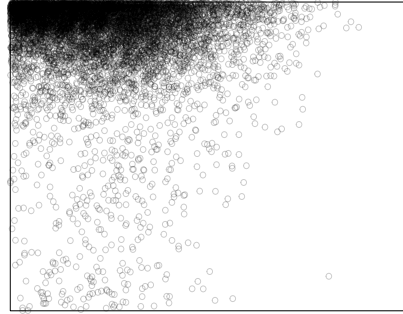


Figure 9.7.8: Inhomogeneous LGCP.

### 9.7.3 Example 3: Inhomogeneous log-Gaussian Cox process

We now consider a log-Gaussian Cox process where the covariance is now of the Matérn form

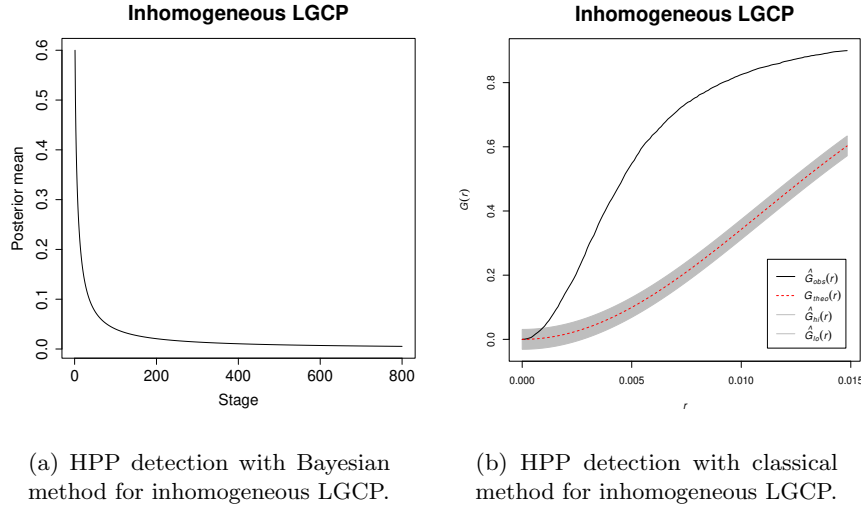
$$\text{Cov}(\log \Lambda(u), \log \Lambda(v)) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \sqrt{2\nu} \frac{\|u-v\|}{\rho} \right)^\nu \mathcal{K}_\nu \left( \sqrt{2\nu} \frac{\|u-v\|}{\rho} \right), \quad (9.7.1)$$

where  $\Gamma$  is the gamma function,  $\mathcal{K}_\nu$  is the modified Bessel function of the second kind of the order  $\nu$ , and  $\rho^{-1}$  is the scale parameter. We chose  $\sigma^2 = 2$ ,  $\rho^{-1} = 0.7$  and  $\nu = 0.5$ . For the mean function, we chose  $\mu(u_1, u_2) = 5 - 1.5(u_1 - 0.5)^2 + 2(u_2 - 0.5)^2$ . Thus, the underlying LGCP is nonstationary. Since the expected intensity is not constant, the point process is inhomogeneous from this perspective.

Using spatstat, we obtained 8814 observations on  $W = [0, 3] \times [0, 2.2]$ , displayed in Figure 9.7.8.

Panel (a) of Figure 9.7.9 shows the result of our Bayesian approach to CSR detection. With  $K = 800$  and  $\hat{C}_1 = 0.24$ , while panel (b) shows the result of the classical method. Both the methods successfully identify that the underlying point process is not CSR.

As shown in panel (a) of Figure 9.7.10, our Bayesian approach captures nonstationarity of the point process. As before, for detection of nonstationarity, we set  $K = 800$  and



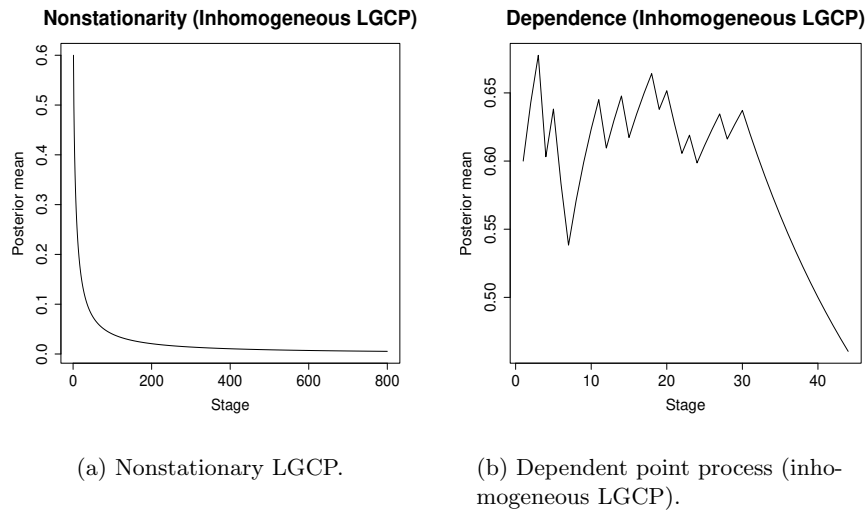
**Figure 9.7.9:** Detection of CSR with our Bayesian method and traditional classical method for LGCP. Both the methods correctly identify that the underlying point process is not CSR.

$$\hat{C}_1 = 0.15.$$

To test mutual independence among  $\mathbf{X}_{C_i}$ , for  $i = 1, \dots, K$ , we set  $K = 45$  (due to reasons of numerical stability) and  $\hat{C}_1 = 0.5$ , as before. Panel (b) of Figure 9.7.10 shows approximately stable behaviour around 0.6 till the last few points, where steady decrease is noticed. The stability around the relatively large value 0.6 for most part of the series indicates mutual independence among most of the random variables  $\mathbf{X}_{C_i}$ , but the last few values of the series suggest that the entire set of random variables  $\mathbf{X}_{C_i}; i = 1, \dots, 45$ , are perhaps not mutually independent. Hence, the entire set of random variables can not be regarded as mutually independent, leading to non-Poisson conclusion.

#### 9.7.4 Example 4: Inhomogeneous log-Gaussian Cox process

In this example, we choose the same Matérn covariance function (9.7.1), with the same values of  $\sigma^2$ ,  $\rho$  and  $\nu$  as before, but now we set  $\mu(u_1, u_2) = 1 - 0.4u_1$ . The resulting inhomogeneous LGCP obtained using spatstat, consisting of 7245 points, is depicted in



**Figure 9.7.10:** Detection of nonstationarity and dependence of inhomogeneous LGCP with our Bayesian method.

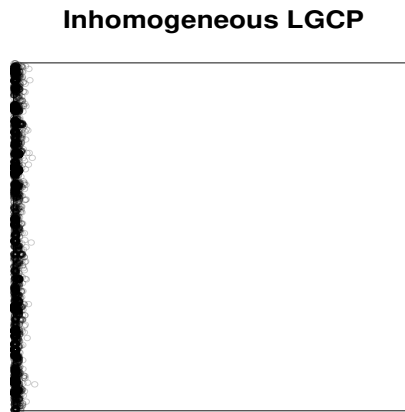
Figure 9.7.11.

With  $K = 800$  and  $\hat{C}_1 = 0.24$ , our Bayesian method successfully identifies the process as not CSR. The classical method is also successful in this regard. The results are shown in Figure 9.7.12.

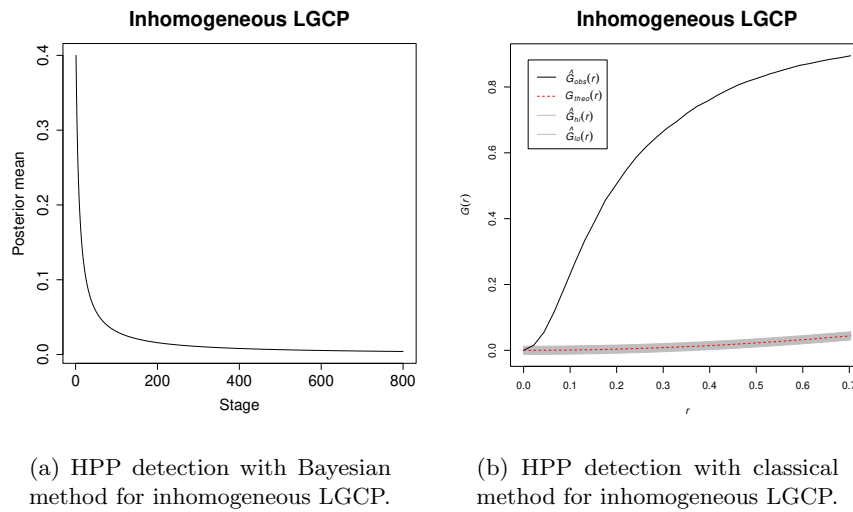
Again with  $K = 800$  and  $\hat{C}_1 = 0.15$ , our Bayesian method detects nonstationarity of the underlying point process. Also, with  $K = 40$  and  $\hat{C}_1 = 0.5$  as before, our method correctly detects dependence among  $\mathbf{X}_{C_i}; i = 1, \dots, K$ .

### 9.7.5 Example 5: Homogeneous Matérn cluster process

The Matérn cluster process is a special case of shot-noise Cox process where the offspring points are distributed uniformly inside a disc around the cluster center. To clarify, first consider a Poisson point process with intensity  $\kappa$ . Then each ‘parent’ point of this Poisson point process is replaced with a random cluster of ‘offspring’ points, where the number of points per cluster is distributed as Poisson with intensity  $\mu$  on a disc

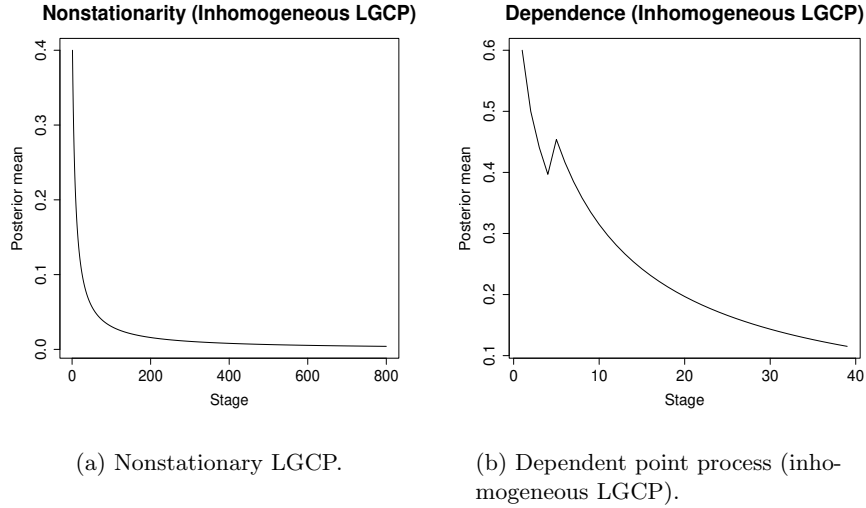


**Figure 9.7.11:** Inhomogeneous LGCP.



**Figure 9.7.12:** Detection of CSR with our Bayesian method and traditional classical method for LGCP. Both the methods correctly identify that the underlying point process is not CSR.





**Figure 9.7.13:** Detection of nonstationarity and dependence of inhomogeneous LGCP with our Bayesian method.

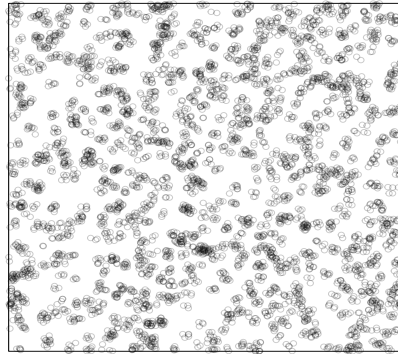
with center being the parent point. This point process is non-Poisson. Mathematically, consider

$$\Lambda(u) = \sum_{(c,\gamma) \in \Phi} \gamma k(c, u), \quad (9.7.2)$$

where  $c \in \mathbb{R}^2$ ,  $\gamma > 0$ ,  $\Phi$  is a Poisson process on  $\mathbb{R}^2 \times (0, \infty)$ , and  $k(c, \cdot)$  is a density for a two-dimensional continuous random variable. Then  $\mathbf{X}$  is a shot noise Cox process if given  $\Lambda$  defined by (9.7.2),  $\mathbf{X}$  is a Poisson process with intensity function  $\Lambda$ . It follows that  $\mathbf{X}$  is the superposition (union) of independent Poisson processes  $\mathbf{X}_{(c,\gamma)}$  with intensity functions  $\gamma k(c, \cdot)$ , where  $(c, \gamma) \in \Phi$ . If  $\gamma$  is a variable (either random or non-random), then  $\mathbf{X}_{(c,\gamma)}$  can be thought of as a cluster with center  $c$  and mean number of points  $\gamma$ . In this sense,  $\mathbf{X}$  is a Poisson cluster process.

The Matérn cluster process is a special case of the above process, where the centre points  $c$  arise from a Poisson process with intensity function  $\kappa$  and  $\gamma \equiv \mu$ , a positive non-random function, and  $k(c, \cdot)$  is the density of the uniform distribution on a disc of

### Matern Cluster Process



**Figure 9.7.14:** Matérn cluster point process pattern.

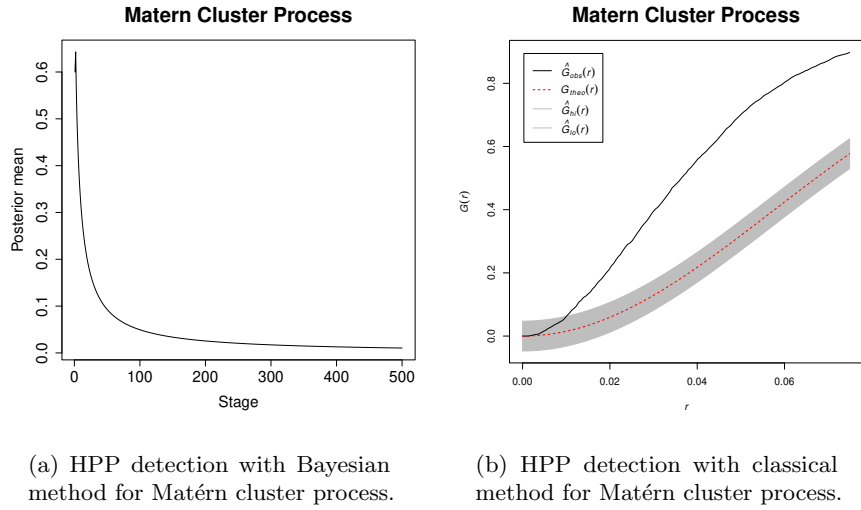
radius  $r$ , with center  $c$ .

In this example, we simulate a Matérn cluster process on a window  $W = [0, 10] \times [0, 10]$ ,  $\kappa = 10$ ,  $\mu = 5$ , and disc radius  $r = 0.1$ , and obtain 4882 points, shown in Figure 9.7.14. As can be easily verified from (9.7.2) and the following expositions, the random intensity function  $\Lambda$  in this case is stationary, and hence,  $\mathbf{X}$  is stationary.

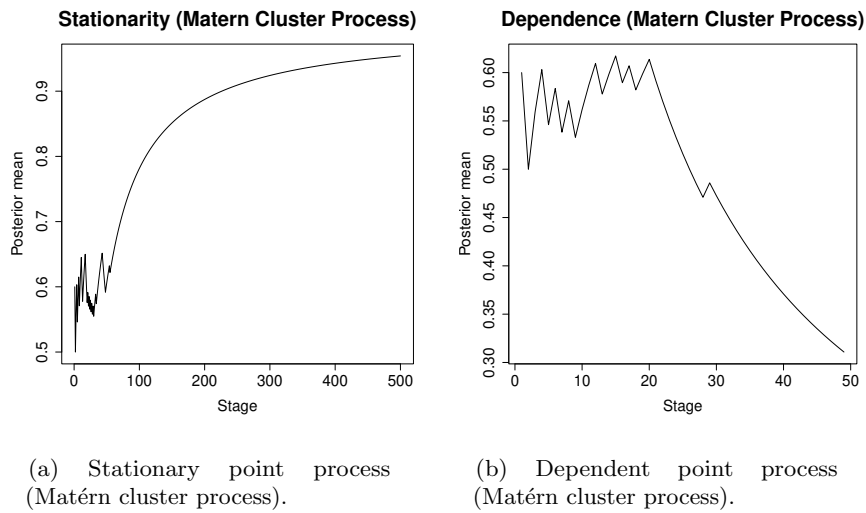
Figure 9.7.15 shows the results of our Bayesian method and the classical method for detecting CSR. Both the methods correctly point out that the underlying point process is not CSR. Here, for the Bayesian method, we set  $K = 500$  and  $\hat{C}_1 = 0.25$ , the maximum value leading to the conclusion of not CSR.

Panel (a) of Figure 9.7.16 shows that stationarity of the point process has been correctly captured by our Bayesian procedure, with  $K = 500$  and  $\hat{C}_1 = 0.06$ , the minimum value of  $\hat{C}_1$  leading to stationarity.

The result of our test for independence is depicted by panel (b) of Figure 9.7.16, for  $K = 50$  and  $\hat{C}_1 = 0.5$  as usual. Dependence is indicated, correctly leading to the non-Poisson conclusion.



**Figure 9.7.15:** Detection of CSR with our Bayesian method and traditional classical method for Matérn cluster process. Both the methods correctly identify that the underlying point process is not CSR.



**Figure 9.7.16:** Detection of stationarity and dependence of Matérn cluster process with our Bayesian method.

### Matern Cluster Process

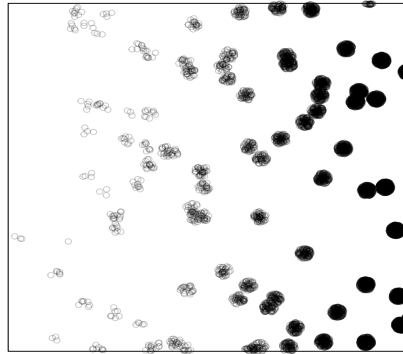


Figure 9.7.17: Inhomogeneous Matérn cluster point process pattern.

#### 9.7.6 Example 6: Inhomogeneous Matérn cluster process with $\mu$ inhomogeneous

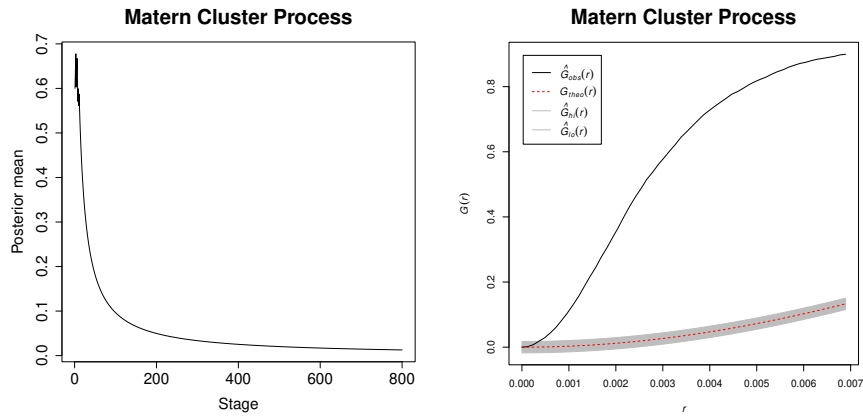
We now consider an inhomogeneous Matérn cluster process with  $\kappa = 10$ , disc radius  $r = 0.05$ , and  $\mu(u_1, u_2) = 2 \exp(2|u_1| - 1)$ , an obtain 8606 points in  $W = [0, 3] \times [0, 3]$ . The points are plotted in Figure 9.7.17.

Figure 9.7.18 shows that both the methods for detecting CSR correctly detect non-CSR. For the Bayesian method, we set  $K = 800$  and  $\hat{C}_1 = 0.6$ , the maximum value leading to the conclusion of not CSR.

With  $K = 800$  and  $\hat{C}_1 = 0.27$ , our Bayesian method correct points out nonstationarity. This value of  $\hat{C}_1$  is the maximum value leading to nonstationarity. As before, the Bayesian method correctly detects dependence with  $K = 50$  and  $\hat{C}_1 = 0.5$ . The results are depicted in Figure 9.7.19.

#### 9.7.7 Example 7: Matérn cluster process with $\kappa$ Inhomogeneous

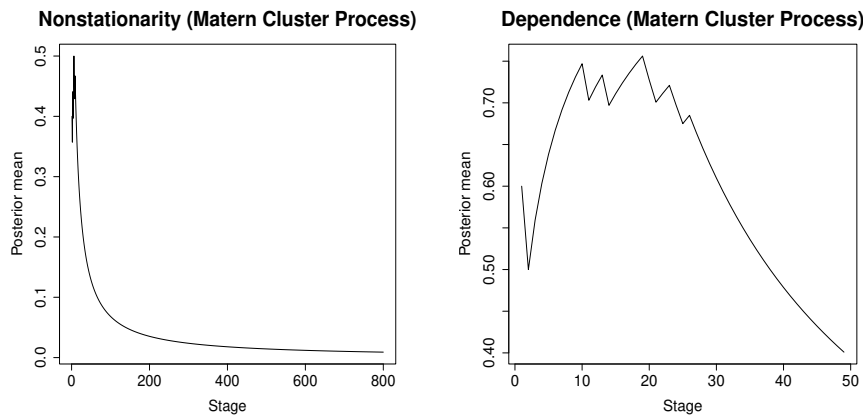
We consider another inhomogeneous Matérn cluster process with  $\kappa(u_1, u_2) = 2 \exp(2|u_1| - 1)$ , disc radius  $r = 0.05$ , and  $\mu = 3$ . The 2625 points that we obtained in  $W = [0, 3] \times [0, 3]$  are displayed in Figure 9.7.20.



(a) HPP detection with Bayesian method for Matérn cluster process.

(b) HPP detection with classical method for Matérn cluster process.

**Figure 9.7.18:** Detection of CSR with our Bayesian method and traditional classical method for inhomogeneous Matérn cluster process. Both the methods correctly identify that the underlying point process is not CSR.

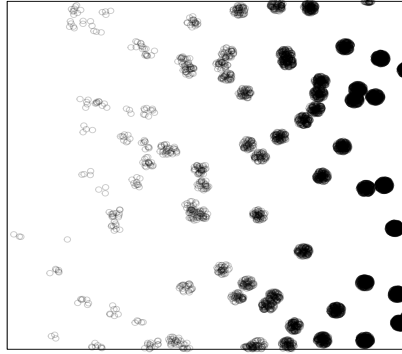


(a) Nonstationary point process (Matérn cluster process).

(b) Dependent point process (Matérn cluster process).

**Figure 9.7.19:** Detection of nonstationarity and dependence of Matérn cluster process with our Bayesian method.

### Matern Cluster Process



**Figure 9.7.20:** Inhomogeneous Matérn cluster point process pattern.

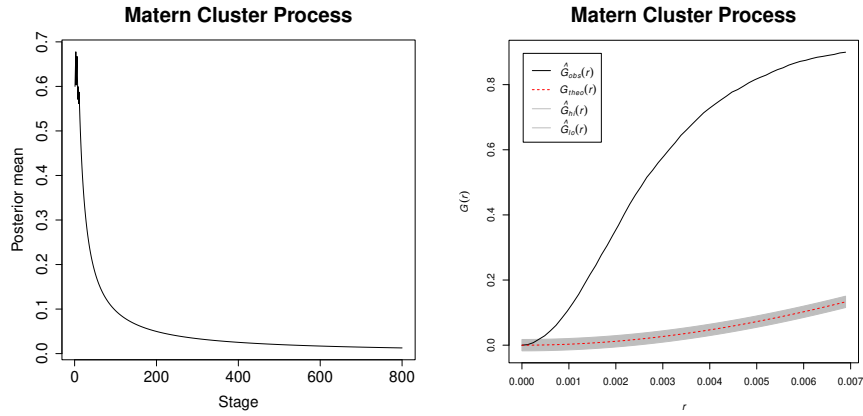
With  $K = 300$  and  $\hat{C}_1 = 0.4$ , the Bayesian algorithm correctly detects non-CSR. The classical method also performs adequately. Figure 9.7.21 shows that both the methods for detecting CSR correctly detect non-CSR.

Nonstationarity is also correctly detected by the Bayesian method with  $K = 300$  and  $\hat{C}_1 = 0.26$ , the maximum value leading to nonstationarity. Correct detection of dependence among  $\mathbf{X}_{C_i}; i = 1, \dots, 50$ , has also been possible with the Bayesian algorithm with  $\hat{C}_1 = 0.5$ . Figure 9.7.22 presents the relevant results.

#### 9.7.8 Example 8: Homogeneous Thomas process

The (modified) Thomas process is a special case of the general shot-noise Cox process in the same way as Matérn cluster process, but where  $k(c, \cdot)$  is the bivariate normal density with mean  $c$  and covariance  $\sigma^2 I$ . From (9.7.2) it is seen that a stationary process  $\mathbf{X}$  results provided  $\kappa$  and  $\mu$  are constants. The intensity after integrating out  $\Lambda$  is constant in this case, leading to homogeneous Thomas process.

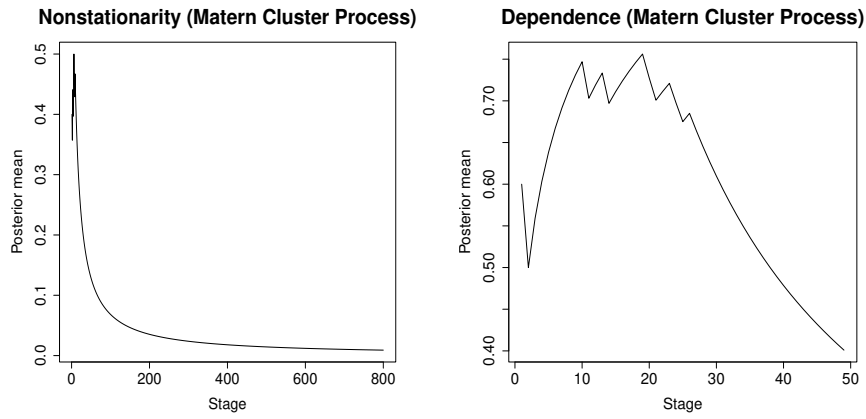
In this example, we first simulate a Thomas process with  $\kappa = 10$ ,  $\mu = 5$ ,  $\sigma^2 = 10$ , on the window  $W = [0, 10] \times [0, 10]$ , and obtained 4858 points. The point pattern for this homogeneous Thomas process is displayed in Figure 9.7.23.



(a) HPP detection with Bayesian method for Matérn cluster process.

(b) HPP detection with classical method for Matérn cluster process.

**Figure 9.7.21:** Detection of CSR with our Bayesian method and traditional classical method for inhomogeneous Matérn cluster process. Both the methods correctly identify that the underlying point process is not CSR.

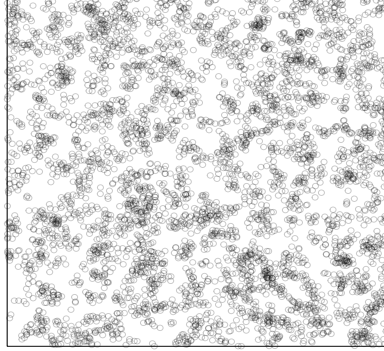


(a) Nonstationary point process (Matérn cluster process).

(b) Dependent point process (Matérn cluster process).

**Figure 9.7.22:** Detection of nonstationarity and dependence of Matérn cluster process with our Bayesian method.

### Homogeneous Thomas Process



**Figure 9.7.23:** Homogeneous Thomas point process pattern.

To test CSR, here we set  $K = 500$  and  $\hat{C}_1 = 0.23$  for the Bayesian method. The Bayesian method, as well as the classical method, correctly indicate that the underlying point process is not CSR. The results are displayed in Figure 9.7.24.

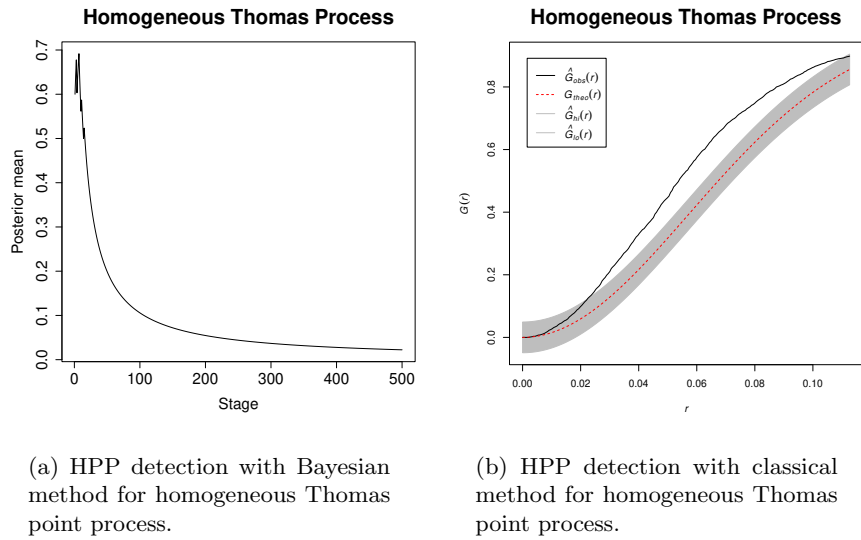
With  $K = 500$  and  $\hat{C}_1 = 0.18$ , we are able to identify stationarity of the underlying homogeneous Thomas point process using our Bayesian method. Also, with  $K = 500$  and  $\hat{C}_1 = 0.5$ , our Bayesian procedure suggests dependence among  $\mathbf{X}_{C_i}; i = 1, \dots, 50$ , leading us to correctly conclude that the point process is not Poisson.

#### 9.7.9 Example 9: Inhomogeneous Thomas process with $\mu$ inhomogeneous

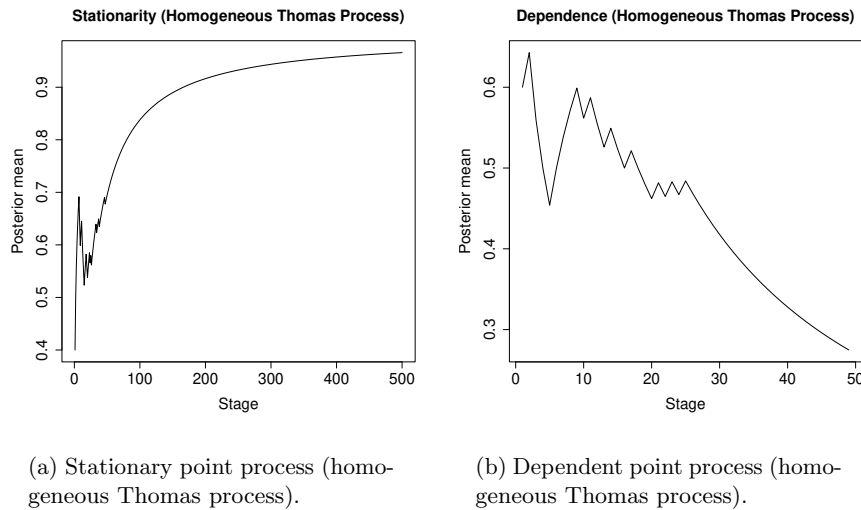
We now test our methods on an inhomogeneous Thomas process in  $W = [0, 3] \times [0, 3]$  with  $\kappa = 10$ ,  $\sigma^2 = 10$ , but  $\mu(u_1, u_2) = 5 \exp(2u_1 - 1)$ . That this process is also nonstationary follows from (9.7.2), since  $\Lambda$  is nonstationary in this case. The 10735 points we obtained using spatstat are shown in Figure 9.7.26.

With  $K = 1000$  and  $\hat{C}_1 = 0.23$ , our Bayesian method correctly identifies non-CSR. The classical method also does as well. The results of both these methods are shown in Figure 9.7.27.





**Figure 9.7.24:** Detection of CSR with our Bayesian method and traditional classical method for homogeneous Thomas point process. Both the methods correctly identify that the underlying point process is not CSR.



**Figure 9.7.25:** Detection of stationarity and dependence of homogeneous Thomas process with our Bayesian method.

### Inhomogeneous Thomas Process

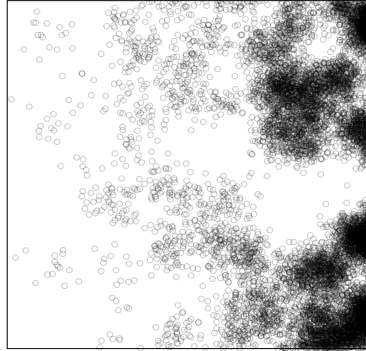


Figure 9.7.26: Inhomogeneous Thomas point process pattern.

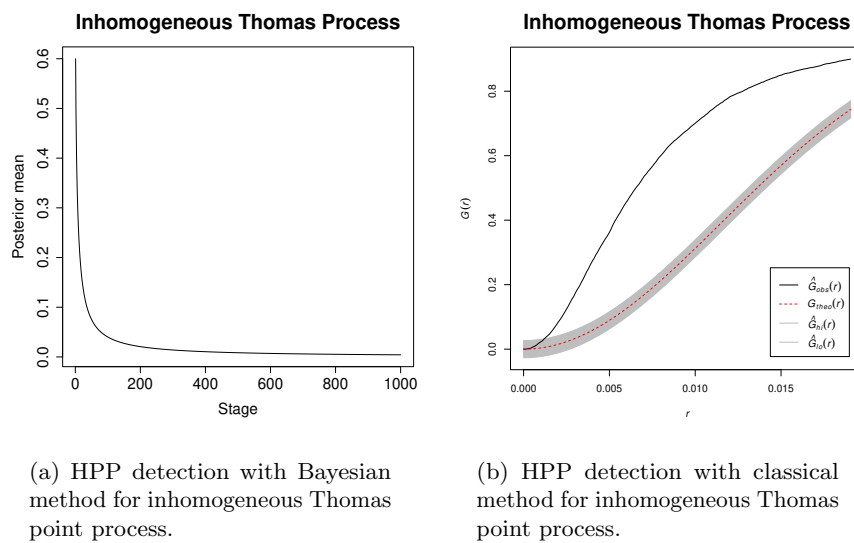
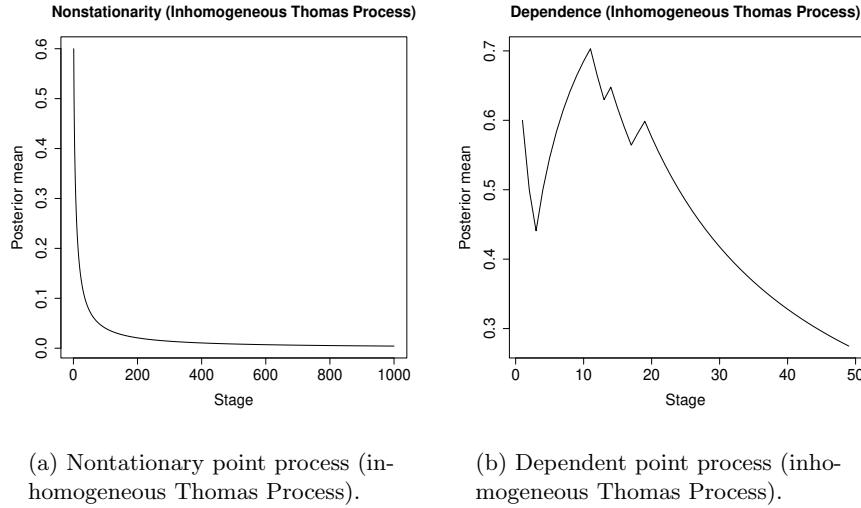


Figure 9.7.27: Detection of CSR with our Bayesian method and traditional classical method for Inhomogeneous Thomas point process. Both the methods correctly identify that the underlying point process is not CSR.



**Figure 9.7.28:** Detection of nonstationarity and dependence of inhomogeneous Thomas process with our Bayesian method.

Our Bayesian algorithm correctly captures nonstationarity with  $K = 1000$  and  $\hat{C}_1 = 0.18$ , the maximum value of  $\hat{C}_1$  leading to nonstationarity. Dependence among  $\mathbf{X}_{C_i}; i = 1, \dots, 50$  is borne out by our Bayesian strategy with  $\hat{C}_1 = 0.5$ . The results are presented in Figure 9.7.28.

### 9.7.10 Example 10: Inhomogeneous Thomas process with $\kappa$ inhomogeneous

We now consider another inhomogeneous Thomas process on  $W = [0, 3] \times [0, 3]$  with  $\mu = 5$ ,  $\sigma^2 = 10$  but  $\kappa(u_1, u_2) = 5 \exp(2x - 1)$ . This is also a nonstationary, non-Poisson, non-homogeneous point process. Figure 9.7.29 displays the 5608 points that we obtained from this process.

With  $K = 500$  and  $\hat{C}_1 = 0.23$ , our Bayesian correctly detected non-CSR. The classical method also performed adequately in this case. The results are shown in Figure 9.7.30.

As before, our Bayesian method correctly detected nonstationarity with  $K = 500$  and

### Inhomogeneous Thomas Process

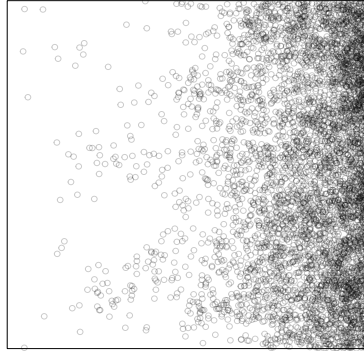


Figure 9.7.29: Inhomogeneous Thomas point process pattern.

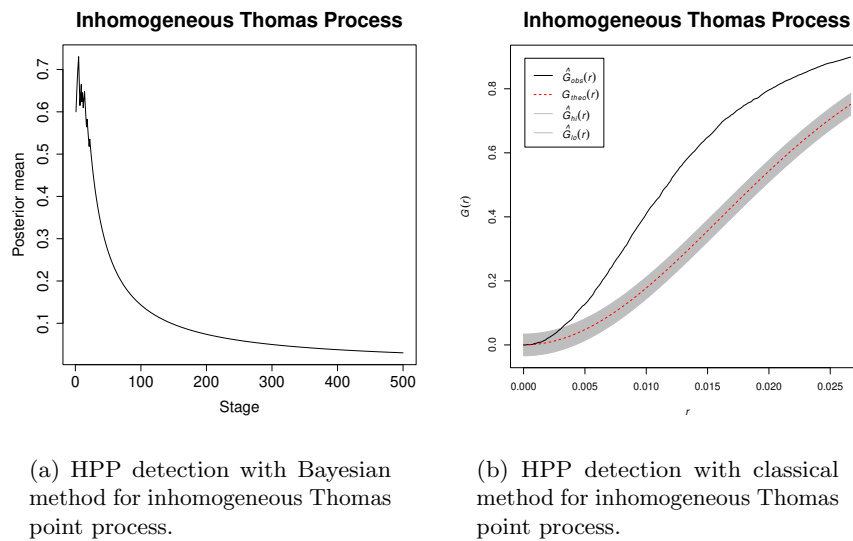
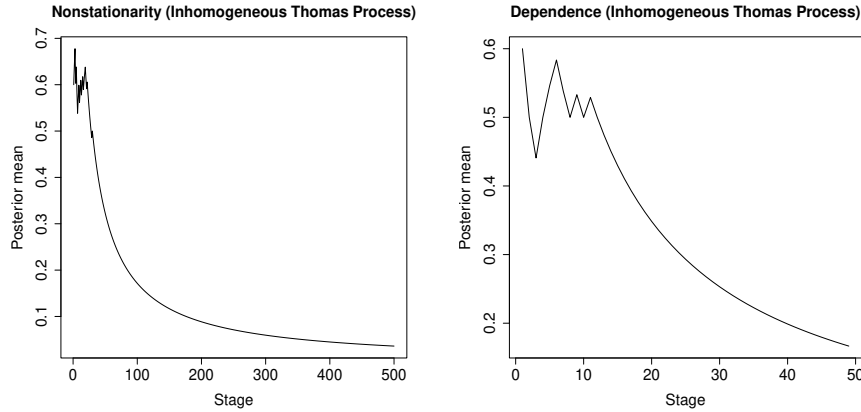


Figure 9.7.30: Detection of CSR with our Bayesian method and traditional classical method for inhomogeneous Thomas point process. Both the methods correctly identify that the underlying point process is not CSR.



(a) Nonstationary point process (inhomogeneous Thomas process).

(b) Dependent point process (inhomogeneous Thomas process).

**Figure 9.7.31:** Detection of nonstationarity and dependence of inhomogeneous Thomas process with our Bayesian method.

$\hat{C}_1 = 0.18$ . Also, as before, dependence among  $\mathbf{X}_{C_i}$ ;  $i = 1, \dots, 50$ , is correctly indicated by our Bayesian method, with  $\hat{C}_1 = 0.5$ .

### 9.7.11 Example 11: Inhomogeneous Thomas process with $\kappa$ and $\mu$ the same inhomogeneous function

Let us consider simulation from another inhomogeneous Thomas process where  $\kappa(u_1, u_2) = \mu(u_1, u_2) = 5 \exp(2u_1 - 1)$ . With  $\sigma^2 = 10$ , we obtained 5302 points on the window  $W = [0, 2] \times [0, 2]$ , displayed in Figure 9.7.32.

Figure 9.7.33 shows the results of Bayesian and classical CSR detection methods; both the methods performed adequately, correctly identifying non-CSR. For the Bayesian method we set  $K = 500$  and  $\hat{C}_1 = 0.23$ .

