

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/360456798>

A BRIEF TREATISE ON BAYESIAN INVERSE REGRESSION

Thesis · May 2022

CITATIONS

0

READS

116

2 authors:



Debashis Chatterjee
Indian Statistical Institute

25 PUBLICATIONS 16 CITATIONS

[SEE PROFILE](#)



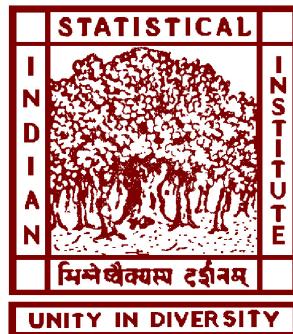
Sourabh Bhattacharya
Indian Statistical Institute

214 PUBLICATIONS 822 CITATIONS

[SEE PROFILE](#)

A BRIEF TREATISE ON BAYESIAN INVERSE REGRESSION

DEBASHIS CHATTERJEE



Indian Statistical Institute, Kolkata

DEBASHIS CHATTERJEE

Thesis submitted to the Indian Statistical Institute, Kolkata
in partial fulfillment of the requirements
for the award of the degree of
Doctor of Philosophy.

Thesis Advisor : Dr. Sourabh Bhattacharya



Indian Statistical Institute
203, B. T. Road, Kolkata, India.

ABSTRACT

Inverse problems, where in a broad sense the task is to learn from the noisy response about some unknown function, usually represented as the argument of some known functional form, has received wide attention in the general scientific disciplines. However, apart from the class of traditional inverse problems, there exists another class of inverse problems, which qualify as more authentic class of inverse problems, but unfortunately did not receive as much attention.

In a nutshell, the other class of inverse problems can be described as the problem of predicting the covariates corresponding to given responses and the rest of the data. Since the model is built for the responses conditional on the covariates, the inverse nature of the prediction problem is evident. Our motivating example in this regard arises in palaeoclimate reconstruction, where the model is built for the multivariate species composition conditional on climate; however, it is of interest to predict past climate given the modern species and climate data and the fossil species data. In the Bayesian context, it is natural to consider a prior for covariate prediction.

In this thesis, we bring to attention such a class of inverse problems, which we refer to as ‘inverse regression problems’ to distinguish them from the traditional inverse problems, which are typically special cases of the former, as we point out. Development of the Bayesian inverse regression setup is the goal of this thesis. We particularly focus on Bayesian model adequacy test and Bayesian model and variable selection in the inverse contexts, proposing new approaches and illuminating their asymptotic properties.

Towards Bayesian model adequacy, we adopt and extend the inverse reference distribution approach of [Bhattacharya \(2013\)](#), proving the convergence properties. Along the way, out of necessity, we develop asymptotic theories for Bayesian covariate consistency

and posterior convergence theories of unknown functions modeled by suitable stochastic processes embedded in normal, double-exponential, binary and Poisson distributions that include rates of convergence and misspecifications.

In the realm of inverse model and variable selection, we first develop an asymptotic theory for Bayes factors in the general setup, and then introduce pseudo-Bayes factors for model selection, showing that the asymptotic properties of the two approaches are in agreement, while the latter is more useful from several theoretical and computational perspectives. Along with the inverse regression setup we also develop the forward regression context, where the aim is to predict new responses given known covariate values, and illustrate the suitability, differences and advantages of the approaches, with various theoretical examples and simulation experiments. We further propose and develop a novel Bayesian multiple testing procedure for model and variable selection in the inverse regression setup, also exploring its elegant asymptotic properties. Our simulation studies demonstrate that this approach outperforms Bayes and pseudo-Bayes factors with respect to inverse model and variable selection.

As an interesting application encompassing most of our developments, we attempt to evaluate if the future world is likely to experience the terrifying global warming projection that has perturbed the scientists and policymakers the world over. Showing that the question falls within the purview of inverse regression problems, we propose a novel nonparametric model for climate dynamics based on Gaussian processes and exploit our inverse regression methodologies to conclude that there is no real threat to the world as far as global warming is concerned.

Contents

ABSTRACT	2
1 Introduction	1
1.1 A brief survey of inverse regression	3
1.2 Relation between traditional forward problems, traditional inverse problems and inverse regression problems	9
1.3 Areas of inverse regression seeking attention for development	12
2 A Statistical Perspective on Inverse and Inverse Regression Problems	17
2.1 Introduction	17
2.2 Traditional inverse problem	21
2.3 Linear inverse problem	25
2.4 Links between Bayesian inverse problems based on Gaussian process prior and deterministic regularizations	32
2.5 Regularization using differential operators and connection with Gaussian process	37
2.6 The Bayesian approach to inverse problems in Hilbert spaces	41
2.7 Conclusion	43
3 Thesis Layout and Overview of Our Contributions	45
3.1 Thesis layout	45
3.2 An overview of our contributions	46

4 Posterior Convergence of Gaussian and General Stochastic Process Regression Under Possible Misspecifications	52
4.1 Introduction	52
4.2 The Gaussian process regression setup	57
4.3 The general nonparametric regression setup	68
4.4 Conclusion	72
APPENDICES	74
Appendix 4.A1 Preliminaries for ensuring posterior consistency under general set-up	74
Appendix 4.A2 Verification of the assumptions of Shalizi for the Gaussian process model with normal errors	78
Appendix 4.A3 Verification of Shalizi's conditions for Gaussian process regression with double exponential error distribution	92
Appendix 4.A4 Verification of the assumptions of Shalizi for the general stochastic process model	104
5 Posterior Convergence of Nonparametric Binary and Poisson Regression Under Possible Misspecifications	114
5.1 Introduction	114
5.2 Model setup and preliminaries of the binary regression	116
5.3 Model setup and preliminaries of Poisson regression	118
5.4 Assumptions and their discussions	119
5.5 Main results on posterior convergence	123
5.6 Rate of convergence	128
5.7 Consequences of model misspecification	129
5.8 Conclusion	131
APPENDICES	132

Appendix 5.A1 Verification of Assumptions (S1) to (S7) of Shalizi for binary regression	132
Appendix 5.A2 Verification of Assumptions (S1) to (S7) of Shalizi for Poisson regression	143
6 Posterior Consistency of Bayesian Inverse Regression and Inverse Reference Distributions	151
6.1 Introduction	151
6.2 Preliminaries and the general setup	153
6.3 Discussion regarding consistency of the LOO-CV and the IRD approach	155
6.4 Prior for \tilde{x}_i	157
6.5 Consistency of the LOO-CV posteriors	160
6.6 Consistency of the IRD approach	162
6.7 Discussion of the applicability of our asymptotic results in the inverse regression contexts	164
6.8 Simulation studies	165
6.9 Conclusion	169
7 A Short Note on Almost Sure Convergence of Bayes Factors in the General Setup	170
7.1 Introduction	170
7.2 The general setup for model comparison using Bayes factors	172
7.3 Convergence of Bayes factors	173
7.4 Conclusion	177
APPENDICES	179
Appendix 7.A1 Illustration of our result on Bayes factor with competing $AR(1)$ models	179
Appendix 7.A2 Verification of Shalizi's conditions for model \mathcal{M}_2	190