Nonstationarity of this point process has been correctly detected by our Bayesian method with  $K = 810$  and  $\hat{C}_1 = 0.18$ . As regards our Bayesian test for mutual independence, we correctly obtained dependence with  $K = 50$  and  $\hat{C}_1 = 0.5$ . The results

### Inhomogeneous Thomas Process

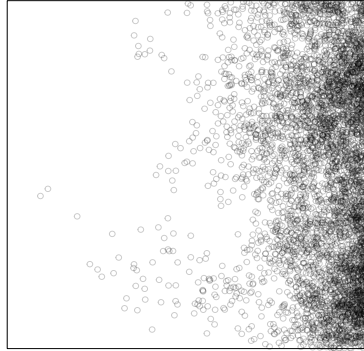
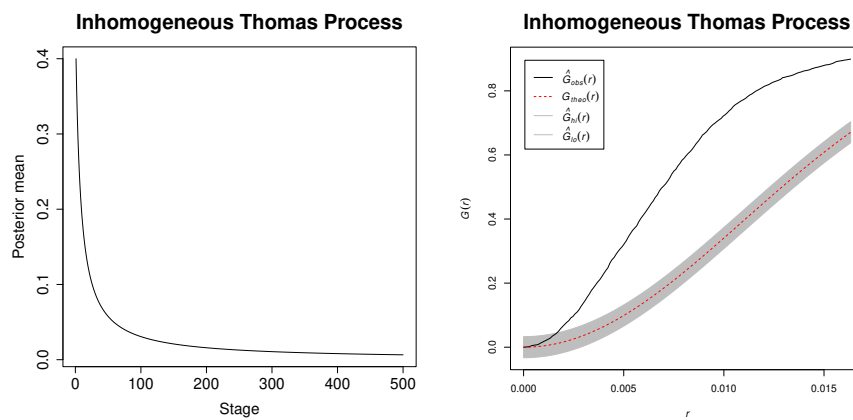


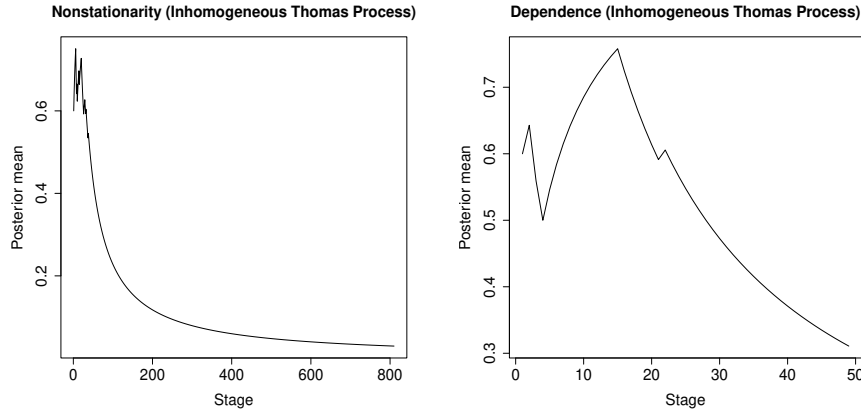
Figure 9.7.32: Inhomogeneous Thomas point process pattern.



(a) HPP detection with Bayesian method for inhomogeneous Thomas point process.

(b) HPP detection with classical method for inhomogeneous Thomas point process.

Figure 9.7.33: Detection of CSR with our Bayesian method and traditional classical method for inhomogeneous Thomas point process. Both the methods correctly identify that the underlying point process is not CSR.



(a) Nonstationary point process (inhomogeneous Thomas process).

(b) Dependent point process (inhomogeneous Thomas process).

**Figure 9.7.34:** Detection of nonstationarity and dependence of inhomogeneous Thomas process with our Bayesian method.

are presented in Figure 9.7.34.

### 9.7.12 Example 12: Inhomogeneous Thomas process with $\kappa$ and $\mu$ different inhomogeneous functions

Let us now consider another inhomogeneous Thomas process, where  $\mu(u_1, u_2) = 5 \exp(2u_1 - 1)$  and  $\kappa(u_1, u_2) = 10(u_1^2 + u_2^2)$ . We obtained 3573 observations with  $\sigma^2 = 10$  on the window  $W = [0, 2] \times [0, 2]$ . The data are displayed in Figure 9.7.35.

With  $K = 500$  and  $\hat{C}_1 = 0.23$ , we correctly obtained non-CSR with our Bayesian method. The classical method also correctly detected non-CSR. The results are presented in Figure 9.7.36.

Our Bayesian algorithm correctly detected nonstationarity with  $K = 500$  and  $\hat{C}_1 = 0.18$ . The Bayesian test for independence also correctly detected dependence with  $K = 50$  and  $\hat{C}_1 = 0.5$ . Both these results are presented in Figure 9.7.37.

### Inhomogeneous Thomas Process

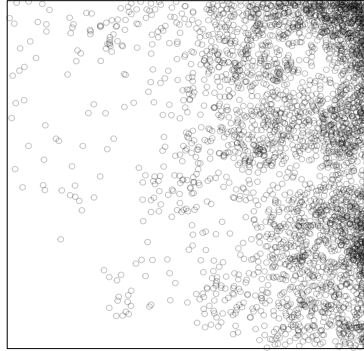


Figure 9.7.35: Inhomogeneous Thomas point process pattern.

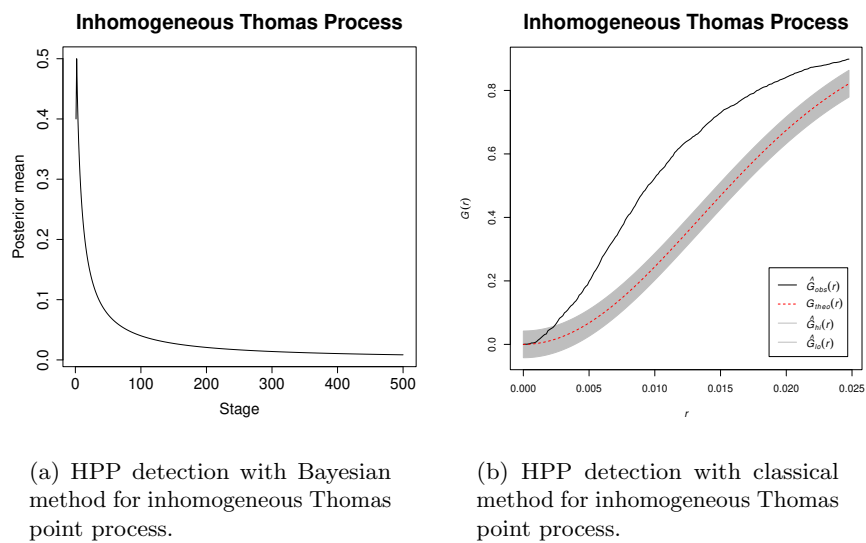
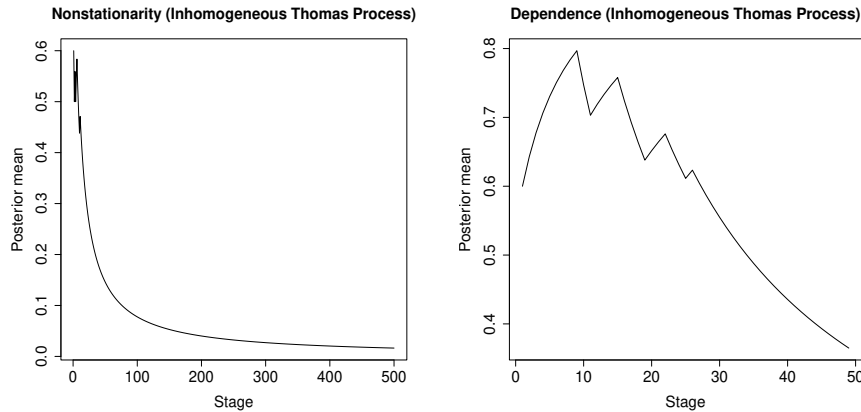


Figure 9.7.36: Detection of CSR with our Bayesian method and traditional classical method for inhomogeneous Thomas point process. Both the methods correctly identify that the underlying point process is not CSR.





(a) Nonstationary point process (inhomogeneous Thomas process).

(b) Dependent point process (inhomogeneous Thomas process).

**Figure 9.7.37:** Detection of nonstationarity and dependence of inhomogeneous Thomas process with our Bayesian method.

### 9.7.13 Example 13: Inhomogeneous Thomas Process with interchanged inhomogeneous $\kappa$ and $\mu$

We consider a final inhomogeneous Thomas process with  $\mu(u_1, u_2) = 10(u_1^2 + u_2^2)$  and  $\kappa(u_1, u_2) = 5 \exp(2u_1 - 1)$ . In this case, we obtained 4008 observations on the window  $W = [0, 2] \times [0, 2]$ , which we display in Figure 9.7.38.

For CSR detection, we set  $K = 500$  and  $\hat{C}_1 = 0.23$  for the Bayesian method. As shown by Figure 9.7.39, both the Bayesian and the classical method successfully detect non-CSR.

Our Bayesian method also successfully detected nonstationarity with  $K = 500$  and  $\hat{C}_1 = 0.18$ , and dependence, with  $K = 27$  (smaller value chosen to ensure numerical stability) and  $\hat{C}_1 = 0.5$ . These results are depicted in Figure 9.7.40.

### Inhomogeneous Thomas Process

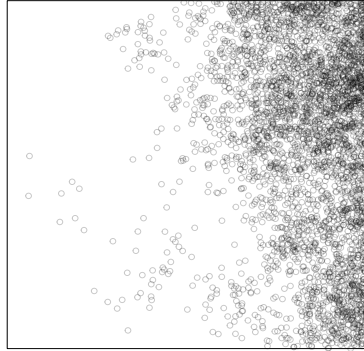


Figure 9.7.38: Inhomogeneous Thomas point process pattern.

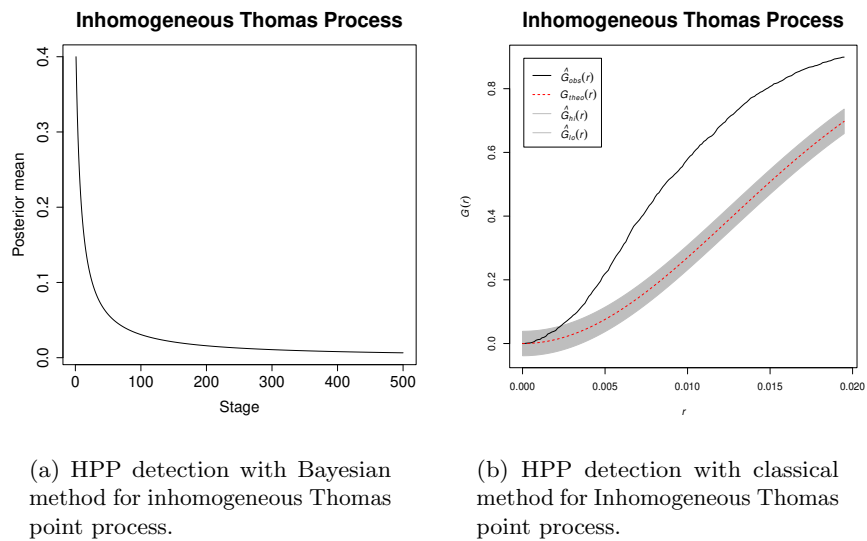
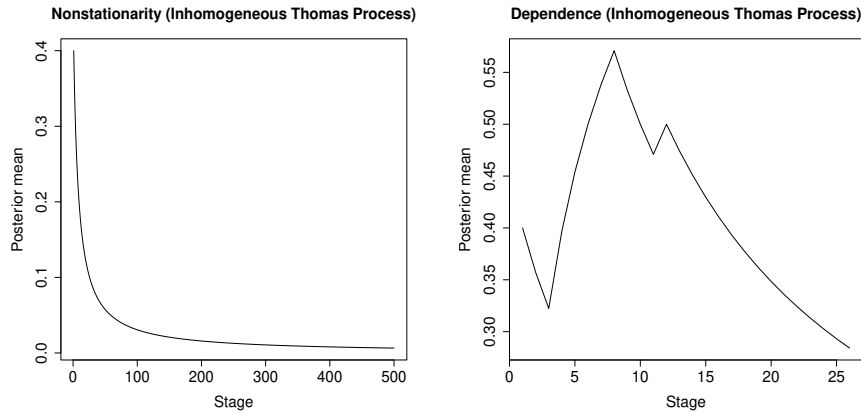


Figure 9.7.39: Detection of CSR with our Bayesian method and traditional classical method for inhomogeneous Thomas point process. Both the methods correctly identify that the underlying point process is not CSR.



(a) Nonstationary point process (inhomogeneous Thomas process).

(b) Dependent point process (inhomogeneous Thomas process).

**Figure 9.7.40:** Detection of nonstationarity and dependence of Inhomogeneous Thomas process with our Bayesian method.

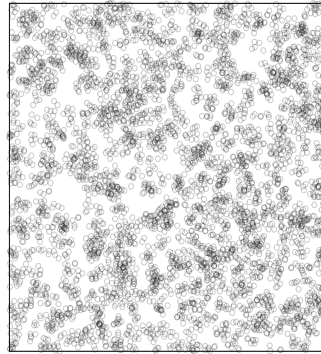
#### 9.7.14 Example 14: Homogeneous Neyman-Scott process

A Neyman-Scott process is a Cox process where the centers  $c$  in (9.7.2) arising from a Poisson process with intensity function  $\kappa$  and  $\gamma \equiv \mu$ , where  $\mu$  is some deterministic function. Note that the Neyman-Scott process is more general than the Thomas process in the sense that the density function  $k(c, \cdot)$  is left unspecified in the Neyman-Scott case, whereas for the Thomas process, this is a specific bivariate normal density.

More generally, the Neyman-Scott process allows a fixed number of offsprings on a disc with the parent point being the center of the disc. Here even though the centers arise from a Poisson process with intensity  $\kappa$ , the offsprings no longer follow the Poisson process, since given the parent points, the number of offsprings given each parent, is non-random. In such a case, the Neyman-Scott process is no longer a Cox process.

In order to test our methods on Neyman-Scott process, we first consider a homogeneous general Neyman-Scott process with  $\kappa = 10$ , with 5 points generated uniformly on each disc of radius 0.2 around the parent centers. The point pattern, simulated on

### Homogeneous Neyman–Scott Process



**Figure 9.7.41:** Homogeneous Neyman-Scott point process pattern.

$W = [0, 10] \times [0, 10]$ , consisting of 4867 observations, is shown in Figure 9.7.41.

Both the Bayesian and the traditional method of checking CSR correctly indicate that the underlying process is not CSR. In the Bayesian case, we set  $K = 500$  and  $\hat{C}_1 = 0.20$ . The results are displayed in Figure 9.7.42.

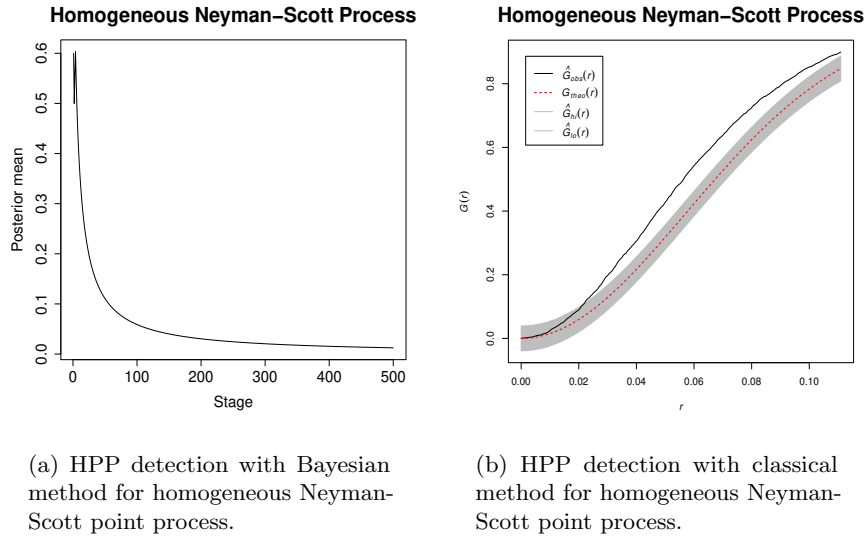
Stationarity is correctly detected by our Bayesian method with  $K = 500$  and  $\hat{C}_1 = 0.23$ . Also, with  $K = 50$  and  $\hat{C}_1 = 0.5$ , Poisson process is correctly ruled out. The results are depicted in Figure 9.7.43.

#### 9.7.15 Example 15: Inhomogeneous Neyman-Scott process

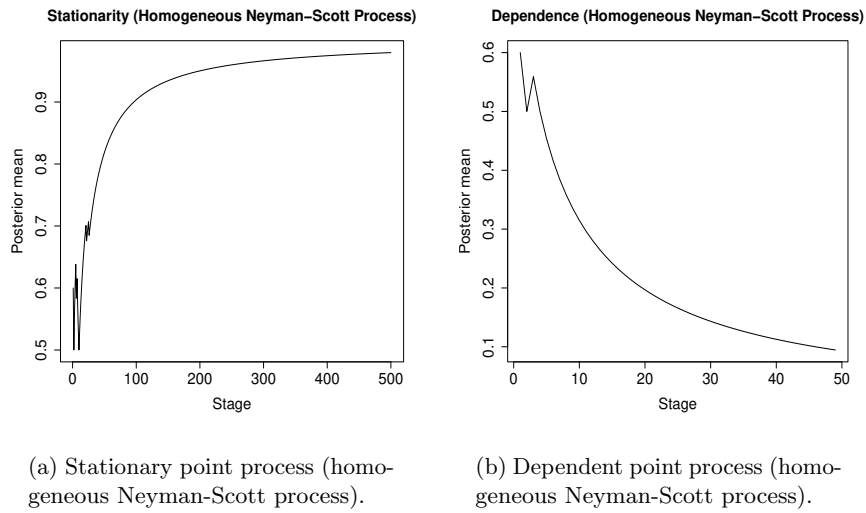
In this case, we generate a sample of size 8358 on  $W = [0, 4] \times [0, 4]$  from a Neyman-Scott process with the same setup as above, but with  $\kappa(u_1, u_2) = 10(u_1^2 + u_2^2)$ . The point pattern thus generated from this inhomogeneous Neyman-Scott process is shown in Figure 9.7.44.

With  $K = 800$  and  $\hat{C}_1 = 0.19$ , we obtain the correct non-CSR conclusion with the Bayesian method. The correct result is also identified by the classical method. Both the results are depicted in Figure 9.7.45.

Nonstationarity of this process is correctly detected by the Bayesian method with



**Figure 9.7.42:** Detection of CSR with our Bayesian method and traditional classical method for homogeneous Neyman-Scott point process. Both the methods correctly identify that the underlying point process is not CSR.



**Figure 9.7.43:** Detection of stationarity and dependence of homogeneous Neyman-Scott process with our Bayesian method.

### Inhomogeneous Neyman–Scott Process

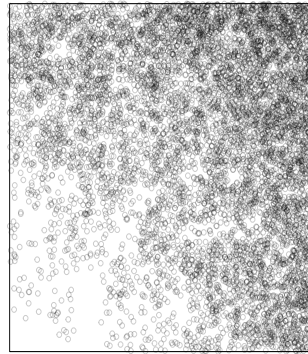
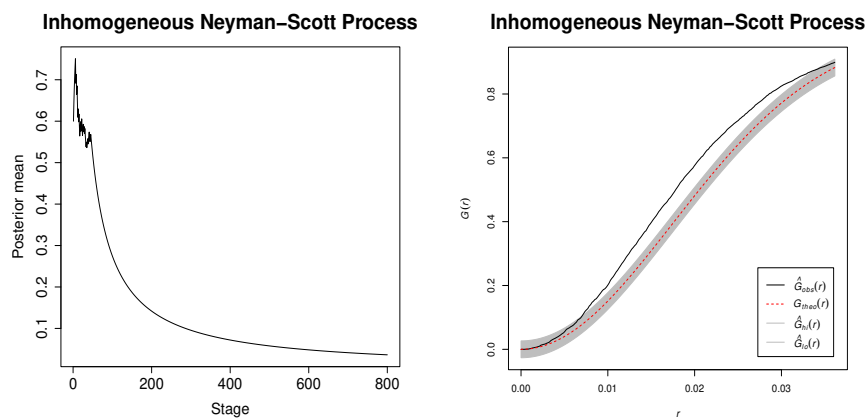


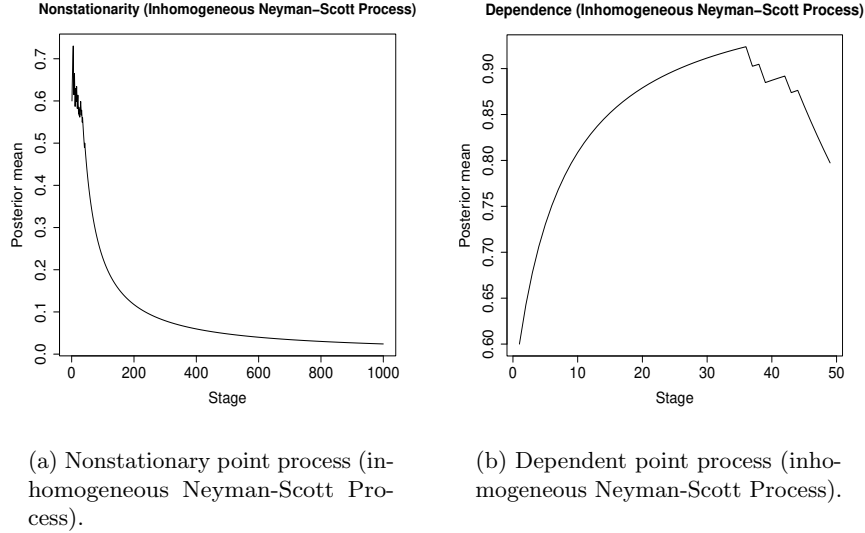
Figure 9.7.44: Inhomogeneous Neyman–Scott point process pattern.



(a) HPP detection with Bayesian method for inhomogeneous Neyman–Scott point process.

(b) HPP detection with classical method for inhomogeneous Neyman–Scott point process.

Figure 9.7.45: Detection of CSR with our Bayesian method and traditional classical method for inhomogeneous Neyman–Scott point process. Both the methods correctly identify that the underlying point process is not CSR.



**Figure 9.7.46:** Detection of nonstationarity and dependence of inhomogeneous Neyman-Scott process with our Bayesian method.

$K = 1000$  and  $\hat{C}_1 = 0.23$ ; this is shown in panel (a) of Figure 9.7.46. For  $K = 50$  and  $\hat{C}_1 = 0.5$ , panel (b) of Figure 9.7.46 shows steady increase for about the first 35 stages, but sharply decreases thenceforward, indicating dependence.

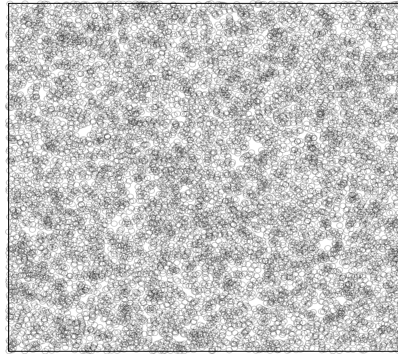
### 9.7.16 Example 16: Strauss process

The Strauss process (Strauss (1975); see also Møller and Waagepetersen (2004)) is an instance of pairwise interaction point process with density (with respect to unit intensity Poisson process)

$$f(x) \propto \beta^{n(x)} \gamma^{s_R(x)}, \quad (9.7.3)$$

where  $\beta > 0$ ,  $n(x)$  is the number of points in  $x$  and  $s_R(x) = \sum_{(\xi, \eta) \subseteq x} I\{\|\xi - \eta\| \leq R\}$  is the number of  $R$ -close pairs of points in  $x$ . Note that if  $\gamma = 1$ , we obtain Poisson process on  $\mathcal{S}$  with intensity  $\beta$ , and if  $\gamma < 1$ , there is repulsion between the  $R$ -close points pairs of points in  $\mathbf{X}$ .

### Strauss Process



**Figure 9.7.47:** Strauss point process pattern.

Using spatstat, we generate 9790 points from a Strauss process with  $\beta = 0.05$ ,  $\gamma = 0.2$  and  $R = 1.5$  on  $W = [0, 500] \times [0, 500]$ . The points are displayed in Figure 9.7.47.

To detect CSR, we set  $K = 800$  and  $\hat{C}_1 = 0.15$  for the Bayesian algorithm. As Figure 9.7.48 shows, both the classical and the Bayesian methods correctly identify that the underlying process is not CSR.

The left panel of Figure 9.7.49 captures the stationarity property of the Strauss process with  $K = 800$  and  $\hat{C}_1 = 0.15$ . As before, larger values of  $\hat{C}_1$  also lead to stationarity. The right panel of Figure 9.7.49 correctly indicates dependence among  $\mathbf{X}_{C_i}$ , for  $i = 1, \dots, 100$ , with  $\hat{C}_1 = 0.5$ .

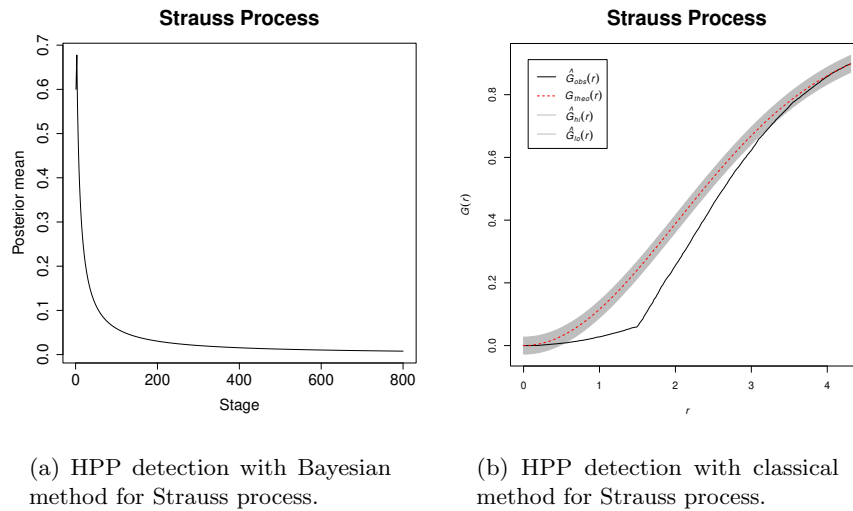
#### 9.7.17 Example 17: Another Strauss process

We now consider simulation from another homogeneous Strauss process with  $\beta = 100$ ,  $\gamma = 0.7$  and  $R = 0.05$  on  $W = [0, 8] \times [0, 8]$ . The 5168 points that we obtained, are plotted in Figure 9.7.50.

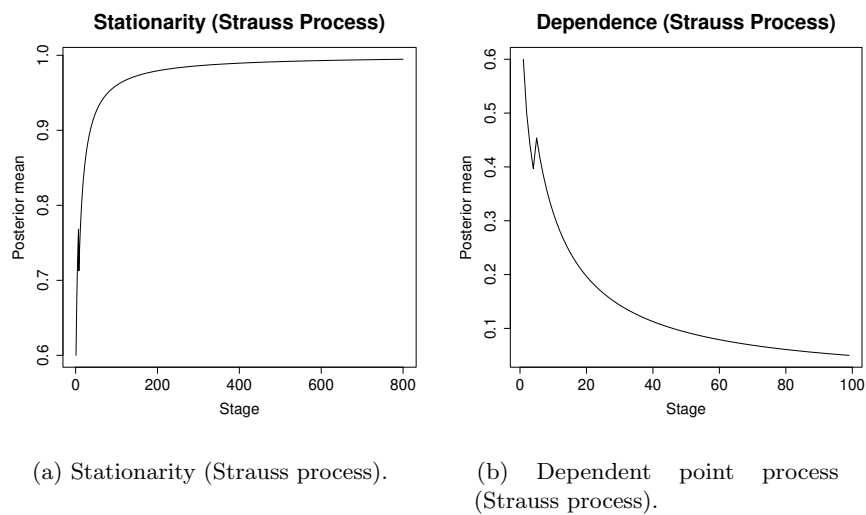
Again, both the Bayesian and classical method correctly detects non-CSR, as shown by Figure 9.7.51. For the Bayesian method, we set  $K = 500$  and  $\hat{C}_1 = 0.15$ .

Again, stationarity of the process is clearly indicated by panel (a) of Figure 9.7.52;





**Figure 9.7.48:** Detection of CSR with our Bayesian method and traditional classical method for Strauss process. Both the methods correctly identify that the underlying point process is not CSR.



**Figure 9.7.49:** Detection of stationarity and dependence of Strauss process with our Bayesian method.

## Strauss Process

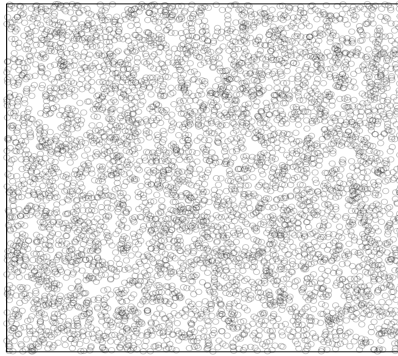


Figure 9.7.50: Strauss Process.

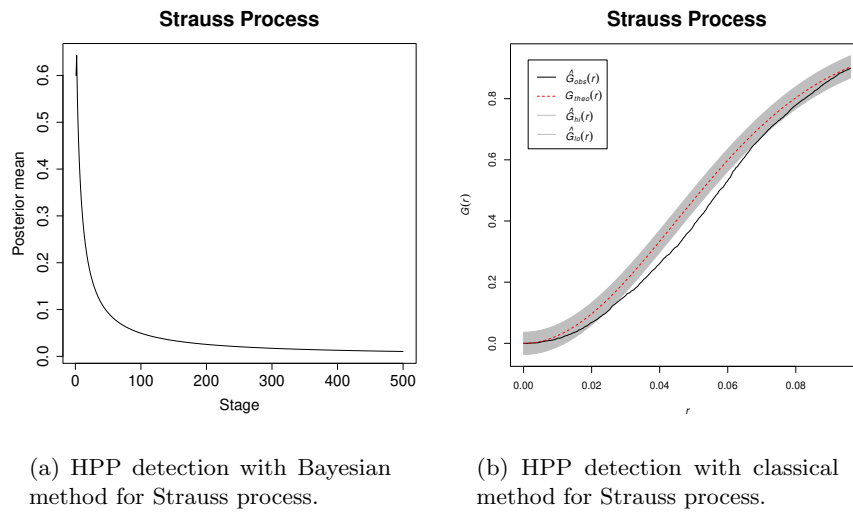
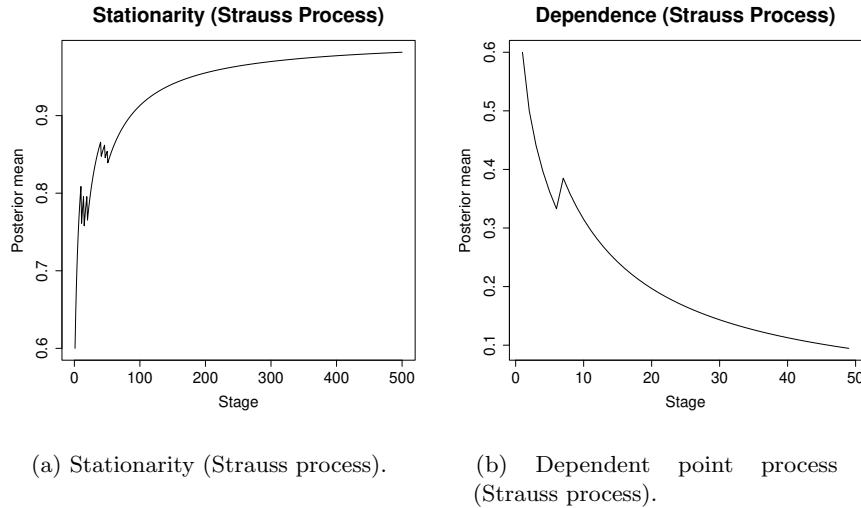


Figure 9.7.51: Detection of CSR with our Bayesian method and traditional classical method for Strauss process. Both the methods correctly identify that the underlying point process is not CSR.



**Figure 9.7.52:** Detection of stationarity and dependence of Strauss process with our Bayesian method.

here  $K = 500$  and  $\hat{C}_1 = 0.15$ . Panel (b) shows dependence with  $K = 50$  and  $\hat{C}_1 = 0.5$ .

## 9.8 Summary and conclusion

For tests for determination of CSR, stationarity, nonstationarity, Poisson or non-Poisson properties of spatial point processes have received almost no attention in the literature. In this chapter, we attempt to contribute to the developments using our principle of Bayesian characterization of stochastic processes detailed in Chapter 6. Using the principle, we characterized complete spatial randomness, using properties of Poisson point process. To characterize Poisson point process, we first characterized mutual independence among a set of random variables. Once we characterized such mutual independence, again using similar principles and the recursive Bayesian concept as before, we showed that how this facilitated characterization of Poisson point process. For mutual independence we made use of simple break-ups of joint distribution of random variables into products of conditional distributions and Bayesian nonparametrics based

on Dirichlet process. The latter particularly improved computational efficiency.

On applications of our Bayesian characterizations of point processes to a large variety of spatial point process examples with respect to simulation experiments, we obtained quite encouraging results that vindicate reliability and effectiveness of our ideas in general spatial point process setups. It is important to mention that extension of our ideas to spatio-temporal point processes is straightforward, and hence do not pursue this in this thesis work for the purpose of brevity. However, we reserve this for our future endeavor, to be communicated elsewhere.

# 10

## Bayesian Determination of Frequencies of Oscillatory Stochastic Processes

### 10.1 Introduction

In this chapter we assume that the underlying stochastic process has single or multiple frequencies of oscillations almost surely, including the possibility that the number of such frequencies is countably infinite. Using our basic principle of Bayesian characterization, we propose a novel Bayesian method of frequency determination, and establish its asymptotic theory. We back up our theory with ample simulation experiments and some real data analyses.

The motivation for this work is derived from periodic, noisy time series, where the goal is to determine the frequencies of oscillations. Indeed, time series with periodic or systematic sinusoidal variations are very common in the time series literature; see,

for example, [Shumway and Stoffer \(2006\)](#), [Montgomery \*et al.\* \(2016\)](#), [Hyndman and Athanasopoulos \(2018\)](#), [Chatfield and Xing \(2002\)](#). Among various such examples provided in [Shumway and Stoffer \(2006\)](#) are time series on computer recognition of speech, El Niño and fish population, functional magnetic resonance imaging (fMRI), economic time series, earthquakes and mining explosions. In all the examples, it is important to detect the frequencies of oscillations of the underlying noisy time series for further analyses and forecasts. In fact, as elucidated in [Shumway and Stoffer \(2006\)](#), the speech recognition example is a complex mixture of frequencies related to opening and closing of the glottis, the El Niño and fish population example is a mixture of two different kinds of frequencies, a seasonal periodic component and an El Niño component. Of fundamental interest in this example is the return period of El Niño as this can immensely influence local climate. Of related interest is whether the periodic components of the new fish population is dependent on the seasonal and El Niño periodicities. Determination of the periodic component of the economic time series is important from the seasonal perspective, and in the fMRI example, it is of importance to determine the periodic component related to the response of the brain to a periodic stimulus. In the earthquake and explosion example, determination of the different frequencies are important for discriminating between earthquakes and nuclear explosions.

Spectral analysis, or the frequency domain approach is the most suited for analysing periodic time series. This proceeds by expressing the underlying time series as Fourier frequencies composed of sines and cosines, with the periodogram analysis providing estimates of the unknown frequencies of oscillations. It is to be noted that there may be multiple frequencies hidden within a single oscillating time series and the frequency domain approach provides a way to estimate all such frequencies. For details, see, for example, [Hamilton \(1994\)](#), [Brockwell and Davis \(2002\)](#), [Shumway and Stoffer \(2006\)](#), [Brockwell and Davis \(2009\)](#).

In contrast with the frequency domain approach, which requires an appropriate model

for the underlying time series for estimating the unknown frequencies, our proposed Bayesian methodology does not require any model specification. Instead, for practical implementation, it requires a suitable transformation to the observed data that renders the oscillations more prominent. We supplement our theory and methods with ample simulation experiments and several real examples, many of which are also analyzed by [Shumway and Stoffer \(2006\)](#) using spectral analysis. Our Bayesian approach yielded encouraging results, and are very much comparable with those reported in [Shumway and Stoffer \(2006\)](#), whenever the comparison is relevant. An important advantage of our Bayesian approach compared to the spectral approach is that the former can readily produce the desired credible regions for the frequencies for any sample size, while the latter requires validation of asymptotic theory with normal approximation, even to approximately obtain the confidence intervals. It is important to remark that validation of asymptotic theory in the frequency domain setup is not straightforward, and requires assumptions that need not always be realistic.

The rest of this chapter is structured as follows. We bring forth our key idea on Bayesian frequency determination in Section 10.2. In Sections 10.3 and 10.4, we develop the Bayesian theory for finite and countably infinite number of frequencies, respectively. Details of our simulation experiments with single and multiple frequencies, as well as in the case of harmonics, are provided in Section 10.5. We apply our Bayesian approach to the real El Niño and fish population example in Section 10.6.

## 10.2 The key idea for Bayesian frequency determination

Let us assume that there are  $N$  ( $\geq 1$ ) frequencies of oscillations of the stochastic process  $\mathbf{X} = \{X_1, X_2, \dots\}$ . Here  $N$  may even be countably infinite. Consider the transformed process  $\mathbf{Z} = \{Z_1, Z_2, \dots\}$ , with  $Z_j = \frac{\exp(X_j)}{1 + \exp(X_j)}$ ;  $j \geq 1$ . Hence,  $Z_j \in [0, 1]$ . Now consider dividing up the interval  $[0, 1]$  into  $\cup_{m=1}^M [\tilde{p}_{m-1}, \tilde{p}_m]$ , for  $M > 1$ , such that  $\tilde{p}_0 = 0$ ,  $\tilde{p}_m = \tilde{p}_{m-1} + q_m$ , where  $\{q_m : m = 1, \dots, M\}$  is some probability distribution satisfying

$0 \leq q_m \leq 1$  for  $m = 1, \dots, M$ , and  $\sum_{m=0}^M q_m = 1$ . Here  $M$  can be even be infinite.

For oscillating stochastic process  $\mathbf{X}$ , for any  $r > 0$ ,  $\mathbf{Z}^r = \{Z_1^r, Z_2^r, \dots\}$  is also an oscillating stochastic process taking values in  $[0, 1]$ . Crucially, when raised to some sufficiently large positive power  $r$ , the originally smaller values of  $\mathbf{Z}$  tend to be much smaller compared to the originally larger values. These larger values of  $\mathbf{Z}^r$  will be contained in  $[\tilde{p}_{m-1}, \tilde{p}_m]$ , for large values of  $m$ . In particular, the largest values of  $\mathbf{Z}^r$  are expected to be contained in  $(\tilde{p}_{M-1}, 1]$ , or in  $[\tilde{p}_{m_0-1}, \tilde{p}_{m_0}]$  for  $1 \leq M_0 < m_0 < M$ . Here  $M_0$  is expected to be reasonably close to  $M$ . In the latter case, intervals of the form  $[\tilde{p}_{m-1}, \tilde{p}_m]$  will remain empty for  $m > m_0$ . The next largest values of  $\mathbf{Z}^r$  will be concentrated in  $[\tilde{p}_{m_1-1}, \tilde{p}_{m_1}]$  for some  $1 \leq M_1 < m_1 < m_0$ . In this case,  $[\tilde{p}_{m-1}, \tilde{p}_m]$  will remain empty for  $m_1 + 1 < m < m_0 - 1$ , and so on.

Note that the proportions of the values contained in the intervals constitute the frequencies of oscillations of the original process  $\mathbf{X}$ . We formalize this key idea into a Bayesian theory, treating  $M$  as finite as well as infinite.

### 10.3 Bayesian theory for finite $M$

To fix ideas, let us define

$$Y_j = m \quad \text{if} \quad \tilde{p}_{m-1} < Z_j^r \leq \tilde{p}_m; \quad m = 1, 2, \dots, M. \quad (10.3.1)$$

We assume that

$$(\mathbb{I}(Y_j = 1), \dots, \mathbb{I}(Y_j = M)) \sim \text{Multinomial}(1, p_{1,j}, \dots, p_{M,j}), \quad (10.3.2)$$

where  $p_{m,j}$  can be interpreted as the probability that  $Z_j^r \in (\tilde{p}_{m-1}, \tilde{p}_m]$ .

Now note that for large  $M$ , the intervals  $(\tilde{p}_{m-1}, \tilde{p}_m]$  correspond to small regions of the index set of the stochastic process  $\mathbf{X}$ , and hence, the part of the process  $\mathbf{Z}^r$



falling in  $(\tilde{p}_{m-1}, \tilde{p}_m]$  can be safely regarded as stationary. Further, assuming ergodicity of the process falling in the interval, it is expected that  $p_{m,j}$  will tend to the correct proportion of the process  $\mathbf{Z}^r$  falling in  $(\tilde{p}_{m-1}, \tilde{p}_m]$ , as  $j \rightarrow \infty$ . Notationally, we let  $\{p_{m,0}; m = 1, \dots, M\}$  denote the actual proportions of the process  $\mathbf{Z}^r$  falling in  $(\tilde{p}_{m-1}, \tilde{p}_m]$ ;  $m = 1, \dots, M$ .

Following the same principle discussed in Section 3.3, and extending the Beta prior to the Dirichlet prior, at the  $k$ -th stage we arrive at the following posterior of  $\{p_{m,k}; m = 1, \dots, M\}$ :

$$\pi(p_{1,k}, \dots, p_{M,k} | y_k) \equiv \text{Dirichlet} \left( \sum_{j=1}^k \frac{1}{j^2} + \sum_{j=1}^k \mathbb{I}(y_j = 1), \dots, \sum_{j=1}^k \frac{1}{j^2} + \sum_{j=1}^k \mathbb{I}(y_j = M) \right). \quad (10.3.3)$$

The posterior mean and posterior variance of  $p_{m,k}$ , for  $m = 1, \dots, M$ , are given by:

$$E(p_{m,k} | y_k) = \frac{\sum_{j=1}^k \frac{1}{j^2} + \sum_{j=1}^k \mathbb{I}(y_j = m)}{M \sum_{j=1}^k \frac{1}{j^2} + k}; \quad (10.3.4)$$

$$\text{Var}(p_{m,k} | y_k) = \frac{\left( \sum_{j=1}^k \frac{1}{j^2} + \sum_{j=1}^k \mathbb{I}(y_j = m) \right) \left( (M-1) \sum_{j=1}^k \frac{1}{j^2} + k - \sum_{j=1}^k \mathbb{I}(y_j = m) \right)}{\left( M \sum_{j=1}^k \frac{1}{j^2} + k \right)^2 \left( M \sum_{j=1}^k \frac{1}{j^2} + k + 1 \right)}. \quad (10.3.5)$$

Since the process  $\mathbf{Z}^r$  falling in  $(\tilde{p}_{m-1}, \tilde{p}_m]$  is stationary and ergodic, it follows from (10.3.4) and (10.3.5) it is easily seen, using  $\frac{\sum_{j=1}^k \mathbb{I}(y_j = m)}{k} \rightarrow p_{m,0}$ , almost surely, as  $k \rightarrow \infty$ , that, almost surely,

$$E(p_{m,k} | y_k) \rightarrow p_{m,0}, \quad \text{and} \quad (10.3.6)$$

$$\text{Var}(p_{m,k} | y_k) = O\left(\frac{1}{k}\right) \rightarrow 0, \quad (10.3.7)$$

as  $k \rightarrow \infty$ .

Theorem 52 formalizes the above arguments in terms of the limits of the marginal posterior probabilities of  $p_{m,k}$ , denoted by  $\pi_m(\cdot|y_k)$ , as  $k \rightarrow \infty$ .