Appendix 7.A3 Convergence of Bayes factor when ρ_1 , ρ_2 , σ_1 and σ_2 are all unknown	194
Appendix 7.A4 A first look at the applicability of our Bayes factor result to some infinite-dimensional models	197
8 Convergence of Pseudo-Bayes Factors in Forward and Inverse Regression Problems	204
8.1 Introduction	204
8.2 Preliminaries and general setup for forward and inverse regression problems	208
8.3 Convergence of PBF in forward problems	213
8.4 Convergence results for PBF in inverse regression: first setup	216
8.5 Convergence results for PBF in inverse regression: second setup	220
8.6 Illustrations of PBF convergence in forward regression problems	224
8.7 Illustrations of PBF convergence in inverse regression problems	238
8.8 Simulation experiments	247
8.9 Summary and future direction	262
APPENDICES	265
Appendix 8.A1 A result on sufficient condition for (S6) of Shalizi	265
Appendix 8.A2 Proof of Theorem 47	266
9 A Bayesian Multiple Testing Paradigm for Model Selection in Inverse Regression Problems	268
9.1 Introduction	268
9.2 A multiple testing framework for model selection in inverse regression problems	273
9.3 Asymptotic properties of the posterior probabilities of the alternative hypotheses	283
9.4 Asymptotic optimality theory for our multiple testing procedure	293

9.5	Asymptotic theory of the error measures	295
9.6	Modification of the multiple testing procedure for practical implementation	299
9.7	First simulation study: selection among Poisson and geometric parametric and nonparametric inverse regression models	300
9.8	Second simulation study: variable selection in Poisson and geometric linear and nonparametric regression models when true model is Poisson linear regression	307
9.9	Summary and discussion	311
10	How Ominous is the Future Global Warming Premonition?	315
10.1	Introduction	315
10.2	Gaussian process based emulation process for nonparametric climate dynamics	324
10.3	Prior distributions for θ_f and σ_ϵ^2	330
10.4	Posterior distributions of current and future time series in our dynamic Gaussian process approach	331
10.5	A Bayesian multiple testing framework for GCM selection in any given climate scenario	334
10.6	Implementation of the Bayesian multiple testing procedure	341
10.7	GCM selection results	344
10.8	GCM simulations as ensembles: extension of our Gaussian process emula- tion approach to the multivariate situation	352
10.9	Results for the multivariate climate dynamics	357
10.10	Future climate forecast with our Bayesian Gaussian process dynamics model	361
10.11	A brief discussion of the existing works on climate model evaluation . .	362
10.12	Summary and discussion	365
11	Summary and Future Directions	369

11.1 Summary	369
11.2 Future directions	371
REFERENCES	392

Listing of figures

1.3.1 Demonstration of posterior inconsistency in inverse regression problems.	
The vertical line denotes the true value.	14
6.8.1 Demonstration of posterior consistency in inverse parametric Poisson regression. The vertical line denotes the true value.	166
6.8.2 Demonstration of posterior consistency in inverse nonparametric Poisson regression. The vertical line denotes the true value.	168
9.7.1 $cFDR_{nm}$ and $cFNR_{nm}$ as functions of β_{nm} in the non-misspecified case.	305
9.7.2 $cFDR_{nm}$ and $cFNR_{nm}$ as functions of β_{nm} in the misspecified case. . .	306
9.8.1 $cFDR_{nm}$ and $cFNR_{nm}$ as functions of β_{nm} in the non-misspecified situation of the model and variable selection problem.	310
9.8.2 $cFDR_{nm}$ and $cFNR_{nm}$ as functions of β_{nm} in the misspecified situation of the model and variable selection problem.	311
10.1. Visualization of the HadCRUT4 data (thick, black line) and the GCM based time series. The temperature is in $^{\circ}\text{C}$	320
10.7.1 FDR and cFNR for GCM selection in the climate scenarios A1B and A2 using Bayesian multiple testing.	345
10.7.2 FDR and cFNR for GCM selection in the climate scenarios B1 and Commitment using Bayesian multiple testing.	346

10.7.3The posteriors corresponding to the HadCRUT4 data or the current global temperature (CGT) conditional on GCM-based average time series are shown as colour plots with progressively higher densities depicted by progressively intense colours. Also shown are the HadCRUT4 data (CGT), GCM based time series (MBGT) and the average of GCM based time series (AMBGT). The temperature is in °C and in the log-scale. . .	349
10.7.4The posteriors corresponding to the HadCRUT4 data or the current global temperature (CGT) conditional on individual best GCM time series are shown as colour plots with progressively higher densities depicted by progressively intense colours. Also shown are the HadCRUT4 data (CGT), GCM based time series (MBGT) and the average of GCM based time series (AMBGT). The temperature is in °C and in the log-scale.	351
10.9.1The posteriors $[\bar{x}_0, \bar{x}_1, \dots, \bar{x}_{T_0} \mathbf{x}_{T_0+1}, \dots, \mathbf{x}_T]$ are shown as colour plots with progressively higher densities depicted by progressively intense colours, along with the HadCRUT4 data (CGT) and the average of GCM based time series (AMBGT). The temperature is in °C and in the log-scale.	358
10.9.2The posteriors $[x_0^{(max)}, x_1^{(max)}, \dots, x_{T_0}^{(max)} \mathbf{x}_{T_0+1}, \dots, \mathbf{x}_T]$ are shown as colour plots with progressively higher densities depicted by progressively intense colours, along with the HadCRUT4 data (CGT) and the maximum of model based global temperature (MMBGT). The temperature is in °C and in the log-scale.	360

10.10 The posteriors $[x_{T_0+1}, \dots, x_T | x_1, \dots, x_{T_0}]$ for future climate prediction are shown as colour plots, along with the posterior modes of the Gaussian process forecasted global temperature (GPFGT), best GCM-specific model based forecasted global temperature (MBFGT) and average model based forecasted global temperature (AMBFGT). The temperature is in °C and in the log-scale. 363

Listing of tables

<p>8.8.1 Results of our simulation study for model selection using FPBF and IPBF.</p> <p>The last two columns show the estimates of $\frac{1}{n} \sum_{i=1}^n \log \pi(y_{i1} \mathbf{Y}_{nm,-i}, \mathbf{X}_n, \mathcal{M})$ and $\frac{1}{n} \sum_{i=1}^n \log \pi(y_{i1} \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M})$, respectively, for forward and inverse setups.</p>	<p>255</p>
<p>8.8.2 Results of our simulation study for model and variable selection using FPBF and IPBF. The last two columns show the estimates of $\frac{1}{n} \sum_{i=1}^n \log \pi(y_{i1} \mathbf{Y}_{nm,-i}, \mathbf{X}_n, \mathcal{M})$ and $\frac{1}{n} \sum_{i=1}^n \log \pi(y_{i1} \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M})$, respectively, for forward and inverse setups.</p>	<p>261</p>
<p>10.7.1 Goodness-of-fit check for the best GCMs with respect to averaged time series. Here 95% BCI stands for 95% Bayesian credible intervals.</p>	<p>350</p>
<p>10.7.2 Goodness-of-fit check for the best GCMs with respect to individual time series. Here 95% BCI stands for 95% Bayesian credible intervals.</p>	<p>352</p>
<p>10.9.1 Goodness-of-fit check for ensembles of GCM time series with respect to $[\bar{x}_0, \bar{x}_1, \dots, \bar{x}_{T_0} \mathbf{x}_{T_0+1}, \dots, \mathbf{x}_T]$. Here 95% BCI stands for 95% Bayesian credible intervals.</p>	<p>359</p>
<p>10.9.2 Goodness-of-fit check for ensembles of GCM time series with respect to $[x_0^{(max)}, x_1^{(max)}, \dots, x_{T_0}^{(max)} \mathbf{x}_{T_0+1}, \dots, \mathbf{x}_T]$. Here 95% BCI stands for 95% Bayesian credible intervals.</p>	<p>361</p>

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 the Interdisciplinary Statistical Research Unit (ISRU) for their continued support.

Finally, I take this opportunity to express my deep sense of gratitude to all my teachers, past and present, who taught me the beautiful subject called Statistics.

1

Introduction

The task in traditional inverse problems is to learn about unknown functions from noisy observations, where the unknown function is typically represented as an argument of some known functional. This paradigm fits a large class of real examples covering various scientific disciplines, and hence, has been able to attract wide attention. Somewhat paradoxically, although such a class of problems seem to be clearly of statistical nature, the statistical literature is not as rich with respect to such inverse problems compared to the other scientific literatures. More unfortunately, there exists another class of statistical problems which, according to us, qualify as *bona fide* inverse problems, yet the statistical literature is almost oblivious of such existence.

For us, the motivating example for the latter class of inverse problems arises in quantitative palaeoclimate reconstruction where ‘modern data’ consisting of multivariate counts of species are available along with the observed climate values. Also available are fossil assemblages of the same species, but deposited in lake sediments for past thousands

of years. This is the fossil species data. However, the past climates corresponding to the fossil species data are unknown, and it is of interest to predict the past climates given the modern data and the fossil species data. Roughly, the species composition are regarded as functions of climate variables, since in general ecological terms, variations in climate drives variations in species, but not vice versa. However, since the interest lies in prediction of climate variables, the inverse nature of the problem is clear. The past climates, which must be regarded as random variables, may also be interpreted as *unobserved covariates*. It is thus natural to put a prior probability distribution on the unobserved covariates. From the nature of the problem, its difference with the traditional inverse problems is evident. Further examples are provided in Section 1.1.1.

Technically, given a data set \mathbf{y} that depends upon covariates \mathbf{x} , having a probability distribution $f(\mathbf{y}|\mathbf{x}, \theta)$ where θ is the model parameter, we call the problem ‘forward’ if it is of interest to predict \tilde{y} for given \tilde{x} . This is the conventional and much-studied statistical paradigm, from both classical and Bayesian perspectives.

On the other hand, we refer to the problem as ‘inverse’ if the goal is to predict the corresponding unknown \tilde{x} given a new observed \tilde{y} and the rest of the data. The literature on such inverse problems is very scarce, in spite of abundance of examples in the inverse problem paradigm. In fact, with respect to predicting unknown covariates from the responses, mostly inverse linear regression, particularly in the classical set-up, has been considered in the literature. The paucity of the literature on inverse problems in the above sense already calls for significant literature development in the subject area. In this regard, in Section 1.3 we shall point out some areas of inverse problems that seek thorough development, which we shall focus on for the theoretical and methodological aspects of this thesis.

To distinguish the traditional inverse problems from the covariate-prediction perspective, we use the phrase ‘inverse regression’ to refer to the latter.

In what follows, we first briefly survey the available literature on inverse regression

in Section 1.1. In Section 1.2 we explore the relationships between traditional forward problems, traditional inverse problems and inverse regression problems and argue that only inverse regression qualifies as *bona fide* inverse problems and include the traditional inverse problems as special cases in the sense that the underlying model may involve unknown functions, which need to be learned about, apart from predicting the unknown covariates. In Section 1.3, we touch upon the development-seeking areas of inverse regression problems.

1.1 A brief survey of inverse regression

We first provide some examples of inverse regression, several of which are based on [Avenhaus *et al.* \(1980\)](#).

1.1.1 Further examples of inverse regression

Example 1: Measurement of nuclear materials

Measurement of the amount of nuclear materials such as plutonium by direct chemical means is an extremely difficult exercise. This motivates model-based methods. For instance, there are physical laws relating heat production or the number of neutrons emitted (the dependent response variable y) to the amount of material present, the latter being the independent variable x . But any measurement instrument based on the physical laws first needs to be calibrated. In other words, the unknown parameters of the model needs to be learned, using known inputs and outputs. However, the independent variables are usually subject to measurement errors, motivating a statistical model. Thus, conditionally on x and parameter(s) θ , $y \sim P(\cdot|x, \theta)$, where $P(\cdot|x, \theta)$ denotes some appropriate probability model. Given \mathbf{y}_n and \mathbf{x}_n , and some specific \tilde{y} , the corresponding \tilde{x} needs to be predicted.

Example 2: Estimation of family incomes

Suppose that it is of interest to estimate the family incomes in a certain city through public opinion poll. Most of the population, however, will be unwilling to provide reliable answers to the questionnaires. One way to extract relatively reliable figures is to consider some dependent variable, say, housing expenses (y), which is supposed to strongly depend on family income (x); see Muth (1960), and such that the population is less reluctant to divulge the correct figures related to y . From past survey data on \mathbf{x}_n and \mathbf{y}_n , and using current data from families who may provide reliable answers related to both x and y , a statistical model may be built, using which the unknown family incomes may be predicted, given their household incomes.

Example 3: Missing variables

In regression problems where some of the covariate values x_i are missing, they may be estimated from the remaining data and the model. In this context, Press and Scott (1975) considered a simple linear regression problem in a Bayesian framework. Under special assumptions about the error and prior distributions, they showed that an optimal procedure for estimating the linear parameters is to first estimate the missing x_i from an inverse regression based only on the complete data pairs.

Example 4: Bioassay

It is usual to investigate the effects of substances (y) given in several dosages on organisms (x) using bioassay methods. In this context it may be of interest to determine the dosage necessary to obtain some interesting effect, making inverse regression relevant (see, for example, Rasch *et al.* (1973)).

Example 5: Learning the Milky Way

The modelling of the Milky Way galaxy is an integral step in the study of galactic dynamics; this is because knowledge of model parameters that define the Milky Way directly influences our understanding of the evolution of our galaxy. Since the nature of the Galaxy's phase space, in the neighbourhood of the Sun, is affected by distinct Milky Way features, measurements of phase space coordinates of individual stars that live in this neighbourhood of the Sun, will bear information about the influence of such features. Then, inversion of such measurements can help us learn the parameters that describe such Milky Way features. In this regard, learning about the location of the Sun with respect to the center of the galaxy, given the two-component velocities of the stars in the vicinity of the Sun, is an important problem. For k such stars, Chakrabarty *et al.* (2015) model the $k \times 2$ -dimensional velocity matrix \mathbf{V} as a function of the galactocentric location (\mathbf{S}) of the Sun, denoted by $\mathbf{V} = \boldsymbol{\xi}(\mathbf{S})$. For a given observed value \mathbf{V}^* of \mathbf{V} , it is then of interest to obtain the corresponding \mathbf{S}^* . Since $\boldsymbol{\xi}$ is unknown, Chakrabarty *et al.* (2015) model $\boldsymbol{\xi}$ as a matrix-variate Gaussian process, and consider the Bayesian approach to learning about \mathbf{S}^* , given data $\{(\mathbf{S}_i, \mathbf{V}_i) : i = 1, \dots, n\}$ simulated from established astrophysical models, and the observed velocity matrix \mathbf{V}^* .