**Theorem 52** *Assume that  $M$  is so large that  $\mathbf{Z}^r$  falling in the intervals  $(\tilde{p}_{m-1}, \tilde{p}_m]$ ;  $m = 1, \dots, M$ , constitute stationary processes, and that such stationary processes are also ergodic.*

*Let  $\mathcal{N}_{p_{m,0}}$  be any neighborhood of  $p_{m,0}$ , with  $p_{m,0}$  satisfying  $0 < p_{m,0} < 1$  for  $m = 1, \dots, M$  such that  $\sum_{m=1}^M p_{m,0} = 1$ . Then*

$$\pi_m(\mathcal{N}_{p_{m,0}}|y_k) \rightarrow 1, \quad (10.3.8)$$

*almost surely as  $k \rightarrow \infty$ .*

**Proof.** For any neighborhood of  $p_{m,0}$ , denoted by  $\mathcal{N}_{p_{m,0}}$ , let  $\epsilon > 0$  be sufficiently small so that  $\mathcal{N}_{p_{m,0}} \supseteq \{|p_{m,k} - p_{m,0}| < \epsilon\}$ . Then by Chebychev's inequality, using (10.3.6) and (10.3.7), it is seen that  $\pi_m(\mathcal{N}_{p_{m,0}}|y_k) \rightarrow 1$ , almost surely, as  $k \rightarrow \infty$ . ■

**Corollary 53** *For adequate choices of  $r$  and  $M$ , the non-zero distinct elements of  $\{p_{m,0}; m = 2, \dots, M\}$  are the desired frequencies of the oscillating stochastic process  $\mathbf{X}$ . Note that for adequately large  $M$ ,  $p_{1,0}$  is associated with the small values of  $Z^r$ , and hence does not correspond to any frequency of the original stochastic process.*

### 10.3.1 Choice of $r$ , $M$ and $\{q_1, \dots, q_M\}$

In principle, the probability distribution  $\{q_1, \dots, q_M\}$  should be chosen based on prior information regarding which intervals contain the desired frequencies. Given sufficiently large  $M$ , the values of  $q_m$  can then be chosen to shorten or widen any given interval. Short intervals are preferable when there is strong prior information of some frequency falling in the vicinity of some point. On the other hand, larger intervals are appropriate in the case of weak prior information. Such prior knowledge may be obtained, say, by periodogram analysis of the underlying time series.

However, in our experiments, the uniform distribution  $q_m = 1/M$ , for  $m = 1, \dots, M$ , yielded excellent results. For the choice of  $r$ , we recommend that value for which the oscillations of  $\mathbf{Z}^r$  as distinctly visible as possible. The choice of  $M$  should be such that  $\{(\tilde{p}_{m-1}, \tilde{p}_m]; m = 1, \dots, M\}$  covers the range of  $\mathbf{Z}^r$  with adequately fine intervals. We discuss these issues in details with simulation studies and real data examples.

## 10.4 Bayesian theory for infinite number of frequencies

We now assume that the number of frequencies,  $m$ , is countably infinite, and that  $\{p_{m,0}; m = 1, 2, 3, \dots\}$ , where  $0 \leq p_{m,0} \leq 1$  and  $\sum_{m=1}^{\infty} p_{m,0} = 1$ , are the true proportions of the process  $\mathbf{Z}^r$  falling in the intervals  $(\tilde{p}_{m-1}, \tilde{p}_m]; m = 1, 2, \dots$

Now we define

$$Y_j = m \text{ if } \tilde{p}_{m-1} < Z_j^r \leq \tilde{p}_m; m = 1, 2, \dots, \infty. \quad (10.4.1)$$

Let  $\mathcal{X} = \{1, 2, \dots\}$  and let  $\mathcal{B}(\mathcal{X})$  denote the Borel  $\sigma$ -field on  $\mathcal{X}$  (assuming every singleton of  $\mathcal{X}$  is an open set). Let  $\mathcal{P}$  denote the set of probability measures on  $\mathcal{X}$ . Then, at the  $j$ -th stage,

$$[Y_j | P_j] \sim P_j, \quad (10.4.2)$$

where  $P_j \in \mathcal{P}$ . We assume that  $P_j$  is the following Dirichlet process (see [Ferguson \(1973\)](#)):

$$P_j \sim DP\left(\frac{1}{j^2}G\right), \quad (10.4.3)$$

where, the probability measure  $G$  is such that, for every  $j \geq 1$ ,

$$G(Y_j = m) = \frac{1}{2^m}. \quad (10.4.4)$$

It then follows using the same previous principles that, at the  $k$ -th stage, the posterior

of  $P_k$  is again a Dirichlet process, given by

$$[P_k|y_k] \sim DP \left( \sum_{j=1}^k \frac{1}{j^2} G + \sum_{j=1}^k \delta_{y_j} \right), \tag{10.4.5}$$

where  $\delta_{y_j}$  denotes point mass at  $y_j$ . It follows from (10.4.5) that

$$E(p_{m,k}|y_k) = \frac{\frac{1}{2^m} \sum_{j=1}^k \frac{1}{j^2} + \sum_{j=1}^k \mathbb{I}(y_j = m)}{\sum_{j=1}^k \frac{1}{j^2} + k}; \tag{10.4.6}$$

$$Var(p_{m,k}|y_k) = \frac{\left( \sum_{j=1}^k \frac{1}{j^2} + \sum_{j=1}^k \mathbb{I}(y_j = m) \right) \left( \left(1 - \frac{1}{2^m}\right) \sum_{j=1}^k \frac{1}{j^2} + k - \sum_{j=1}^k \mathbb{I}(y_j = m) \right)}{\left( \sum_{j=1}^k \frac{1}{j^2} + k \right)^2 \left( \sum_{j=1}^k \frac{1}{j^2} + k + 1 \right)}. \tag{10.4.7}$$

As before, it easily follows from (10.4.6) and (10.4.7) that for  $m = 1, 2, 3, \dots$ ,

$$E(p_{m,k}|y_k) \rightarrow p_{m,0}, \quad \text{and} \tag{10.4.8}$$

$$Var(p_{m,k}|y_k) = O\left(\frac{1}{k}\right) \rightarrow 0, \tag{10.4.9}$$

almost surely, as  $k \rightarrow \infty$ .

The theorem below formalizes the above arguments in the infinite number of frequency situation in terms of the limit of the marginal posterior probabilities of  $p_{m,k}$ , as  $k \rightarrow \infty$ .

**Theorem 54** *Assume that  $Z^r$  falling in the intervals  $(\tilde{p}_{m-1}, \tilde{p}_m]$ ;  $m = 1, 2, \dots$ , constitute stationary processes, and that such stationary processes are also ergodic.*

*Let  $\mathcal{N}_{p_{m,0}}$  be any neighborhood of  $p_{m,0}$ , with  $p_{m,0}$  satisfying  $0 \leq p_{m,0} \leq 1$  for  $m = 1, 2, \dots$  such that  $\sum_{m=1}^{\infty} p_{m,0} = 1$ , with at most finite number of  $m$  such that  $p_{m,0} = 0$ . Then with  $Y_j$  defined as in (10.4.1),*

$$\pi_m(\mathcal{N}_{p_{m,0}}|y_k) \rightarrow 1, \tag{10.4.10}$$

almost surely, as  $k \rightarrow \infty$ .

**Proof.** Follows using the same ideas as the proof of Theorem 12. ■

**Corollary 55** *The non-zero distinct elements of  $\{p_{m,0}; m = 1, 2, \dots\}$  are the desired frequencies of the oscillating stochastic process  $\mathbf{X}$ . Again,  $p_{1,0}$  does not correspond to any frequency of the original stochastic process.*

**Remark 56** *As regards the choice of the quantities  $q_m$ , we suggest setting  $q_m = 2^{-m}$ , for  $m \geq 1$ , which is the same as the base measure for the Dirichlet process prior. For countably infinite number of frequencies, the choice of  $r$  is difficult to decide. But we hope that selecting  $r$  such that most of the oscillations are visible as much as possible, will work even in this situation.*

**Remark 57** *It is useful to remark that our theory with countably infinite number of frequencies is readily applicable to situations where the number of frequencies is finite but unknown. In such cases, only a finite number of the probabilities  $\{p_{m,j}; m = 2, 3, \dots\}$  will have posterior probabilities around positive quantities, while the rest will concentrate around zero. For known finite number of limit points, it is only required to specify  $G$  such that it gives positive mass to only a specific finite set.*

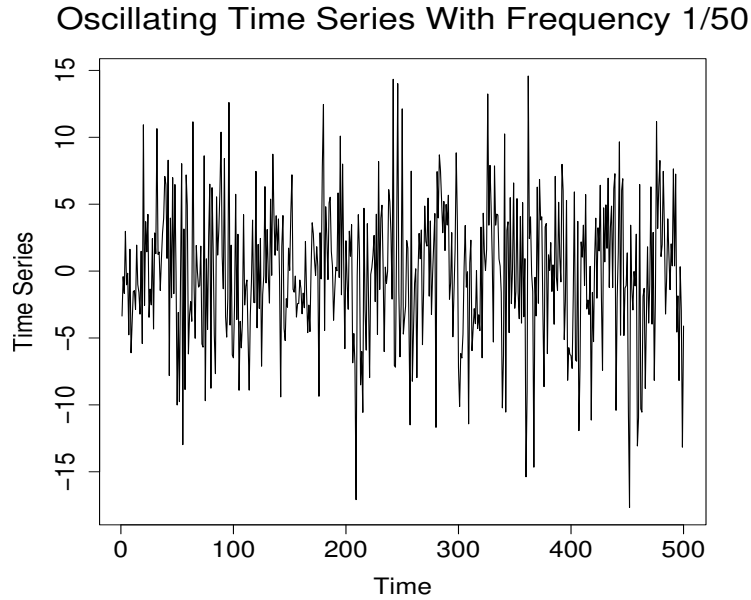
We now illustrate our Bayesian theory for detecting frequencies using simulation studies.

## 10.5 Simulation experiments

### 10.5.1 Simulation study with a single frequency

Following Example 2.8 of [Shumway and Stoffer \(2006\)](#), we generate  $T = 500$  observations from the model

$$x_t = A \cos(2\pi\omega t + \varphi) + \epsilon_t, \quad (10.5.1)$$

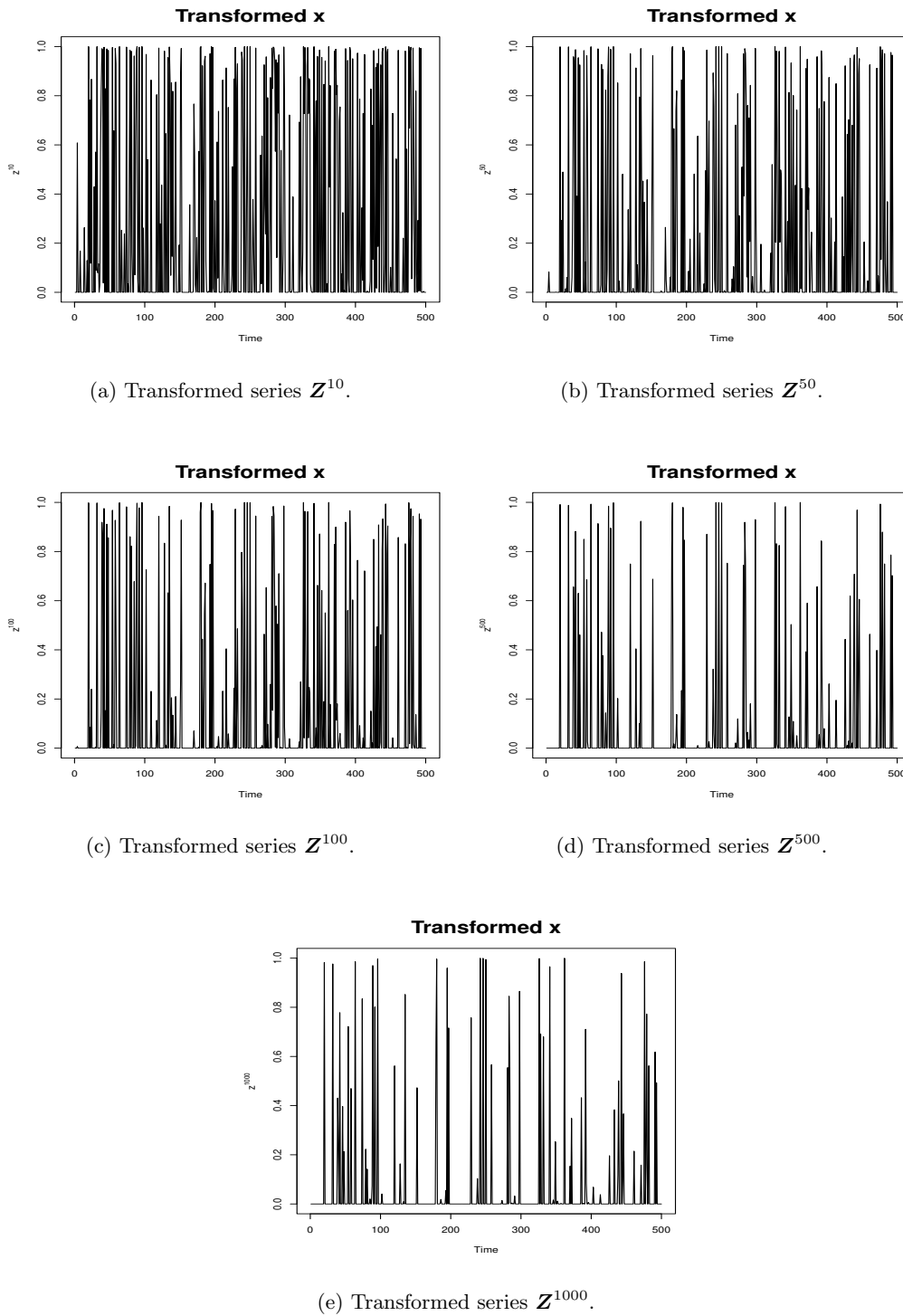


**Figure 10.5.1:** Simulated oscillating time series with true frequency 0.02.

where  $\omega = 1/50$ ,  $A = 2$ ,  $\varphi = 0.6\pi$ , and  $\epsilon_t \stackrel{iid}{\sim} N(0, \sigma^2)$ , with  $\sigma = 5$ . Figure 10.5.1 displays the generated time series. Observe that due to the relatively large  $\sigma$ , the true frequency is blurred in the observed time series. Our goal is to recover the frequency  $\omega = 1/50$  using our Bayesian method, pretending that the true frequency is unknown.

We apply our Bayesian technique based on Dirichlet process, but with the base measure  $G_0$  giving probability  $1/M$  to each of the values  $1, \dots, M$ . Since our method depends crucially on the choices of  $r$  and  $M$ , it is important to carefully choose these quantities. As we had already prescribed,  $r$  should be so chosen that the oscillations of  $\mathbf{Z}^r$  are easy to visualize. Figure 10.5.2 shows the transformed time series  $\mathbf{Z}^r$  for different values of  $r$ . In this example we see that as  $r$  is increased, the oscillations tend to be more and more explicit. Thus, it seems that  $r = 1000$  is the best choice among those experimented with.

For the choice of  $M$  we need to select a large enough value such that the range of  $\mathbf{Z}^r$  gets adequately partitioned within  $\{(\tilde{p}_{m-1}, \tilde{p}_m]; m = 1, \dots, M\}$ . In other words,



**Figure 10.5.2:** Illustration of effects of  $r$  in  $Z^r$  in determining single frequency in (10.5.1). Here the true frequency is 0.02.

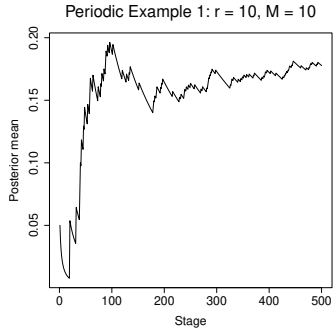
relatively large values of  $r$  and  $M$  are expected to yield good Bayesian results. We investigate this by implementing our Bayesian method for different values of  $r$  and  $M$  and comparing the results.

Figures 10.5.3 and 10.5.4 depict the results of our Bayesian method for various choices of  $r$  and  $M$ . As shown by the figures, for increasing values of  $r = 10, 50, 100, 500, 1000$ , and  $M = 10, 50, 100$ , the posterior of  $p_{M,j}$  associated with the interval  $(\tilde{p}_{M-1}, \tilde{p}_M]$ , increasingly converges to the true frequency 0.02. Note that for relatively small values of either  $r$  or  $M$ , the relevant posteriors fail to converge. Thus, the results are in keeping with our expectation of obtaining superior results for large values of  $r$  and  $M$ . Note that the rate of convergence of the posterior seems to be faster with respect to increasing values of  $r$  compared to increasing values of  $M$ . Thus, appropriate choice of  $r$  seems to be more important than  $M$ .

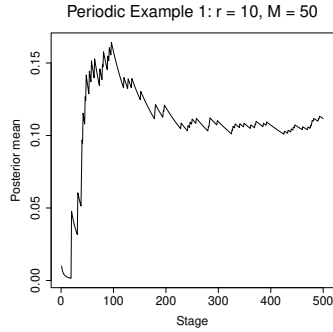
Following [Shumway and Stoffer \(2006\)](#) we have generated only 500 observations from (10.5.1) for inference, due to reasons of comparability with the results obtained by [Shumway and Stoffer \(2006\)](#). If large enough datasets are not available in reality, our Bayesian inference needs to be as accurate as possible based on the available data, and our analyses indeed provide glimpses of such reliable Bayesian inference. But in the current “big data” era large datasets are making their appearances, and it is important to weigh our inference with respect to large datasets, which also provide opportunities to properly validate our convergence theory, which is usually not viable for small datasets.

We thus generate a dataset from (10.5.1) with  $T = 5 \times 10^5$ , and apply our Bayesian procedure with  $r = 1000$  and  $M = 10, 50, 100$ , in order to detect the true frequency 0.02. The results are displayed in Figure 10.5.5. Observe that for  $M = 10$ , the true frequency is overestimated, as shown in panel (a) associated with convergence of  $p_{10,j}$  as  $j \rightarrow \infty$ , and for  $M = 100$ , underestimation occurs, as captured by panel (c) associated with convergence of  $p_{100,j}$  as  $j \rightarrow \infty$ . Panel (b) shows convergence of  $p_{50,j}$  as  $j \rightarrow \infty$ , where convergence occurs around 0.019, quite close to the truth. Panel (d) displays the result

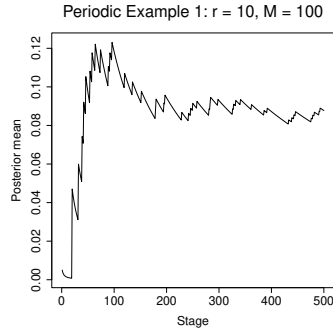




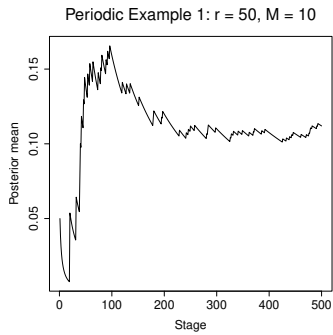
(a)  $r = 10, M = 10$ .



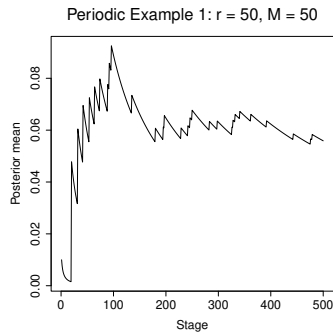
(b)  $r = 10, M = 50$ .



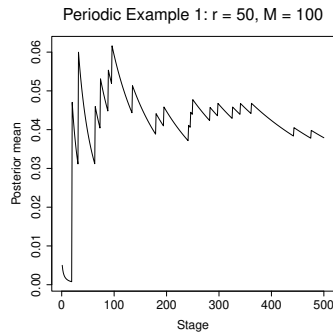
(c)  $r = 10, M = 100$ .



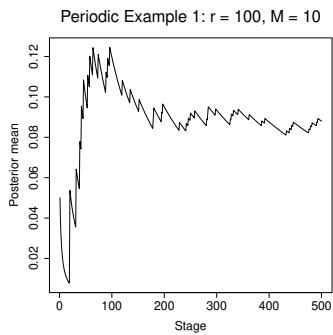
(d)  $r = 50, M = 10$ .



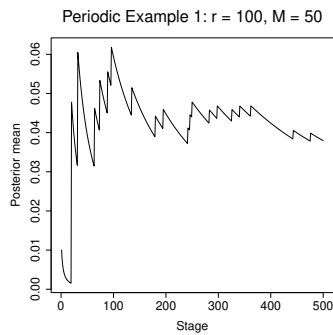
(e)  $r = 50, M = 50$ .



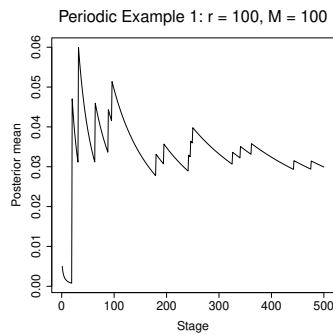
(f)  $r = 50, M = 100$ .



(g)  $r = 100, M = 10$ .

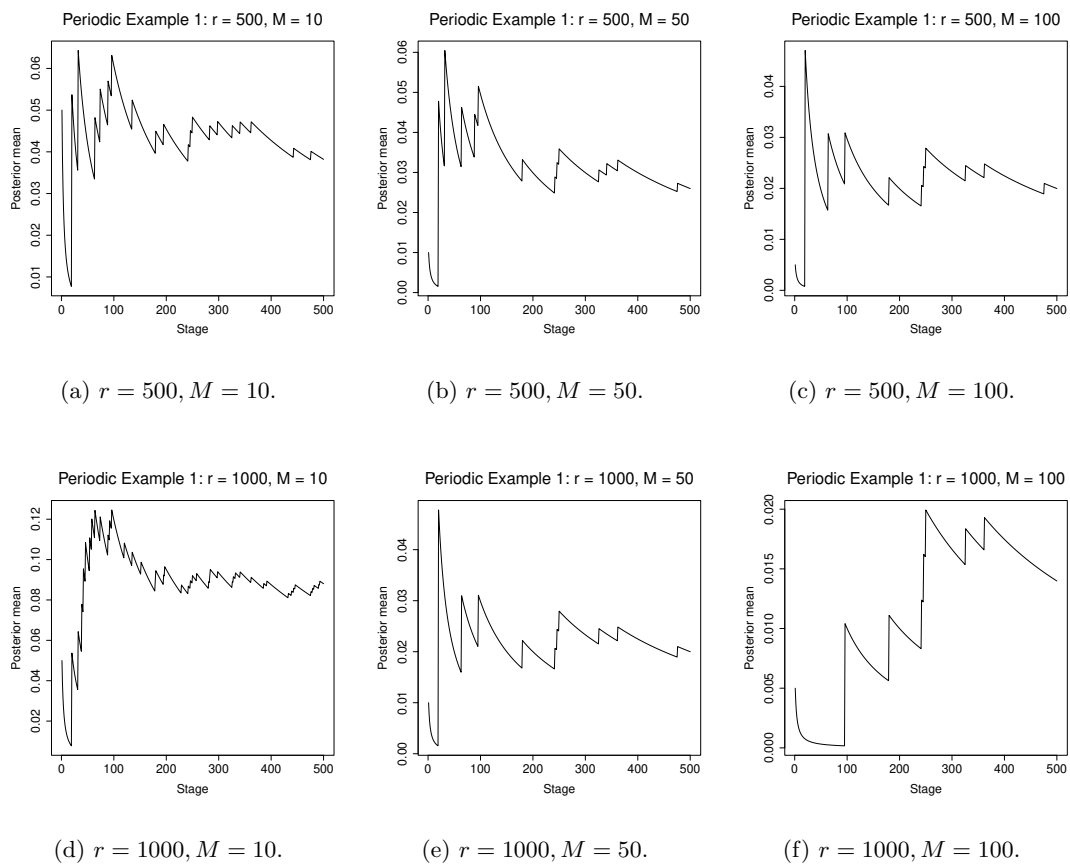


(h)  $r = 100, M = 50$ .



(i)  $r = 100, M = 100$ .

**Figure 10.5.3:** Illustration of our Bayesian method for determining single frequency. Here the true frequency is 0.02.



**Figure 10.5.4:** Illustration of our Bayesian method for determining single frequency. Here the true frequency is 0.02.

of convergence of  $p_{100,j} + p_{99,j}$ , as  $j \rightarrow \infty$ . This sum converges around 0.019. The reason for over and under estimation for  $M = 10$  and  $100$  can be attributed to too coarse and too fine partitions of  $[0, 1]$  via the choice of  $M$ , while for  $M = 50$ , the partitioning seems more reasonable in comparison. Adding up  $p_{100,j}$  and  $p_{99,j}$  compensates for the too fine partitioning of  $[0, 1]$  in this case.

The effects of partitioning also points towards another issue – even  $p_{50,j}$  and  $p_{100,j} + p_{99,j}$  fail to capture the true frequency as  $j \rightarrow \infty$ , since the posterior variance becomes negligibly small as  $j \rightarrow \infty$ . In principle, it is possible to partition  $[0, 1]$  appropriately (perhaps, using good choices of  $q_m$ ), such that convergence to the exact true frequency is achieved. In this example, setting  $M = 40$  is enough, as depicted in Figure 10.5.6. Note that such subtle issues can not be detected or analyzed for sample size as small as 500. Nevertheless, our final Bayesian results do convey very reliable analysis even for such small dataset.

### 10.5.2 Simulation study with multiple frequencies

As in Example 4.1 of [Shumway and Stoffer \(2006\)](#), for  $t = 1, \dots, 100$ , first we generate the following three series:

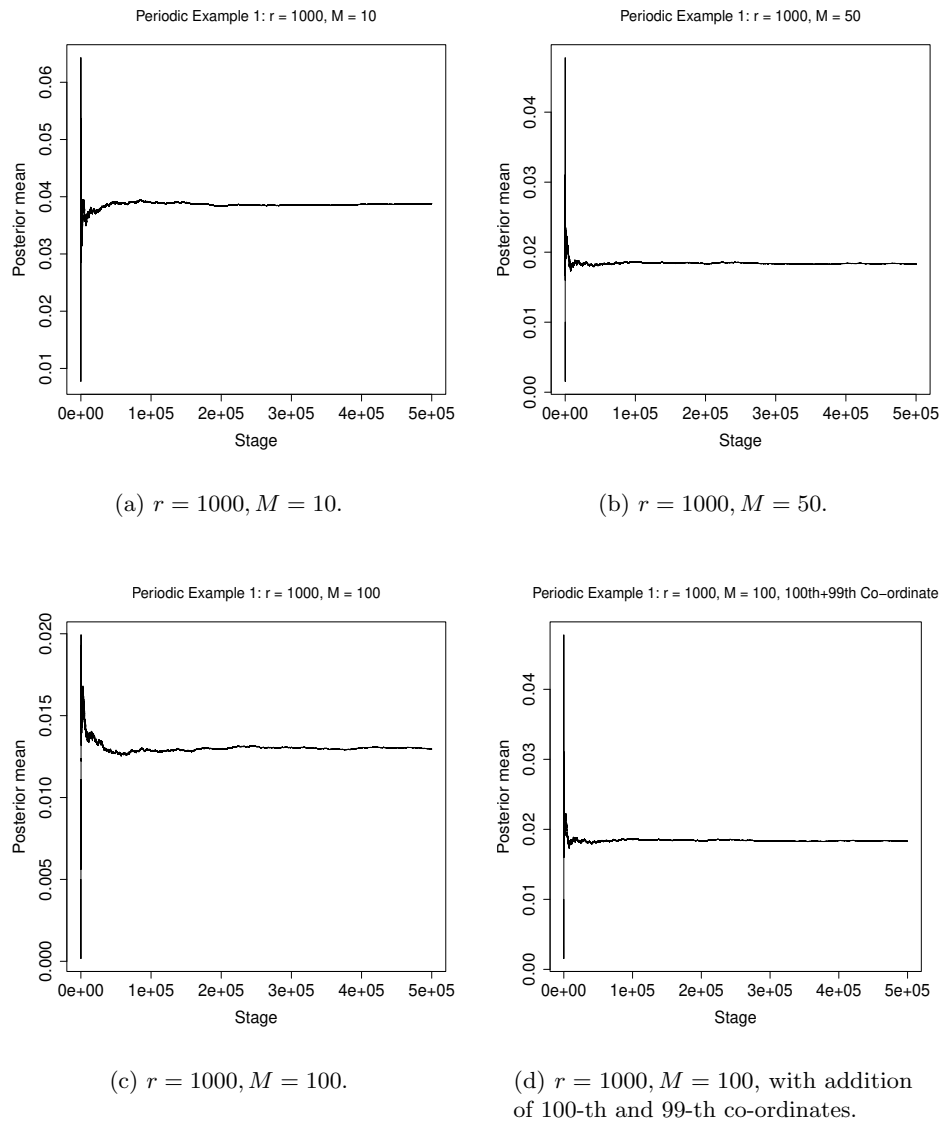
$$\begin{aligned}x_{t_1} &= 2 \cos(2\pi t 6/100) + 3 \sin(2\pi t 6/100); \\x_{t_2} &= 4 \cos(2\pi t 10/100) + 5 \sin(2\pi t 10/100); \\x_{t_3} &= 6 \cos(2\pi t 40/100) + 7 \sin(2\pi t 40/100),\end{aligned}$$

and set

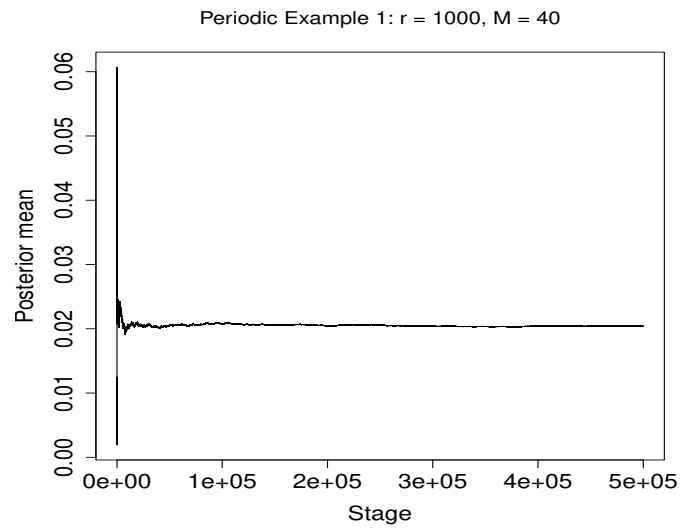
$$x_t = x_{t_1} + x_{t_2} + x_{t_3}. \tag{10.5.2}$$

The series  $x_t$ , which consists of the three frequencies 0.4, 0.1 and 0.06, is shown in Figure 10.5.7.

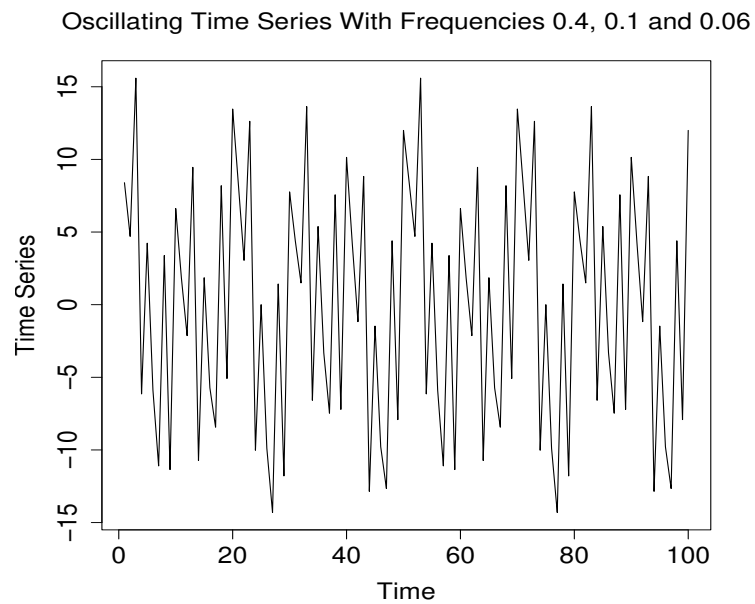
Before applying our Bayesian method based on Dirichlet process to this example,



**Figure 10.5.5:** Illustration of our Bayesian method for determining single frequency for long enough time series. Here the true frequency is 0.02.



**Figure 10.5.6:** Convergence of our Bayesian method to the true frequency 0.02 for long enough time series with  $r = 1000$  and  $M = 40$ .



**Figure 10.5.7:** Simulated oscillating time series with true frequencies 0.4, 0.1 and 0.06.

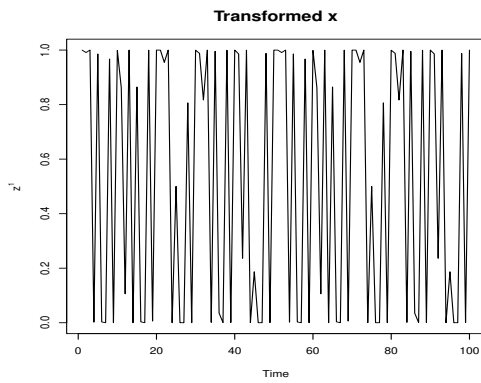
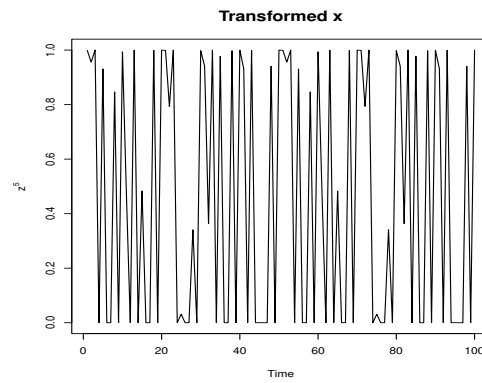
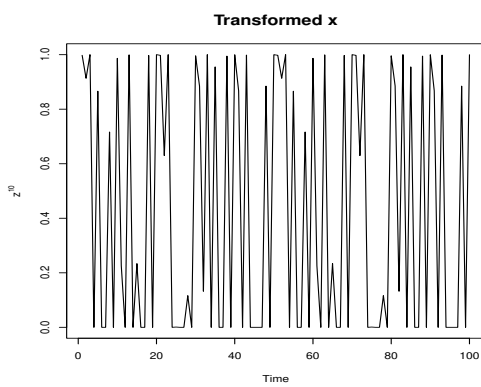
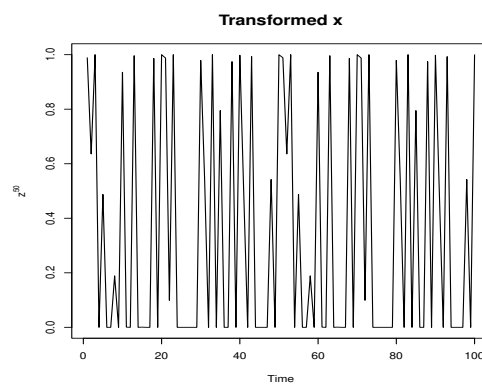
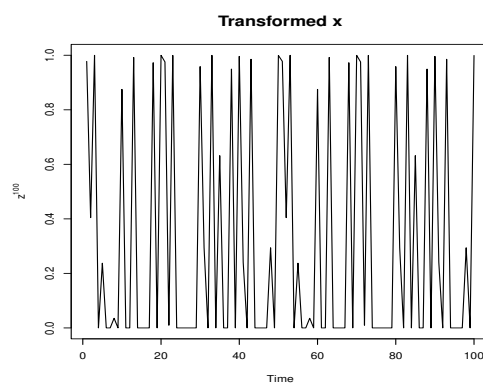
we again need to choose  $r$  and  $M$  properly. Regarding the choice of  $r$ , Figure 10.5.8 depicts the process  $\mathbf{Z}^r$  for  $r = 1, 5, 10, 50, 100$ . Here although it seems at first glance that increasing  $r$  leads to increasing isolation of the oscillations, actually, it is evident from closer look that increasing the power here has the effect of reducing the peaks of many relevant oscillations quite close to the highest peaks that are present in panel (a) of the figure, corresponding to  $r = 1$ . Thus, in this example, large values of  $r$  are inappropriate, unlike in the first example on single frequency. Here  $r = 1$  seems more appropriate compared to the other values of  $r$ .

Regarding adequacy of the choice of  $r$  and  $M$ , a detailed analysis of our Bayesian results for this multiple frequency example is provided by Figures 10.5.9, 10.5.10, 10.5.11, 10.5.12 and 10.5.13. Most of these diagrams, for given  $r$  and  $M$ , are obtained by summing up the  $p_{m,j}$  for nearby values of  $m$ . These yielded the three frequencies associated with our Bayesian technique. The values of  $m$  that are summed up, are provided on the top of each panel. Indeed, for relatively larger values of  $M$ , the frequencies are divided up into several nearby intervals  $(\tilde{p}_{m-1}, \tilde{p}_m]$ .

Recall that we do not consider the first interval  $(\tilde{p}_0, \tilde{p}_1]$  at all as it is a small interval around zero for relatively large  $M$  and hence not associated with any true frequency significantly different from zero. The proportions of the intervals that converged to zero, are not considered either.

Figures 10.5.9, 10.5.10 and 10.5.11 depict the details of our results for  $r = 1, 5, 10$  and  $M = 10, 50, 100$ . Observe that  $r = 1$  gives the best performance, while the performance deteriorates for  $r = 5$  is also close. But observe that for  $r = 5, M = 10$ , the frequency 0.06 seems to be somewhat underestimated. However importantly, for  $r = 10$ , while the frequencies 0.4 and 0.1 are correctly converged to for these values of  $r$ , the frequency 0.06 seems to be significantly underestimated, for  $M = 10, 50, 100$ .

As seen in Figures 10.5.12 and 10.5.13, for  $r = 50$  and 100, although the frequency 0.06 is underestimated in some cases, the most conspicuous is the case of underestimation

(a) Transformed series  $Z^1$ .(b) Transformed series  $Z^5$ .(c) Transformed series  $Z^{10}$ .(d) Transformed series  $Z^{50}$ .(e) Transformed series  $Z^{100}$ .

**Figure 10.5.8:** Illustration of effects of  $r$  in  $Z^r$  in determining multiple frequencies in (10.5.2). Here the true frequencies are 0.4, 0.1 and 0.06.

of the highest frequency 0.4. This is due to the fact that for relatively large values of  $r$ , about half of the peaks of the original process close to the highest peaks, die down. Since half of these peaks close to the highest peaks contribute half of the total frequency 0.4 (obvious from direct counting of the highest and second highest peaks in Figure 10.5.7, this results in significant underestimation of the highest frequency.

Hence, consistent from the insight gained from Figure 10.5.8,  $r = 1$  yields the best performance. The choice of  $M$  seems to be less important compared to that of  $r$ , as in the previous example with single frequency.

### 10.5.3 Harmonics

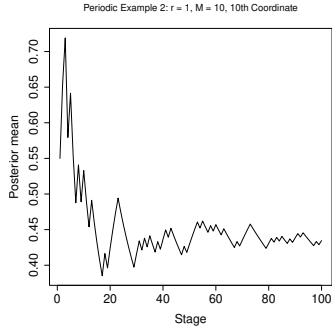
Since in reality most signals are not sinusoidal, it is preferable to use harmonics to model such signals. In this respect, we consider Example 4.12 of [Shumway and Stoffer \(2006\)](#) where a signal is constructed using a sinusoid oscillating at two cycles per unit time, and 5 harmonics obtained from the sinusoid oscillating at decreasing amplitudes. Specifically, their signal is given by

$$x_t = \sin(2\pi 2t) + 0.5 \sin(2\pi 4t) + 0.4 \sin(2\pi 6t) + 0.3 \sin(2\pi 8t) + 0.2 \sin(2\pi 10t) + 0.1 \sin(2\pi 12t), \quad (10.5.3)$$

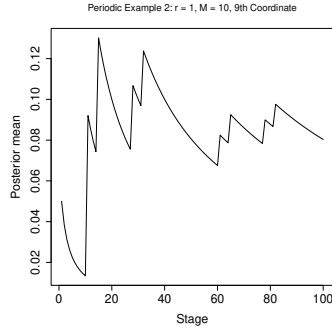
for  $0 \leq t \leq 1$ . The original signal  $\mathbf{X}$  and the transformation  $\mathbf{Z}^2$  are displayed in Figure 10.5.14, after considering 201 equidistant points in the time interval  $[0, 1]$ . Note that the original signal is not even close to sinusoidal. For the transformation  $\mathbf{Z}^r$ , we chose  $r = 2$  such that the structure of  $\mathbf{X}$  is essentially retained, but the gaps between the oscillations are increased to facilitate detection of the frequencies.

Since  $\mathbf{Z}^2$  suggests multiple frequencies that are likely to be close to each other, we chose  $M = 150$  to divide  $[0, 1]$  into larger number of finer sub-intervals compared to the previous synthetic examples to properly detect the oscillations. Application of our Bayesian procedure revealed 6 distinct values out of  $M = 150$  at the end of the

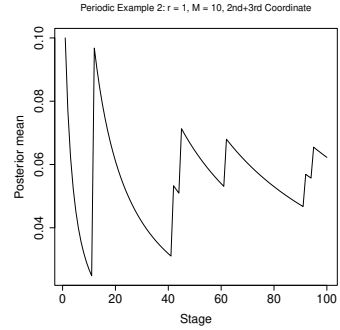




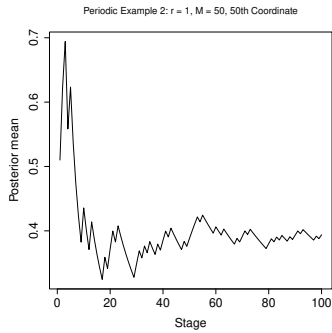
(a)  $r = 1, M = 10$ . True frequency = 0.4.



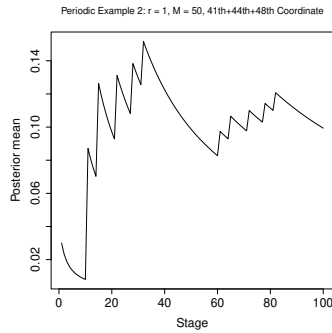
(b)  $r = 1, M = 10$ . True frequency = 0.1.



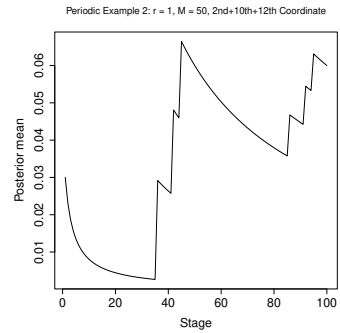
(c)  $r = 1, M = 10$ . True frequency = 0.06.



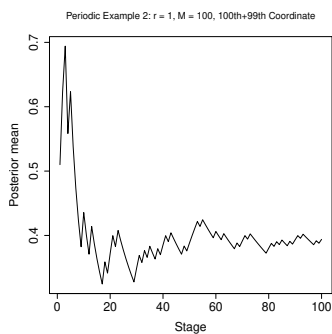
(d)  $r = 1, M = 50$ . True frequency = 0.4.



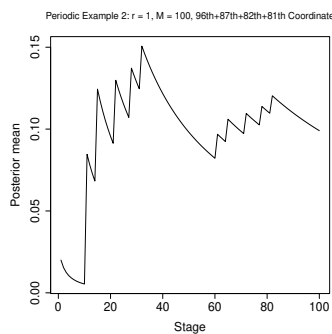
(e)  $r = 1, M = 50$ . True frequency = 0.1.



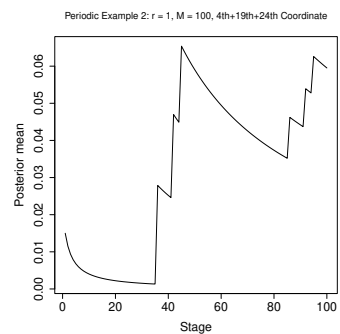
(f)  $r = 1, M = 50$ . True frequency = 0.06.



(g)  $r = 1, M = 100$ . True frequency = 0.4.

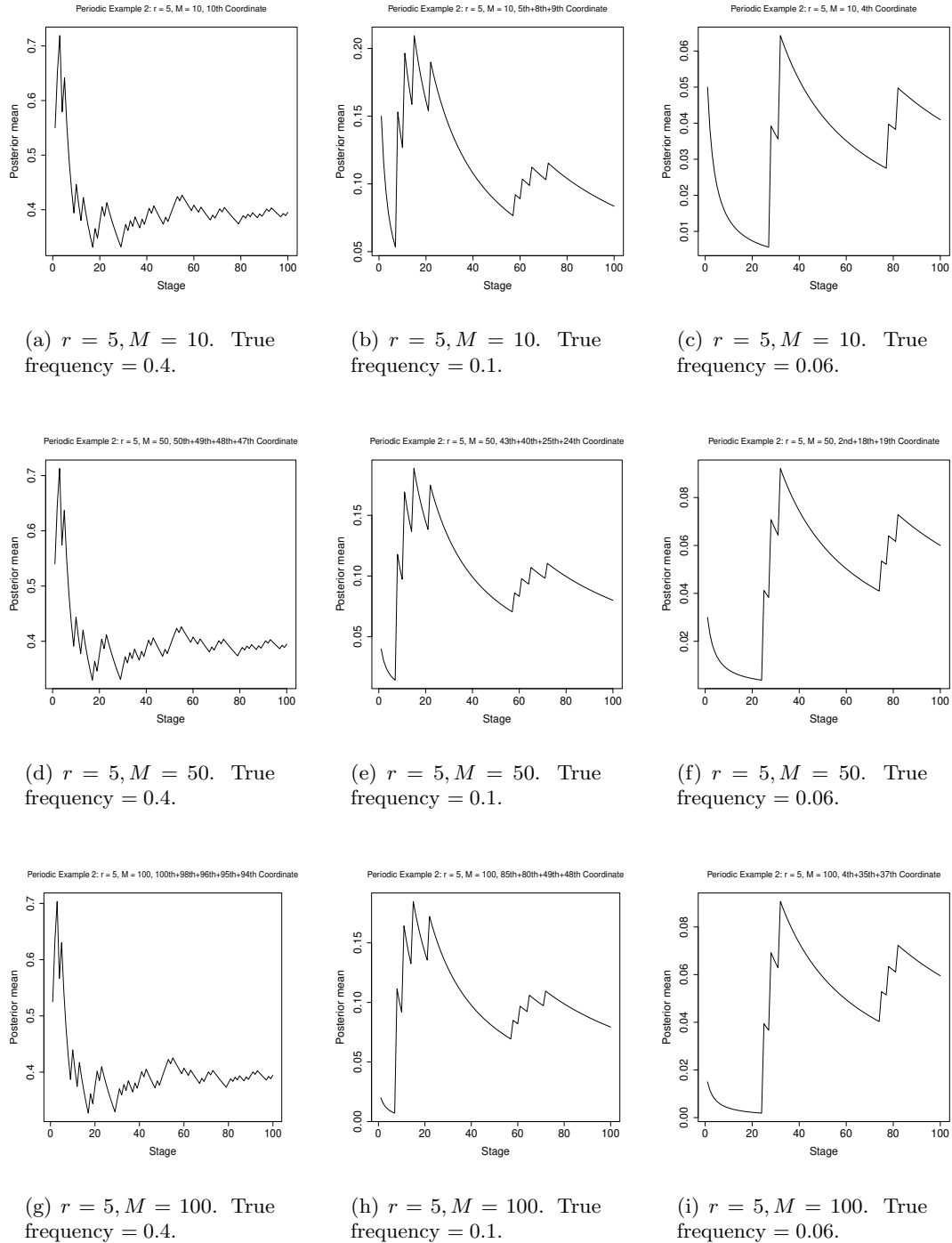


(h)  $r = 1, M = 100$ . True frequency = 0.1.

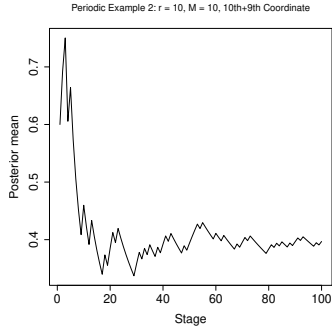


(i)  $r = 1, M = 100$ . True frequency = 0.06.

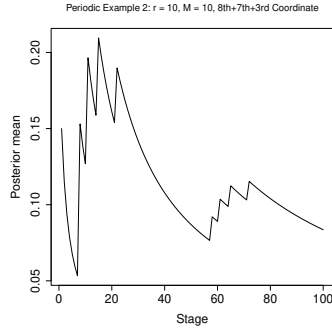
**Figure 10.5.9:** Illustration of our Bayesian method for determining multiple frequencies. Here the true frequencies are 0.4, 0.1 and 0.06.



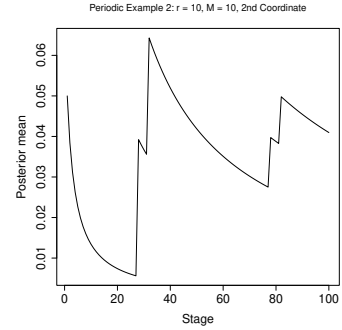
**Figure 10.5.10:** Illustration of our Bayesian method for determining multiple frequencies. Here the true frequencies are 0.4, 0.1 and 0.06.



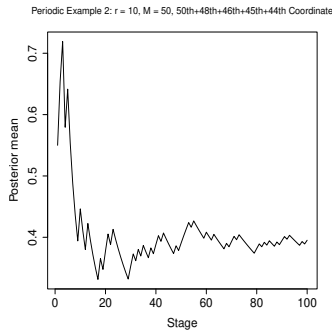
(a)  $r = 10, M = 10$ . True frequency = 0.4.



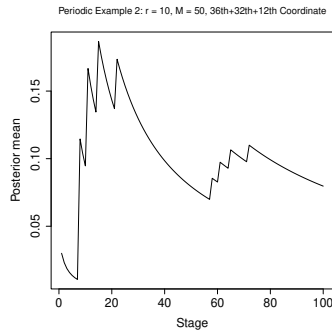
(b)  $r = 10, M = 10$ . True frequency = 0.1.



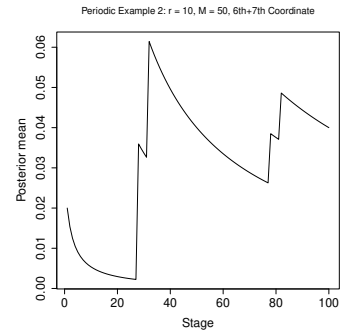
(c)  $r = 10, M = 10$ . True frequency = 0.06.



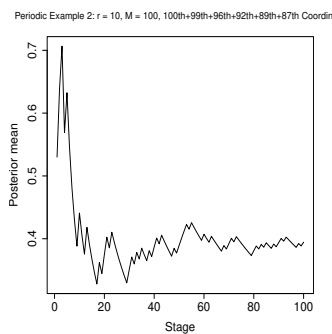
(d)  $r = 10, M = 50$ . True frequency = 0.4.



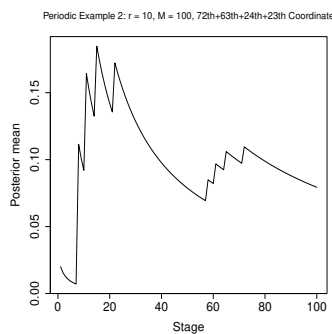
(e)  $r = 10, M = 50$ . True frequency = 0.1.



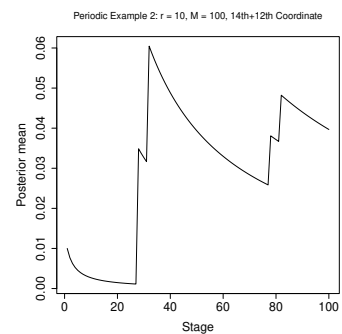
(f)  $r = 10, M = 50$ . True frequency = 0.06.



(g)  $r = 10, M = 100$ . True frequency = 0.4.

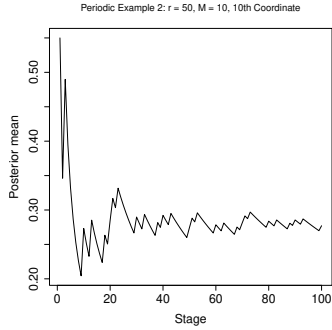


(h)  $r = 10, M = 100$ . True frequency = 0.1.

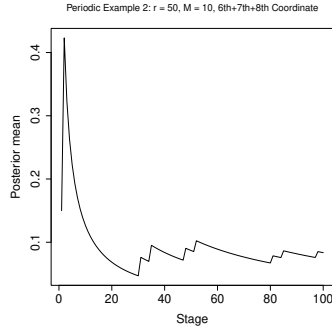


(i)  $r = 10, M = 100$ . True frequency = 0.06.

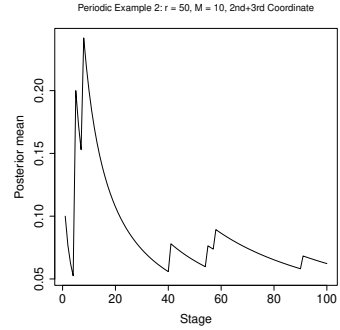
**Figure 10.5.11:** Illustration of our Bayesian method for determining multiple frequencies. Here the true frequencies are 0.4, 0.1 and 0.06.



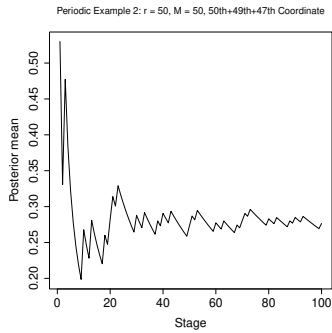
(a)  $r = 50, M = 10$ . True frequency = 0.4.



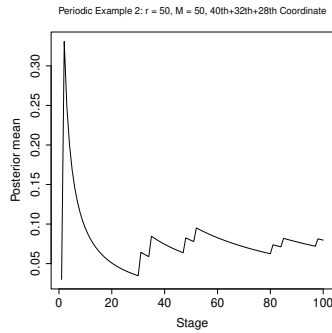
(b)  $r = 50, M = 10$ . True frequency = 0.1.



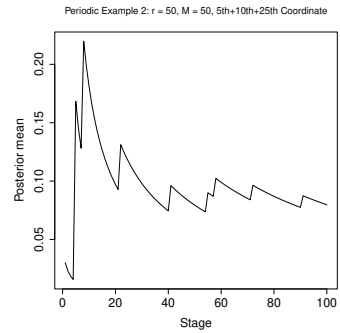
(c)  $r = 50, M = 10$ . True frequency = 0.06.



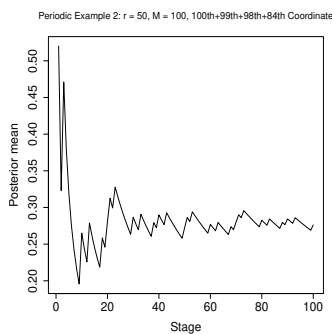
(d)  $r = 50, M = 50$ . True frequency = 0.4.



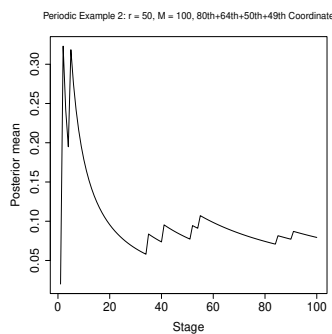
(e)  $r = 50, M = 50$ . True frequency = 0.1.



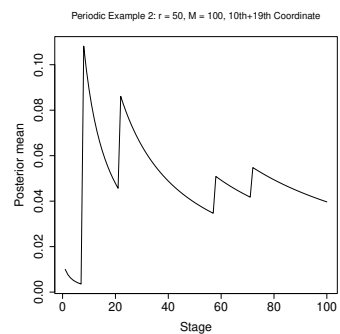
(f)  $r = 50, M = 50$ . True frequency = 0.06.



(g)  $r = 50, M = 100$ . True frequency = 0.4.

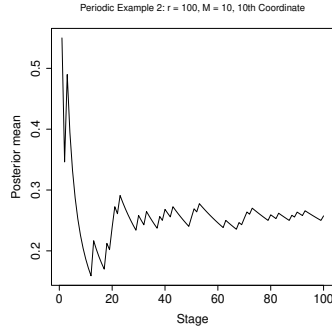


(h)  $r = 50, M = 100$ . True frequency = 0.1.

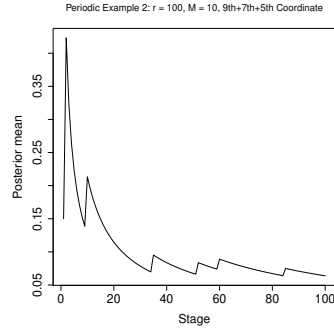


(i)  $r = 50, M = 100$ . True frequency = 0.06.

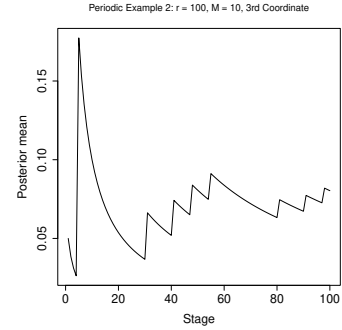
**Figure 10.5.12:** Illustration of our Bayesian method for determining multiple frequencies. Here the true frequencies are 0.4, 0.1 and 0.06.



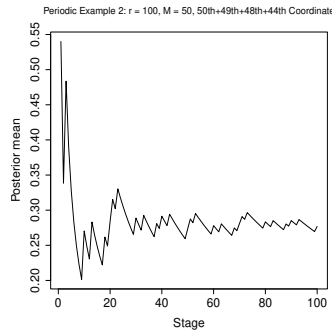
(a)  $r = 100, M = 10$ . True frequency = 0.4.



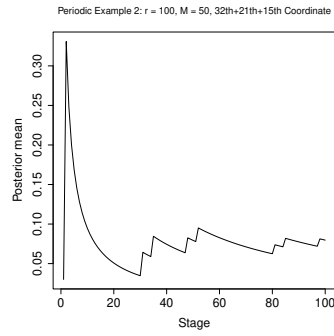
(b)  $r = 100, M = 10$ . True frequency = 0.1.



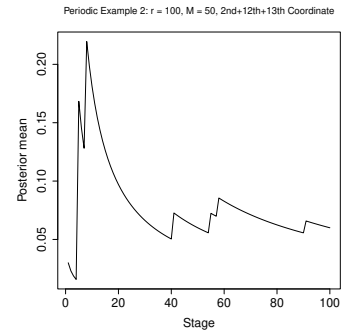
(c)  $r = 100, M = 10$ . True frequency = 0.06.



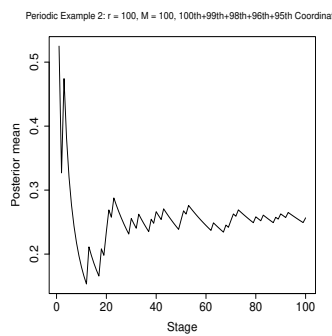
(d)  $r = 100, M = 50$ . True frequency = 0.4.



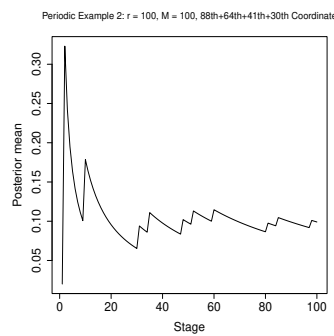
(e)  $r = 100, M = 50$ . True frequency = 0.1.



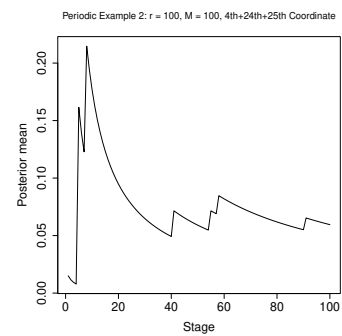
(f)  $r = 100, M = 50$ . True frequency = 0.06.



(g)  $r = 100, M = 100$ . True frequency = 0.4.



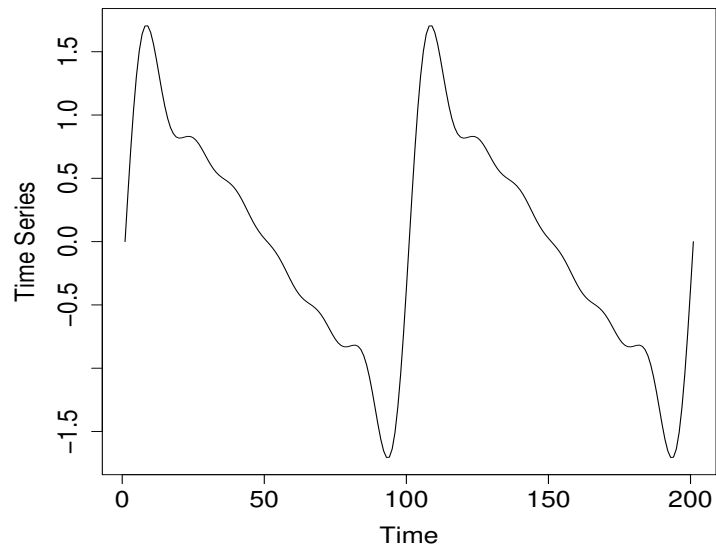
(h)  $r = 100, M = 100$ . True frequency = 0.1.



(i)  $r = 100, M = 100$ . True frequency = 0.06.

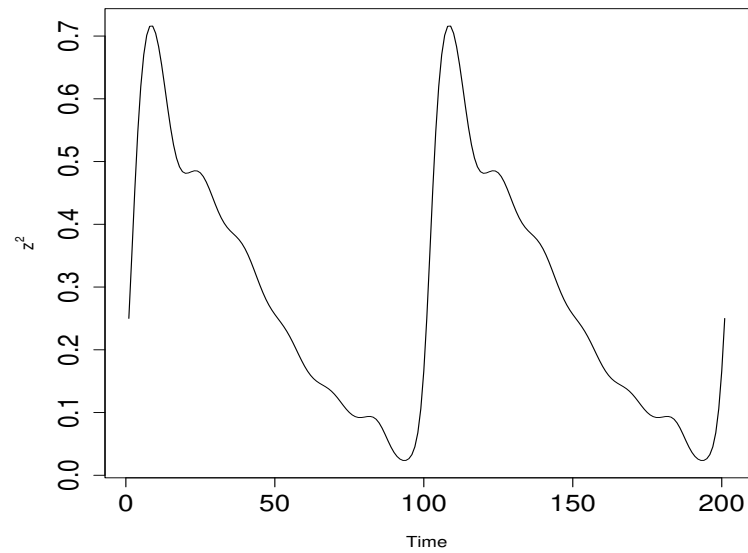
**Figure 10.5.13:** Illustration of our Bayesian method for determining multiple frequencies. Here the true frequencies are 0.4, 0.1 and 0.06.

### Oscillating Time Series With Six Harmonics



(a) The original signal with 6 harmonics.

### Transformed x



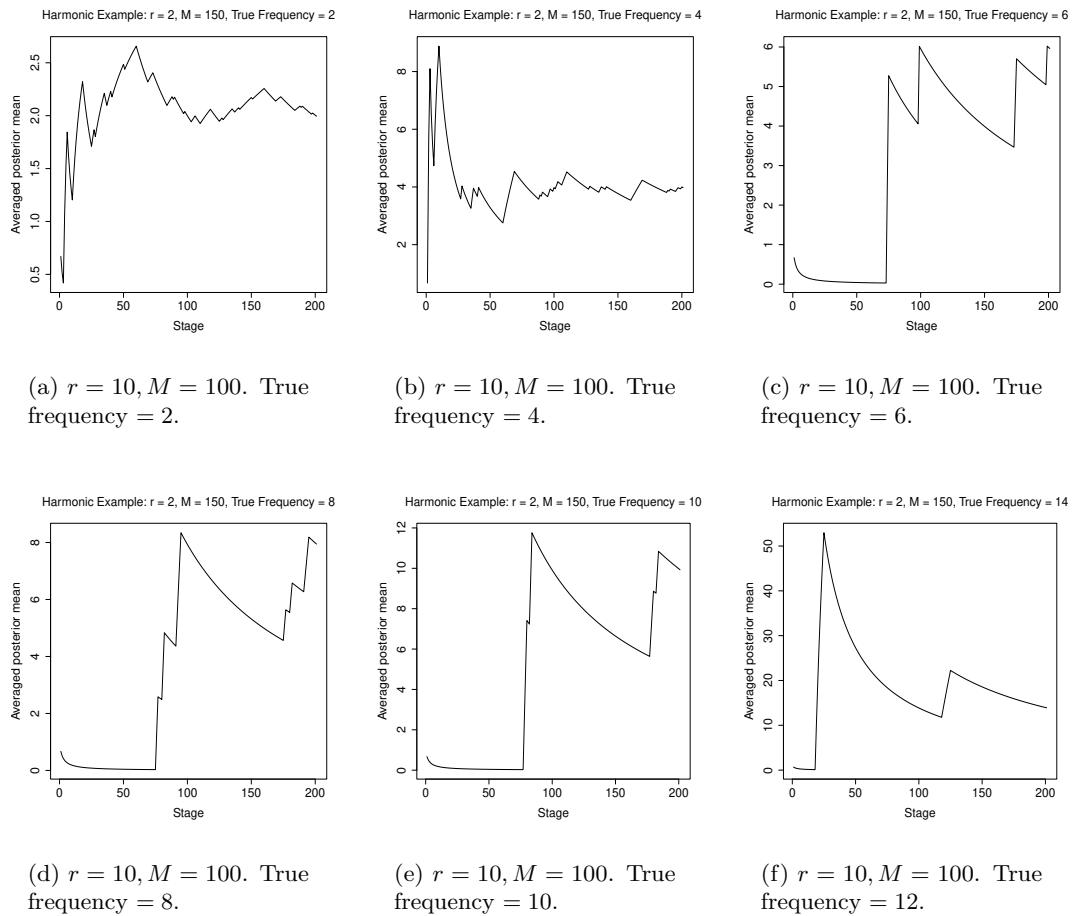
(b) The transformed signal with 6 harmonics.

**Figure 10.5.14:** The original and the transformed signal with 6 harmonics.

201-th iteration, while the rest converged to zero. We take the averages of the coordinates yielding the same distinct value, and present the results in Figure 10.5.15, after multiplication by 201, to yield the Bayesian results on frequencies per unit time. As is evident from the diagrams, the final iterations produced the frequencies 2, 4, 6, 8, 10, 14, obtained after rounding off the values. Except the frequency 14, which somewhat overestimates the true frequency 12, the others are indeed the true frequencies. That so accurate results are obtained by our Bayesian method even for a challenging time series with small length, is really encouraging.

## 10.6 Real data example: El Niño and fish population

Based on data provided by Dr. Roy Mendelsohn of the Pacific Environmental Fisheries Group, [Shumway and Stoffer \(2006\)](#) analyse two oscillating time series on monthly values of an environmental series called the Southern Oscillation Index (SOI) and associated Recruitment (number of new fish), available for a period of 453 months, ranging over the years 1950–1987. The plots are provided in [Shumway and Stoffer \(2006\)](#); see also panel (a) of Figure 10.6.1 and panel (a) of Figure 10.6.3. The quantity SOI is a measurement of air pressure change associated with sea surface temperatures in the central Pacific Ocean. The El Niño effect is considered to cause warming of the central Pacific every three to seven years, which in turn, is presumed to be responsible for causing floods in the midwestern portions of the United States in the year 1997. It is thus important to identify the frequency of oscillation of the SOI series and the associated dependent Recruitment series, which seem to have slightly slower frequency of oscillation in comparison to the SOI series. At first glance, both the series seem to have two significant frequencies of oscillations. For instance, the Recruitment series seems to oscillate once in every 12 months and also once in every 50 months. Slightly faster frequencies can be expected of the SOI series. The periodogram analyses provided in [Shumway and Stoffer \(2006\)](#) indeed give weight to these frequencies.



**Figure 10.5.15:** Illustration of our Bayesian method for determining multiple frequencies in non-sinusoidal signals. Here the true frequencies are 2, 4, 6, 8, 10 and 12 oscillations per unit time.

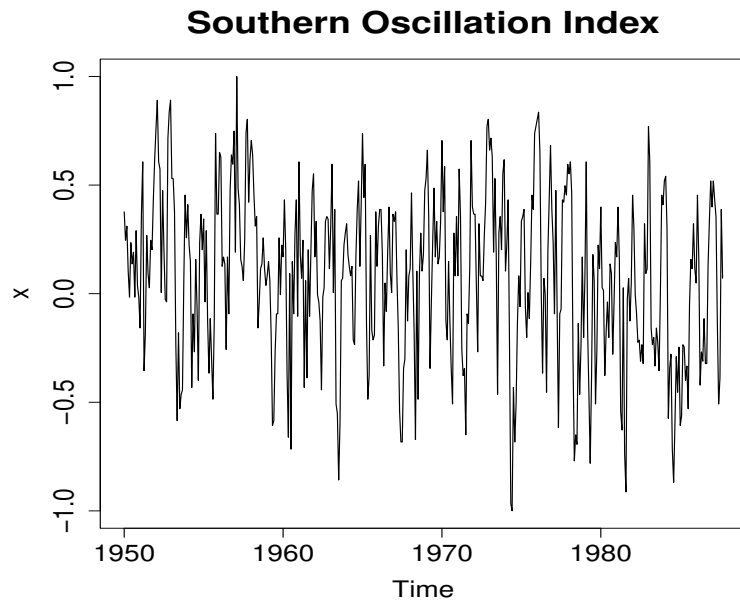


We now apply our Bayesian method to investigate the frequencies hidden in the two underlying time series. Although the two series seem to be dependent, we consider their analyses one by one. In the case of dependence, the frequencies in this situation are expected to be close.

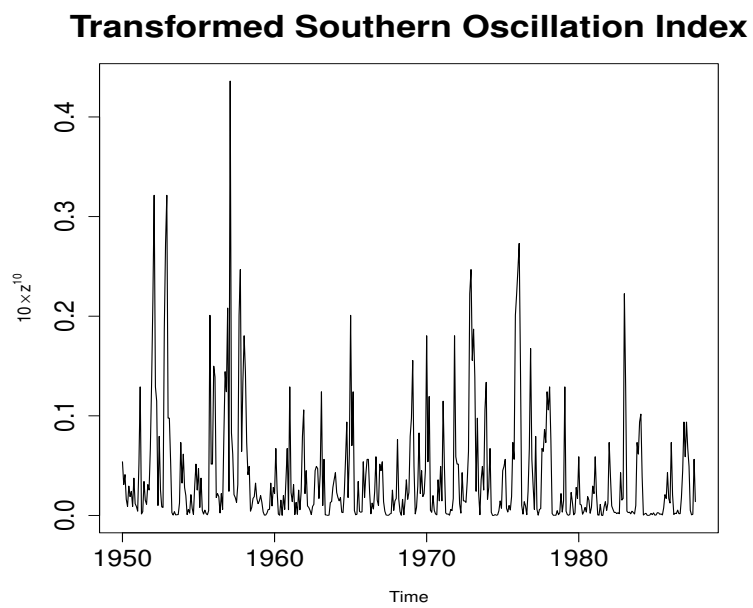
### 10.6.1 SOI series

We first take up the case of the SOI series. Denoting the series by  $X_t$ , for our purpose, we need to consider a transformation of the series to  $Z_t^r$ , with  $Z^t = \exp(X_t) / (1 + \exp(X_t))$ . We choose  $r (> 0)$  such that the oscillations in the process  $\mathbf{Z}^r = \{Z_t^r\}$  become as explicit as possible. With  $r = 10$ , this goal seems to be achieved. However, multiplying the aforementioned transformed series with 10 increased the range of the transformed time series while preserving easy visualization of the oscillations. Increasing the range decreased the possibility of too finely partitioning the interval  $[0, 1]$ . Note that too fine partitions contain too many sub-intervals that do not contribute to the frequencies, but slows down implementation of the Bayesian code. Thus, increasing the range can prevent wastefulness and improve run-time of the computer code. The original SOI time series and the transformed time series  $10 \times \mathbf{Z}^{10}$  are shown in Figure 10.6.1.

Although we increased the range of the transformed time series by multiplying it with 10, still a relatively fine partition is required in this case, as the maximum of the range is still quite less than 1. Hence, we implement our Dirichlet process based Bayesian method with  $r = 10$  and  $M = 1000$ . Figure 10.6.2 shows the results of our implementation. Panel (a) of the figure shows convergence of the relevant posterior of  $p_{25,j} + p_{27,j}$  approximately to a slightly lesser frequency than 0.02, while panel (b) shows convergence of  $p_{16,j} + p_{18,j} + p_{19,j} + p_{21,j} + p_{22,j}$  approximately to a slightly higher frequency than 0.08. The relatively fine partition of  $[0, 1]$  is the reason for dissipating of the proportions to many intervals  $(\tilde{p}_{m-1,j}, \tilde{p}_{m,j}]$ . Other than the aforementioned  $p_{m,j}$ 's contributing to the frequencies, the rest of the  $p_{m,j}$ 's, except  $p_{1,j}$ , converged to zero.

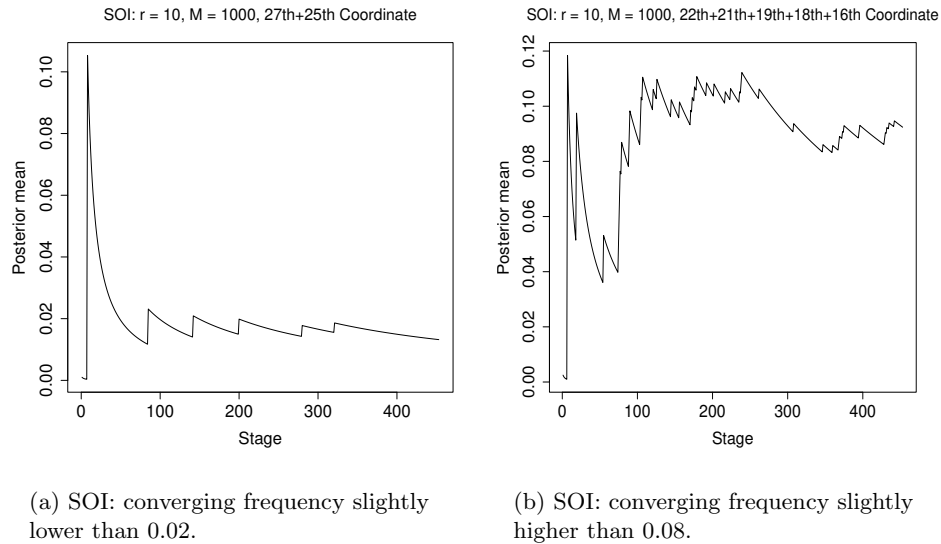


(a) The original SOI time series.



(b) The transformed SOI time series.

**Figure 10.6.1:** The original and the transformed SOI time series.

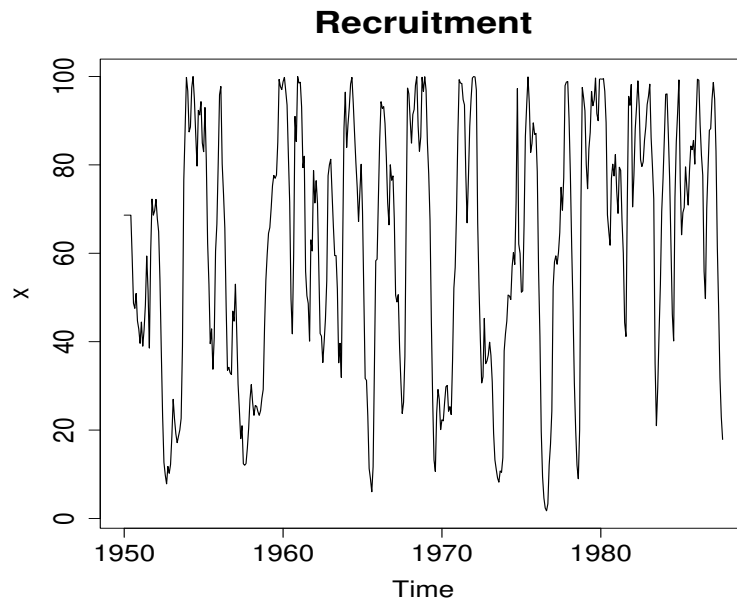


**Figure 10.6.2:** Bayesian results for frequency determination of the SOI time series.

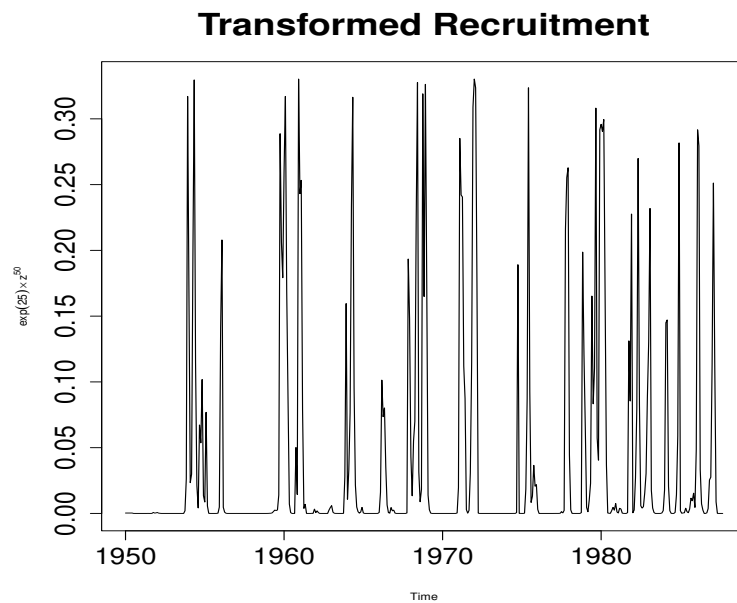
Thus, our results are consistent with the periodogram analysis of [Shumway and Stoffer \(2006\)](#).

### 10.6.2 Recruitment series

We now turn to the Recruitment time series. The original Recruitment series and the transformation  $\exp(25) \times \mathbf{Z}^{50}$  are displayed in Figure 10.6.3. This transformation enabled the most explicit visualization of the oscillations, among those that we experimented with. The multiplicative factor  $\exp(25)$  raises the range to a reasonable limit. We consider  $M = 1000$  for our Bayesian implementation based on Dirichlet process. Figure 10.6.4 depicts the posterior convergence path to the relevant frequencies. Note that the convergences in panel (a) occurs towards slightly lower than 0.02, while that in panel (b) occurs towards slightly higher than 0.08. These are consistent with the results associated with the SOI series.

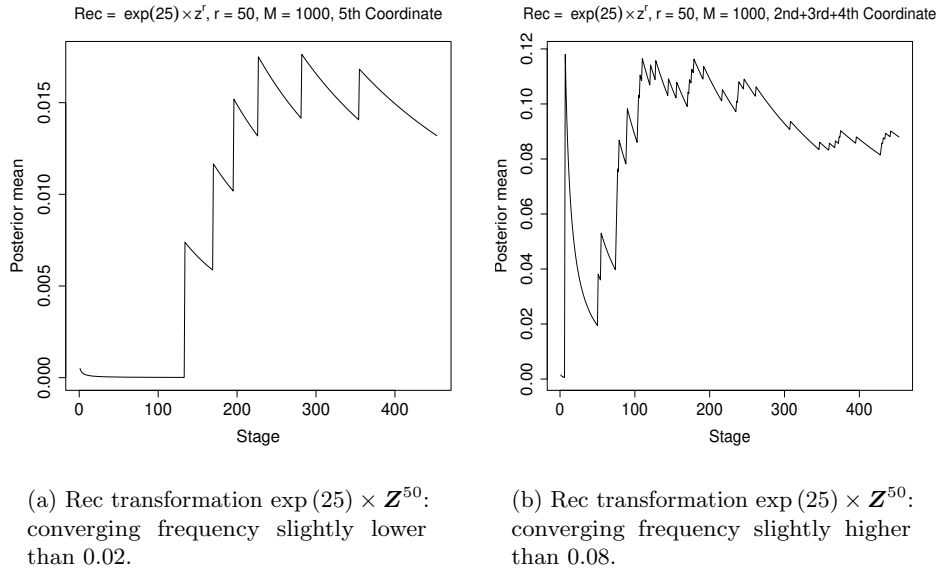


(a) The original Recruitment time series.



(b) The transformed Recruitment time series.

**Figure 10.6.3:** The original and the transformed Recruitment time series.



**Figure 10.6.4:** Bayesian results for frequency determination of the Recruitment time series with transformed time series  $\exp(25) \times \mathbf{Z}^{50}$ .

## 10.7 Summary and conclusion

It is interesting to see that our Bayesian characterization idea can be fruitfully adopted for the purpose of frequency determination in oscillating stochastic processes. The basic idea here is to first provide an appropriate bijective transformation to the data such that the transformed process takes values on  $[0, 1]$ . The transformed process can then be raised to some appropriate power such that the oscillations become as explicit as possible. Dividing up the interval  $[0, 1]$  into appropriate sub-intervals, we consider the proportions of oscillations contained in the sub-intervals. These can then be related to the frequencies of oscillation of the underlying stochastic process, and again facilitates characterization with our recursive Bayesian principle. We characterize single and multiple frequencies, as well as infinite number of frequencies of oscillation.

The results of our ideas applied to both simulated and real examples once again bring out the effectiveness of our Bayesian thought. We have also clarified that our Bayesian

approach has advantages over the classical spectral analysis in the sense of yielding desired credible regions for any sample size without requiring any approximation of assumption to be validated. Moreover, ours is a model-free approach, in contrast with the classical frequency domain methodology.

# 11

## Function Optimization with Posterior Gaussian Derivative Process

### 11.1 Introduction

There is no scientific discipline that does not require function optimization. Hence, it is needless to mention that there exists an enormous literature on the topic, with a plethora of optimization techniques and algorithms. Most of the existing algorithms are deterministic and heuristic in nature, focussing on quick computations via popular software usage. The solutions provided by most such algorithms are often found to be reasonable in practical applications, even though there need not be any guarantee of convergence to the true optima. Indeed, optimization algorithms are generally designed depending upon the problem at hand, and it seems to be almost impossible to pinpoint towards any optimization method that works for a even a relatively large class of

problems. Among stochastic optimization methods, we consider the simulated annealing methodology to be a general purpose procedure, but it is crucially important to very carefully choose the sequence of “temperatures” to converge to the optima in practice, and even in moderately high dimensions this often turns out to be an extremely difficult exercise. Effective choice of proposal distributions also plays a major role in practical convergence, which is again a difficult issue in high dimensions.

In this chapter, in accordance with the Bayesian embedding theme underlying this thesis, we propose and develop a novel Bayesian algorithm for optimization of functions whose first and second partial derivatives are available.

Our approach is to embed the function of interest, along with its derivatives, in a random function scenario, driven by Gaussian processes and the induced derivative Gaussian processes, the latter forming the crux of our methodology. In a nutshell, with data consisting of suitable choices of input points in the function domain and their function values, we first obtain the posterior derivative process corresponding to the original Gaussian process. Then we construct the posterior distribution of the solutions corresponding to setting random partial derivative functions to the null vector. This posterior is expected to emulate the stationary points of the objective function. Now consider a uniform prior on the function domain having the constraints that the first partial derivatives are reasonably close to the null vector and that the matrix of second order partial derivatives is positive definite (for minimization problem, and negative definite for maximization problem). Due to the prior constraints, the resultant posterior solutions may be expected to approximately emulate the true optima even if the dataset is not large enough.

However, for small datasets, the prior constraints will be usually too restrictive to let any posterior simulation method progress. Hence, we shall begin with posterior solution simulations associated with less restrictions, such as that the Euclidean norm of the partial derivative vector is bounded above by some reasonably large constant. Once



adequate posterior solution simulations are obtained, we shall consider iterative stages, where we shall progressively simulate from the posterior with finer prior restrictions, and at each stage augmenting realizations (and their function values) that meet the finer restrictions, to the original dataset. Thus, as the iterations tend to infinity, the resulting posteriors are expected to converge to the true optima. These key concepts lead to an effective, general-purpose optimization algorithm.

As can be anticipated from the above intuitions, convergence of the algorithm depends crucially on convergence of the posterior derivative process to the true function derivatives. And such convergence depends upon appropriate design of the input points of the function domain. As such, under appropriate fixed-domain infill asymptotics setups, we prove almost sure uniform convergence of the posteriors corresponding to Gaussian and Gaussian derivative processes to the objective function and its derivatives. Interestingly, we are also able to obtain rates of convergence under a particular infill asymptotics setup. To our knowledge, these results are new and are of independent interest. Utilizing these results, we prove almost sure convergence of our optimization algorithm to the true optima as the number of iterations tends to infinity. As an aside, we also provide Bayesian characterization of the number of optima, borrowing information from our optimization algorithm.

To illustrate our ideas, we consider five different optimization problems involving maxima, minima, saddle points and even inconclusiveness. The problems vary from simple, one-dimensional to challenging 50 and 100-dimensional situations. On application of our Bayesian optimization algorithm with increased sophistication demanded by the increasingly challenging examples, we obtain encouraging and insightful results in each case. We elucidate various issues on accuracy and computation as we proceed with the applications.

A general and important feature of our Bayesian optimization algorithm is that it is able to recognize significantly more accurate solutions than the existing optimization

algorithms. This interesting ability can be attributed to the posterior simulation approach ingrained in our Bayesian optimization concept. Indeed, the posterior simulation approach ensures that our algorithm can explore regions of the input space around any given solution obtained by any other approach, and neighborhoods of such solution must contain at least one new solution which is at least as close to the true optimum compared to the given solution.