We now provide a brief overview of the methods of inverse linear regression, which is the most popular among inverse regression problems. Our discussion is generally based on Hoadley (1970) and Avenhaus *et al.* (1980).

1.1.2 Inverse linear regression

Let us consider the following simple linear regression model: for $i = 1, \dots, n$,

$$y_i = \alpha + \beta x_i + \sigma \epsilon_i, \quad (1.1.1)$$

where $\epsilon_i \stackrel{iid}{\sim} N(0, 1)$.

For simplicity, let us consider a single unknown \tilde{x} , associated with a further set of m responses $\{\tilde{y}_1, \dots, \tilde{y}_m\}$, related by

$$\tilde{y}_i = \alpha + \beta \tilde{x} + \tau \tilde{\epsilon}_i, \quad (1.1.2)$$

for $i = 1, \dots, m$, where $\tilde{\epsilon}_i \stackrel{iid}{\sim} N(0, 1)$ and are independent of the ϵ_i 's associated with (1.1.1).

The interest in the above problem is inference regarding the unknown x . Based on (1.1.1), first least squares estimates of α and β are obtained as

$$\hat{\beta} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}; \quad (1.1.3)$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}, \quad (1.1.4)$$

where $\bar{y} = \sum_{i=1}^n y_i/n$ and $\bar{x} = \sum_{i=1}^n x_i/n$. Then, letting $\tilde{\bar{y}} = \sum_{i=1}^n \tilde{y}_i/n$, a ‘classical’ estimator of x is given by

$$\hat{x}_C = \frac{\tilde{\bar{y}} - \hat{\alpha}}{\hat{\beta}}, \quad (1.1.5)$$

which is also the maximum likelihood estimator for the likelihood associated with (1.1.1) and (1.1.2), assuming known σ and τ . However,

$$E \left[(\hat{x}_C - x)^2 | \alpha, \beta, \sigma, \tau, x \right] = \infty, \quad (1.1.6)$$

which prompted Krutchkoff (1967) to propose the following ‘inverse’ estimator:

$$\hat{x}_I = \hat{\gamma} + \hat{\delta} \tilde{\bar{y}}, \quad (1.1.7)$$

where

$$\hat{\delta} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (y_i - \bar{y})^2}; \quad (1.1.8)$$

$$\hat{\gamma} = \bar{x} - \hat{\delta}\bar{y}, \quad (1.1.9)$$

are the least squares estimators of the slope and intercept when the x_i are regressed on the y_i . It can be shown that the mean square error of this inverse estimator is finite. However, Williams (1969) showed that if $\sigma^2 = \tau^2$ and if the sign of β is known, then the unique unbiased estimator of x has infinite variance. Williams advocated the use of confidence limits instead of point estimators.

Hoadley (1970) derive confidence limits setting $\sigma = \tau$ and assuming without loss of generality that $\sum_{i=1}^n x_i = 0$. Under these assumptions, the maximum likelihood estimators of σ^2 with \mathbf{x}_n and \mathbf{y}_n only, $\tilde{\mathbf{y}}_n = (\tilde{y}_1, \dots, \tilde{y}_n)^T$ only, and with the entire available data set are, respectively,

$$\hat{\sigma}_1^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2; \quad (1.1.10)$$

$$\hat{\sigma}_2^2 = \frac{1}{m-1} \sum_{i=1}^n (\tilde{y}_i - \bar{\tilde{y}})^2; \quad (1.1.11)$$

$$\hat{\sigma}^2 = \frac{1}{n-2+m-1} [(n-2)\sigma_1^2 + (m-1)\sigma_2^2]. \quad (1.1.12)$$

Now consider the F -statistic $F = \frac{n\hat{\beta}^2}{\hat{\sigma}^2}$ for testing the hypothesis $\beta = 0$. Note that under the null hypothesis this statistic has the F distribution with 1 and $n+m$ degrees of freedom. For $m = 1$,

$$\hat{\beta}(\hat{x}_C - x) \sqrt{\frac{n}{\sigma^2(n+1+x^2)}}$$

has a t distribution with $n-2$ degrees of freedom. Letting $F_{\alpha;1,\nu}$ denote the upper α point of the F distribution with 1 and ν degrees of freedom, a confidence set S can be

derived as follows:

$$S = \begin{cases} \{x : x_L \leq x \leq x_U\} & \text{if } F > F_{\alpha;1,n-2}; \\ \{x : x \leq x_L\} \cup \{x \geq x_U\} & \text{if } \frac{n+1}{n+1+\hat{x}_C^2} F_{\alpha;1,n-2} \leq F < F_{\alpha;1,n-2}; \\ (-\infty, \infty) & \text{if } F < \frac{n+1}{n+1+\hat{x}_C^2} F_{\alpha;1,n-2}, \end{cases} \quad (1.1.13)$$

where x_L and x_U are given by

$$\frac{F\hat{x}_C}{F - F_{\alpha;1,n-1}} \pm \frac{\{F_{\alpha;1,n-2} [(n+1)(F - F_{\alpha;1,n-2}) + F\hat{x}_C^2]\}^{1/2}}{F - F_{\alpha;1,n-2}}.$$

Hence, if $F < \frac{n+1}{n+1+\hat{x}_C^2} F_{\alpha;1,n-2}$, then the associated confidence interval is $S = (-\infty, \infty)$, which is of course useless.

[Hoadley \(1970\)](#) present a Bayesian analysis of this problem, presented below in the form of the following two theorems.

Theorem 1 (Hoadley (1970)) *Assume that $\sigma = \tau$, and let x be independent of $(\alpha, \beta, \sigma^2)$ a priori. With any prior $\pi(x)$ on x and the prior*

$$\pi(\alpha, \beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

on $(\alpha, \beta, \sigma^2)$, the posterior density of x given by

$$\pi(x | \mathbf{y}_n, \mathbf{x}_n, \tilde{\mathbf{y}}_n) \propto \pi(x) L(x),$$

where

$$L(x) = \frac{\left(1 + \frac{n}{m} + x^2\right)^{\frac{m+n-3}{2}}}{\left[1 + \frac{n}{m} + R\hat{x}_C^2 + \left(\frac{F}{m+n-3} + 1\right) (x - R\hat{x}_C)^2\right]^{\frac{m+n-2}{2}}},$$

where

$$R = \frac{F}{F + m + n - 3}.$$

1.2. RELATION BETWEEN TRADITIONAL FORWARD PROBLEMS, TRADITIONAL INVERSE PROBLEMS AND INVERSE REGRESSION PROBLEMS

For $m = 1$, Hoadley (1970) present the following result characterizing the inverse estimator \hat{x}_I :

Theorem 2 (Hoadley (1970)) *Consider the following informative prior on x :*

$$x = t_{n-3} \frac{n+1}{n-3},$$

where t_ν denotes the t distribution with ν degrees of freedom. Then the posterior distribution of x given \mathbf{y}_n , \mathbf{x}_n and $\tilde{\mathbf{y}}_n$ has the same distribution as

$$\hat{x}_I + t_{n-2} \sqrt{\frac{n+1 + \frac{\hat{x}_I^2}{R}}{F+n-2}}.$$

In particular, it follows from Theorem 2 that the posterior mean of x is \hat{x}_I when $m = 1$. In other words, the inverse estimator \hat{x}_I is Bayes with respect to the squared error loss and a particular informative prior distribution for x .

Since the goal of Hoadley (1970) was to provide a theoretical justification of the inverse estimator, he had to choose a somewhat unusual prior so that it leads to \hat{x}_I as the posterior mean. In general it is not necessary to confine ourselves to any specific prior for Bayesian analysis of inverse regression. It is also clear that the Bayesian framework is appropriate for any inverse regression problem, not just linear inverse regression; indeed, the palaeoclimate reconstruction problem (Haslett *et al.* (2006)) and the Milky Way problem (Chakrabarty *et al.* (2015)) are examples of very highly non-linear inverse regression problems.

1.2 Relation between traditional forward problems, traditional inverse problems and inverse regression problems

The similarities and dissimilarities between inverse problems and the more traditional forward problems are usually not clearly explained in the literature, and often “ill-posed”

1.2. RELATION BETWEEN TRADITIONAL FORWARD PROBLEMS, TRADITIONAL INVERSE PROBLEMS AND INVERSE REGRESSION PROBLEMS

is the term used to loosely characterize inverse problems. We point out that these two problems may have the same goal or different goal, while both consider the same model given the data. We first elucidate using the traditional case of deterministic differential equations, that the goals of the two problems may be the same. Consider a dynamical system

$$\frac{dx_t}{dt} = G(t, x_t, \theta), \quad (1.2.1)$$

where G is a known function and θ is a parameter. In the forward problem the goal is to obtain the solution $x_t \equiv x_t(\theta)$, given θ and the initial conditions, whereas, in the inverse problem, the aim is to obtain θ given the solution process x_t . Realistically, the differential equation would be perturbed by noise, and so, one observes the data $\mathbf{y} = (y_1, \dots, y_T)^T$, where

$$y_t = x_t(\theta) + \epsilon_t, \quad (1.2.2)$$

for noise variables ϵ_t having some suitable independent and identical (*iid*) error distribution q , which we assume to be known for simplicity of illustration. A typical method of estimating θ , employed by the scientific community, is the method of calibration, where the solution of (2.1.1) would be obtained for each θ -value on a proposed grid of plausible values, and a set $\tilde{\mathbf{y}}(\theta) = (\tilde{y}_1(\theta), \dots, \tilde{y}_T(\theta))^T$ is generated from the model (2.1.2) for every such θ after simulating, for $i = 1, \dots, T$, $\tilde{\epsilon}_t \stackrel{iid}{\sim} q$; then forming $\tilde{y}_t(\theta) = x_t(\theta) + \tilde{\epsilon}_t$, and finally reporting that value θ in the grid as an estimate of the true values for which $\|\mathbf{y} - \tilde{\mathbf{y}}(\theta)\|$ is minimized, given some distance measure $\|\cdot\|$; maximization of the correlation between \mathbf{y} and $\tilde{\mathbf{y}}(\theta)$ is also considered. In other words, the calibration method makes use of the forward technique to estimate the desired quantities of the model. On the other hand, the inverse problem paradigm attempts to directly estimate θ from the observed data \mathbf{y} usually by minimizing some discrepancy measure between \mathbf{y} and $\mathbf{x}(\theta)$, where $\mathbf{x}(\theta) = (x_1(\theta), \dots, x_T(\theta))^T$. Hence, from this perspective the goals of both forward and inverse approaches are the same, that is, estimation of θ . However, the forward

1.2. RELATION BETWEEN TRADITIONAL FORWARD PROBLEMS, TRADITIONAL INVERSE PROBLEMS AND INVERSE REGRESSION PROBLEMS

approach is well-posed, whereas, the inverse approach is often ill-posed. To clarify, note that within a grid, there always exists some $\hat{\theta}$ that minimizes $\|\mathbf{y} - \tilde{\mathbf{y}}(\theta)\|$ among all the grid-values. In this sense the forward problem may be thought of as well-posed. However, direct minimization of the discrepancy between \mathbf{y} and $\mathbf{x}(\theta)$ with respect to θ is usually difficult and for high-dimensional θ , the solution to the minimization problem is usually not unique, and small perturbations of the data causes large changes in the possible set of solutions, so that the inverse approach is usually ill-posed. Of course, if the minimization is sought over a set of grid values of θ only, then the inverse problem becomes well-posed.

From the statistical perspective, the unknown parameter θ of the model needs to be learned, in either the classical or the Bayesian way, and hence, in this sense there is no real distinction between forward and inverse problems. Indeed, statistically, since the data are modeled conditionally on the parameters, all problems where learning the model parameter given the data is the goal, are inverse problems. We remark that the literature usually considers learning unknown functions from the data in the realm of inverse problems, but a function is nothing but an infinite-dimensional parameter, which constitutes a very common learning problem in statistics.

We now explain when forward and inverse problems can differ in their aims, and are significantly different even from the statistical perspective. In this regard, consider Example 1 of Chapter 1.1.1, namely, the palaeoclimate reconstruction problem. Recall that the inverse nature of the problem is associated with prediction of the fossil climate values, given the pollen assemblages. The forward problem would result, if given the fossil climate values (if known), the fossil pollen abundances (if unknown), were to be predicted.

Note that the class of inverse regression problems includes the class of traditional inverse problems. The Milky Way problem (Example 5 of Chapter 1.1.1) is an example where learning the unknown, matrix-variate function ξ (inverse problem) was required,

even though learning about \mathbf{S} , the galactocentric location of the sun (inverse regression problem) was the primary goal. The Bayesian approach allowed learning both \mathbf{S} and ξ simultaneously and coherently.

In the palaeoclimate models proposed in Haslett *et al.* (2006), Bhattacharya (2006) and Mukhopadhyay and Bhattacharya (2013), although species assemblages are modeled conditionally on climate variables, the functional relationship between species and climate are not even approximately known. In all these works, it is of interest to learn about the functional relationship as well as to predict the unobserved climate values, the latter being the main aim. Again, the Bayesian approach facilitated appropriate learning of both the unknown quantities.

Our discussion shows that statistically, there is nothing special about the existing literature on inverse problems that considers estimation of unknown (perhaps, infinite-dimensional) parameters, and the only class of problems that can be truly regarded as inverse problems as distinguished from forward problems are those which consider prediction of unknown covariates from the dependent response data. It is, however, important to point out that in our thesis, (asymptotic) posterior learning of the unknown covariates and its ramifications require (asymptotic) posterior learning of the associated unknown functions, establishing the connection between inverse regression problems and traditional inverse problems.