It must be mentioned that function optimization methods using traditional Gaussian process posteriors do exist in the literature (see, for example, [Frazier \(2018\)](#) and the references therein), but these methods consider the objective function to be a “black box” and assume that the derivatives are unavailable. These methods would naturally be far less accurate compared to ours when the derivatives are available. We are also not aware of any convergence result for such “derivative-free” Gaussian process methods which make use of so-called “acquisition functions” to progress. Since there are many choices of acquisition functions, each with its own merits and demerits, it seems doubtful if such methods can have solid foundation, let alone reliability in practical implementations. In this work, we shall not concern ourselves with the existing Bayesian optimization methods.

The rest of this chapter is structured as follows. In Section 11.2 we provide details on derivation of the posterior associated with Gaussian derivative process and in Section 11.3 we derive the form of the posterior for the random optima associated with the solutions of the Gaussian derivative process set equal to zero, along with other desirable restrictions using objective function derivatives. Almost sure uniform convergence results for posteriors associated with Gaussian process and Gaussian process derivatives are presented in Section 11.4, along with the proofs. In Section 11.5 we introduce our general-purpose Bayesian optimization algorithm, and in Section 11.6 we establish Bayesian characterization of the number of optima of the objective function. Illustration of our Bayesian optimization algorithm with various optimization problems is taken up in

Section 11.7. Summarization of our contribution, along with concluding remarks, are provided in Section 3.7.

## 11.2 Posterior Gaussian derivative process

### 11.2.1 Details of the objective function

Consider any function  $f : \mathbb{R}^d \mapsto \mathbb{R}$ , where  $\mathbb{R}$  is the real line and  $d (\geq 1)$  is the dimension of the input space, assumed to be finite. We further assume that the second order partial derivatives of  $f$ , namely,  $\partial^2 f(\mathbf{x})/\partial x_i \partial x_j$  exist and are continuous for all  $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$  for  $i, j = 1, \dots, d$ . The objective is to optimize the function  $f(\mathbf{x})$  with respect to  $\mathbf{x} \in \mathcal{X}$ . For theoretical purposes, we assume that  $\mathcal{X}$  is compact; the assumptions is not required for implementation of our methodology.

### 11.2.2 Data from the objective function

Assume that corresponding to arbitrary inputs  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \in \mathcal{X}$ , where  $n > 1$ , the output vector  $\mathbf{f}_n = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))^T$  is available. Here  $T$  denotes transpose. Let  $\mathbf{D}_n = \{(\mathbf{x}_i, f(\mathbf{x}_i)) : i = 1, \dots, n\}$ .

### 11.2.3 Gaussian process representation of the objective function

Let  $g : \mathbb{R}^d \mapsto \mathbb{R}$  denote a random function such that given  $\mathbf{D}_n$ ,  $g(\mathbf{x}_i) = f(\mathbf{x}_i)$  for  $i = 1, \dots, n$ .

Since Gaussian processes have the above interpolation property, we model  $g(\cdot)$  by a Gaussian process with mean function  $\mu(\cdot)$  and covariance function  $\sigma^2 c(\cdot, \cdot)$ , where  $\sigma^2$  is the process variance. In other words,  $E[g(\mathbf{x})] = \mu(\mathbf{x})$  for any  $\mathbf{x} \in \mathcal{X}$  and  $Cov(g(\mathbf{x}), g(\mathbf{y})) = \sigma^2 c(\mathbf{x}, \mathbf{y})$ , for all  $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ . Let  $\mu(\cdot)$  be continuous in  $\mathcal{X}$  and  $c(\cdot, \cdot)$  be Lipschitz continuous on  $\mathcal{X} \times \mathcal{X}$ . Then  $g(\cdot)$  is actually a continuous-path Gaussian process with mean function  $\mu(\cdot)$  and covariance function  $\sigma^2 c(\cdot, \cdot)$ .

We shall use the notation  $\mathbf{g}_n$  to denote  $(g(\mathbf{x}_1), \dots, g(\mathbf{x}_n))^T$  when distribution of this vector will be considered and  $\mathbf{f}_n$  when this vector is conditioned upon.

#### 11.2.4 Gaussian derivative process

For  $\mathbf{x} = (x_1, \dots, x_d)$  and  $\mathbf{y} = (y_1, \dots, y_d)$ , let us assume that the second order mixed partial derivatives

$$\frac{\partial^2 c(\mathbf{x}^*, \mathbf{y}^*)}{\partial x_i \partial y_i} = \frac{\partial^2 c(\mathbf{x}, \mathbf{y})}{\partial x_i \partial y_i} \Big|_{\mathbf{x}=\mathbf{x}^*, \mathbf{y}=\mathbf{y}^*}$$

are Lipschitz continuous on  $\mathcal{X} \times \mathcal{X}$  for  $i = 1, \dots, d$ .

With the above assumption on the covariance function and with the further assumption that  $\mu(\cdot)$  is twice continuously differentiable with continuous mixed second order partial derivatives, for  $\mathbf{x} = (x_1, \dots, x_d)$ ,

$$g'_i(\mathbf{x}^*) = \frac{\partial g(\mathbf{x}^*)}{\partial x_i} = \frac{\partial g(\mathbf{x})}{\partial x_i} \Big|_{\mathbf{x}=\mathbf{x}^*},$$

corresponding to the original continuous-path Gaussian process  $g(\cdot)$  exists for  $i = 1, \dots, d$ , for all  $\mathbf{x}^* \in \mathcal{X}$ . Specifically, for  $i = 1, \dots, d$ ,  $g'_i(\cdot) = \partial g(\cdot) / \partial x_i$  is a continuous-path Gaussian process with mean function  $\mu'_i(\cdot) = \partial \mu(\cdot) / \partial x_i$  and covariance function  $\sigma^2 \partial^2 c(\cdot, \cdot) / \partial x_i \partial y_i$ .

For general details on Gaussian and Gaussian derivative processes, see, for example, [Adler \(1981\)](#), [Adler and Taylor \(2007\)](#).

### 11.2.5 Joint distribution of Gaussian variables and Gaussian derivative variables

Note that, given  $\mathbf{x}^*, \mathbf{y}^* \in \mathcal{X}$ ,

$$Cov(g'_i(\mathbf{x}^*), g'_j(\mathbf{x}^*)) = \sigma^2 \frac{\partial^2 c(\mathbf{x}, \mathbf{y})}{\partial x_i \partial y_j} \Big|_{\mathbf{x}=\mathbf{x}^*, \mathbf{y}=\mathbf{x}^*}; \quad (11.2.1)$$

$$Cov(g'_i(\mathbf{x}^*), g(\mathbf{y}^*)) = \sigma^2 \frac{\partial c(\mathbf{x}, \mathbf{y})}{\partial x_i} \Big|_{\mathbf{x}=\mathbf{x}^*, \mathbf{y}=\mathbf{y}^*}. \quad (11.2.2)$$

With the above covariance forms, for given  $\mathbf{x}^* \in \mathcal{X}$ , we have

$$(g'_1(\mathbf{x}^*), \dots, g'_d(\mathbf{x}^*), g(\mathbf{x}_1), \dots, g(\mathbf{x}_n))^T \sim N_{d+n}(\boldsymbol{\nu}_{d+n}, \sigma^2 \boldsymbol{\Sigma}^{\overline{d+n} \times \overline{d+n}}), \quad (11.2.3)$$

that is, the vector on the left hand side of (11.2.3) has the  $(d+n)$ -variate normal distribution with mean vector  $\boldsymbol{\nu}_{d+n}$  and covariance matrix  $\sigma^2 \boldsymbol{\Sigma}^{\overline{d+n} \times \overline{d+n}}$ . Here

$$\boldsymbol{\nu}_{d+n}(\mathbf{x}^*) = (\boldsymbol{\mu}'_d(\mathbf{x}^*)^T, \boldsymbol{\mu}_n^T)^T, \quad (11.2.4)$$

where  $\boldsymbol{\mu}'_d(\mathbf{x}^*) = (\partial \mu(\mathbf{x}^*)/\partial x_1, \dots, \partial \mu(\mathbf{x}^*)/\partial x_d)^T$  with  $\partial \mu(\mathbf{x}^*)/\partial x_i = \frac{\partial \mu(\mathbf{x})}{\partial x_i} \Big|_{\mathbf{x}=\mathbf{x}^*}$  for  $i = 1, \dots, d$ , and  $\boldsymbol{\mu}_n = (\mu(\mathbf{x}_1), \dots, \mu(\mathbf{x}_n))^T$ . Also,

$$\boldsymbol{\Sigma}^{\overline{d+n} \times \overline{d+n}}(\mathbf{x}^*) = \begin{pmatrix} \boldsymbol{\Sigma}_{11}^{d \times d}(\mathbf{x}^*) & \boldsymbol{\Sigma}_{12}^{d \times n}(\mathbf{x}^*) \\ \boldsymbol{\Sigma}_{21}^{n \times d}(\mathbf{x}^*) & \boldsymbol{\Sigma}_{22}^{n \times n} \end{pmatrix}, \quad (11.2.5)$$

where  $\boldsymbol{\Sigma}_{11}^{d \times d}$  is the  $d$ -th order correlation matrix with  $(i, j)$ -th element  $\sigma^{-2} Cov(g'_i(\mathbf{x}^*), g'_j(\mathbf{x}^*))$  where the covariance term is given by (11.2.1),  $\boldsymbol{\Sigma}_{12}^{d \times n}(\mathbf{x}^*)$  is the  $d \times n$  matrix with  $(i, j)$ -th element  $\sigma^{-2} Cov(g'_i(\mathbf{x}^*), g(\mathbf{x}_j))$  where the covariance term is of the same form as (11.2.2) with  $\mathbf{y}^*$  replaced with  $\mathbf{x}_j$ ,  $\boldsymbol{\Sigma}_{21}^{n \times d}(\mathbf{x}^*)$  is the transpose of  $\boldsymbol{\Sigma}_{12}^{d \times n}(\mathbf{x}^*)$ , and  $\boldsymbol{\Sigma}_{22}^{n \times n}$  is the  $n \times n$  matrix with  $(i, j)$ -th element  $c(\mathbf{x}_i, \mathbf{x}_j)$ , the correlation between  $g(\mathbf{x}_i)$  and  $g(\mathbf{x}_j)$ . In our examples, we shall consider the squared exponential correlation function having the form

$$c(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{y})^T \mathbf{\Lambda}^{-1}(\mathbf{x} - \mathbf{y})\right), \quad (11.2.6)$$

for  $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ , where  $\mathbf{\Lambda}$  is a  $d \times d$  diagonal matrix with positive diagonal elements  $\lambda_i$ ;  $i = 1, \dots, d$ . It follows that  $\mathbf{\Sigma}_{11} = \mathbf{\Lambda}^{-1}$  and for  $j = 1, \dots, n$ , the  $j$ -th column of  $\mathbf{\Sigma}_{12}(\mathbf{x}^*)$  is  $-\mathbf{\Lambda}^{-1}(\mathbf{x}^* - \mathbf{x}_j)c(\mathbf{x}^*, \mathbf{x}_j)$ .

### 11.2.6 Posterior distribution of the Gaussian derivatives given the data and parameters

From (11.2.3) it follows that the joint posterior distribution of  $\mathbf{g}'(\mathbf{x}^*) = (g'_1(\mathbf{x}^*), \dots, g'_d(\mathbf{x}^*))^T$  given  $\mathbf{x}^*$ ,  $\mathbf{D}_n$ ,  $\sigma^2$  and other parameters  $\boldsymbol{\theta}$ , is the following  $d$ -variate normal distribution:

$$\pi(\mathbf{g}'(\mathbf{x}^*) | \sigma^2, \boldsymbol{\theta}, \mathbf{D}_n) \equiv N_d\left(\tilde{\boldsymbol{\mu}}(\mathbf{x}^*), \sigma^2 \tilde{\boldsymbol{\Sigma}}(\mathbf{x}^*)\right), \quad (11.2.7)$$

where

$$\tilde{\boldsymbol{\mu}}(\mathbf{x}^*) = \boldsymbol{\mu}'_d(\mathbf{x}^*) + \mathbf{\Sigma}_{12}(\mathbf{x}^*) \mathbf{\Sigma}_{22}^{-1}(\mathbf{f}_n - \boldsymbol{\mu}_n); \quad (11.2.8)$$

$$\tilde{\boldsymbol{\Sigma}}(\mathbf{x}^*) = \mathbf{\Sigma}_{11}(\mathbf{x}^*) - \mathbf{\Sigma}_{12}(\mathbf{x}^*) \mathbf{\Sigma}_{22}^{-1} \mathbf{\Sigma}_{21}(\mathbf{x}^*). \quad (11.2.9)$$

### 11.2.7 Prior and posterior distributions of the parameters

In our examples we assume that  $\boldsymbol{\mu}(\mathbf{x}) = \mathbf{h}(\mathbf{x})^T \boldsymbol{\beta}$ , where  $\mathbf{h}(\mathbf{x})^T = (1, x_1, \dots, x_d)$  and  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_d)^T \in \mathbb{R}^d$ . Let  $\mathbf{H}^{n \times d+1} = (\mathbf{h}(\mathbf{x}_1), \dots, \mathbf{h}(\mathbf{x}_n))^T$ .

#### Priors for the parameters

We assume that *a priori*

$$\pi(\boldsymbol{\beta} | \sigma^2) \equiv N_{d+1}(\boldsymbol{\beta}_0, \sigma^2 \boldsymbol{\Sigma}_0), \quad (11.2.10)$$

where  $\beta_0$  is the mean vector and  $\Sigma_0$  is the positive definite covariance matrix, and

$$\pi(\sigma^{-2}) \equiv \mathcal{G}(a, b), \quad (11.2.11)$$

the gamma distribution with mean  $a/b$  and variance  $a/b^2$ , where  $a, b > 0$ .

### Posteriors for the parameters

Let us first obtain the posterior distribution of  $\sigma^{-2}$  given  $\mathbf{D}_n$ . Note that

$$\begin{aligned} \pi(\sigma^{-2} | \mathbf{D}_n) &\propto \pi(\sigma^{-2}) \pi(\mathbf{g}_n | \sigma^{-2}) \\ &= \pi(\sigma^{-2}) \int \pi(\mathbf{g}_n | \sigma^{-2}, \beta) \pi(\beta | \sigma^2) d\beta. \end{aligned} \quad (11.2.12)$$

To obtain  $\pi(\mathbf{g}_n | \sigma^{-2}) = \int \pi(\mathbf{g}_n | \sigma^{-2}, \beta) \pi(\beta | \sigma^2) d\beta$  note that

$$\pi(\mathbf{g}_n | \sigma^{-2}, \beta) \equiv N_n(\mathbf{H}\beta, \sigma^2 \Sigma_{22}) \quad (11.2.13)$$

and since  $\pi(\beta | \sigma^2)$  has the normal distribution (11.2.10), it follows that

$$\pi(\mathbf{g}_n | \sigma^{-2}) \equiv N_n(\mathbf{H}\beta_0, \sigma^2 (\mathbf{H}\Sigma_0\mathbf{H}^T + \Sigma_{22})). \quad (11.2.14)$$

Combining (11.2.14) with (11.2.12) and (11.2.11) it follows that

$$\pi(\sigma^{-2} | \mathbf{D}_n) \equiv \mathcal{G}\left(a + \frac{d}{2}, b + \frac{1}{2} (\mathbf{f}_n - \mathbf{H}\beta_0)^T (\mathbf{H}\Sigma_0\mathbf{H}^T + \Sigma_{22})^{-1} (\mathbf{f}_n - \mathbf{H}\beta_0)\right). \quad (11.2.15)$$

Also, combining (11.2.13) and (11.2.10) it is easy to see that

$$\pi(\beta | \mathbf{D}_n, \sigma^2) \equiv N_{d+1}\left((\mathbf{H}^T \Sigma_{22}^{-1} \mathbf{H} + \Sigma_0^{-1})^{-1} (\mathbf{H}^T \Sigma_{22}^{-1} \mathbf{f}_n + \Sigma_0^{-1} \beta_0), \sigma^2 (\mathbf{H}^T \Sigma_{22}^{-1} \mathbf{H} + \Sigma_0^{-1})^{-1}\right). \quad (11.2.16)$$

### 11.2.8 Marginal posterior distribution of the derivative process

Now, from (11.2.7) it follows that

$$\pi(\mathbf{g}'(\mathbf{x}^*)|\sigma^2, \boldsymbol{\beta}, \mathbf{D}_n) \equiv N_d\left(\mathbf{A}\boldsymbol{\beta} + \boldsymbol{\Sigma}_{12}(\mathbf{x}^*)\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{f}_n - \mathbf{H}\boldsymbol{\beta}), \sigma^2\tilde{\boldsymbol{\Sigma}}(\mathbf{x}^*)\right), \quad (11.2.17)$$

where  $\mathbf{A}^{d \times \overline{d+1}} = \begin{pmatrix} d \times 1 & \mathbb{I}_d \end{pmatrix}$ . Here  $\mathbf{0}^{d \times 1}$  is the  $d$ -dimensional null vector and  $\mathbb{I}_d$  is the identity matrix of order  $d$ . Integrating (11.2.17) with respect to (11.2.16) we obtain

$$\pi(\mathbf{g}'(\mathbf{x}^*)|\sigma^2, \mathbf{D}_n) \equiv N_d\left(\hat{\boldsymbol{\mu}}'(\mathbf{x}^*), \sigma^2\hat{\boldsymbol{\Sigma}}(\mathbf{x}^*)\right), \quad (11.2.18)$$

where  $\hat{\boldsymbol{\mu}}'(\mathbf{x}^*)$  and  $\hat{\boldsymbol{\Sigma}}(\mathbf{x}^*)$  are given by

$$\hat{\boldsymbol{\mu}}'(\mathbf{x}^*) = \mathbf{A}\hat{\boldsymbol{\beta}} + \boldsymbol{\Sigma}_{12}(\mathbf{x}^*)\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{f}_n - \mathbf{H}\hat{\boldsymbol{\beta}}); \quad (11.2.19)$$

$$\hat{\boldsymbol{\Sigma}}(\mathbf{x}^*) = \tilde{\boldsymbol{\Sigma}}(\mathbf{x}^*) + (\mathbf{A} - \boldsymbol{\Sigma}_{12}(\mathbf{x}^*)\boldsymbol{\Sigma}_{22}^{-1}\mathbf{H}) (\mathbf{H}^T\boldsymbol{\Sigma}_{22}^{-1}\mathbf{H} + \boldsymbol{\Sigma}_0^{-1})^{-1} (\mathbf{A} - \boldsymbol{\Sigma}_{12}(\mathbf{x}^*)\boldsymbol{\Sigma}_{22}^{-1}\mathbf{H})^T, \quad (11.2.20)$$

with

$$\hat{\boldsymbol{\beta}} = (\mathbf{H}^T\boldsymbol{\Sigma}_{22}^{-1}\mathbf{H} + \boldsymbol{\Sigma}_0^{-1})^{-1} (\mathbf{H}^T\boldsymbol{\Sigma}_{22}^{-1}\mathbf{f}_n + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0). \quad (11.2.21)$$

Integrating (11.2.18) with respect to (11.2.15) we obtain

$$\pi(\mathbf{g}'(\mathbf{x}^*)|\mathbf{D}_n) \equiv t_d\left(\hat{\boldsymbol{\mu}}'(\mathbf{x}^*), \frac{(a + \frac{d}{2})\hat{\boldsymbol{\Sigma}}(\mathbf{x}^*)^{-1}}{b + \frac{1}{2}(\mathbf{f}_n - \mathbf{H}\boldsymbol{\beta}_0)^T(\mathbf{H}\boldsymbol{\Sigma}_0\mathbf{H}^T + \boldsymbol{\Sigma}_{22})^{-1}(\mathbf{f}_n - \mathbf{H}\boldsymbol{\beta}_0)}, 2\left(a + \frac{d}{2}\right)\right), \quad (11.2.22)$$

where for any  $d$ -dimensional vector  $\boldsymbol{\mu}$ ,  $d$ -th order covariance matrix  $\boldsymbol{\Sigma}$ , and  $\alpha > 0$ ,  $t_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}, \alpha)$  is a  $d$ -variate Student's  $t$  distribution with density at  $\mathbf{x} \in \mathbb{R}^d$  given by

$$t_d(\mathbf{x} : \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}, \alpha) = C [1 + \alpha^{-1}(\mathbf{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})]^{-\frac{(\alpha+d)}{2}},$$



where

$$C = \frac{\Gamma\left(\frac{\alpha+d}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)(\alpha\pi)^{\frac{d}{2}}},$$

with  $\Gamma(\cdot)$  denoting the gamma function.

### 11.3 Posterior distribution of random optima corresponding to the posterior derivative process

From (11.2.22) we obtain the following posterior density at  $\mathbf{g}'(\mathbf{x}^*) = \mathbf{0}$ :

$$\pi(\mathbf{g}'(\mathbf{x}^*) = \mathbf{0} | \mathbf{D}_n) \propto \left[ 1 + \frac{\hat{\boldsymbol{\mu}}'(\mathbf{x}^*)^T \hat{\boldsymbol{\Sigma}}(\mathbf{x}^*)^{-1} \hat{\boldsymbol{\mu}}'(\mathbf{x}^*)}{2b + (\mathbf{f}_n - \mathbf{H}\boldsymbol{\beta}_0)^T (\mathbf{H}\boldsymbol{\Sigma}_0\mathbf{H}^T + \boldsymbol{\Sigma}_{22})^{-1} (\mathbf{f}_n - \mathbf{H}\boldsymbol{\beta}_0)} \right]^{-(a+d)}. \quad (11.3.1)$$

Now, with prior  $\pi(\mathbf{x}^*)$  on  $\mathbf{x}^*$ , the posterior of  $\mathbf{x}^*$ , given  $\mathbf{g}'(\mathbf{x}^*)$  and  $\mathbf{D}_n$  can be obtained as follows:

$$\begin{aligned} \pi(\mathbf{x}^* | \mathbf{g}'(\mathbf{x}^*), \mathbf{D}_n) &\propto \pi(\mathbf{x}^*) \pi(\mathbf{g}'(\mathbf{x}^*), \mathbf{g}_n | \mathbf{x}^*) \\ &= \pi(\mathbf{x}^*) \pi(\mathbf{g}'(\mathbf{x}^*) | \mathbf{D}_n, \mathbf{x}^*) \pi(\mathbf{g}_n | \mathbf{x}^*) \\ &= \pi(\mathbf{x}^*) \pi(\mathbf{g}'(\mathbf{x}^*) | \mathbf{D}_n, \mathbf{x}^*) \pi(\mathbf{g}_n) \\ &\propto \pi(\mathbf{x}^*) \pi(\mathbf{g}'(\mathbf{x}^*) | \mathbf{D}_n, \mathbf{x}^*). \end{aligned} \quad (11.3.2)$$

In the second step of (11.3.2),  $\pi(\mathbf{g}_n | \mathbf{x}^*)$  is the marginal distribution of  $\mathbf{g}_n$ , integrated over the parameters. Since this does not depend upon  $\mathbf{x}^*$ , we denoted this as  $\pi(\mathbf{g}_n)$  in the third step of (11.3.2). From (11.3.2) it then follows that

$$\pi(\mathbf{x}^* | \mathbf{g}'(\mathbf{x}^*) = \mathbf{0}, \mathbf{D}_n) \propto \pi(\mathbf{x}^*) \pi(\mathbf{g}'(\mathbf{x}^*) = \mathbf{0} | \mathbf{D}_n, \mathbf{x}^*). \quad (11.3.3)$$

Now,  $\pi(\mathbf{g}'(\mathbf{x}^*) = \mathbf{0} | \mathbf{D}_n, \mathbf{x}^*)$  will also depend upon parameters of the covariance function, which will be unknown generally. It is legitimate to estimate them using the maximum likelihood estimation method (see, for example, Santner *et al.* (2003)) and treat them as fixed.

The formula (11.3.3) holds for any prior  $\pi(\mathbf{x}^*)$  for  $\mathbf{x}^*$ . However, we shall consider a uniform prior on  $\mathcal{X}$  constrained by the first and second derivatives of  $f(\cdot)$ . The details are presented below.

### 11.3.1 Prior for $\mathbf{x}^*$

Without loss of generality, let us assume that our objective is to obtain the minima of the function  $f(\cdot)$  on  $\mathcal{X}$ . For  $i, j = 1, \dots, d$ , let  $f''_{ij}(\mathbf{x}^*) = \left. \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \right|_{\mathbf{x}=\mathbf{x}^*}$  denote the second order partial derivatives of the objective function  $f(\cdot)$  at any  $\mathbf{x}^* \in \mathcal{X}$ . Let  $\Sigma''(\mathbf{x}^*)$  stand for the  $d \times d$  matrix of such second order partial derivatives at  $\mathbf{x}^*$  with  $(i, j)$ -th element  $f''_{ij}(\mathbf{x}^*)$ . Let  $\Sigma''(\mathbf{x}^*) > 0$  denote that  $\Sigma''(\mathbf{x}^*)$  is positive definite. Then we consider the following prior for  $\mathbf{x}^*$ :

$$\pi(\mathbf{x}^*) \propto I_{B(\epsilon)}(\mathbf{x}^*), \quad (11.3.4)$$

where, for any set  $A$  and vector  $\mathbf{x}$ ,  $I_A(\mathbf{x}) = 1$  if  $\mathbf{x} \in A$  and zero otherwise. Also, for any  $\mathbf{x}$  and  $\epsilon > 0$ ,

$$B(\epsilon) = \mathcal{X} \cap \{\mathbf{x} : \|\mathbf{f}'(\mathbf{x})\|_d < \epsilon\} \cap \{\mathbf{x} : \Sigma''(\mathbf{x}) > 0\}, \quad (11.3.5)$$

where  $\|\cdot\|_d$  denotes the Euclidean norm in the  $d$ -dimensional Euclidean space.

## 11.4 Almost sure uniform convergence of posterior Gaussian and Gaussian derivative processes

Consider the joint posterior distribution of

$$\pi(g(\cdot), \mathbf{g}'(\cdot) | \mathbf{D}_n) = \pi(g(\cdot) | \mathbf{D}_n) \pi(\mathbf{g}'(\cdot) | g(\cdot), \mathbf{D}_n). \quad (11.4.1)$$

Then the marginal posterior  $\pi(\mathbf{g}'(\cdot) | \mathbf{D}_n)$  is of the same form as (11.2.22), and the marginal posterior distribution  $\pi(g(\cdot) | \mathbf{D}_n)$  in (11.4.1) corresponds to the  $t_1$  process, the form of which is not relevant for our purpose.

For  $n \geq 1$ , let  $\mathbf{X}_n$  denote the  $n$  input points in  $\mathbf{D}_n$ . Note that even after marginalizing out the parameters of the Gaussian process with respect to their posteriors (here  $\beta$  and  $\sigma^2$ ), the interpolation property of  $g(\cdot)$  given  $\mathbf{D}_n$  is preserved. That is, the marginal posterior  $\pi(g(\mathbf{x}^*) | \mathbf{D}_n)$  gives full posterior mass to  $f(\mathbf{x}^*)$  if  $\mathbf{x}^* \in \mathbf{X}_n$ .

Let  $g_n(\cdot)$  denote any random function associated with any non-null set of the marginalized posterior measure of  $g(\cdot)$  given  $\mathbf{D}_n$  (here, the  $t_1$  posterior measure). Also, let  $\mathbf{g}'_n(\cdot)$  denote any  $d$ -dimensional random function associated with any non-null set of the marginalized posterior measure of  $\mathbf{g}'(\cdot)$  given  $\mathbf{D}_n$ , the form of which is provided explicitly by (11.2.22). Theorems 59, 60, 61 and 62 prove almost sure uniform convergence of  $g_n(\cdot)$  and  $\mathbf{g}'_n(\cdot)$  to  $f(\cdot)$  and  $\mathbf{f}'(\cdot)$  respectively, as  $n \rightarrow \infty$ . In particular, Theorems 61 and 62 also provide rates of such convergences. Before introducing the theorems, we first state and prove a lemma that will aid in proving the theorems.

**Lemma 58** *Consider a sequence of real-valued continuous functions  $\{f_n\}_{n=1}^{\infty}$  on any compact set  $\mathcal{X}$  such that  $f_n(\mathbf{x}) \rightarrow f(\mathbf{x})$  for all  $\mathbf{x} \in \mathcal{X}$ , where  $f$  is some real-valued continuous function on  $\mathcal{X}$ . Then*

$$\limsup_{n \rightarrow \infty} \sup_{\mathbf{x} \in \mathcal{X}} |f_n(\mathbf{x})| < \infty.$$

**Proof.** Note that  $f_n$ , for  $n \geq 1$ , and  $f$  are actually uniformly continuous since  $\mathcal{X}$  is compact. Now let us first consider an arbitrary  $\mathbf{x}_1 \in \mathcal{X}$ . Then due to pointwise convergence of  $f_n$  to  $f$ , for any  $\epsilon > 0$ , there exists  $n_1 \geq 1$  such that for  $n \geq n_1$ ,  $|f_n(\mathbf{x}_1) - f(\mathbf{x}_1)| < \epsilon$ . Moreover, due to uniform continuity of  $f_n$  and  $f$ , there exists an open neighborhood  $\mathcal{N}(\mathbf{x}_1)$  of  $\mathbf{x}_1$  such that  $|f_n(\mathbf{x}) - f(\mathbf{x})| < \epsilon_1$  for all  $\mathbf{x} \in \mathcal{N}(\mathbf{x}_1)$ , where  $\epsilon_1$  is some positive finite constant. Since  $f$  is continuous on the compact set  $\mathcal{X}$ , it is uniformly bounded. Hence,  $\sup_{\mathbf{x} \in \mathcal{N}(\mathbf{x}_1)} |f_n(\mathbf{x})| < M_1$  for all  $n \geq n_1$ , where  $M_1$  is some positive finite constant.

Now consider another point  $\mathbf{x}_2 \in \mathcal{X} \setminus \mathcal{N}(\mathbf{x}_1)$ . Then similar argument shows that  $\sup_{\mathbf{x} \in \mathcal{N}(\mathbf{x}_2)} |f_n(\mathbf{x})| < M_2$  for all  $n \geq n_2 \geq n_1$ , where  $\mathcal{N}(\mathbf{x}_2)$  is some appropriate open neighborhood of  $\mathbf{x}_2$  and  $M_2$  is some positive finite constant.

Thus, starting with  $\mathcal{N}(\mathbf{x}_1)$  and the associated bound  $\sup_{\mathbf{x} \in \mathcal{N}(\mathbf{x}_1)} |f_n(\mathbf{x})| < M_1$ , continuing the procedure for  $i \geq 2$ , we can construct neighborhoods  $\mathcal{N}(\mathbf{x}_i)$  with  $\mathbf{x}_i \in \mathcal{X} \setminus \cup_{j=1}^{i-1} \mathcal{N}(\mathbf{x}_j)$  and bounds  $M_i$  such that for all  $n \geq n_i \geq n_{i-1} \geq \dots \geq n_2 \geq n_1$ ,  $\sup_{\mathbf{x} \in \mathcal{N}(\mathbf{x}_i)} |f_n(\mathbf{x})| < M_i$ . Note that  $\mathcal{X} \subseteq \cup_{i=1}^{\infty} \mathcal{N}(\mathbf{x}_i)$ . That is, the set of neighborhoods  $\{\mathcal{N}(\mathbf{x}_i) : i = 1, 2, \dots\}$  constitutes an open cover for  $\mathcal{X}$ . Since  $\mathcal{X}$  is compact, there exists a finite sub-cover for  $\mathcal{X}$ , say,  $\{\mathcal{N}(\mathbf{x}_{i_j}) : j = 1, 2, \dots, K\}$ , where  $K$  is finite. Now, by our construction, for  $n \geq n_{i_j}$ ,  $\sup_{\mathbf{x} \in \mathcal{N}(\mathbf{x}_{i_j})} |f_n(\mathbf{x})| < M_{i_j}$ , for  $j = 1, \dots, K$ . Let  $n_0 = \max\{n_{i_j} : j = 1, \dots, K\}$  and  $M = \max\{M_{i_j} : j = 1, \dots, K\}$ . Then for all  $n \geq n_0$ ,  $\sup_{\mathbf{x} \in \mathcal{X}} |f_n(\mathbf{x})| < M < \infty$ . ■

**Theorem 59** Consider a fixed-domain infill asymptotics framework such that for any  $\mathbf{x} \in \mathcal{X}$ , there exists  $\mathbf{x}_n \in \mathbf{X}_n$  for  $n \geq 1$  satisfying

$$\lim_{n \rightarrow \infty} \|\mathbf{x}_n - \mathbf{x}\|_d = 0. \tag{11.4.2}$$

Also assume that points of the form  $\mathbf{x}_n + \mathbf{h}_n \in \mathbf{X}_n$ , where  $\mathbf{h}_n \rightarrow \mathbf{0}$ , as  $n \rightarrow \infty$ .

For  $\mathbf{x} = (x_1, \dots, x_d)$  and  $\mathbf{y} = (y_1, \dots, y_d)$ , let the correlation function be such that

$$\frac{\partial^2 c(\mathbf{x}^*, \mathbf{y}^*)}{\partial x_i \partial y_i} = \frac{\partial^2 c(\mathbf{x}, \mathbf{y})}{\partial x_i \partial y_i} \Big|_{\mathbf{x}=\mathbf{x}^*, \mathbf{y}=\mathbf{y}^*}$$

exists for all  $\mathbf{x}^*, \mathbf{y}^* \in \mathcal{X}$  and is Lipschitz continuous on  $\mathcal{X} \times \mathcal{X}$  for  $i = 1, \dots, d$ . Then, for almost all sequences  $\{g_n(\cdot)\}_{n=1}^\infty$ ,

$$\sup_{\mathbf{x} \in \mathcal{X}} |g_n(\mathbf{x}) - f(\mathbf{x})| \rightarrow 0, \text{ as } n \rightarrow \infty. \quad (11.4.3)$$

**Proof.** Consider any  $\mathbf{x} \in \mathcal{X}$ . Then there exists  $\mathbf{x}_n \in \mathbf{X}_n$  for  $n \geq 1$  satisfying (11.4.2). Now by Taylor's series expansion up to the first order,

$$g_n(\mathbf{x}_n) = g_n(\mathbf{x}) + (\mathbf{x}_n - \mathbf{x})^T \mathbf{g}'_n(\mathbf{c}_n), \quad (11.4.4)$$

where  $\mathbf{c}_n$  lies on the line joining  $\mathbf{x}$  and  $\mathbf{x}_n - \mathbf{x}$ .

Now, for  $i = 1, \dots, d$ , consider the  $i$ -th partial derivative  $g'_{in}(\cdot)$  of  $g_n(\cdot)$ . With any sequence  $h_{in} \rightarrow 0$  as  $n \rightarrow \infty$ , we have

$$\frac{g_n(x_{1n}, \dots, x_{i-1,n}, x_{in} + h_{in}, x_{i+1,n}, \dots, x_{dn}) - g_n(\mathbf{x}_n)}{h_{in}} = g'_{in}(\mathbf{x}_n^*), \quad (11.4.5)$$

where  $\mathbf{x}_n^* = (x_{1n}, \dots, x_{i-1,n}, x_{in}^*, x_{i+1,n}, \dots, x_{dn})$ ; here  $x_{in}^*$  lies between  $x_{in}$  and  $x_{in} + h_{in}$ . Since  $(x_{1n}, \dots, x_{i-1,n}, x_{in} + h_{in}, x_{i+1,n}, \dots, x_{dn})^T \in \mathbf{X}_n$  and  $\mathbf{x}_n \in \mathbf{X}_n$ ,  $g_n(x_{1n}, \dots, x_{i-1,n}, x_{in} + h_{in}, x_{i+1,n}, \dots, x_{dn}) = f(x_{1n}, \dots, x_{i-1,n}, x_{in} + h_{in}, x_{i+1,n}, \dots, x_{dn})$  and  $g_n(\mathbf{x}_n) = f(\mathbf{x}_n)$ , almost surely. Hence, from (11.4.5) it follows that

$$g'_{in}(\mathbf{x}_n^*) = f'_i(\mathbf{z}_n), \text{ almost surely,} \quad (11.4.6)$$

with  $\mathbf{z}_n = (x_{1n}, \dots, x_{i-1,n}, z_{in}, x_{i+1,n}, \dots, x_{dn})$ , where  $z_{in}$  lies between  $x_{in}$  and  $x_{in} + h_{in}$ . Clearly,  $\mathbf{z}_n \rightarrow \mathbf{x}$ , as  $n \rightarrow \infty$ . Hence, taking limits of both sides of (11.4.6) as  $n \rightarrow \infty$ ,

and using continuity of  $f'_i(\cdot)$ , yields

$$\lim_{n \rightarrow \infty} g'_{in}(\mathbf{x}_n^*) = f'_i(\mathbf{x}), \text{ almost surely.} \quad (11.4.7)$$

Now, by the hypothesis of Lipschitz continuity of the second order mixed partial derivatives of the correlation function ensures existence and sample path continuity of the partial derivatives  $g'_{in}(\cdot)$ , for  $i = 1, \dots, d$ , for any  $n \geq 1$ . Since  $\mathcal{X}$  is compact,  $g'_{in}(\cdot)$  are uniformly continuous on  $\mathcal{X}$ , for  $i = 1, \dots, d$ , for any  $n \geq 1$ . Uniform continuity of  $g'_{in}(\cdot)$  for all  $n \geq 1$  implies that for any  $\epsilon > 0$ ,  $|g'_{in}(\mathbf{x}_n^*) - g'_{in}(\mathbf{x})| < \epsilon$ , whenever  $\|\mathbf{x}_n^* - \mathbf{x}\|_d < \delta$ , where  $\delta (> 0)$  depends upon  $\epsilon$  only. Now, since  $\mathbf{x}_n^* \rightarrow \mathbf{x}$  as  $n \rightarrow \infty$ , there exists  $n_0 (\geq 1)$  depending upon  $\delta$  such that  $\|\mathbf{x}_n^* - \mathbf{x}\|_d < \delta$  for  $n \geq n_0$ . Further, using (11.4.7) we obtain for any  $\mathbf{x} \in \mathcal{X}$ , the following:

$$\lim_{n \rightarrow \infty} g'_{in}(\mathbf{x}) = \lim_{n \rightarrow \infty} g'_{in}(\mathbf{x}_n^*) + \lim_{n \rightarrow \infty} (g'_{in}(\mathbf{x}) - g'_{in}(\mathbf{x}_n^*)) = f'_i(\mathbf{x}), \text{ almost surely.} \quad (11.4.8)$$

That is,  $g'_{in}(\cdot)$  converges pointwise to  $f'_i(\cdot)$  almost surely, as  $n \rightarrow \infty$ . Moreover,  $g'_{in}(\cdot)$  is almost surely continuous on  $\mathcal{X}$  for all  $n \geq 1$  and  $f'_i(\cdot)$  is continuous on  $\mathcal{X}$ . Since  $\mathcal{X}$  is compact, we invoke Lemma 58 to conclude that there exists a positive, finite constant  $M$  depending upon  $f'_i(\cdot)$ ;  $i = 1, \dots, d$  such that

$$\lim_{n \rightarrow \infty} \sup_{\mathbf{x} \in \mathcal{X}} |g'_{in}(\mathbf{x})| < M, \text{ almost surely, for } i = 1, \dots, d. \quad (11.4.9)$$

Hence, using the Cauchy-Schwartz inequality in (11.4.4), boundedness of the partial derivatives  $g'_{in}(\cdot)$  for  $i = 1, \dots, d$  for large enough  $n$  and (11.4.2), we obtain

$$|g_n(\mathbf{x}_n) - g_n(\mathbf{x})| = |(\mathbf{x}_n - \mathbf{x})^T \mathbf{g}'_n(\mathbf{c}_n)| \leq \|\mathbf{x}_n - \mathbf{x}\|_d \times \|\mathbf{g}'_n(\mathbf{c}_n)\|_d \rightarrow 0, \text{ as } n \rightarrow \infty. \quad (11.4.10)$$

Hence, using (11.4.10) and continuity of  $f(\cdot)$  we obtain, for any  $\mathbf{x} \in \mathcal{X}$ ,

$$\lim_{n \rightarrow \infty} g_n(\mathbf{x}) = \lim_{n \rightarrow \infty} g_n(\mathbf{x}_n) = \lim_{n \rightarrow \infty} f(\mathbf{x}_n) = f(\lim_{n \rightarrow \infty} \mathbf{x}_n) = f(\mathbf{x}), \quad (11.4.11)$$

proving pointwise convergence of  $g_n(\cdot)$  to  $f(\cdot)$ .

Thus, we have shown that  $g_n(\cdot)$  converges pointwise to  $f(\cdot)$  on  $\mathcal{X}$  almost surely, as  $n \rightarrow \infty$  (equation (11.4.11)), and also that the partial derivatives of  $g_n(\cdot)$  are uniformly bounded in the limit almost surely (equation (11.4.9)). The latter also implies that  $g_n(\cdot)$  is almost surely Lipschitz continuous on  $\mathcal{X}$ . Since  $\mathcal{X}$  is compact, by the stochastic Ascoli lemma (see, for example, Billingsley (2013)), it follows that (11.4.3) holds. ■

**Theorem 60** Consider a fixed-domain infill asymptotics framework such that for any  $\mathbf{x} \in \mathcal{X}$ , there exists  $\mathbf{x}_n \in \mathbf{X}_n$  for  $n \geq 1$  satisfying (11.4.2) and that points of the form  $\mathbf{x}_n + \mathbf{h}_n \in \mathbf{X}_n$ , where  $\mathbf{h}_n \rightarrow \mathbf{0}$ , as  $n \rightarrow \infty$ .

For  $\mathbf{x} = (x_1, \dots, x_d)$  and  $\mathbf{y} = (y_1, \dots, y_d)$ , let the correlation function be such that

$$\frac{\partial^4 c(\mathbf{x}^*, \mathbf{y}^*)}{\partial x_i \partial y_i \partial x_j \partial y_j} = \frac{\partial^4 c(\mathbf{x}, \mathbf{y})}{\partial x_i \partial y_i \partial x_j \partial y_j} \Big|_{\mathbf{x}=\mathbf{x}^*, \mathbf{y}=\mathbf{y}^*}$$

exists for all  $\mathbf{x}^*, \mathbf{y}^* \in \mathcal{X}$  and is Lipschitz continuous on  $\mathcal{X} \times \mathcal{X}$  for  $i, j = 1, \dots, d$ .