1.3 Areas of inverse regression seeking attention for development

1.3.1 Consistency of covariates in inverse regression problems

In the above linear inverse regression, notice that if $\tau > 0$, then the variance of the estimator of x can not tend to zero, even as the data size tends to infinity. This shows that no estimator of x can be consistent. The same argument applies even to Bayesian

approaches; for any sensible prior on x that does not give point mass to the true value of x , the posterior of x will not converge to the point mass at the true value of x as the data size increases indefinitely. The arguments remain valid for any inverse regression problem where the response variable y probabilistically depends upon the independent variable x . Not only in inverse regression problems, even in forward regression problems where the interest is in prediction of y given x , any estimate of y or any posterior predictive distribution y will be inconsistent.

To give an example of inconsistency in non-linear and non-normal inverse problem, consider the following set-up: $y_i \stackrel{iid}{\sim} \text{Poisson}(\theta x_i)$, for $i = 1, \dots, n$, where $\theta > 0$ and $x_i > 0$ for each i . Let us consider the prior $\pi(\theta) \equiv 1$ for all $\theta > 0$. For some $i^* \in \{1, \dots, n\}$ let us assume the leave-one-out cross-validation set-up in that we wish to learn $x = x_{i^*}$ assuming it is unknown, from the rest of the data. Putting the prior $\pi(x) \equiv 1$ for $x > 0$, the posterior of x is given by (see [Bhattacharya and Haslett \(2007\)](#), [Bhattacharya \(2013\)](#))

$$\pi(x|\mathbf{x}_n \setminus x_i, \mathbf{y}_n) \propto \frac{x^{y_i}}{(x + \sum_{j \neq i} x_j)^{(\sum_{j=1}^n y_j + 1)}}. \quad (1.3.1)$$

Figure 1.3.1 displays the posterior of x when $i^* = 10$, for increasing sample size. Observe that the variance of the posterior does not decrease even with sample size as large as 100,000, clearly demonstrating inconsistency. Hence, special, innovative priors are necessary for consistency in such cases.

Recall that the underlying model may involve unknown functions as well, which may be modeled nonparametrically using appropriate stochastic processes. In such situations, conceiving of appropriate consistent priors for the unknown covariates may be rendered a far more difficult problem.

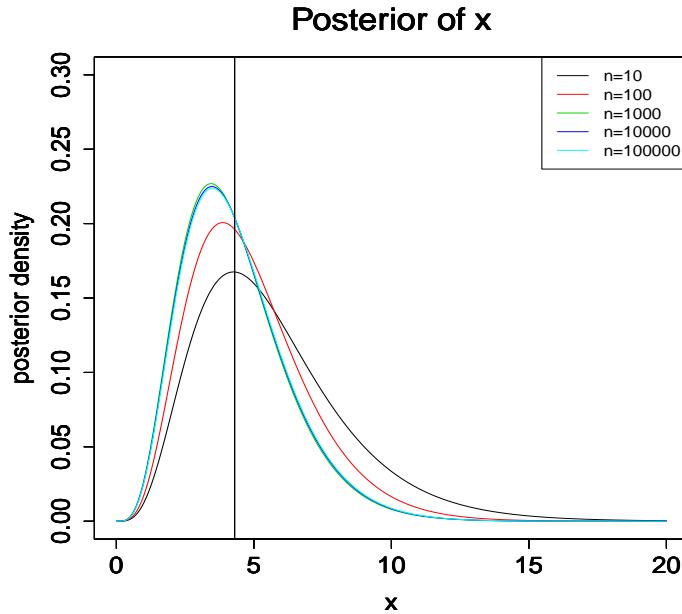


Figure 1.3.1: Demonstration of posterior inconsistency in inverse regression problems. The vertical line denotes the true value.

1.3.2 Model adequacy tests in inverse regression problems

Assessment of model adequacy is always fundamental in statistics – this basic realization has given rise to a huge literature on testing goodness of model fit. However, compared to the classical literature, the Bayesian literature on model adequacy test is much scarce. A comprehensive overview of the existing approaches is provided in [Vehtari and Ojanen \(2012\)](#). Two relatively prominent existing formal and general approaches in this direction are those of [Gelman *et al.* \(1996\)](#) and [Bayarri and Berger \(2000\)](#). The former relies on posterior predictive P -value associated with a discrepancy measure that is a function of the data as well as the parameters. The latter criticize this approach on account of ‘double use of the data’ and come up with two alternative P -values, demonstrating their advantages over the posterior predictive P -value. Indeed, double use of the data prevents the posterior predictive P -value to have uniform distribution on $[0, 1]$, while the P -values of [Bayarri and Berger \(2000\)](#) at least asymptotically has the desired uniform

distribution on $[0, 1]$.

[Bhattacharya \(2013\)](#) introduced a different approach to Bayesian model assessment in inverse regression problems. Broadly, the model assessment method of [Bhattacharya \(2013\)](#) is based on the simple idea that the model fits the data if the posterior distribution of the random variables corresponding to the covariates capture the observed values of the covariates. Assuming that the covariates are unobserved, one can predict these values in terms of the posterior distribution of the random quantities standing for the (assumed) missing covariates. [Bhattacharya \(2013\)](#) demonstrated that it makes more sense to consider leave-one-out cross-validation (LOO-CV) of the covariates particularly when some of the model parameters are given improper prior. From the traditional statistical perspective, LOO-CV is also a very natural method in model assessment. Briefly, based on the LOO-CV posteriors of the covariates, some appropriate ‘inverse reference distribution’ (IRD) is constructed. This IRD can be viewed as a distribution of some appropriate statistic associated with the unobserved covariates. If the distribution captures the observed statistic associated with the observed covariates, then the model is said to fit the data. Otherwise, the model does not fit the data. [Bhattacharya \(2013\)](#) provided a Bayesian decision theoretic justification of the key idea and show that the relevant IRD based posterior probability analogue of the aforementioned P -values have the uniform distribution on $[0, 1]$. Furthermore, ample simulation studies and successful applications to several real, palaeoclimate models and data sets reported in [Bhattacharya \(2006\)](#), [Bhattacharya \(2013\)](#) and [Mukhopadhyay and Bhattacharya \(2013\)](#), vindicate the practicality and usefulness of the IRD approach.

It is however, important to establish the asymptotic validity of the test proposed by [Bhattacharya \(2013\)](#), which is again clearly related to establishment of consistency of covariates in inverse regression problems discussed in Section 1.3.1.

1.3.3 Model selection in inverse regression problems

Comparison of different inverse models given the same data, or covariate selection in inverse models, seems to be non-existent in the literature, either classical or Bayesian. Given the abundance of inverse regression problems, this seems to be somewhat surprising. Although the IRD approach of [Bhattacharya \(2013\)](#) seems to be useful for evaluating adequacy of any given inverse Bayesian model, it does not seem to be straightforward to extend the method to the model selection paradigm. That is, if several models pass the IRD based model adequacy test, the question of selecting the best model among them for final inference, remains.

It is hence essential to develop theories and methods for model selection in inverse problems. Here it is useful to remark that although there exists a plethora of approaches to model and covariate selection in the forward context, they do not necessarily admit easy generalization to inverse regression setups. The lack of covariate consistency in inverse regression setups with general priors also demonstrate that even if new model and variable selection approaches may be constructed for inverse setups, asymptotic validation of such approaches is likely to be highly non-trivial.

2

A Statistical Perspective on Inverse and Inverse Regression Problems

2.1 Introduction

The similarities and dissimilarities between inverse problems and the more traditional forward problems are usually not clearly explained in the literature, and often “ill-posed” is the term used to loosely characterize inverse problems. We point out that these two problems may have the same goal or different goal, while both consider the same model given the data. We first elucidate using the traditional case of deterministic differential equations, that the goals of the two problems may be the same. Consider a dynamical system

$$\frac{dx_t}{dt} = G(t, x_t, \theta), \quad (2.1.1)$$

where G is a known function and θ is a parameter. In the forward problem the goal is to obtain the solution $x_t \equiv x_t(\theta)$, given θ and the initial conditions, whereas, in the inverse problem, the aim is to obtain θ given the solution process x_t . Realistically, the differential equation would be perturbed by noise, and so, one observes the data $\mathbf{y} = (y_1, \dots, y_T)^T$, where

$$y_t = x_t(\theta) + \epsilon_t, \quad (2.1.2)$$

for noise variables ϵ_t having some suitable independent and identical (*iid*) error distribution q , which we assume to be known for simplicity of illustration. A typical method of estimating θ , employed by the scientific community, is the method of calibration, where the solution of (2.1.1) would be obtained for each θ -value on a proposed grid of plausible values, and a set $\tilde{\mathbf{y}}(\theta) = (\tilde{y}_1(\theta), \dots, \tilde{y}_T(\theta))^T$ is generated from the model (2.1.2) for every such θ after simulating, for $i = 1, \dots, T$, $\tilde{\epsilon}_t \stackrel{iid}{\sim} q$; then forming $\tilde{y}_t(\theta) = x_t(\theta) + \tilde{\epsilon}_t$, and finally reporting that value θ in the grid as an estimate of the true values for which $\|\mathbf{y} - \tilde{\mathbf{y}}(\theta)\|$ is minimized, given some distance measure $\|\cdot\|$; maximization of the correlation between \mathbf{y} and $\tilde{\mathbf{y}}(\theta)$ is also considered. In other words, the calibration method makes use of the forward technique to estimate the desired quantities of the model. On the other hand, the inverse problem paradigm attempts to directly estimate θ from the observed data \mathbf{y} usually by minimizing some discrepancy measure between \mathbf{y} and $\mathbf{x}(\theta)$, where $\mathbf{x}(\theta) = (x_1(\theta), \dots, x_T(\theta))^T$. Hence, from this perspective the goals of both forward and inverse approaches are the same, that is, estimation of θ . However, the forward approach is well-posed, whereas, the inverse approach is often ill-posed. To clarify, note that within a grid, there always exists some $\hat{\theta}$ that minimizes $\|\mathbf{y} - \tilde{\mathbf{y}}(\theta)\|$ among all the grid-values. In this sense the forward problem may be thought of as well-posed. However, direct minimization of the discrepancy between \mathbf{y} and $\mathbf{x}(\theta)$ with respect to θ is usually difficult and for high-dimensional θ , the solution to the minimization problem is usually not unique, and small perturbations of the data causes large changes in the possible set of solutions, so that the inverse approach is usually ill-posed. Of course, if

the minimization is sought over a set of grid values of θ only, then the inverse problem becomes well-posed.

From the statistical perspective, the unknown parameter θ of the model needs to be learned, in either the classical or the Bayesian way, and hence, in this sense there is no real distinction between forward and inverse problems. Indeed, statistically, since the data are modeled conditionally on the parameters, all problems where learning the model parameter given the data is the goal, are inverse problems. We remark that the literature usually considers learning unknown functions from the data in the realm of inverse problems, but a function is nothing but an infinite-dimensional parameter, which constitutes a very common learning problem in statistics.

We now explain when forward and inverse problems can differ in their aims, and are significantly different even from the statistical perspective. In this regard, consider Example 1 of Chapter 1.1.1, namely, the palaeoclimate reconstruction problem. Recall that the inverse nature of the problem is associated with prediction of the fossil climate values, given the pollen assemblages. The forward problem would result, if given the fossil climate values (if known), the fossil pollen abundances (if unknown), were to be predicted.

Note that the class of inverse regression problems includes the class of traditional inverse problems. The Milky Way problem (Example 5 of Chapter 1.1.1) is an example where learning the unknown, matrix-variate function ξ (inverse problem) was required, even though learning about S , the galactocentric location of the sun (inverse regression problem) was the primary goal. The Bayesian approach allowed learning both S and ξ simultaneously and coherently.

In the palaeoclimate models proposed in [Haslett et al. \(2006\)](#), [Bhattacharya \(2006\)](#) and [Mukhopadhyay and Bhattacharya \(2013\)](#), although species assemblages are modeled conditionally on climate variables, the functional relationship between species and climate are not even approximately known. In all these works, it is of interest to learn about the

functional relationship as well as to predict the unobserved climate values, the latter being the main aim. Again, the Bayesian approach facilitated appropriate learning of both the unknown quantities.

Our discussion shows that statistically, there is nothing special about the existing literature on inverse problems that considers estimation of unknown (perhaps, infinite-dimensional) parameters, and the only class of problems that can be truly regarded as inverse problems as distinguished from forward problems are those which consider prediction of unknown covariates from the dependent response data. Nevertheless, due to its importance, as well as for the sake of completeness, we shall attempt to provide a comprehensive review of the traditional inverse problems related to learning of unknown functions. Indeed, in our thesis, (asymptotic) posterior learning of the unknown covariates and its ramifications require (asymptotic) posterior learning of the associated unknown functions, rendering such a review all the more important.

The rest of this chapter is structured as follows. In Section 2.2 we discuss the general inverse model, providing several examples. In Section 2.3 we focus on linear inverse problems, which constitute the most popular class of inverse problems, and review the links between the Bayesian approach based on simple finite difference priors and the deterministic Tikhonov regularization. Connections between Gaussian process based Bayesian inverse problems and deterministic regularizations are reviewed in Section 2.4. In Section 2.5 we provide an overview of the connections between the Gaussian process based Bayesian approach and regularization using differential operators, which generalizes the discussion of Section 2.3 on the connection between finite difference priors and the Tikhonov regularization. The Bayesian approach to inverse problems in Hilbert spaces is discussed in Section 2.6. Finally, we make concluding remarks in Section 2.7.

2.2 Traditional inverse problem

Suppose that one is interested in learning about the function θ given the noisy observed responses $\mathbf{y}_n = (y_1, \dots, y_n)^T$, where the relationship between θ and \mathbf{y}_n is governed by following equation (2.2.1) :

$$y_i = G(x_i, \theta) + \epsilon_i, \quad (2.2.1)$$

for $i = 1, \dots, n$, where x_i are known covariates or design points, ϵ_i are errors associated with the i -th observation and G is a forward operator defined appropriately, which is usually allowed to be non-injective.