Then, for almost all sequences  $\{\mathbf{g}'_n(\cdot)\}_{n=1}^\infty$ ,

$$\sup_{\mathbf{x} \in \mathcal{X}} \|\mathbf{g}'_n(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\|_d \rightarrow 0, \text{ as } n \rightarrow \infty. \quad (11.4.12)$$

**Proof.** Note that for  $i = 1, \dots, d$ , pointwise convergence of  $g'_{in}(\cdot)$  to  $f'_i(\cdot)$  as  $n \rightarrow \infty$ , is already shown by (11.4.8), in the proof of Theorem 59. Hence, if we can show that for  $i, j = 1, \dots, d$ , the second order partial derivatives  $|g''_{ijn}(\cdot)|$  are uniformly bounded on  $\mathcal{X}$  as  $n \rightarrow \infty$ , then this would imply that  $g'_{in}(\cdot)$  are almost surely Lipschitz continuous on  $\mathcal{X}$  for large enough  $n$ . Since  $\mathcal{X}$  is compact, this would then imply by the stochastic Ascoli result that  $\sup_{\mathbf{x} \in \mathcal{X}} |g'_{in}(\mathbf{x}) - f'_i(\mathbf{x})| \rightarrow 0$ , almost surely, as  $n \rightarrow \infty$ , for

$i = 1, \dots, d$ , which is equivalent to (11.4.12). Hence, in the rest of the proof, we show that  $\limsup_{n \rightarrow \infty} \sup_{\mathbf{x} \in \mathcal{X}} |g''_{ijn}(\mathbf{x})| < \infty$ .

As before, let us fix any  $\mathbf{x} \in \mathcal{X}$ . Hence, by hypothesis, there exists  $\mathbf{x}_n \in \mathbf{X}_n$  for  $n \geq 1$ , such that  $\mathbf{x}_n \rightarrow \mathbf{x}$  as  $n \rightarrow \infty$ . Also let  $\mathbf{h}_n \rightarrow \mathbf{0}$ , as  $n \rightarrow \infty$ . Using Taylor's series expansion of  $g_n(\mathbf{x}_n + \mathbf{h}_n)$  up to the second order we obtain

$$g_n(\mathbf{x}_n + \mathbf{h}_n) = g_n(\mathbf{x}_n) + \mathbf{h}_n^T \mathbf{g}'_n(\mathbf{x}_n) + \frac{\mathbf{h}_n^T \mathbf{g}''_n(\mathbf{x}_n^*) \mathbf{h}_n}{2}, \quad (11.4.13)$$

where  $\mathbf{x}_n^*$  lies on the line joining  $\mathbf{x}_n$  and  $\mathbf{x}_n + \mathbf{h}_n$ . Existence and sample path continuity of the second order partial derivatives  $g''_{ijn}(\cdot)$  for  $i, j = 1, \dots, d$ , which constitute the elements of the matrix  $\mathbf{g}''_n(\cdot)$ , are guaranteed by the hypothesis of existence and Lipschitz continuity of  $\frac{\partial^4 c(\mathbf{x}^*, \mathbf{y}^*)}{\partial x_i \partial y_i \partial x_j \partial y_j}$ .

Now, given  $i \in \{1, \dots, d\}$ , let  $\mathbf{h}_{in}$  denote the vector with  $h_n$  at the  $i$ -th co-ordinate and 0 at the remaining co-ordinates. Then (11.4.13) reduces to

$$g_n(\mathbf{x}_n + \mathbf{h}_{in}) = g_n(\mathbf{x}_n) + h_n g'_{in}(\mathbf{x}_n) + \frac{h_n^2}{2} g''_{iin}(\mathbf{x}_{i1n}^*), \quad (11.4.14)$$

where  $\mathbf{x}_{i1n}^*$  lies on the line joining  $\mathbf{x}_n$  and  $\mathbf{x}_n + \mathbf{h}_{in}$ . Similar arguments also yield

$$g_n(\mathbf{x}_n - \mathbf{h}_{in}) = g_n(\mathbf{x}_n) - h_n g'_{in}(\mathbf{x}_n) + \frac{h_n^2}{2} g''_{iin}(\mathbf{x}_{i2n}^*), \quad (11.4.15)$$

where  $\mathbf{x}_{i2n}^*$  lies on the line joining  $\mathbf{x}_n - \mathbf{h}_{in}$  and  $\mathbf{x}_n$ . From (11.4.15) we obtain  $h_n g'_{in}(\mathbf{x}_n) = g_n(\mathbf{x}_n) - g_n(\mathbf{x}_n - \mathbf{h}_{in}) + \frac{h_n^2}{2} g''_{iin}(\mathbf{x}_{i2n}^*)$ , which we substitute in (11.4.14) to obtain  $g_n(\mathbf{x}_n + \mathbf{h}_{in}) = 2g_n(\mathbf{x}_n) - g_n(\mathbf{x}_n - \mathbf{h}_{in}) + \frac{h_n^2}{2} (g''_{iin}(\mathbf{x}_{i1n}^*) + g''_{iin}(\mathbf{x}_{i2n}^*))$ . Thus, denoting the first



and second order partial derivatives of  $f(\cdot)$  by  $f'_i(\cdot)$  and  $f''_{ij}(\cdot)$ , we obtain, almost surely,

$$\begin{aligned} \frac{1}{2} (g''_{iin}(\mathbf{x}_{i1n}^*) + g''_{iin}(\mathbf{x}_{i2n}^*)) &= \frac{g_n(\mathbf{x}_n + \mathbf{h}_{in}) - 2g_n(\mathbf{x}_n) + g_n(\mathbf{x}_n - \mathbf{h}_{in})}{h_n^2} \\ &= \frac{f(\mathbf{x}_n + \mathbf{h}_{in}) - 2f(\mathbf{x}_n) + f(\mathbf{x}_n - \mathbf{h}_{in})}{h_n^2} \\ &= \frac{1}{2} (f''_{ii}(\mathbf{x}_{i3n}^*) + f''_{ii}(\mathbf{x}_{i4n}^*)), \end{aligned} \quad (11.4.16)$$

where  $\mathbf{x}_{i3n}^*$  lies on the line joining  $\mathbf{x}_n$  and  $\mathbf{x}_n + \mathbf{h}_{in}$  and  $\mathbf{x}_{i4n}^*$  lies on the line joining  $\mathbf{x}_n - \mathbf{h}_{in}$  and  $\mathbf{x}_n$ . Now, since for  $k = 1, 2, 3, 4$ ,  $\mathbf{x}_{ikn}^* \rightarrow \mathbf{x}$  as  $n \rightarrow \infty$ , and since  $g''_{iin}(\cdot)$  is continuous for all  $n \geq 1$  and  $f''_{ii}(\cdot)$  is also continuous, it is easy to see that for  $k = 1, 2$ ,  $g''_{iin}(\mathbf{x}_{ikn}^*) \sim g''_{iin}(\mathbf{x})$  and  $f''_{ii}(\mathbf{x}_{ikn}^*) \sim f''_{ii}(\mathbf{x})$  for  $k = 3, 4$ , as  $n \rightarrow \infty$ , where for any two sequences  $a_n$  and  $b_n$ ,  $a_n \sim b_n$  stands for  $a_n/b_n \rightarrow 1$  as  $n \rightarrow \infty$ . An implicit assumption in the above arguments on asymptotic equivalence is that  $f''_{ii}(\mathbf{x}) \neq 0$  and  $g''_{iin}(\mathbf{x}) \rightarrow 0$  almost surely, as  $n \rightarrow \infty$ . Hence, taking limits of both sides of (11.4.16) yields, for each  $\mathbf{x} \in \mathcal{X}$ ,

$$\lim_{n \rightarrow \infty} g''_{iin}(\mathbf{x}) = f''_{ii}(\mathbf{x}) \text{ almost surely,} \quad (11.4.17)$$

proving pointwise convergence of  $g''_{iin}(\cdot)$  to  $f''_{ii}(\cdot)$  almost surely as  $n \rightarrow \infty$ , for each  $i = 1, \dots, d$  when  $f''_{ii}(\mathbf{x}) \neq 0$  and  $g''_{iin}(\mathbf{x}) \rightarrow 0$  almost surely, as  $n \rightarrow \infty$ . Note that if  $f''_{ii}(\mathbf{x}) = 0$  and  $g''_{iin}(\mathbf{x}) \rightarrow 0$  almost surely as  $n \rightarrow \infty$ , then (11.4.17) holds trivially. In other words, (11.4.17) holds for all  $\mathbf{x} \in \mathcal{X}$ .

Now, in (11.4.13), let  $\mathbf{h}_{ijn}$  be the vector with  $h_n$  at the  $i$ -th and  $j$ -th co-ordinates and zero elsewhere. Then (11.4.13) boils down to

$$g_n(\mathbf{x}_n + \mathbf{h}_{ijn}) = g_n(\mathbf{x}_n) + h_n (g'_{in}(\mathbf{x}_n) + g'_{jn}(\mathbf{x}_n)) + \frac{h_n^2}{2} (g''_{iin}(\mathbf{x}_{ijn}^*) + g''_{jjn}(\mathbf{x}_{ijn}^*) + 2g''_{ijn}(\mathbf{x}_{ijn}^*)), \quad (11.4.18)$$

where  $\mathbf{x}_{ijn}^*$  lies on the line joining  $\mathbf{x}_n$  and  $\mathbf{x}_n + \mathbf{h}_{ijn}$ . Substituting  $h_n g'_{in}(\mathbf{x}_n) = g_n(\mathbf{x}_n) - g_n(\mathbf{x}_n - \mathbf{h}_{in}) + \frac{h_n^2}{2} g''_{iin}(\mathbf{x}_{in}^*)$  and  $h_n g'_{jn}(\mathbf{x}_n) = g_n(\mathbf{x}_n) - g_n(\mathbf{x}_n - \mathbf{h}_{jn}) + \frac{h_n^2}{2} g''_{jjn}(\mathbf{x}_{jn}^*)$  in

(11.4.18), where  $\mathbf{x}_{in}^*$  lies on the line joining  $\mathbf{x}_n - \mathbf{h}_{in}$  and  $\mathbf{x}_n$ , and  $\mathbf{x}_{jn}^*$  lies on the line joining  $\mathbf{x}_n - \mathbf{h}_{jn}$  and  $\mathbf{x}_n$ , we obtain

$$\begin{aligned} g_n(\mathbf{x}_n + \mathbf{h}_{ijn}) &= 3g_n(\mathbf{x}_n) - g_n(\mathbf{x}_n - \mathbf{h}_{in}) - g_n(\mathbf{x}_n - \mathbf{h}_{jn}) \\ &+ \frac{h_n^2}{2} [(g''_{iin}(\mathbf{x}_{in}^*) + g''_{iin}(\mathbf{x}_{ijn}^*)) + (g''_{jjn}(\mathbf{x}_{jn}^*) + g''_{jjn}(\mathbf{x}_{ijn}^*))] + h_n^2 g''_{ijn}(\mathbf{x}_{ijn}^*). \end{aligned} \quad (11.4.19)$$

Now, continuity of  $g''_{iin}(\cdot)$  for  $n \geq 1$ , continuity of  $f''_{ii}(\cdot)$ , and (11.4.17) imply that as  $n \rightarrow \infty$ ,  $g''_{iin}(\mathbf{x}_{in}^*) \sim g''_{iin}(\mathbf{x}_n) \sim g''_{iin}(\mathbf{x}) \sim f''_{ii}(\mathbf{x}) \sim f''_{ii}(\mathbf{x}_n)$ , almost surely. Similarly,  $g''_{iin}(\mathbf{x}_{ijn}^*) \sim f''_{ii}(\mathbf{x}_n)$ ,  $g''_{jjn}(\mathbf{x}_{jn}^*) \sim f''_{jj}(\mathbf{x}_n)$  and  $g''_{jjn}(\mathbf{x}_{ijn}^*) \sim f''_{jj}(\mathbf{x}_n)$ , almost surely, as  $n \rightarrow \infty$ . These, applied to (11.4.19), yield

$$\begin{aligned} g''_{ijn}(\mathbf{x}_{ijn}^*) &= \frac{g_n(\mathbf{x}_n + \mathbf{h}_{ijn}) - 3g_n(\mathbf{x}_n) + g_n(\mathbf{x}_n - \mathbf{h}_{in}) + g_n(\mathbf{x}_n - \mathbf{h}_{jn})}{h_n^2} \\ &\quad - \frac{1}{2} [(g''_{iin}(\mathbf{x}_{in}^*) + g''_{iin}(\mathbf{x}_{ijn}^*)) + (g''_{jjn}(\mathbf{x}_{jn}^*) + g''_{jjn}(\mathbf{x}_{ijn}^*))] \\ &= \frac{f(\mathbf{x}_n + \mathbf{h}_{ijn}) - 3f(\mathbf{x}_n) + f(\mathbf{x}_n - \mathbf{h}_{in}) + f(\mathbf{x}_n - \mathbf{h}_{jn})}{h_n^2} \\ &\quad - \frac{1}{2} [(g''_{iin}(\mathbf{x}_{in}^*) + g''_{iin}(\mathbf{x}_{ijn}^*)) + (g''_{jjn}(\mathbf{x}_{jn}^*) + g''_{jjn}(\mathbf{x}_{ijn}^*))] \quad (11.4.20) \\ &\sim \frac{f(\mathbf{x}_n + \mathbf{h}_{ijn}) - 3f(\mathbf{x}_n) + f(\mathbf{x}_n - \mathbf{h}_{in}) + f(\mathbf{x}_n - \mathbf{h}_{jn})}{h_n^2} \\ &\quad - (f''_{ii}(\mathbf{x}_n) + f''_{jj}(\mathbf{x}_n)), \quad (11.4.21) \end{aligned}$$

almost surely, as  $n \rightarrow \infty$ . Again, implicit in (11.4.21) is the assumption that  $f''_{ii}(\mathbf{x}) \neq 0$ ,  $f''_{jj}(\mathbf{x}) \neq 0$ ,  $g''_{iin}(\mathbf{x}) \rightarrow 0$  and  $g''_{jjn}(\mathbf{x}) \rightarrow 0$  almost surely, as  $n \rightarrow \infty$ . However, if either or both of  $f''_{ii}(\mathbf{x}) = 0$  and  $g''_{iin}(\mathbf{x}) \rightarrow 0$  and  $f''_{jj}(\mathbf{x}) = 0$  and  $g''_{jjn}(\mathbf{x}) \rightarrow 0$ , then the relevant expressions in (11.4.20) and (11.4.21) converge to zero almost surely, as  $n \rightarrow \infty$ . Hence, the above asymptotic equivalence for  $g''_{ijn}(\mathbf{x}_{ijn}^*)$  remains valid for all  $\mathbf{x} \in \mathcal{X}$ .

Taylor's series expansion of  $f(\mathbf{x}_n + \mathbf{h}_{ijn})$  in the same way as (11.4.18), where  $\mathbf{x}_{ijn}^*$  must be replaced with some  $\mathbf{x}_{ijn}^{**}$  lying on the line joining  $\mathbf{x}_n$  and  $\mathbf{x}_n + \mathbf{h}_{ijn}$ , yields the

same asymptotic expression (11.4.21) for  $f''_{ij}(\mathbf{x}_{ijn}^{**})$ . In other words, we have

$$g''_{ijn}(\mathbf{x}_{ijn}^*) \sim f''_{ij}(\mathbf{x}_{ijn}^{**}), \text{ almost surely, as } n \rightarrow \infty. \quad (11.4.22)$$

Hence, taking limit of both sides of (11.4.22), using continuity of  $g''_{ijn}(\cdot)$  for  $n \geq 1$  and continuity of  $f''_{ij}(\cdot)$  gives

$$\lim_{n \rightarrow \infty} g''_{ijn}(\mathbf{x}) = f''_{ij}(\mathbf{x}), \text{ almost surely, for all } \mathbf{x} \in \mathcal{X}, \quad (11.4.23)$$

Since  $g''_{ijn}(\cdot)$  is almost surely continuous for all  $n \geq 1$  and  $f''_{ij}(\cdot)$  is continuous, with  $\mathcal{X}$  being compact, the pointwise convergence result (11.4.23) lets us conclude, using Lemma 58, that  $\limsup_{n \rightarrow \infty} \sup_{\mathbf{x} \in \mathcal{X}} |g''_{ijn}(\mathbf{x})| < \infty$ . ■

Theorems 59 and 60 prove almost sure uniform convergence of  $g_n(\cdot)$  and  $\mathbf{g}'_n(\cdot)$  to  $f(\cdot)$  and  $\mathbf{f}'(\cdot)$ , respectively. However, the rates of convergence are not provided by these theorems. Further fine-tuning the structure of the set of input points  $\mathbf{X}_n$  helps achieve desired rates of convergence, as we show next in Theorems 61 and 62.

**Theorem 61** *Let  $\mathcal{X} = \prod_{i=1}^d \mathcal{X}_i$ , where, for  $i = 1, \dots, d$ ,  $\mathcal{X}_i$  are compact subsets of  $\mathbb{R}$ . For each  $i \in \{1, \dots, d\}$ , let  $x_{1i} < x_{2i} < \dots < x_{\tilde{n}_i i}$  be an ordered set of points partitioning  $\mathcal{X}_i$ , with  $h_i = \max_{1 \leq j \leq \tilde{n}_i - 1} (x_{j+1i} - x_{ji})$ . For  $i = 1, \dots, d$ , and for  $j = 1, \dots, \tilde{n}_i$ , let input points of the form  $(x_1^*, \dots, x_{i-1}^*, x_{ji}, x_{i+1}^*, \dots, x_d^*)$  belong to  $\mathbf{X}_n$ , where  $(x_1^*, \dots, x_{i-1}^*, x_{i+1}^*, \dots, x_d^*) \in \prod_{j \neq i} \mathcal{X}_j$  may be arbitrary.*

For  $\mathbf{x} = (x_1, \dots, x_d)$  and  $\mathbf{y} = (y_1, \dots, y_d)$ , let the correlation function be such that

$$\frac{\partial^4 c(\mathbf{x}^*, \mathbf{y}^*)}{\partial x_i \partial y_i \partial x_j \partial y_j} = \frac{\partial^4 c(\mathbf{x}, \mathbf{y})}{\partial x_i \partial y_i \partial x_j \partial y_j} \Big|_{\mathbf{x}=\mathbf{x}^*, \mathbf{y}=\mathbf{y}^*}$$

exists for all  $\mathbf{x}^*, \mathbf{y}^* \in \mathcal{X}$  and is Lipschitz continuous on  $\mathcal{X} \times \mathcal{X}$  for  $i, j = 1, \dots, d$ .

Then, letting  $h = \max_{1 \leq i \leq d} h_i$ , the following holds for almost all sequences  $\{\mathbf{g}'_n(\cdot)\}_{n=1}^\infty$

with the above forms of the input points:

$$\sup_{\mathbf{x} \in \mathcal{X}} \|\mathbf{g}'_n(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\|_d = O\left(\sqrt{h}\right), \text{ as } n \rightarrow \infty. \quad (11.4.24)$$

The constant associated with  $O\left(\sqrt{h}\right)$  depends only upon  $d$  and  $f(\cdot)$ .

**Proof.** For any  $i \in \{1, \dots, d\}$ , and for arbitrary  $\mathbf{X}_{-i}^* = (x_1^*, \dots, x_{i-1}^*, x_{i+1}^*, \dots, x_d^*)^T \in \prod_{j \neq i} \mathcal{X}_j$ , for any  $x \in \mathcal{X}_i$ , let

$$g_{in}(x|\mathbf{X}_{-i}^*) = g_n(x_1^*, \dots, x_{i-1}^*, x, x_{i+1}^*, \dots, x_d^*); \quad (11.4.25)$$

$$f_i(x|\mathbf{X}_{-i}^*) = f(x_1^*, \dots, x_{i-1}^*, x, x_{i+1}^*, \dots, x_d^*). \quad (11.4.26)$$

Since  $x \in \mathcal{X}_i$ , it must belong to some interval of the form  $[x_{ji}, x_{\overline{j+1}i}]$ , for some  $j \in \{1, 2, \dots, \tilde{n}_i - 1\}$ . Let us fix that  $j$ . For  $y = x_{ji}$  and  $y = x_{\overline{j+1}i}$ ,  $g_{in}(y|\mathbf{X}_{-i}^*) = f_i(y|\mathbf{X}_{-i}^*)$  by interpolation property of the posterior Gaussian process, assuming that  $(x_1^*, \dots, x_{i-1}^*, y, x_{i+1}^*, \dots, x_d^*) \in \mathbf{X}_n$  for  $y = x_{ji}$  and  $y = x_{\overline{j+1}i}$ . That is,  $g_i(y|\mathbf{X}_{-i}^*) - f_i(y|\mathbf{X}_{-i}^*) = 0$  for  $y = x_{ji}$  and  $y = x_{\overline{j+1}i}$ . Hence, by Rolle's theorem,  $g'_{in}(u|\mathbf{X}_{-i}^*) - f'_i(u|\mathbf{X}_{-i}^*) = 0$ , for some  $u \in (x_{ji}, x_{\overline{j+1}i})$ . This permits the following representation:

$$g'_{in}(x|\mathbf{X}_{-i}^*) - f'_i(x|\mathbf{X}_{-i}^*) = \int_u^x (g''_{in}(v|\mathbf{X}_{-i}^*) - f''_i(v|\mathbf{X}_{-i}^*)) dv, \quad (11.4.27)$$

The hypothesis of Lipschitz continuity of the 4-th order mixed partial derivatives of the correlation function ensures existence and sample path continuity of  $g''_{in}(v|\mathbf{X}_{-i}^*)$ . Hence,

by the Cauchy-Schwartz inequality we obtain from (11.4.27), the following:

$$\begin{aligned}
 & |g'_{in}(x|\mathbf{X}_{-i}^*) - f'_i(x|\mathbf{X}_{-i}^*)| \\
 & \leq \left[ \int_u^x (g''_{in}(v|\mathbf{X}_{-i}^*) - f''_i(v|\mathbf{X}_{-i}^*))^2 dv \right]^{1/2} \times |x - u|^{1/2} \\
 & \leq \left[ \int_{\mathcal{X}_i} (g''_{in}(v|\mathbf{X}_{-i}^*) - f''_i(v|\mathbf{X}_{-i}^*))^2 dv \right]^{1/2} \times h_i^{1/2} \\
 & \leq \sup_{\mathbf{X}_{-i}^* \in \prod_{j \neq i} \mathcal{X}_j} \left[ \int_{\mathcal{X}_i} (g''_{in}(v|\mathbf{X}_{-i}^*) - f''_i(v|\mathbf{X}_{-i}^*))^2 dv \right]^{1/2} \times h^{1/2}. \quad (11.4.28)
 \end{aligned}$$

Now, since the hypotheses of this theorem constitute a special case of Theorem 60, the result of almost sure uniform boundedness of  $|g''_{in}(\cdot)|$  as  $n \rightarrow \infty$ , is valid here. Specifically, from the proof of Theorem 60 in this case it holds that  $\lim_{n \rightarrow \infty} \sup_{(u, \mathbf{X}_{-i}^*) \in \mathcal{X}} |g''_{in}(v|\mathbf{X}_{-i}^*)| < \infty$ . This, along with continuity of  $f''_i(v|\mathbf{X}_{-i}^*)$  and compactness of  $\mathcal{X}$ , shows that (11.4.28) is  $O(\sqrt{h})$ . Hence, switching to our usual notation, it follows that

$$\sup_{\mathbf{x} \in \mathcal{X}} (g'_{in}(\mathbf{x}) - f'_i(\mathbf{x}))^2 = O(h), \text{ almost surely, as } n \rightarrow \infty. \quad (11.4.29)$$

Since  $\sup_{\mathbf{x} \in \mathcal{X}} \sum_{i=1}^d (g'_{in}(\mathbf{x}) - f'_i(\mathbf{x}))^2 \leq \sum_{i=1}^d \sup_{\mathbf{x} \in \mathcal{X}} (g'_{in}(\mathbf{x}) - f'_i(\mathbf{x}))^2$ , it follows from (11.4.29) that

$$\sup_{\mathbf{x} \in \mathcal{X}} \|g'_n(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\|_d = O(\sqrt{h}), \text{ almost surely, as } n \rightarrow \infty,$$

proving (11.4.24). Note that the constant associated with  $O(\sqrt{h})$  above depends only upon  $d$  and  $f$ . ■

The following result holds as a consequence of Theorem 61.

**Theorem 62** *Under the conditions of Theorem 61, the following holds with the forms of the input points as specified in Theorem 61: for almost all sequences  $\{g_n(\cdot)\}_{n=1}^\infty$ ,*

$$\sup_{\mathbf{x} \in \mathcal{X}} |g_n(\mathbf{x}) - f(\mathbf{x})| = O(h^{3/2}), \text{ as } n \rightarrow \infty. \quad (11.4.30)$$

The constant associated with  $O(h^{3/2})$  depends only upon  $d$  and  $f$ .

**Proof.** As in the proof of Theorem 61, for any  $x \in \mathcal{X}_i$ , it must belong to some interval of the form  $[x_{ji}, x_{\overline{j+1}i}]$ , for some  $j \in \{1, 2, \dots, \tilde{n}_i - 1\}$ . Let us fix that  $j$ . Now,  $g_{in}(x_{ji}|\mathbf{X}_{-i}^*) = f_i(x_{ji}|\mathbf{X}_{-i}^*)$  almost surely, by interpolation property of the posterior process  $g_n(\cdot)$ , assuming that  $(x_1^*, \dots, x_{i-1}^*, x_{ji}, x_{i+1}^*, \dots, x_d^*) \in \mathbf{X}_n$ . Hence,

$$g_{in}(x|\mathbf{X}_{-i}^*) - f_i(x|\mathbf{X}_{-i}^*) = \int_{x_{ji}}^x (g'_{in}(v|\mathbf{X}_{-i}^*) - f'_i(v|\mathbf{X}_{-i}^*)) dv. \quad (11.4.31)$$

The Cauchy-Schwartz inequality applied to (11.4.31) gives

$$|g_{in}(x|\mathbf{X}_{-i}^*) - f_i(x|\mathbf{X}_{-i}^*)| \leq \left[ \int_{x_{ji}}^x (g'_{in}(v|\mathbf{X}_{-i}^*) - f'_i(v|\mathbf{X}_{-i}^*))^2 dv \right]^{1/2} \times |x - x_{ji}|^{1/2}. \quad (11.4.32)$$

From (11.4.28) it follows that the integral on the right hand side of (11.4.32) is  $O(h) \times |x - x_{ji}|$ , almost surely, as  $n \rightarrow \infty$ . Recall that the constant associated with  $O(h) \times |x - x_{ji}|$  depends only upon  $d$  and  $f$ . Since  $|x - x_{ji}|$  is bounded above by  $h$ , it follows that the right hand side of (11.4.32) is  $O(h^{3/2})$ , almost surely, as  $n \rightarrow \infty$ . Switching to the usual notation, it is seen that (11.4.30) holds. ■

**Remark 63** As  $h \rightarrow 0$ ,  $\mathbf{g}'(\cdot)$  uniformly converges to  $\mathbf{f}'(\cdot)$  at the rate  $h^{1/2}$  and  $g(\cdot)$  uniformly converges to  $f(\cdot)$  at the rate  $h^{3/2}$ , almost surely with respect to their posteriors.

**Remark 64** In Theorems 59, 60, 61 and 62 we have referred to the posterior (11.2.22), which corresponds to a linear mean structure of the Gaussian process prior and conjugate priors for  $\beta$  and  $\sigma^2$ . However, as can be seen from the proofs, both the theorems continue to hold for any mean function that has continuous second order mixed partial derivatives and any prior on the parameters, including the parameters of the correlation function such that the posteriors of the parameters are proper.

**Remark 65** *The hypotheses of Theorems 61 and 62 require input points of the form  $(x_1^*, \dots, x_{i-1}^*, x_{ji}, x_{i+1}^*, \dots, x_d^*)$  to belong to  $\mathbf{D}_n$  for  $i = 1, \dots, d$ , and for  $j = 1, \dots, \tilde{n}_i$ , where  $(x_1^*, \dots, x_{i-1}^*, x_{i+1}^*, \dots, x_d^*) \in \prod_{j \neq i} \mathcal{X}_j$  may be chosen arbitrarily. Now observe that if  $\mathcal{X}_i = [a, b]$ , for  $i = 1, \dots, d$ , for some  $a < b$ , then we can set  $\tilde{n}_i = n$  and  $h_i = h$ , for  $i = 1, \dots, d$ . In such cases, inclusion of the set of  $n^d$  points  $\{(x_{j_1}, \dots, x_{j_d}) : j_1, \dots, j_d \in \{1, \dots, n\}\}$  in  $\mathbf{D}_n$  is sufficient for Theorems 61 and 62 to hold. However, when  $d$  and  $n$  are even moderately large,  $n^d$  is an extremely large number, which would prohibit computation of  $\Sigma_{22}^{-1}$ , and hence computation of the posterior of  $\mathbf{g}'(\cdot)$ . Hence, for practical purposes it makes sense to refer to the general setup of Theorems 59 and 60.*

## 11.5 Algorithm for optimization with the Gaussian process derivative method

We now propose a general methodology for function optimization, which judiciously exploits the posterior form (11.3.3). Without loss of generality, we consider the minimization problem for notational convenience. In a nutshell, the initial stage (say, the 0-th stage) of the methodology involves simulations from  $\pi(\mathbf{x}^* | \mathbf{g}'(\mathbf{x}^*) = \mathbf{0}, \mathbf{D}_n)$  satisfying  $\|\mathbf{f}'(\cdot)\|_d < \epsilon$  for some  $\epsilon > 0$  and  $\Sigma''(\cdot) > 0$ . In the subsequent stages  $k = 1, 2, \dots$ , previous stage realizations satisfying  $\|\mathbf{f}'(\cdot)\|_d < \eta_k$ , where  $\eta_k \rightarrow 0$  as  $k \rightarrow \infty$ , are successively augmented with  $\mathbf{D}_n$  and realizations from the posterior associated with the augmented data are generated at each stage  $k$  by a judicious importance resampling strategy. As  $k \rightarrow \infty$ , the posteriors given the successively augmented data converge to the true optima.

The importance of the 0-th stage simulation algorithm for generating realizations from  $\pi(\mathbf{x}^* | \mathbf{g}'(\mathbf{x}^*) = \mathbf{0}, \mathbf{D}_n)$  satisfying the restrictions  $\|\mathbf{f}'(\cdot)\|_d < \epsilon$  for some  $\epsilon > 0$  and  $\Sigma''(\cdot) > 0$ , is enormous, particularly because the entire  $d$ -dimensional random variable  $\mathbf{x}^*$  must be updated in a single block to meet the restrictions. Traditional MCMC

algorithms are not known to be efficient in such problems as even for moderately large dimensions good proposal distributions are difficult to devise, and the acceptance rates can be poor, along with poor mixing properties. In this regard, the transformation based Markov Chain Monte Carlo (TMCMC) proposed by [Dutta and Bhattacharya \(2014\)](#) is an effective methodology. Indeed, TMCMC is designed to update all (or most of) the components of the high-dimensional random variable in a single block using appropriate deterministic transformations of some single (or low-dimensional) random variable. As such, this strategy drastically reduces effective dimensionality, which is responsible for maintaining good acceptance rates in spite of high dimensions. Good mixing properties can also be ensured by judiciously choosing the relevant “move-types”, and judicious mixtures of additive and multiplicative transformations usually lead to desired mixing properties. For details on TMCMC and its properties, see [Dutta and Bhattacharya \(2014\)](#), [Dey and Bhattacharya \(2016\)](#), [Dey and Bhattacharya \(2017\)](#), [Dey and Bhattacharya \(2019\)](#). As such, we recommend TMCMC for our optimization methodology.

We provide the detailed Bayesian optimization methodology below as Algorithm 1.



---

**Algorithm 1** Optimization with Gaussian process derivatives

---

- (1) First simulate  $N$  realizations  $\{\mathbf{x}_1^*, \dots, \mathbf{x}_N^*\}$  from  $\pi(\mathbf{x}^* | \mathbf{g}'(\mathbf{x}^*) = \mathbf{0}, \mathbf{D}_n)$  given by (11.3.3) using TMCMC, where the prior for  $\mathbf{x}^*$  is given by (11.3.4) for some pre-fixed  $\epsilon > 0$ . This includes simulations from posteriors associated with most plausible functions  $g(\cdot)$  satisfying  $\mathbf{g}'(\mathbf{x}^*) = \mathbf{0}$  for  $\mathbf{x}^* \in \mathcal{X}$ , given  $\mathbf{D}_n$ . That is,  $\{\mathbf{x}_i^*; i = 1, \dots, N\}$ , represents the set of solutions for  $\mathbf{g}'(\mathbf{x}^*) = \mathbf{0}$  for functions  $g(\cdot)$  that satisfy  $g(\mathbf{x}_i) = f(\mathbf{x}_i); i = 1, \dots, n$ . Thanks to the prior (11.3.4), these solutions further satisfy  $\|\mathbf{f}'(\mathbf{x}_i^*)\|_d < \epsilon$  and  $\mathbf{\Sigma}''(\mathbf{x}_i^*) > 0$ , for  $i = 1, \dots, N$ . Note that  $\mathbf{\Sigma}''(\mathbf{x}^*) > 0$  can be checked by computing the eigenvalues and checking if all the eigenvalues are positive. But a more efficient alternative is to check if Cholesky decomposition of  $\mathbf{\Sigma}''(\mathbf{x}^*)$  is possible, the information of which is provided by the subroutines of the BLAS and LAPACK libraries. We exploit the latter for our implementation.
- (2) For stages  $k = 1, 2, 3, \dots$ ,
- (i) For  $i = 1, \dots, N$ , compute importance weights proportional to

$$w_k(\mathbf{x}_i^*) = \begin{cases} 1 & \text{if } k = 1; \\ w_{k-1}(\mathbf{x}_i^*) \times \frac{\pi(\mathbf{g}'(\mathbf{x}_i^*) = \mathbf{0} | \mathbf{D}_{n+\sum_{j=0}^{k-1} n_j}, \mathbf{x}_i^*)}{\pi(\mathbf{g}'(\mathbf{x}_i^*) = \mathbf{0} | \mathbf{D}_{n+\sum_{j=0}^{k-2} n_j}, \mathbf{x}_i^*)} & \text{if } k \geq 2, \end{cases} \quad (11.5.1)$$

where  $n_0 = 0$ .

- (ii) Select a subsample  $\{\mathbf{x}_{i_1}^*, \dots, \mathbf{x}_{i_M}^*\}$  from  $\{\mathbf{x}_1^*, \dots, \mathbf{x}_N^*\}$  with probabilities proportional to  $w_k(\mathbf{x}_i^*); i = 1, \dots, N$ . Note that, as  $M \rightarrow \infty$  and  $N \rightarrow \infty$  such that  $M/N \rightarrow 0$ ,  $\mathbf{x}_{i_j}^*; j = 1, \dots, M$ , follow the distribution  $\pi(\mathbf{x}^* | \mathbf{g}'(\mathbf{x}^*) = \mathbf{0}, \mathbf{D}_{n+\sum_{j=1}^{k-1} n_j})$ . The recursively computed importance weights  $w_k(\mathbf{x}_i^*); i = 1, \dots, N$ , are expected to be stable, since for each stage  $k$ , for  $i = 1, \dots, N$ , the factors  $\frac{\pi(\mathbf{g}'(\mathbf{x}_i^*) = \mathbf{0} | \mathbf{D}_{n+\sum_{j=1}^{k-1} n_j}, \mathbf{x}_i^*)}{\pi(\mathbf{g}'(\mathbf{x}_i^*) = \mathbf{0} | \mathbf{D}_{n+\sum_{j=1}^{k-2} n_j}, \mathbf{x}_i^*)}$  in (11.5.1) are not expected to be very different from 1 if  $n_{k-1}$  is not significantly greater than zero. Note that the importance weights  $w_k(\mathbf{x}_i^*)$ , for  $i = 1, \dots, N$ , can be computed simultaneously on parallel processors. This, along with stability of the recursive formulation (11.5.1), is expected to make for an efficient computational strategy.
- (iii) For  $j = 1, \dots, M$ , check if  $\|\mathbf{f}'(\mathbf{x}_{i_j}^*)\|_d < \eta_k$ , where  $\eta_k \rightarrow 0$  as  $k \rightarrow \infty$ . If  $\mathbf{x}_{i_j}^*$  satisfies this condition, then  $\mathbf{x}_{i_j}^*$  is a realization from  $\pi(\mathbf{x}^* | \mathbf{g}'(\mathbf{x}^*) = \mathbf{0}, \mathbf{D}_{n+\sum_{j=1}^{k-1} n_j})$ , where the prior for  $\mathbf{x}^*$  is uniform on  $B(\eta_k)$ , the form of which is given by (11.3.5).
- (iv) Let  $n_k (\geq 0)$  realizations among the  $M$  realizations satisfy the condition  $\|\mathbf{f}'(\mathbf{x}_{i_j}^*)\|_d < \eta_k$ . Without loss of generality, assume that  $\mathbf{x}_{i_j}^*; j = 1, \dots, n_k$  are such realizations. Compute  $f(\mathbf{x}_{i_j}^*); j = 1, \dots, n_k$ , and augment  $(\mathbf{x}_{i_j}^*, f(\mathbf{x}_{i_j}^*)); j = 1, \dots, n_k$ , with  $\mathbf{D}_{n+\sum_{j=0}^{k-1} n_j}$  to form  $\mathbf{D}_{n+\sum_{j=0}^k n_j}$ .
- (v) Store the realizations  $\{\mathbf{x}_{i_j}^* : j = 1, \dots, n_k\}$ .
-

### 11.5.1 Further discussion of Algorithm 1

Algorithm 1 begins by simulating from the posterior of  $\mathbf{x}^*$  satisfying  $\|\mathbf{f}'(\mathbf{x}^*)\|_d < \epsilon$  and  $\Sigma''(\mathbf{x}^*) > 0$ . In the subsequent steps  $k \geq 1$ , the set of realizations  $\{\mathbf{x}_{i_j}^* : j = 1, \dots, n_k\}$  generated by importance resampling further satisfy  $\|\mathbf{f}'(\mathbf{x}_{i_j}^*)\|_d < \eta_k$  along with  $\Sigma''(\mathbf{x}_{i_j}^*) > 0$ , for  $j = 1, \dots, n_k$ . The implication is that,  $\epsilon$  may be chosen somewhat larger to achieve reasonably good TMCMC mixing acceptance rates. Indeed, if  $n$  is not large enough, then  $\mathbf{g}'$  is not expected to be sufficiently close to  $\mathbf{f}'$ , and hence for too small  $\epsilon$ ,  $\{\mathbf{x}^* : \|\mathbf{f}'(\mathbf{x}^*)\|_d < \epsilon\}$  would be too small a region to contain the solutions  $\{\mathbf{x}^* : \|\mathbf{g}'(\mathbf{x}^*)\|_d = 0\}$ , given  $\mathbf{D}_n$ . This would result in poor TMCMC mixing.

Once adequate mixing with reasonable acceptance rates are achieved with relatively small  $n$  and relatively large  $\epsilon$ , the subsequent steps increase the data size by augmenting the data with those values that satisfy  $\|\mathbf{f}'(\mathbf{x}_{i_j}^*)\|_d < \eta_k$ . Thus, in the subsequent steps, these data points help better approximate the region around the stationary points of  $f$  by the posterior, and enables more simulations from the region  $\|\mathbf{f}'(\mathbf{x}_{i_j}^*)\|_d < \eta_k$ , finally leading to convergence of the solutions  $\{\mathbf{x}^* : \mathbf{g}'_k(\mathbf{x}^*) = \mathbf{0}, \Sigma''(\mathbf{x}^*) > 0\}$  to  $\{\mathbf{x}^* : \mathbf{f}'(\mathbf{x}^*) = \mathbf{0}, \Sigma''(\mathbf{x}^*) > 0\}$ , almost surely, as  $k \rightarrow \infty$ , where  $\mathbf{g}'_k(\cdot)$  denotes any realization from the posterior of  $\mathbf{g}'(\cdot)$  given  $\mathbf{D}_{n+\sum_{j=0}^{k-1} n_j}$ . This intuition is formalized below as Theorem 66.

**Theorem 66** *Consider the setup of Theorem 60 (or more specifically, that of Theorem 61). Then, as  $k \rightarrow \infty$ , the set  $\{\mathbf{x}_{i_j}^* : j = 1, \dots, n_k\}$  of Algorithm 1 almost surely contains all the local minima of the objective function  $f(\cdot)$ , as  $M \rightarrow \infty$  and  $N \rightarrow \infty$  such that  $M/N \rightarrow 0$ .*

**Proof.** Note that at stage  $k$ , as  $M \rightarrow \infty$  and  $N \rightarrow \infty$  such that  $M/N \rightarrow 0$ ,  $\mathbf{x}_{i_j}^*$ ;  $j = 1, \dots, n_k$ , arise from  $\pi(\mathbf{x}^* | \mathbf{g}'(\mathbf{x}^*) = \mathbf{0}, \mathbf{D}_{n+\sum_{j=0}^{k-1} n_j})$ , subject to  $\|\mathbf{f}'(\mathbf{x}_{i_j}^*)\|_d < \eta_k$  and  $\Sigma''(\mathbf{x}_{i_j}^*) > 0$  for  $j = 1, \dots, n_k$ . These realizations are solutions of  $\mathbf{g}'_k(\mathbf{x}^*) = \mathbf{0}$  and  $\Sigma''(\mathbf{x}^*) > 0$  when the data observed is  $\mathbf{D}_{n+\sum_{j=0}^{k-1} n_j}$ . By Theorem 60 (or more specifically

by Theorem 61), as  $k \rightarrow \infty$  (equivalently, as  $h \rightarrow 0$  in Theorem 61),  $\mathbf{g}'_k(\cdot)$  uniformly converges to  $\mathbf{f}'(\cdot)$  almost surely. Hence, as  $k \rightarrow \infty$ ,

$$\{\mathbf{x}^* : \mathbf{g}'_k(\mathbf{x}^*) = \mathbf{0}, \|\mathbf{f}'(\mathbf{x}^*)\|_d < \eta_k, \Sigma''(\mathbf{x}^*) > 0\} \rightarrow \{\mathbf{x}^* : \mathbf{f}'(\mathbf{x}^*) = \mathbf{0}, \Sigma''(\mathbf{x}^*) > 0\}, \quad (11.5.2)$$

almost surely.

Due to (11.5.2), as  $k \rightarrow \infty$ , the set  $\{\mathbf{x}_{i_j}^* : j = 1, \dots, n_k\}$  contains all the local minima of the objective function  $f(\cdot)$ , as  $M \rightarrow \infty$  and  $N \rightarrow \infty$  such that  $M/N \rightarrow 0$ . ■

**Remark 67** *In step (2) (iv) of Algorithm 1 we have suggested augmentation of all realizations  $(\mathbf{x}_{i_j}^*, f(\mathbf{x}_{i_j}^*))$  satisfying  $\|\mathbf{f}'(\mathbf{x}_{i_j}^*)\|_d < \eta_k$  to the existing data  $\mathbf{D}_{n+\sum_{j=0}^{k-1} n_j}$ . In practice, augmentation of all such realizations may enlarge the dataset to such an extent that invertibility of the resultant  $\Sigma_{22}$  may be infeasible or numerically unstable, so that computation of the corresponding posterior densities of  $\mathbf{g}'(\mathbf{x}_i^*) = \mathbf{0}$ , and hence the importance weights (11.5.1), may not yield reliable results. Hence in practice, as a rule of thumb, we recommend augmenting at most 5 realizations satisfying  $\|\mathbf{f}'(\mathbf{x}_{i_j}^*)\|_d < \eta_k$ , which we consider in all our applications.*

**Remark 68** *As the stage number  $k$  in step (2) of Algorithm 1 increases,  $n_k$  decreases. Hence, in practice,  $n_k$  will be zero after some large enough  $k$ . When  $d$  is large, due to the curse of dimensionality, only the first few stages are expect to yield positive  $n_k$ .*

**Remark 69** *In step (1) of Algorithm 1, that is, in the TMCMC step, as well as in any stage  $k$  of step (2) of the algorithm provided that  $n_k$  is sufficiently large, desired credible regions of the respective posterior distributions of  $\mathbf{x}^*$  can be obtained. These quantify the uncertainty in a posteriori learning about the optima, given  $\|\mathbf{f}'(\cdot)\|_d < \epsilon$  or  $\|\mathbf{f}'(\cdot)\|_d < \eta_k$ . As  $k \rightarrow \infty$ , the uncertainty decreases, and the credibility regions shrink to the points representing the true optima. However, as mentioned in Remark 63, in practice, particularly for large  $d$ ,  $n_k$  would be zero for most stages  $k$ , which would preclude computation of credible regions for most stages.*

**Remark 70** *Note that if it is known beforehand that there is a single global minimum of  $f(\cdot)$  on  $\mathcal{X}$ , then step (2) of Algorithm 1 is not required. It is then sufficient to report  $\mathbf{x}_{i^*}^*$  as the (approximate) minimizer of  $f(\cdot)$ , where  $i^* = \min\{f(\mathbf{x}_i^*) : i = 1, \dots, N\}$ .*

## 11.6 Bayesian characterization of the number of local minima of the objective function with recursive posteriors

Steps (2) (iii) and (2) (iv) can be combined to obtain a Bayesian characterization of the number of local minima of the objective function. In this regard, for stage  $j$ , let us define  $Y_j = \sum_{r=1}^M I_{B(\eta_j)}(\mathbf{x}_{i_r}^*)$ , where  $B(\eta_j)$  is given by (11.3.5). Thus,  $\{Y_j = m\}$  with probability  $p_{mj}$ , for  $m = 0, 1, 2, \dots, M$ . Since  $M \rightarrow \infty$ , we allow  $Y_j$  to take values on the entire set of non-negative integers. That is, we set

$$P(Y_j = m) = p_{mj}; \quad m = 0, 1, 2, \dots, \quad (11.6.1)$$

the infinite-dimensional multinomial distribution, where  $0 \leq p_{mj} \leq 1$  for  $m = 0, 1, 2, \dots$  and  $j \geq 1$ . Further,  $\sum_{m=0}^{\infty} p_{mj} = 1$  for all  $j \geq 1$ . We assume that the true probabilities  $p_{m0} \in [0, 1]$ ;  $m = 0, 1, 2, \dots$ , such that  $\sum_{m=0}^{\infty} p_{m0} = 1$ , are unknown. Indeed, if  $f(\cdot)$  has finite number of local minima, then there must exist  $\tilde{m} \geq 0$  such that  $p_{\tilde{m}0} = 1$  and  $p_{m0} = 0$  for  $m \neq \tilde{m}$ . For infinite number of local minima, we must have  $p_{m0} = 0$  for any finite integer  $m \geq 0$ .

We adopt the approach of Section 4.3 based on Dirichlet process to obtain the posterior distribution of the infinite set of parameters  $\{p_{mk}; m = 0, 1, 2, \dots\}$ . In particular, we obtain a recursive Bayesian methodology of the same form as in Section 4.3, albeit with a definition of  $Y_j$  that is different from that of Section 4.3.

Continuing the recursive process as before we obtain that, at the  $k$ -th stage, the

posterior of  $P_k$  is a Dirichlet process, given by

$$\pi(P_k|y_k) \sim DP\left(\sum_{j=1}^k \frac{1}{j^2} G + \sum_{j=1}^k \delta_{y_j}\right). \quad (11.6.2)$$

It follows from (11.6.2) that

$$E(p_{mk}|y_k) = \frac{\frac{1}{2^m} \sum_{j=1}^k \frac{1}{j^2} + \sum_{j=1}^k I(y_j = m)}{\sum_{j=1}^k \frac{1}{j^2} + k}; \quad (11.6.3)$$

$$Var(p_{mk}|y_k) = \frac{\left(\sum_{j=1}^k \frac{1}{j^2} + \sum_{j=1}^k I(y_j = m)\right) \left(\left(1 - \frac{1}{2^m}\right) \sum_{j=1}^k \frac{1}{j^2} + k - \sum_{j=1}^k I(y_j = m)\right)}{\left(\sum_{j=1}^k \frac{1}{j^2} + k\right)^2 \left(\sum_{j=1}^k \frac{1}{j^2} + k + 1\right)}. \quad (11.6.4)$$

The theorem below characterizes the number of local minima of the objective function  $f(\cdot)$  in terms of the limit of the marginal posterior probabilities of  $p_{mk}$ , denoted by  $\pi_m(\cdot|y_k)$ , as  $k \rightarrow \infty$ .

**Theorem 71** *Assume the conditions of Theorem 61, and in Algorithm 1, assume that  $M \rightarrow \infty$  and  $N \rightarrow \infty$  such that  $M/N \rightarrow 0$ . Then  $f(\cdot)$  has  $\tilde{m} (\geq 0)$  local minima if and only if*

$$\pi_{\tilde{m}}(\mathcal{N}_1|y_k) \rightarrow 1, \text{ almost surely with respect to the posterior (11.3.3),} \quad (11.6.5)$$

as  $k \rightarrow \infty$ . In the above,  $\mathcal{N}_1$  is any neighborhood of 1 (one).

**Proof.** First, let us assume that  $f(\cdot)$  has  $\tilde{m} (\geq 0)$  local minima. Then, by Theorem 66, Algorithm 1 converges to the  $\tilde{m}$  local minima almost surely, as  $k \rightarrow \infty$ , provided that  $M \rightarrow \infty$  and  $N \rightarrow \infty$  such that  $M/N \rightarrow 0$ . Hence, there exists  $j_0 \geq 1$ , such that for  $j \geq j_0$ ,  $y_j = \tilde{m}_0$ . Thus, almost surely,  $I(y_j = \tilde{m}) = 1$ , for  $j \geq j_0$ . Consequently, it easily

follows from the forms (4.3.6) and (4.3.7) that almost surely, as  $k \rightarrow \infty$ ,

$$E(p_{\tilde{m}k}|y_k) \rightarrow 1, \quad \text{and} \quad (11.6.6)$$

$$\text{Var}(p_{\tilde{m}k}|y_k) = O\left(\frac{1}{k}\right) \rightarrow 0. \quad (11.6.7)$$

Now let  $\mathcal{N}_1$  denote any neighborhood of 1, and let  $\epsilon (> 0)$  be sufficiently small such that  $\mathcal{N}_1 \supseteq \{1 - p_{\tilde{m}k} < \epsilon\}$ . Then using Markov's inequality we obtain

$$\begin{aligned} \pi_{\tilde{m}}(\mathcal{N}_1|y_k) &\geq \pi_{\tilde{m}}(1 - p_{\tilde{m}k} < \epsilon|y_k) \\ &= 1 - \pi_{\tilde{m}}(1 - p_{\tilde{m}k} \geq \epsilon|y_k) \\ &\geq 1 - \frac{E(1 - p_{\tilde{m}k}|y_k)^2}{\epsilon^2} \\ &= 1 - \frac{1 - 2E(p_{\tilde{m}k}|y_k) + E(p_{\tilde{m}k}^2|y_k)}{\epsilon^2}. \end{aligned} \quad (11.6.8)$$

Now, as  $k \rightarrow \infty$ ,  $E(p_{\tilde{m}k}|y_k) \rightarrow 1$  by (11.6.6), and  $E(p_{\tilde{m}k}^2|y_k) = \text{Var}(p_{\tilde{m}k}|y_k) + [E(p_{\tilde{m}k}|y_k)]^2 \rightarrow 1$  by (11.6.6) and (11.6.7). Hence, the right hand side of (11.6.8) converges to 1 almost surely, as  $k \rightarrow \infty$ . This proves (11.6.5).

Now assume that (11.6.5) holds for any neighborhood  $\mathcal{N}_1$  of 1. Let us fix  $\eta \in (0, 1)$ . Then given any  $\epsilon \in (0, 1 - \eta)$ ,

$$\pi_{\tilde{m}}(1 - p_{\tilde{m}k} < \epsilon|y_k) \rightarrow 1, \quad (11.6.9)$$

almost surely as  $k \rightarrow \infty$ . The left hand side of (11.6.9) admits the following Markov's inequality:

$$\pi_{\tilde{m}}(1 - p_{\tilde{m}k} < \epsilon|y_k) = \pi_{\tilde{m}}(p_{\tilde{m}k} > 1 - \epsilon|y_k) < \frac{E(p_{\tilde{m}k}|y_k)}{1 - \epsilon}. \quad (11.6.10)$$

Due to (11.6.9), validity of (11.6.10) for all  $\epsilon \in (0, 1 - \eta)$ , and almost sure upper

boundedness of  $p_{\tilde{m}k}$  by 1, it follows that

$$E(p_{\tilde{m}k}|y_k) \rightarrow 1, \text{ almost surely, as } k \rightarrow \infty. \quad (11.6.11)$$

Now, if  $f(\cdot)$  is assumed to have  $m^*$  local minima where  $m^* \neq \tilde{m}$ , then due to (11.6.6) we must have

$$E(p_{m^*k}|y_k) \rightarrow 1, \text{ almost surely, as } k \rightarrow \infty. \quad (11.6.12)$$

Also note that since  $0 \leq p_{mk} \leq 1$  for all  $m$  and  $k$ , the dominated convergence theorem ensures the following, almost surely:

$$1 = \lim_{k \rightarrow \infty} \sum_{m=1}^{\infty} E(p_{mk}|y_k) = \sum_{m=1}^{\infty} \lim_{k \rightarrow \infty} E(p_{mk}|y_k). \quad (11.6.13)$$

Hence, (11.6.12) implies that  $E(p_{\tilde{m}k}|y_k) \rightarrow 0$ , almost surely, as  $k \rightarrow \infty$ . But this would contradict (11.6.11). Hence,  $f(\cdot)$  must have  $\tilde{m}$  local minima. ■

## 11.7 Experiments

We consider application of Algorithm 1 to 5 different optimization problems, ranging from simple to challenging, several of which are problems of finding both maxima and minima, and one is concerned with saddle points and inconclusiveness in addition to maximum. Encouragingly, all our experiments bring forth the versatility of Algorithm 1 in capturing all the optima, saddlepoints, as well as inconclusiveness of the problems.

As regards TMCMC, in our examples, we expectedly find that in the less challenging, low-dimensional problems, additive TMCMC is sufficient, while in the more challenging cases, we consider appropriate mixtures of additive and multiplicative moves, followed by a further move of a specialized mixture of additive and multiplicative transformations to improve mixing.

In most of our experiments, we discard the first  $10^5$  TMCMC iterations as burn-in

and store every 10-th TMCMC realization in the next  $5 \times 10^5$  iterations, to obtain 50000 realizations before proceeding to step (2) of Algorithm 1. However, in the 4-th example, due to inadequate mixing, we discard the first  $10^6$  iterations and store every 10-th realization in the next  $5 \times 10^6$  iterations to obtain  $5 \times 10^5$  TMCMC realizations. For the resample size in step (2), we set  $M = 1000$ .

For the posterior of  $\mathbf{g}'(\cdot) = \mathbf{0}$  given by (11.3.1), we set  $a = b = 0.1$ ,  $\beta_0 = \mathbf{0}$  and  $\Sigma_0 = \mathbb{I}_d$ , for all the examples, where  $d$  is the dimension relevant to the problem. In all the examples, we also set  $\mathcal{X} = [-10, 10]^d$  and the initial input size  $n = 10$ . With  $\mathbf{x}_i = (x_{i1}, \dots, x_{id})$  for  $i = 1, \dots, n$ , we choose the inputs points as  $x_{ik} = -10 + 2(i - 1)$ , for  $i = 1, \dots, n$  and for  $k = 1, \dots, d$ . We then evaluate  $f(\cdot)$  at each  $\mathbf{x}_i$ ;  $i = 1, \dots, n$ , to form  $\mathbf{D}_n$  with  $n = 10$ . As we shall demonstrate, this strategy, in conjunction with the rest of the methodology, leads to adequate estimation of the optima in our examples.

### 11.7.1 Example 1

We begin with a simple example, where the goal is to obtain the maxima and minima of the function

$$f(x) = 2x^3 - 3x^2 - 12x + 6. \quad (11.7.1)$$

Here  $f'(x) = 6(x - 2)(x + 1)$  and  $f''(x) = 6(2x - 1)$ . Hence, this function has a maximum at  $x = -1$  and minimum at  $x = 2$ .

We apply step (1) of Algorithm 1 with  $\epsilon = 1$ , implementing additive TMCMC with equal move-type probabilities for forward and backward transformations. Specifically, at iteration  $t = 1, 2, \dots$ , letting  $x^{(t-1)}$  denote the TMCMC realization at iteration  $t - 1$ , we draw  $\varepsilon \sim N(0, 1)$  and consider the transformation  $y = x^{(t-1)} + b|\varepsilon|$ , where  $b$  takes the values 1 and  $-1$  with equal probabilities. We set  $x^{(t)} = y$  with probability

$$\alpha = \min \left\{ 1, \frac{\pi(y)\pi(g'(y) = 0 | \mathbf{D}_n, y)}{\pi(x^{(t-1)})\pi(g'(x^{(t-1)}) = 0 | \mathbf{D}_n, x^{(t-1)})} \right\},$$

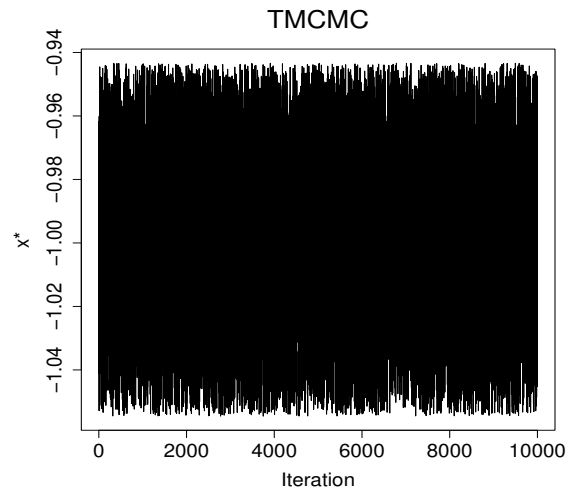


and set  $x^{(t)} = x^{(t-1)}$  with the remaining probability. This is of course same as the ordinary random walk Metropolis algorithm since the dimension here is  $d = 1$ ; see [Dutta and Bhattacharya \(2014\)](#) for ramifications and detailed discussions. Figure 11.7.1 shows the trace plots of  $x^*$  for maximum and minimum, where to reduce the figure file size, we thinned the original TMCMC sample of size  $N = 50000$  to 10000 by displaying every 5-th realization. The trace plots indicate adequate mixing properties of TMCMC. We run step (2) of Algorithm 1 for  $k = 1, \dots, S$  stages with  $S = 40$ , setting  $\eta_k = 1/(10 + k - 1)^2$ , computing the importance weights for the  $N$  TMCMC realizations at each stage on 100 64-bit cores in a VMWare parallel computing environment. The cores have 2.80 GHz speed, and have access to 1 TB memory. All our codes are written in C, using the Message Passing Interface (MPI) protocol for parallel processing. As such, our entire exercise is completed in about 2 minutes. We obtain  $\hat{x}_{\max} = -0.999995$  and  $\hat{x}_{\min} = 2.000023$ , as our estimates of the maximum and the minimum, respectively, which are quite accurate.

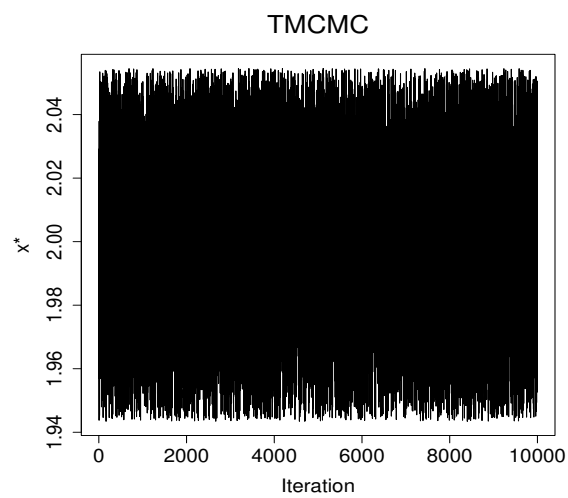
### 11.7.2 Example 2

We now consider maximization and minimization of the function  $f(x) = \sin(x)$  for  $x \in [-10, 10]$ . Here, the true maxima are  $x_{\max} = \{\frac{\pi}{2} - 2\pi = -4.712389, \frac{\pi}{2} = 1.570796, \frac{\pi}{2} + 2\pi = 7.853982\}$ , and the minima are  $x_{\min} = \{-7.853982, -1.570796, 4.712389\}$ .

We implement Algorithm 1 in the same way as in Example 1. Figure 11.7.2 displays the TMCMC trace plots, thinned to 10000 realizations to reduce figure file sizes. Clear tri-modality can be visualized from both the trace plots. After implementing step (2) of Algorithm 1 for  $S = 40$  stages on VMWare parallel computing architecture, we obtain  $\hat{x}_{\max} = \{-4.712581, 1.570655, 7.860879\}$  and  $\hat{x}_{\min} = \{-7.854051, -1.570423, 4.713222\}$ , which turn out to be adequate approximations to the truths. Again, the entire exercise takes about 2 minutes.

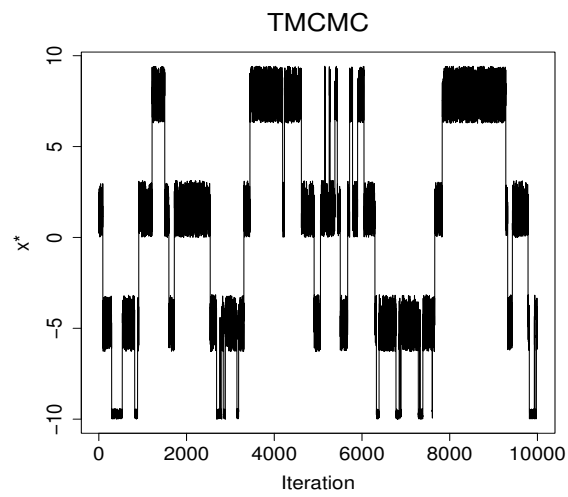


(a) TMCMC for maximum.

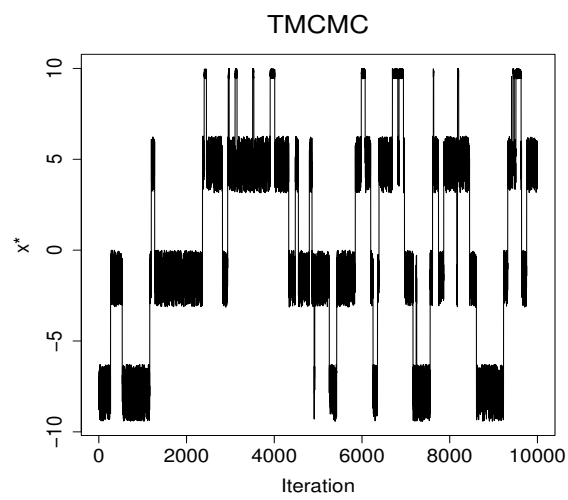


(b) TMCMC for minimum.

**Figure 11.7.1:** TMCMC trace plots for Example 1.



(a) TMCMC for maximum.



(b) TMCMC for minimum.

**Figure 11.7.2:** TMCMC trace plots for Example 2.

### 11.7.3 Example 3

Let us now consider a two-dimensional example, given by  $f(x_1, x_2) = x_1 x_2 (x_1 + x_2)(1 + x_2)$ .

The first derivatives are given by  $f'_1(x_1, x_2) = x_2(2x_1 + x_2)(x_2 + 1)$  and  $f'_2(x_1, x_2) = x_1(3x_2^2 + 2x_2(x_1 + 1) + x_1)$ .

The second derivatives are  $f''_{11}(x_1, x_2) = 2x_2(x_2 + 1)$ ,  $f''_{12}(x_1, x_2) = f''_{21}(x_1, x_2) = 4x_1 x_2 + 3x_2^2 + 2(x_1 + x_2)$  and  $f''_{22}(x_1, x_2) = 2x_1(3x_2 + x_1 + 1)$ .

Consider the determinant  $D(x_1, x_2) = f''_{11}(x_1, x_2)f''_{22}(x_1, x_2) - [f''_{12}(x_1, x_2)]^2$ . Now, if  $(a, b)$  is any critical point of  $f(\cdot)$  satisfying  $f'_1(a, b) = 0$  and  $f'_2(a, b) = 0$ , then  $(a, b)$  is a local maximum if  $D(a, b) > 0$  and  $f''_{11}(a, b) < 0$ ;  $(a, b)$  is a local minimum if  $D(a, b) > 0$  and  $f''_{11}(a, b) > 0$ ;  $(a, b)$  is a saddle point if  $D(a, b) < 0$ . Furthermore, if  $D(a, b) = 0$ , then  $(a, b)$  may be either maximum, minimum or even a saddle point, that is, the derivative test remains inconclusive in such cases.

In this example, it is easy to verify that there are four critical points  $(0, 0)$ ,  $(0, -1)$ ,  $(1, -1)$  and  $(\frac{3}{8}, -\frac{3}{4})$ . The last point is a local maximum;  $(0, -1)$  and  $(1, -1)$  are saddle points, and the derivative test remains inconclusive about  $(0, 0)$ .

We implement Algorithm 1 using the above conditions on  $D(\mathbf{x}^*)$  and  $f''_{11}(\mathbf{x}^*) > 0$ , along with the condition  $\|\mathbf{f}'(\mathbf{x}^*)\|_2 < \epsilon$ , with  $\epsilon = 1$ , in the prior for  $\mathbf{x}^* = (x_1^*, x_2^*)$ , for detection of maxima, minima, saddle points and inconclusiveness.

We implement additive TMCMC with equal move-type probabilities for forward and backward transformations. In these cases, at iteration  $t = 1, 2, \dots$ , letting  $\mathbf{x}^{(t-1)}$  denote the TMCMC realization at iteration  $t - 1$ , we draw  $\varepsilon \sim N(0, 1)$  and consider the transformation  $\mathbf{y} = \mathbf{x}^{(t-1)} + \mathbf{b}|\varepsilon|$ , where,  $\mathbf{b} = (b_1, b_2)^T$  and each of  $b_1$  and  $b_2$  independently takes the values 1 and  $-1$  with equal probabilities. We set  $\mathbf{x}^{(t)} = \mathbf{y}$  with probability

$$\alpha = \min \left\{ 1, \frac{\pi(\mathbf{y})\pi(\mathbf{g}'(\mathbf{y}) = \mathbf{0} | \mathbf{D}_n, \mathbf{y})}{\pi(\mathbf{x}^{(t-1)})\pi(\mathbf{g}'(\mathbf{x}^{(t-1)}) = \mathbf{0} | \mathbf{D}_n, \mathbf{x}^{(t-1)})} \right\},$$

and set  $\mathbf{x}^{(t)} = \mathbf{x}^{(t-1)}$  with the remaining probability. Note that unlike the one-dimensional setup, this additive TMCMC is no longer equivalent to random walk Metropolis (see [Dutta and Bhattacharya \(2014\)](#) for details).

Implementation of Algorithm 1 for obtaining the maximum and the saddle points took about 7 minutes to complete on our VMWare, implemented on 100 cores.

### Maximum

Figure 11.7.3 displaying the TMCMC trace plots for maxima finding, indicates quite adequate mixing. Running step (2) of Algorithm 1 for  $S = 40$  steps yields  $\hat{\mathbf{x}}_{max} = (0.376858, -0.752406)$ , which is reasonably close to the true maximum.

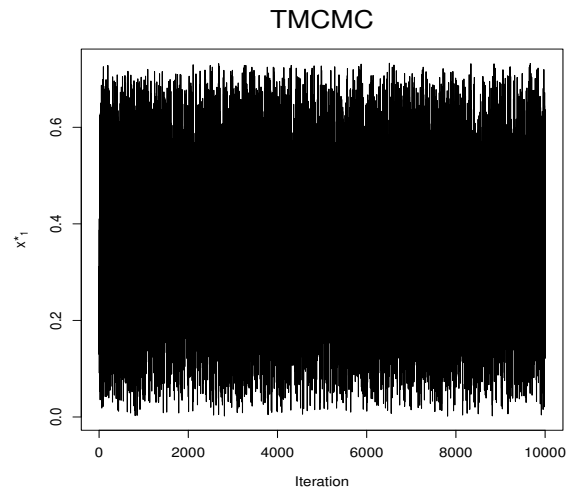
### Saddle points

Our initial TMCMC investigations revealed two modal regions roughly around  $(0.1, -1.1)$  and  $(1.1, -1.1)$ . For better exploration of the two modal regions, we implemented two separate TMCMC runs beginning at the above two points, and continue Algorithm 1 to ultimately obtain two separate results after  $S = 40$  stages at step (2), associated with the two different starting points. Figures 11.7.4 and 11.7.5 display the TMCMC trace plots associated with the two different starting points.

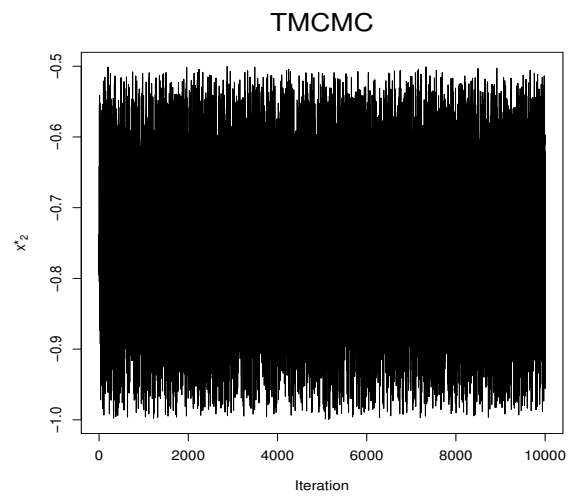
Running step (2) of Algorithm 1 for  $S = 40$  steps yields  $\hat{\mathbf{x}}_{saddle}^{(1)} = (-0.000309, -0.999580)$  and  $\hat{\mathbf{x}}_{saddle}^{(2)} = (0.999433, -0.999915)$  as estimates of two saddle points, which are both reasonably close to the true saddle points.

### Inconclusiveness

Investigation of situations where  $D(a, b) = 0$  for any critical point  $(a, b)$  yielded the TMCMC trace plots displayed as Figure 11.7.6. On completion of step (2) of Algorithm 1, we obtain  $\hat{\mathbf{x}}_{incon} = (-0.000087, -0.000265)$  as the estimate of the stationary point

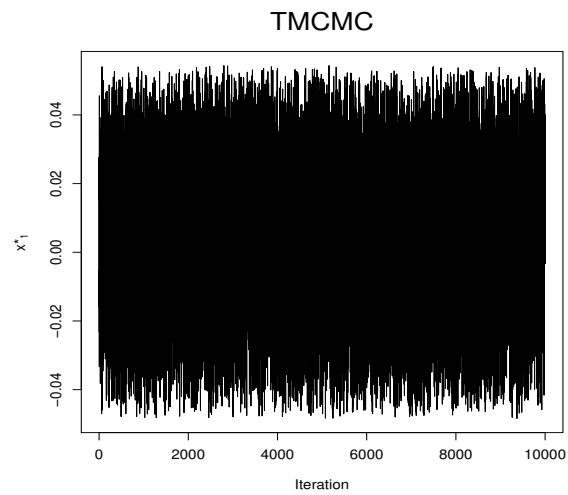


(a) TMCMC for maximum: first co-ordinate.

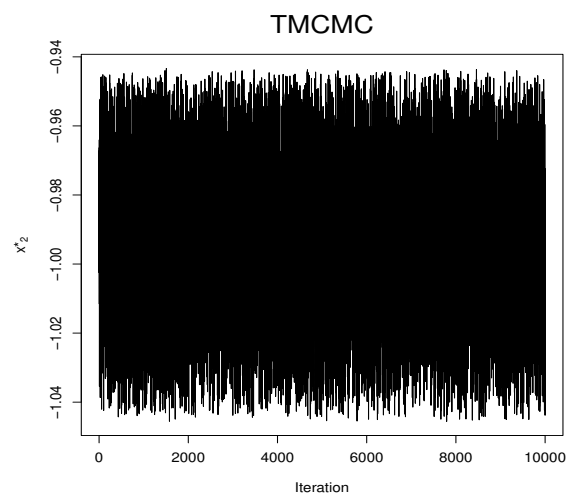


(b) TMCMC for maximum: second co-ordinate.

**Figure 11.7.3:** TMCMC trace plots for Example 3 for finding maxima.

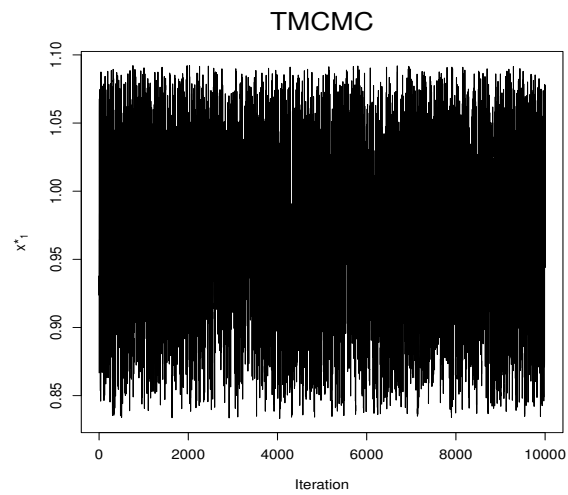


(a) TMCMC for first saddle point: first coordinate.

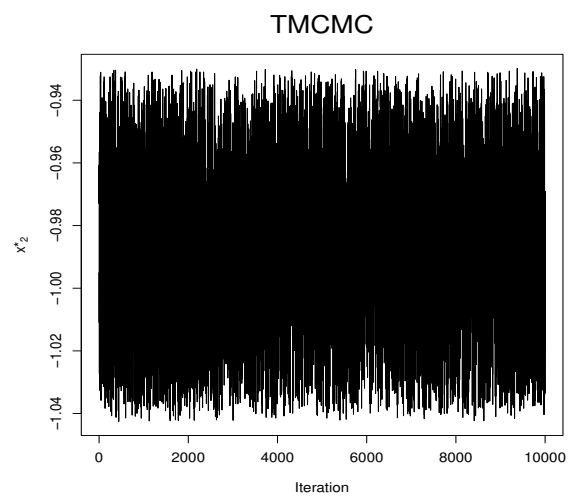


(b) TMCMC for first saddle point: second coordinate.

**Figure 11.7.4:** TMCMC trace plots for Example 3 for finding the first saddle point.



(a) TMCMC for second saddle point: first coordinate.



(b) TMCMC for second saddle point: second coordinate.

**Figure 11.7.5:** TMCMC trace plots for Example 3 for finding the second saddle point.



regarding which conclusion can not be drawn using the second derivatives. Observe that  $\hat{\mathbf{x}}_{incon}$  quite adequately estimates the actual point  $(0, 0)$  where conclusion fails.

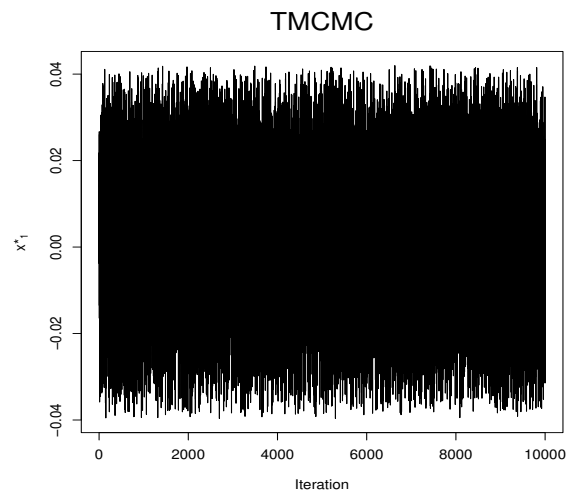
### Minimum

Our attempt to implement TMCMC with the restrictions  $D(\mathbf{x}^*) > 0$  and  $f''_{11}(\mathbf{x}^*) > 0$  with  $\|\mathbf{f}'(\mathbf{x}^*)\|_2 < \epsilon$  did not yield any acceptance, even for arbitrary initial values. In other words, we could not obtain any solution that satisfies all the above restrictions, and hence conclude that there is no critical point on  $\mathcal{X}$  that satisfy the above restrictions.

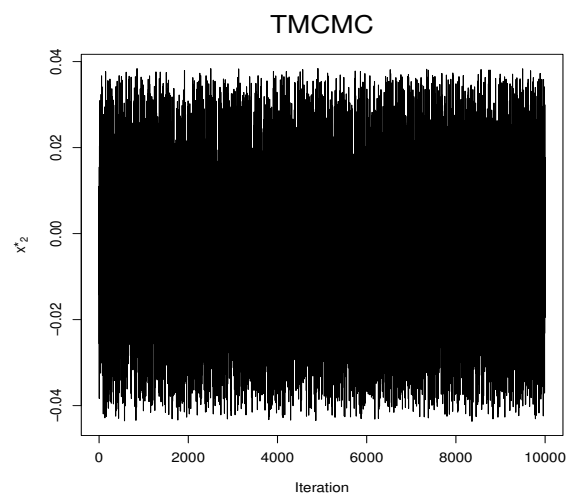
#### 11.7.4 Example 4

Table 14.2 of [Lange \(2010\)](#) reports quarterly data on AIDS deaths in Australia during 1983 – 1986, which is considered for illustration of fitting Poisson regression model. Specifically, for  $i = 1, \dots, 14$ , [Lange \(2010\)](#) considers the model  $Y_i \sim Poisson(\lambda_i)$ , with  $\lambda_i = \exp(\beta_0 + i\beta_1)$ . [Lange \(2010\)](#) computed the maximum likelihood estimate (MLE) of  $\boldsymbol{\beta} = (\beta_0, \beta_1)$  using Fisher’s scoring method, which is equivalent to Newton’s method in this case. The final estimate obtained is  $\hat{\boldsymbol{\beta}}_{MLE} = (0.3396, 0.2565)$ .

In our notation, the function to maximize is  $f(x_1, x_2) = -\sum_{i=1}^{14} \exp(x_1 + ix_2) + \sum_{i=1}^{14} y_i(x_1 + ix_2)$ , with respect to  $\mathbf{x} = (x_1, x_2)$ . Note that this is a concave maximization problem and hence the second derivative is irrelevant. We thus consider the only constraint  $\|\mathbf{f}'(\mathbf{x}^*)\|_2 < \epsilon$  for our implementation, with  $\epsilon = 1$ . However, additive TMCMC did not exhibit adequate mixing properties in this example, and hence we consider a mixture of additive and multiplicative TMCMC that is expected to improve mixing by using a mixture of localised moves of additive TMCMC and non-local (“random dive”) moves of multiplicative TMCMC (see [Dutta \(2012\)](#), [Dey and Bhattacharya \(2016\)](#) for details). We strengthen the mixture TMCMC strategy with a further step of a mixture of specialized additive and multiplicative moves, which has parallels with [Liu and Sabatti \(2000\)](#). The detailed TMCMC algorithm, for general dimension  $d$ , is provided below as



(a) TMCMC for inconclusiveness: first coordinate.



(b) TMCMC for inconclusiveness: second coordinate.

**Figure 11.7.6:** TMCMC trace plots for Example 3 for investigating inconclusiveness.

Algorithm 2.

In our case,  $d = 2$ , and we choose  $a_j^{(1)} = a_j^{(2)} = 0.05$ , for  $j = 1, 2$ ; we also set  $p = q = 1/2$ . However, in spite of such a sophisticated TMCMC algorithm, we failed to achieve excellent mixing in this example, even with long TMCMC runs, discarding the first  $10^6$  iterations and storing every 10-th realization in the next  $5 \times 10^6$  iterations. The trace plots of all stored  $5 \times 10^5$  realizations shown in Figure 11.7.7 indeed demonstrate that the TMCMC chain does not have excellent mixing properties. In fact, the trace plots correspond to a reasonable initial value, chosen to be the 4-th iterate of the Fisher scoring method (see Table 14.3 of Lange (2010)). The subsequent Fisher scoring iterates as initial values led to increasingly improved performance. But here our goal is to demonstrate that even when the mixing is less adequate, the estimates of the optima obtained by our method can still significantly outperform the existing techniques.

As it is known that this example is a concave maximization problem, step (2) of Algorithm 1 is unnecessary. Following Remark 70, we set  $i^* = \min\{f(\mathbf{x}_i^*) : i = 1, \dots, N\}$  and report  $\mathbf{x}_{i^*}^*$  as the (approximate) maximizer of  $f(\cdot)$ . Thus, with the  $N = 5 \times 10^5$  TMCMC realizations shown in Figure 11.7.7, we obtain the estimate  $\hat{\mathbf{x}}_{MLE} = (0.364422, 0.254428)$ . For this value,  $\|\mathbf{f}'(\hat{\mathbf{x}}_{MLE})\|_2 = 0.395978$ . On the other hand, the MLE obtained by Lange (2010) yields  $\|\mathbf{f}'(\hat{\boldsymbol{\beta}}_{MLE})\|_2 = 0.755344$ , which is much away from zero compared to our method. In other words,  $\hat{\mathbf{x}}_{MLE}$  is much more reliable compared to  $\hat{\boldsymbol{\beta}}_{MLE}$ , indicating that our method significantly outperforms the existing popular methods of MLE computation. This is indeed a general statement, since with good choices of initial values for TMCMC, which may typically be optima obtained by existing, popular optimization methods, we can explore regions in  $\mathcal{X}$  which almost surely contain values that yield smaller  $\|\mathbf{f}'\|_d$  compared to the optima obtained by other numerical methods. It is also encouraging to note that for this example TMCMC takes much less than a minute to yield  $6 \times 10^6$  iterations on our VMWare, implemented on a single processor.

---

**Algorithm 2** Mixture TMCMC
 

---

(1) Fix  $p, q \in (0, 1)$ . Set an initial value  $\mathbf{x}^{(0)}$ .

(2) For  $t = 1, \dots, N$ , do the following:

1. Generate  $U \sim U(0, 1)$ .

(a) If  $U < p$ , then do the following:

(i) Generate  $\varepsilon \sim N(0, 1)$ ,  $b_j \stackrel{iid}{\sim} U(\{-1, 1\})$  for  $j = 1, \dots, d$ , and set  $y_j = x_j^{(t-1)} + b_j a_j^{(1)} |\varepsilon|$ , for  $j = 1, \dots, d$ . Here  $a_j^{(1)}$  are positive scaling constants.

(ii) Evaluate

$$\alpha_1 = \min \left\{ 1, \frac{\pi(\mathbf{y})\pi(\mathbf{g}'(\mathbf{y}) = \mathbf{0} | \mathbf{D}_n, \mathbf{y})}{\pi(\mathbf{x}^{(t-1)})\pi(\mathbf{g}'(\mathbf{x}^{(t-1)}) = \mathbf{0} | \mathbf{D}_n, \mathbf{x}^{(t-1)})} \right\}.$$

(iii) Set  $\tilde{\mathbf{x}}^{(t)} = \mathbf{y}$  with probability  $\alpha_1$ , else set  $\tilde{\mathbf{x}}^{(t)} = \mathbf{x}^{(t-1)}$ .

(b) If  $U \geq p$ , then

(i) Generate  $\varepsilon \sim U(-1, 1)$ ,  $b_j \stackrel{iid}{\sim} U(\{-1, 0, 1\})$  for  $j = 1, \dots, d$ , and set  $y_j = x_j^{(t-1)} \varepsilon$  if  $b_j = 1$ ,  $y_j = x_j^{(t-1)} / \varepsilon$  if  $b_j = -1$  and  $y_j = x_j^{(t-1)}$  if  $b_j = 0$ , for  $j = 1, \dots, d$ . Calculate  $|J| = |\varepsilon|^{\sum_{j=1}^d b_j}$ .

(ii) Evaluate

$$\alpha_2 = \min \left\{ 1, \frac{\pi(\mathbf{y})\pi(\mathbf{g}'(\mathbf{y}) = \mathbf{0} | \mathbf{D}_n, \mathbf{y})}{\pi(\mathbf{x}^{(t-1)})\pi(\mathbf{g}'(\mathbf{x}^{(t-1)}) = \mathbf{0} | \mathbf{D}_n, \mathbf{x}^{(t-1)})} \times |J| \right\}.$$

(iii) Set  $\tilde{\mathbf{x}}^{(t)} = \mathbf{y}$  with probability  $\alpha_2$ , else set  $\tilde{\mathbf{x}}^{(t)} = \mathbf{x}^{(t-1)}$ .

2. Generate  $U \sim U(0, 1)$ .

(a) If  $U < q$ , then do the following

(i) Generate  $\tilde{U} \sim U(0, 1)$  and  $\varepsilon \sim N(0, 1)$ . If  $\tilde{U} < 1/2$ , set  $y_j = \tilde{x}_j^{(t)} + a_j^{(2)} |\varepsilon|$ , for  $j = 1, \dots, d$ ; else, set  $y_j = \tilde{x}_j^{(t)} - a_j^{(2)} |\varepsilon|$ , for  $j = 1, \dots, d$ . Here  $a_j^{(2)}$  are positive scaling constants.

(ii) Evaluate

$$\alpha_3 = \min \left\{ 1, \frac{\pi(\mathbf{y})\pi(\mathbf{g}'(\mathbf{y}) = \mathbf{0} | \mathbf{D}_n, \mathbf{y})}{\pi(\tilde{\mathbf{x}}^{(t)})\pi(\mathbf{g}'(\tilde{\mathbf{x}}^{(t)}) = \mathbf{0} | \mathbf{D}_n, \tilde{\mathbf{x}}^{(t)})} \right\}.$$

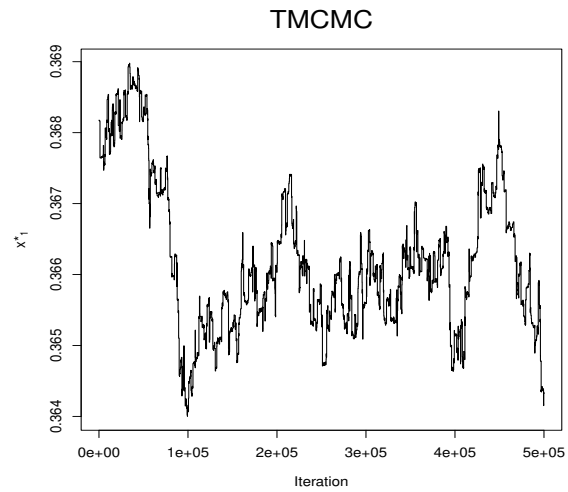
(iii) Set  $\mathbf{x}^{(t)} = \mathbf{y}$  with probability  $\alpha_3$ , else set  $\mathbf{x}^{(t)} = \tilde{\mathbf{x}}^{(t)}$ .

(b) If  $U \geq q$ , then

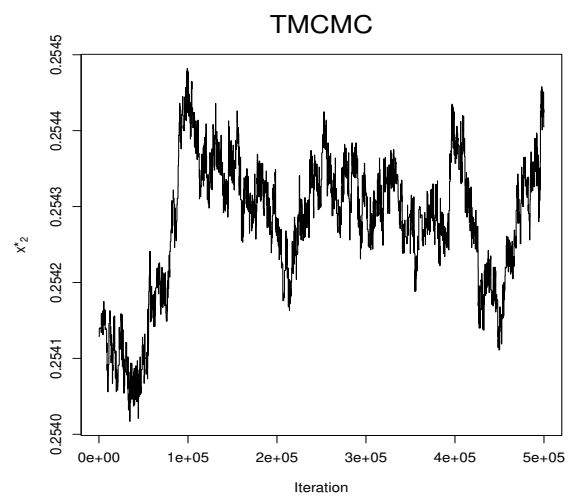
(i) Generate  $\varepsilon \sim U(-1, 1)$  and  $\tilde{U} \sim U(0, 1)$ . If  $\tilde{U} < 1/2$ , set  $y_j = \tilde{x}_j^{(t)} \varepsilon$  for  $j = 1, \dots, d$  and  $|J| = |\varepsilon|^d$ , else set  $y_j = \tilde{x}_j^{(t)} / \varepsilon$  for  $j = 1, \dots, d$  and  $|J| = |\varepsilon|^{-d}$ .

(ii) Evaluate

$$\alpha_4 = \min \left\{ 1, \frac{\pi(\mathbf{y})\pi(\mathbf{g}'(\mathbf{y}) = \mathbf{0} | \mathbf{D}_n, \mathbf{y})}{\pi(\tilde{\mathbf{x}}^{(t)})\pi(\mathbf{g}'(\tilde{\mathbf{x}}^{(t)}) = \mathbf{0} | \mathbf{D}_n, \tilde{\mathbf{x}}^{(t)})} \times |J| \right\}.$$



(a) TMCMC for MLE: first co-ordinate.



(b) TMCMC for MLE: second co-ordinate.

**Figure 11.7.7:** TMCMC trace plots for Example 4 for finding MLE.

### 11.7.5 Example 5

Hunter and Lange (2000) refer to a nonlinear optimization problem of the form  $Y_i \sim N(\mu_i, \sigma^2)$  for  $i = 1, \dots, m$ , where

$$\mu_i = \sum_{j=1}^d [\exp(-z_{ij}\theta_j^2) + z_{ij}\theta_{d-j+1}];$$

$z_{ij}$  being the  $i$ -th observation of the  $j$ -th covariate, for  $i = 1, \dots, m$  and  $j = 1, \dots, d$ . The goal is to compute the MLE of  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)$ , assuming that  $\sigma$  is known. In our notation, the objective is to minimize

$$f(\boldsymbol{x}) = \sum_{i=1}^m \left( y_i - \sum_{j=1}^d [\exp(-z_{ij}x_j^2) + z_{ij}x_{d-j+1}] \right)^2,$$

with respect to  $\boldsymbol{x}$ .

We consider 5 simulation experiments in this regard for  $d = 2, 5, 10, 50, 100$ . In each case we generate  $\theta_{0j} \sim U(-1, 1)$  independently for  $j = 1, \dots, d$ ,  $z_{ij} \sim N(0, 1)$  independently, for  $i = 1, \dots, m$  and  $j = 1, \dots, d$ , and set, for  $i = 1, \dots, m$ ,  $\mu_{0i} = \sum_{j=1}^d [\exp(-z_{ij}\theta_{0j}^2) + z_{ij}\theta_{0,d-j+1}]$ . We finally generate the response data by simulating  $Y_i \sim N(\mu_{0i}, \sigma_0^2)$  independently for  $i = 1, \dots, m$ , where we set  $\sigma_0^2 = 0.1$ . For  $d = 2, 5, 10, 50, 100$ , we generate datasets of sizes  $m = 10, 10, 20, 75, 200$ .

Note that this is not a convex minimization problem and the matrix of second derivatives  $\boldsymbol{\Sigma}''(\cdot)$  plays an important role, along with  $\|\boldsymbol{f}'(\cdot)\|_d$ . Thus it is important to check positive definiteness of  $\boldsymbol{\Sigma}''(\cdot)$  for any dimension  $d$ . We use the LAPACK library function “*dpotrf*” for Cholesky decomposition of  $\boldsymbol{\Sigma}''(\cdot)$ , which contains a parameter “*info*”. Given any  $\boldsymbol{x}$ , *info* = 0 indicates positive definiteness of  $\boldsymbol{\Sigma}''(\boldsymbol{x})$ , while other values of *info* rules out positive definiteness. Note that since this is not a convex minimization problem, step (2) of Algorithm 1 is necessary, unlike in Example 4.

We implement the mixture TMCMC algorithm 2 for all values of  $d$ , with  $p = q = 1/2$

and set  $a_j^{(1)} = a_j^{(2)} = 0.05$  for  $j = 1, \dots, d$ .

**Case 1:  $d = 2$**

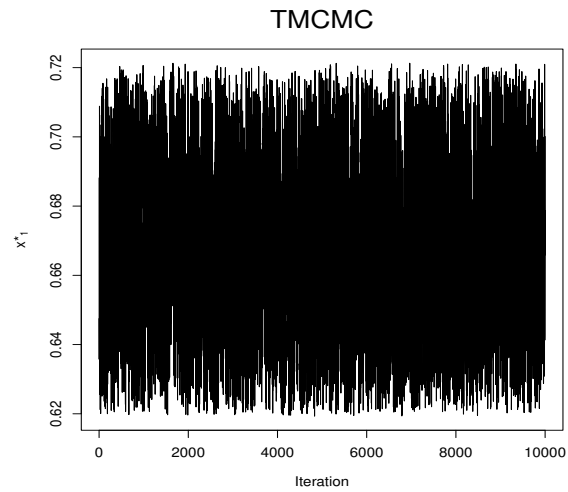
In the TMCMC step, we set  $\epsilon = 1$  in the restriction  $\|\mathbf{f}'(\cdot)\|_2 < \epsilon$ . The trace plots, shown in Figure 11.7.8, exhibit adequate mixing. Running step (2) of Algorithm 1 till  $S = 40$  stages with  $\eta_k = 1/(10 + k - 1)$  for  $k = 1, \dots, S$ , yielded the estimate of the MLE to be  $\hat{\mathbf{x}}_{MLE} = (0.678854, 0.293575)$ , for which  $\|\mathbf{f}'(\hat{\mathbf{x}}_{MLE})\|_2 = 0.012514$ . The exercise 3 minutes on our VMWare, implemented in parallel on 100 cores.

However, examination of the samples obtained by importance resampling at the different stages of step (2) of Algorithm 1 did not reveal any evidence of multimodality, and hence it is pertinent to consider that estimate which corresponds to the minimum of  $\{\|\mathbf{f}'(\mathbf{x}_i^*)\|_2 : i = 1, \dots, N\}$ , where  $\mathbf{x}_i^*$  are the original TMCMC samples, with  $N = 50000$ . As such, we modify the previous estimate to  $\hat{\mathbf{x}}_{MLE} = (0.678912, 0.293809)$ , which yields  $\|\mathbf{f}'(\hat{\mathbf{x}}_{MLE})\|_2 = 0.007221$ , which is somewhat closer to zero compared to that for the previous estimate. Note however, that the two estimates of MLE are quite close to each other.

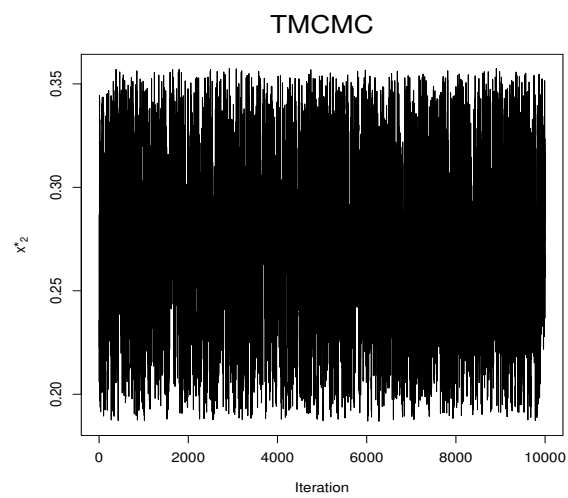
**Case 2:  $d = 5$**

Here, for the 5-dimensional TMCMC, we set  $\epsilon = 3$  in the restriction  $\|\mathbf{f}'(\cdot)\|_5 < \epsilon$ , as smaller values of  $\epsilon$  led to poor convergence. Note that the Euclidean norm increases with dimension (see, for example, Giraud (2015)), and so it is a natural requirement to increase  $\epsilon$  as dimension increases. Similarly, we had to increase  $\eta_k$  to  $\eta_k = 1.5/\log(10+k)$  for implementing step (2) of Algorithm 1. The rest of the parameters of Algorithm 2 remain the same as for  $d = 2$ .

The trace plots exhibited reasonable mixing; those for the first and the last (5-th) co-ordinate of  $\mathbf{x}^*$  are depicted in Figure 11.7.9. Running step (2) of Algorithm 1 till  $S = 40$  stages we obtain  $\hat{\mathbf{x}}_{MLE} = (0.663327, 0.431669, 0.045598, 0.239091, 0.301665)$ ,



(a) TMCMC for MLE: first co-ordinate.



(b) TMCMC for MLE: second co-ordinate.

**Figure 11.7.8:** TMCMC trace plots for Example 5 for finding MLE for dimension  $d = 2$ .



which corresponds to  $\min\{\|\mathbf{f}'(\mathbf{x}_i^*)\|_5 : i = 1, \dots, N\}$   
 $= 0.254217$ , with  $N = 50000$ . Note the increase in  $\|\mathbf{f}'(\cdot)\|_5$  compared to those of the smaller dimensions. Again, this is clearly to be expected because of the curse of dimensionality. The exercise takes 4 minutes to complete on our VMWare.

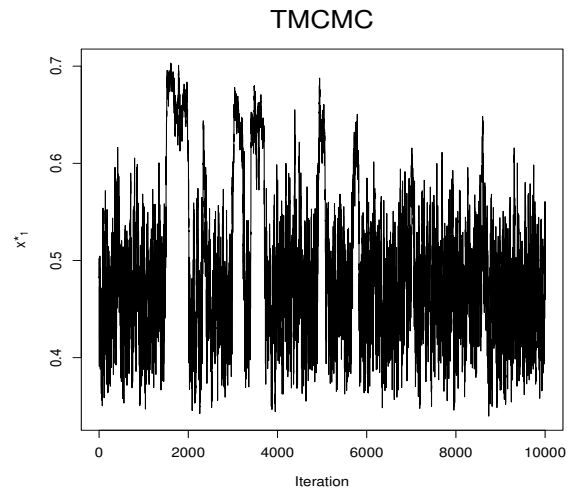
**Case 3:**  $d = 10$

Here, for  $d = 10$ , we had to set  $\epsilon = 6$  for the restriction  $\|\mathbf{f}'(\cdot)\|_{10} < \epsilon$  in TMCMC for adequate convergence. We also set  $\eta_k = 7/\log(10 + k - 1)$  for implementing step (2) of Algorithm 1.

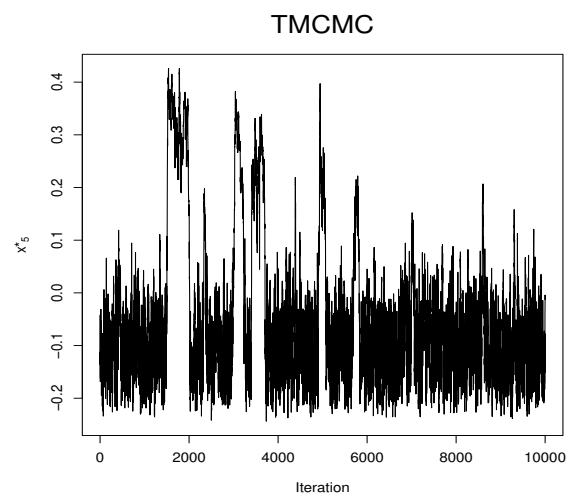
Our investigation shows that the mixing of TMCMC is not inadequate. The trace plots of the first and the last co-ordinate of  $\mathbf{x}^*$  shown in Figure 11.7.10 also bear evidence to this. After implementing step (2) of Algorithm 1 till  $S = 40$  stages, we obtain  $\hat{\mathbf{x}}_{MLE} = (0.472309, 0.10124, 0.079194, 0.108072, 0.096287, -0.511566, 0.517567, -0.887624, 0.637796, -0.317721)$  with  $\|\mathbf{f}'(\hat{\mathbf{x}}_{MLE})\|_{10} = 1.917746$ . On the other hand,  $\min\{\|\mathbf{f}'(\mathbf{x}_i^*)\|_{10} : i = 1, \dots, 50000\} = 1.902788$ , which corresponds to  $\hat{\mathbf{x}}_{MLE} = (0.471200, 0.100131, 0.078085, 0.101934, 0.095178, -0.504813, 0.516459, -0.888733, 0.631659, -0.323859)$ . Thus, both the estimates as well as the corresponding gradients are quite close to each other. Again note the increase in  $\|\mathbf{f}'(\cdot)\|_{10}$  compared to those of the smaller dimensions. The entire exercise takes 17 minutes on our VMWare to complete.

**Case 4:**  $d = 50$

For this somewhat large dimension, we had to set  $\epsilon = 100$  and  $\eta_k = 200/\log(10 + k - 1)$ . Figure 11.7.11, which displays all stored 50000 TMCMC realizations for  $x_1^*$  and  $x_{50}^*$  do not indicate excellent mixing, in spite of the sophistication of Algorithm 2. However, due to the high-dimension and complexity of the posterior this is not unexpected. As demonstrated by Example 4, we can still expect to get closer to the MLE compared to other optimization methods. In this case, we obtain  $\min\{\|\mathbf{f}'(\mathbf{x}_i^*)\|_{50} : i = 1, \dots, 50000\} =$

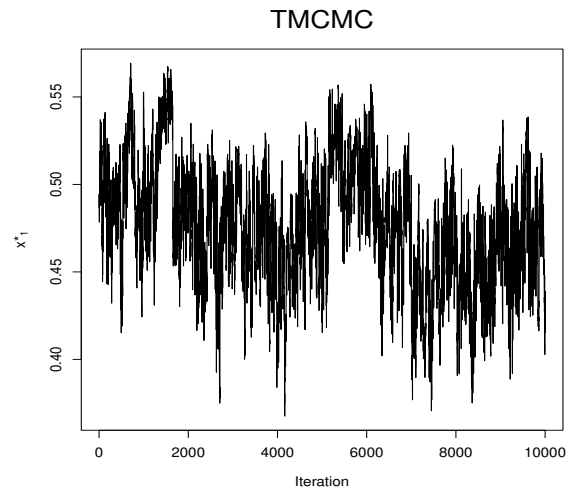


(a) TMCMC for MLE: first co-ordinate.

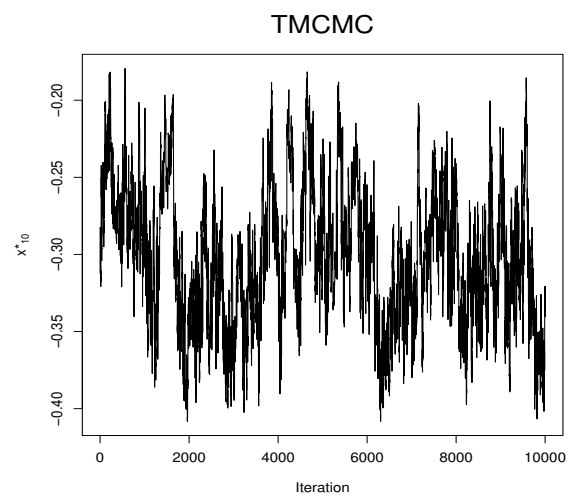


(b) TMCMC for MLE: fifth co-ordinate.

**Figure 11.7.9:** TMCMC trace plots for Example 5 for finding MLE for dimension  $d = 5$ .

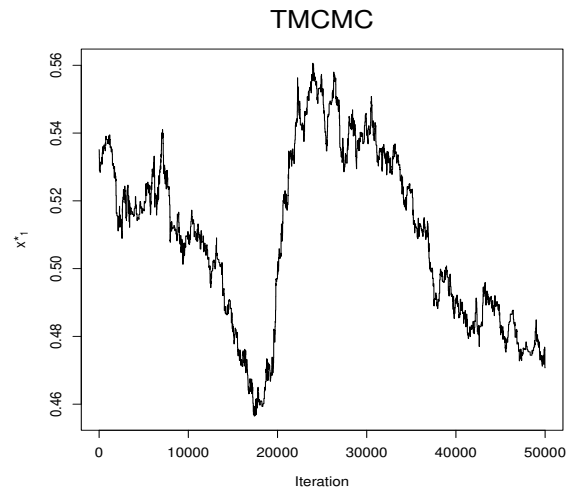


(a) TMCMC for MLE: first co-ordinate.

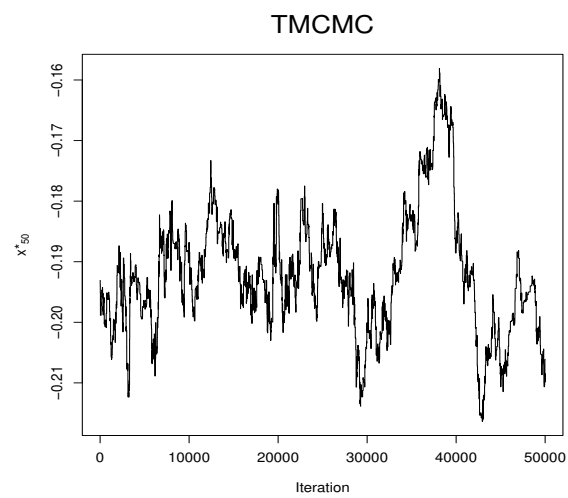


(b) TMCMC for MLE: 10-th co-ordinate.

**Figure 11.7.10:** TMCMC trace plots for Example 5 for finding MLE for dimension  $d = 10$ .



(a) TMCMC for MLE: first co-ordinate.



(b) TMCMC for MLE: 50-th co-ordinate.

**Figure 11.7.11:** TMCMC trace plots for Example 5 for finding MLE for dimension  $d = 50$ .

73.05261 while step (2) of Algorithm 1 implemented for till  $S = 100$  stages, of which only the first four stages consisted of positive  $n_k$ , yielded  $\|\mathbf{f}'(\hat{\mathbf{x}}_{MLE})\|_{50} = 76.28244$ . Thus, the norms of the gradients have increased considerably in this high dimension, compared to the previous  $d = 2, 5, 10$ . Given the high dimension in this problem, the above gradient values 73.05261 and 76.28244 are quite close.

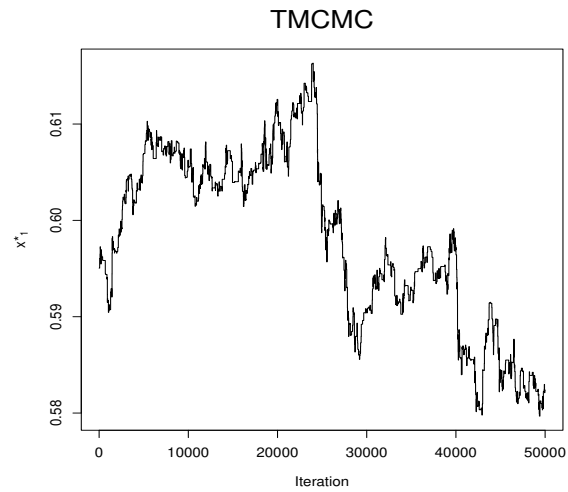
The TMCMC implementation took about 12 hours on a single core in our VMWare and step (2) of Algorithm 1 took additional 2 hours on 100 cores.

**Case 5:**  $d = 100$

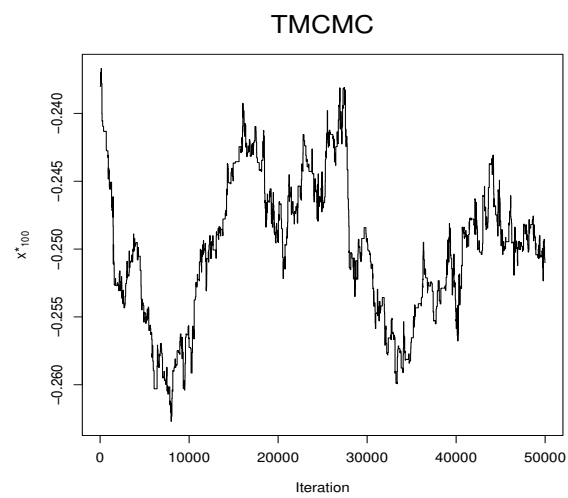
The curse of dimensionality now forced us to set  $\epsilon = 400$  for TMCMC convergence and  $\eta_k = 850/\log(10 + k - 1)$ . It took around 15 hours to complete the TMCMC run; the trace plots shown in Figure 11.7.12 do not bear evidence of non-convergence, even though mixing is expectedly inadequate in such high dimension. Here we obtain  $\min\{\|\mathbf{f}'(\mathbf{x}_i^*)\|_{50} : i = 1, \dots, 50000\} = 330.4483$ , while step (2) of Algorithm 1 implemented for till  $S = 100$  stages, of which only the first three stages yielded positive  $n_k$ , produced  $\hat{\mathbf{x}}_{MLE}$  such that  $\|\mathbf{f}'(\hat{\mathbf{x}}_{MLE})\|_{100} = 341.2009$ . Given the high dimension, this value is quite close to the above minimum gradient. Implementation of step (2) of Algorithm 1 took 2 hours 36 minutes on 100 cores on our VMWare.

## 11.8 Summary and conclusion

In this chapter, we have proposed and developed a novel Bayesian algorithm for general function optimization, judiciously exploiting its derivatives in conjunction with posterior Gaussian derivative process given data consisting of input points from the function domain and their function evaluations. The posterior simulation approach inherent in our method ensures improved accuracy of our results compared to existing optimization algorithms. Another important feature of our algorithm is that for any desired degree of



(a) TMCMC for MLE: first co-ordinate.



(b) TMCMC for MLE: 50-th co-ordinate.

**Figure 11.7.12:** TMCMC trace plots for Example 5 for finding MLE for dimension  $d = 100$ .

accuracy, Bayesian credible regions of the optima of any desired level, become readily available after implementation of the algorithm.

Under appropriate fixed-domain infill asymptotics setup, we prove almost sure convergence of the algorithm to the true optima. Along the way, we establish almost sure uniform convergence of the posteriors corresponding to Gaussian and Gaussian derivative processes to the objective function and its derivatives, under the fixed-domain infill asymptotics setup, providing rates of convergence under a specific setup. We also establish Bayesian characterization of the number of optima of the objective function by exploiting the information existing in our algorithm.

Applications of our Bayesian optimization algorithm to various examples ranging from simple to challenging, led to encouraging and insightful results. Choice of initial values for starting TMCMC of our algorithm is seen to affect mixing, but as we argued, optima yielded by good existing optimization algorithms can act as good initial values for TMCMC. Moreover, we have also demonstrated that even with less adequate mixing, our Bayesian algorithm can still significantly outperform popular optimization methods. Thus, the TMCMC mixing issue is perhaps not too important, at least for low dimensional problem.

Dimensionality of the problem seems to be a far more serious issue. Indeed, our experimental results demonstrate that as dimension increases, accuracy of our algorithm deteriorates. High dimensionality also seriously affects TMCMC mixing by excessively restricting the input space through the prior constraints. These are only natural, but need to be dealt with seriously. Our future endeavors will address these issues.

## References

- Adler, R. J. (1981). *The Geometry of Random Fields*. John Wiley and Sons, New York.
- Adler, R. J. and Taylor, J. E. (2007). *Random Fields and Geometry*. Springer, New York.
- Alekseyev, M. A. (2011). On Convergence of the Flint Hills Series. Available at “<http://arxiv.org/pdf/1104.5100v1.pdf>”.
- Baddeley, A. and Turner, R. (2005). Spatstat: an R Package for Analyzing Spatial Point Patterns. *Journal of Statistical Software*, **12**, 1–42. URL: [www.jstatsoft.org](http://www.jstatsoft.org), ISSN: 1548–7660.
- Bandopadhyay, S. and Rao, S. S. (2017). A Test for Stationarity for Irregularly Spaced Spatial Data. *Journal of the Royal Statistical Society. Series B*, **79**, 95–123.
- Bandopadhyay, S., Jentsch, C., and Rao, S. S. (2017). A Spectral Domain Test for Stationarity of Spatio-Temporal Data. *Journal of Time Series Analysis*, **38**, 326–351.
- Basu, P., Rudoy, D., and Wolfe, P. J. (2009). A Nonparametric Test for Stationarity Based on Local Fourier Analysis. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3005–3008.
- Bera, A. K. and Higgins, M. L. (1993). ARCH Models: Properties, Estimation and Testing. *Journal of Economic Surveys*, **7**, 305–366.
- Beran, R., Bilodeau, M., and L de Micheaux, P. (2007). Nonparametric Tests of Independence Between Random Vectors. *Journal of Multivariate Analysis*, **98**, 1805–1824.



- Berkes, I., Horváth, L., and Kokoszka, P. (2003). GARCH Processes: Structure and Estimation. *Bernoulli*, **9**, 201–227.
- Billingsley, P. (2013). *Convergence of Probability Measures*. John Wiley & Sons Ltd., New York.
- Bilodeau, M. and L de Micheaux, P. (2005). A Multivariate Empirical Characteristic Function Test of Independence With Normal Marginals. *Journal of Multivariate Analysis*, **95**, 345–369.
- Bilodeau, M. and Nangué, A. G. (2017). Tests of Mutual or Serial Independence of Random Vectors with Applications. *Journal of Machine Learning Research*, **18**, 1–40.
- Bivand, R. S., Pebesma, E. J., and Gomez-Rubio, V. (2008). *Applied Spatial Data Analysis with R*. Springer, New York.
- Black, F. and Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, **81**, 637–654.
- Blum, J. R., Kiefer, J., and Rosenblatt, M. (1961). Distribution Free Tests of Independence Based on the Sample Distribution Function. *Annals of Mathematical Statistics*, **32**, 485–498.
- Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, **31**, 307–327.
- Borwein, P., Choi, S., Rooney, B., and Weirathmueller, A. (2006). *The Riemann Hypothesis: For the Aficionado and Virtuoso Alike*. Springer, New York.
- Bougerol, P. and Picard, N. (1992). Stationarity of GARCH Processes and of Some Nonnegative Time Series. *Journal of Econometrics*, **52**, 115–127.

- Bourchtein, L., Bourchtein, A., Nornberg, G., and Venzke, C. (2011). A Hierarchy of the Convergence Tests for Numerical Series Based on Kummer's Theorem. *Bulletin of the Paranaense Society of Mathematics*, **29**, 83–107.
- Bourchtein, L., Bourchtein, A., Nornberg, G., and Venzke, C. (2012). A Hierarchy of the Convergence Tests Related to Cauchy's Test. *International Journal of Mathematical Analysis*, **6**, 1847–1869.
- Breitung, J. (2002). Nonparametric Tests for Unit Roots and Cointegration. *Journal of Econometrics*, **10**, 343–363.
- Brockwell, P. J. and Davis, R. A. (2002). *Introduction to Time Series and Forecasting*. Springer, New York.
- Brockwell, P. J. and Davis, R. A. (2009). *Time Series: Theory and Methods*. Springer, New York.
- Bromwich, T. J. I. (2005). *An introduction to the theory of infinite series*. AMS, Providence.
- Brooks, S., Gelman, A., Jones, G. L., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. Interdisciplinary Statistics. Chapman and Hall/CRC, London.
- Brooks, S. P. and Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–455.
- Brooks, S. P. and Roberts, G. O. (1998). Assessing Convergence of Markov Chain Monte Carlo Algorithms. *Statistics and Computing*, **8**, 319–335.
- Cardinali, A. and Nason, G. P. (2018). Practical Powerful Wavelet Packet Tests for Second-Order Stationarity. *Applied and Computational Harmonic Analysis*, **44**, 558–583.

- Carey, J. C. (2003). The Riemann Hypothesis and Hardy Spaces. Available at “<http://jcarey.best.vwh.net/RHHardy.pdf>”.
- Chatfield, C. and Xing, H. (2002). *The Analysis of Time Series: An Introduction with R*. Chapman and Hall/CRC, Boca Raton, FL.
- Chatterjee, D. and Bhattacharya, S. (2020). How Ominous is the Future Global Warming Premonition? Available at <https://arxiv.org/abs/2008.11175>.
- Cléroux, R., Lazraq, A., and Lepage, Y. (1995). Vector Correlation Based on Ranks and a Nonparametric Test of No Association Between Vectors. *Communications in Statistics. Theory and Methods*, **24**, 713–733.
- Cowles, M. K. and Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, **91**, 883–904.
- Cressie, N. A. C. (1993a). *Hierarchical Modeling and Analysis of Spatial Data*. Chapman & Hall/CRC Press, Boca Raton, FL.
- Cressie, N. A. C. (1993b). *Statistics for Spatial Data*. Wiley, New York.
- Cressie, N. A. C. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley, New York.
- Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes*. Springer-Verlag, New York. Second Edition.
- Das, M. (2018). *A Novel Nonstationary Bayesian Space-Time Model with a New Transdimensional Transformation Based Markov Chain Monte Carlo*. Doctoral thesis, Indian Statistical Institute.

- Das, M. and Bhattacharya, S. (2020). Nonstationary Nonparametric Bayesian Spatio-Temporal Modeling Using Kernel Convolution of Order Based Dependent Dirichlet Process. ArXiv preprint.
- Derbyshire, J. (2004). *Prime Obsession: Bernhard Riemann and the Greatest Unsolved Problem in Mathematics*. Penguin, New York.
- Dey, K. K. and Bhattacharya, S. (2016). On Geometric Ergodicity of Additive and Multiplicative Transformation Based Markov Chain Monte Carlo in High Dimensions. *Brazilian Journal of Probability and Statistics*, **30**, 570–613.
- Dey, K. K. and Bhattacharya, S. (2017). A Brief Tutorial on Transformation Based Markov Chain Monte Carlo and Optimal Scaling of the Additive Transformation. *Brazilian Journal of Probability and Statistics*, **31**, 509–617.
- Dey, K. K. and Bhattacharya, S. (2019). A Brief Review of Optimal Scaling of the Main MCMC Approaches and Optimal Scaling of Additive TMCMC Under Non-Regular Cases. *Brazilian Journal of Probability and Statistics*, **33**, 222–266.
- Dickey, D. and Fuller, W. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association*, **74**, 427–431.
- Driver, B. (2010). Probability Tools with Examples. Available at [http://www.math.ucsd.edu/~bdriver/Cornell%20Summer%20Notes%202010/Lecture\\_Notes/Probability%20Tools%20with%20Examples.pdf](http://www.math.ucsd.edu/~bdriver/Cornell%20Summer%20Notes%202010/Lecture_Notes/Probability%20Tools%20with%20Examples.pdf).
- Dutta, S. (2012). Multiplicative Random Walk Metropolis-Hastings on the Real Line. *Sankhya. Series B*, **74**, 315–342. Also available at arXiv.
- Dutta, S. and Bhattacharya, S. (2014). Markov Chain Monte Carlo Based on Deterministic Transformations. *Statistical Methodology*, **16**, 100–

116. Also available at <http://arxiv.org/abs/1106.5850>. Supplement available at <http://arxiv.org/abs/1306.6684>.

Engle, R. F. (1982). Autoregressive Conditional Heteroskedasticity With Estimates of the Variance of U.K. Inflation. *Econometrica*, **50**, 987–1008.

Ephraty, A., Tabrikian, J., and Messer, H. (2001). Underwater Source Detection Using a Spatial Stationary Test. *The Journal of the Acoustical Society of America*, **109**, 1053–1063.

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, **1**, 209–230.

Fichtenholz, G. M. (1970). *Infinite Series: Rudiments*. Gordon and Breach Publishing, New York.

Frazier, P. I. (2018). A Tutorial on Bayesian Optimization. arXiv preprint.

Fuentes, M. (2002). Spectral Methods for Nonstationary Spatial Processes. *Biometrika*, **89**, 197–210.

Fuentes, M. (2005). A Formal Test for Non-Stationarity of Spatial Stochastic Processes. *Journal of Multivariate Analysis*, **96**, 30–54.

Gelfand, A. E., Diggle, P., Fuentes, M., and Guttorp, P. (2011). *Handbook of Spatial Statistics*. Chapman & Hall/CRC Press, Boca Raton, FL.

Gelman, A. and Rubin, D. B. (1992). Inference From Iterative Simulation Using Multiple Sequences. *Statistical Science*, **7**, 457–472.

Geweke, J. (1992). Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 169–193, Oxford. Clarendon Press.

- Ghoudi, K., Kulperger, R. J., and Rémillard, B. (2001). A Nonparametric Test of Serial Independence for Time Series and Residuals. *Journal of Multivariate Analysis*, **79**, 191–218.
- Gieser, P. W. and Randles, R. H. (1997). A Nonparametric Test of Independence Between Two Vectors. *Journal of the American Statistical Association*, **92**, 561–567.
- Gilks, W. R. and Roberts, G. O. (1996). Strategies for improving MCMC. In W. Gilks, S. Richardson, and D. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, Interdisciplinary Statistics, pages 89–114, London. Chapman and Hall.
- Giraitis, L., Leipus, R., and Surgailis, D. (2005). Recent Advances in ARCH Modelling. In *Long Memory in Economics*, pages 3–39, Berlin. Springer.
- Giraud, C. (2015). *Introduction to High-Dimensional Statistics*. CRC Press, New York.
- Guan, Y., Sherman, M., and Calvin, J. A. (2004). A Nonparametric Test For Spatial Isotropy Using Subsampling. *Journal of the American Statistical Association*, **99**, 810–821.
- Guha, S. (2020). *Some Theoretical and Methodological Contributions to the Dynamic Modeling of Discrete-Time Spatial Time Series Data*. Doctoral thesis, Indian Statistical Institute.
- Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, Princeton, New Jersey.
- Hoeffding, W. (1948). A Nonparametric Test of Independence. *Annals of Mathematical Statistics*, **19**, 546–557.
- Horsley, S. (1772). ΚΟΣ ΚΙΝΟΝ ΕΠΑΤΟΣ Θ ΕΝΟΣ. or, The Sieve of Eratosthenes. Being an Account of his Method of Finding all the Prime Numbers by the Rev. Samuel Horsley, F. R. S. *Philosophical Transactions (1683–1775)*, **62**, 327–347.

- Hunter, D. R. and Lange, K. (2000). Quantile Regression via an MM Algorithm. *Journal of Computational and Graphical Statistics*, **9**(1), 60–77.
- Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. Otexts, Online.
- Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. John Wiley & Sons, Ltd, Chichester, UK.
- Ilyin, V. A. and Poznyak, E. G. (1982). *Fundamentals of Mathematical Analysis, Vol.1*. Mir Publishers, Moscow.
- Jones, P. D., New, M., Parker, D. E., Martin, S., and Rigor, I. G. (1999). Surface Air Temperature and its Variations Over the Last 150 Years. *Reviews of Geophysics*, **37**, 173–199.
- Jun, M. and Genton, M. (2012). A Test For Stationarity of Spatio-Temporal Random Fields On Planar and Spherical Domains. *Statistica Sinica*, **22**, 1737–1764.
- Kaufman, D., McKay, N., Routson, C., M.Erb, Dätwyler, C., Sommer, P. S., Heiri, O., and Davis, B. (2020). A Global Database of Holocene Paleotemperature Records. *Scientific Data*, **7**(115), 1–13. Available at <https://doi.org/10.1038/s41597-020-0445-3>.
- Kawata, T. (1972). *Fourier Analysis in Probability Theory*. Academic Press, New York.
- Knopp, K. (1990). *Theory and Application of Infinite Series*. Dover Publishers, New York.
- Kwiatkowski, D., Schmidt, P., and Shin, Y. (1992). Testing the Null Hypothesis of Atationarity Against the Alternative of a Unit Root. *Journal of Econometrics*, **54**, 159–178.

- Landau, E. (1906). Über den Zusammenhang einiger neuer Sätze der analytischen Zahlentheorie. *Wiener Sitzungberichte, Math. Klasse*, **115**, 589–632.
- Lange, K. (2010). *Numerical Analysis for Statisticians*. Springer-Verlag, New York.
- Lawson, A. (2009). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*. Chapman & Hall/CRC Press, Boca Raton, FL.
- Li, B., Genton, M. G., and Sherman, M. (2008). Testing the Covariance Structure of Multivariate Random Fields. *Biometrika*, **95**, 813–829.
- Lifyand, E., Tikhonov, S., and Zeltser, M. (2011). Extending Tests for Convergence of Number Series. *Journal of Mathematical Analysis and Applications*, **377**, 194–206.
- Lioen, W. M. and van de Lune, J. (1994). Systematic Computations on Mertens' Conjecture and Dirichlet's Divisor Problem by Vectorized Sieving. In K. Apt, L. Schrijver, and N. Temme, editors, *From Universal Morphisms to Megabytes: a Baayen Space Odyssey*, pages 421–432, CWI, Amsterdam.
- Liu, J. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, New York.
- Liu, J. S. and Sabatti, S. (2000). Generalized Gibbs Sampler and Multigrid Monte Carlo for Bayesian Computation. *Biometrika*, **87**, 353–369.
- Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, London.
- Møller, J. and Waagepetersen, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall/CRC, Boca Raton, Florida.
- Montgomery, D. C., Jennings, C. L., and Kulahci, M. (2016). *Introduction to Time Series Analysis and Forecasting*. John Wiley and Sons, New Jersey.



- Mukhopadhyay, S. and Bhattacharya, S. (2012). Perfect Simulation for Mixtures with Known and Unknown Number of Components. *Bayesian Analysis*, **7**, 675–714.
- Øksendal, B. (2000). *Stochastic Differential Equations*. Springer-Verlag, Hiedelberg, New York. 5th Edition.
- Ornstein, L. S. and Uhlenbeck, G. E. (1930). On the Theory of Brownian Motion. *Physical Review*, **36**, 823–841.
- O’Sullivan, D. and Unwin, D. J. (2003). *Geographical Information Analysis*. Wiley, Hoboken, NJ.
- Paciorek, C. J., Yanosky, J. D., and Puett, R. C. (2009). Practical Large-Scale Spatio-Temporal Modeling of Particulate Matter Concentrations. *The Annals of Applied Statistics*, **3**, 370–397.
- Pakes, A. G. (2004). Convergence and Divergence of Random Series. *Australia and New Zealand Journal of Statistics*, **46**, 29–40.
- Philips, P. C. B. and Perron, P. (1988). Testing for a Unit Root in Time Series Regression. *Biometrika*, **75**, 335–346.
- Pickover, C. A. (2002). *The Mathematics of Oz: Mental Gymnastics from Beyond the Edge*. Cambridge University Press, U. K.
- Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley, New York.
- Raftery, A. E. and Lewis, S. M. (1992). How Many Iterations in the Gibbs Sampler? In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 763–773, Oxford. Clarendon Press.
- Resnick, S. I. (2014). *A Probability Path*. Springer-Verlag, New York.

- Riemann, B. (1859). Ueber die Anzahl der Primzahlen unter einer gegebenen Grösse. Monatsberichte der Berliner Akademie. In *Gesammelte Werke*, Teubner, Leipzig (1892), Reprinted by Dover, New York (1953). Original manuscript (with English translation). Reprinted in (Borwein et al. 2008) and (Edwards 1974).
- Robert, C. P. (1995). Convergence control methods for Markov chain Monte Carlo algorithms. *Statistical Science*, **10**, 231–253.
- Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer-Verlag, New York.
- Roy, V. (2019). Convergence Diagnostics for Markov Chain Monte Carlo. *Annual Review of Statistics and its Application*. To appear. Available at “<https://arxiv.org/pdf/1909.11827.pdf>”.
- Rudin, W. (1976). *Principles of Mathematical Analysis*. McGraw-Hill, New York.
- Santner, T. J., Williams, B. J., and I. Notz, W. (2003). *The Design and Analysis of Computer Experiments*. Springer-Verlag, New York.
- Schabenberger, D. and Gotway, C. A. (2005). *Statistical Methods for Spatial Data Analysis*. Chapman and Hall, London.
- Schervish, M. J. (1995). *Theory of Statistics*. Springer-Verlag, New York.
- Shumway, R. H. and Stoffer, D. S. (2006). *Time Series Analysis and Its Applications With R Examples*. Springer, New York.
- Spivak, M. (1994). *Calculus, Publish or Perish*.
- Straumann, D. (2005). Estimation in Conditionally Heteroscedastic Time Series Models. In *Volume 181 of Lecture Notes in Statistics*, Berlin. Springer-Verlag.
- Strauss, D. J. (1975). A Model for Clustering. *Biometrika*, **63**, 467–475.

- Stroock, D. (1999). *Probability Theory: An Analytic View*. Cambridge University Press, U. K.
- Stute, W. and Schumann, G. (1980). A General Glivenko-Cantelli Theorem for Stationary Sequences of Random Observations. *Scandinavian Journal of Statistics*, **7**, 102–104.
- Sury, B. (2003). Bernoulli Numbers and the Riemann Zeta Function. *Resonance*, **8**, 54–62.
- Taylor, S. J. (1986). *Modelling Financial Time Series*. Wiley, Chichester.
- Um, Y. and Randles, R. H. (2001). A Multivariate Nonparametric Test of Independence Among Many Vectors. *Journal of Nonparametric Statistics*, **13**, 699–708.
- van Delft, A., Characiejus, V., and Dette, H. (2018). A Nonparametric Test for Stationarity in Functional Time Series. arXiv preprint arXiv:1708.05248.
- Waller, L. A. and Gotway, C. A. (2004). *Applied Spatial Statistics for Public Health Data*. Wiley, Hoboken, NJ.