Note that since $\boldsymbol{\epsilon}_n = (\epsilon_1, \dots, \epsilon_n)^T$ is unknown, the noisy observation vector \mathbf{y}_n itself may not be in the image set of G . If θ is a p -dimensional parameter, then there will often be situations when the number of equations is smaller than the number of unknowns, in the sense that $p > n$ (see, for example, Dashti and Stuart (2015)). Modern statistical research is increasingly coming across such inverse problems termed as “ill-posed” which are not in the exact domain of statistical estimation procedures (O’Sullivan (1986)) where the maximum likelihood solution or classical least squares may not be uniquely defined and with very bad perturbation sensitivity of the classical solution. However, although such problematic issues are said to characterize inverse problems, the problems in fact fall in the so-called “large p small n ” paradigm and has received wide attention in statistics; see, for example, Bühlmann and van de Geer (2011), Giraud (2015). A key concept involved in handling such problems is inclusion of some appropriate penalty term in the discrepancy to be minimized with respect to θ . Such regularization methods are initiated by Tikhonov (1963) and Tikhonov and Arsenin (1977). Under this method, usually a criterion of the following form is chosen for the minimization purpose:

$$\frac{1}{n} \sum_{i=1}^n [y_i - G(x_i, \theta)]^2 + \lambda J(\theta), \quad \lambda > 0. \quad (2.2.2)$$

The functional J is chosen such that highly implausible or irregular values of θ has large values (O’Sullivan (1986)). Thus, depending on the problem at hand, $J(\theta)$ can be used to induce “sparsity” in an appropriate sense so that the minimization problem may be well-defined. We next present several examples of classical inverse problems based on Aster *et al.* (2013).

2.2.1 Examples of inverse problems

Vertical seismic profiling

In this scientific field, one wishes to learn about the vertical seismic velocity of the material surrounding a borehole. A source generates downward-propagating seismic wavefront at the surface, and in the borehole, a string of seismometers sense these seismic waves. The arrival times of the seismic wavefront at each instrument are measured from the recorded seismograms. These times provide information on the seismic velocity for vertically traveling waves as a function of depth. The problem is nonlinear if it is expressed in terms of seismic velocities. However, we can linearize this problem via a simple change of variables, as follows. Letting z denote the depth, it is possible to parameterize the seismic structure in terms of slowness, $s(z)$, which is the reciprocal of the velocity $v(z)$. The observed travel time at depth z can then be expressed as:

$$t(z) = \int_0^z s(u)du = \int_0^\infty s(u)H(z-u)du, \quad (2.2.3)$$

where H is the Heaviside step function. The interest is to learn about $s(z)$ given observed $t(z)$. Theoretically, $s(z) = \frac{dt(z)}{dz}$, but in practice, simply differentiating the observations need not lead to useful solutions because noise is generally present in the observed times $t(z)$, and naive differentiation may lead to unrealistic features of the solution.

Estimation of buried line mass density from vertical gravity anomaly

Here the problem is to estimate an unknown buried line mass density $m(x)$ from data on vertical gravity anomaly, $d(x)$, observed at some height, h . The mathematical relationship between $d(x)$ and $m(x)$ is given by

$$d(x) = \int_{-\infty}^{\infty} \frac{h}{[(u - x)^2 + h^2]^{\frac{3}{2}}} m(u) du.$$

As before, noise in the data renders the above linear inverse problem difficult. Variations of the above example has been considered in [Aster et al. \(2013\)](#).

Estimation of incident light intensity from diffracted light intensity

Consider an experiment in which an angular distribution of illumination passes through a thin slit and produces a diffraction pattern, for which the intensity is observed. The data, $d(s)$, are measurements of diffracted light intensity as a function of the outgoing angle $-\pi/2 \leq s \leq \pi/2$. The goal here is to obtain the intensity of incident light on the slit, $m(\theta)$, as a function of the incoming angle $-\pi/2 \leq \theta \leq \pi/2$, using the following mathematical relationship:

$$d(s) = \int_{-\pi/2}^{\pi/2} (\cos(s) + \cos(\theta))^2 \left(\frac{\sin(\pi(\sin(s) + \sin(\theta)))}{\pi(\sin(s) + \sin(\theta))} \right)^2 m(\theta) d\theta.$$

Groundwater pollution source history reconstruction problem

Consider the problem of recovering the history of groundwater pollution at a source site from later measurements of the contamination at downstream wells to which the contaminant plume has been transported by advection and diffusion. The mathematical model for contamination transport is given by the following advection-diffusion equation

with respect to t and transported site x :

$$\begin{aligned}\frac{\partial C}{\partial t} &= D \frac{\partial^2 C}{\partial x^2} - \nu \frac{\partial C}{\partial x} \\ C(0, t) &= C_{in}(t) \\ C(x, t) &\rightarrow 0 \text{ as } x \rightarrow \infty.\end{aligned}$$

In the above, D is the diffusion coefficient, ν is the velocity of the groundwater flow, and $C_{in}(t)$ is the time history of contaminant injection at $x = 0$. The solution to the above advection-diffusion equation is given by

$$C(x, T) = \int_0^T C_{in}(t) f(x, T-t) dt,$$

where

$$f(x, T-t) = \frac{x}{2\sqrt{\pi D(T-t)^3}} \exp\left[\frac{(x-\nu(T-t))^2}{4D(T-t)}\right].$$

It is of interest to learn about $C_{in}(t)$ from data observed on $C(x, T)$.

Transmission tomography

The most basic physical model for tomography assumes that wave energy traveling between a source and receiver can be considered to be propagating along infinitesimally narrow ray paths. In seismic tomography, if the slowness at a point x is $s(x)$, and the ray path is known, then the travel time for seismic energy transiting along that ray path is given by the line integral along ℓ :

$$t = \int_{\ell} s(x(l)) dl. \quad (2.2.4)$$

Learning of $s(x)$ from t is required. Note that (2.2.4) is a high-dimensional generalization of (2.2.3). In reality, seismic ray paths will be bent due to refraction and/or reflection,

resulting in nonlinear inverse problem.

The above examples demonstrate the ubiquity of linear inverse problems. As a result, in the next section we take up the case of linear inverse problems and illustrate the Bayesian approach in details, also investigating connections with the deterministic approach employed by the general scientific community.

2.3 Linear inverse problem

The motivating examples and discussions in this section are based on [Bui-Thanh \(2012\)](#).

Let us consider the following one-dimensional integral equation on a finite interval as in equation (2.3.1):

$$G(x, \theta) = \int K(x, t) \theta(t) dt, \quad (2.3.1)$$

where $K(x, \cdot)$ is some appropriate, known, real-valued function given x . Now, let the dataset be $\mathbf{y}_n = (y_1, y_2, \dots, y_n)^T$. Then for a known system response $K(x_i, t)$ for the dataset, the equation can be written as follows:

$$y_i = \int G(x_i, \theta) + \epsilon_i ; \quad i \in \{1, 2, \dots, n\} \quad (2.3.2)$$

As a particular example, let $G(x, \theta) = \int_0^1 K(x, t) \theta(t) dt$, where $K(x, t) = \frac{1}{\sqrt{2\pi\psi^2}} \exp\left\{-\frac{(x-t)^2}{2\psi^2}\right\}$ is the Gaussian kernel and $\theta : [0, 1] \mapsto \mathbb{R}$ is to be learned given the data \mathbf{y}_n and $\mathbf{x}_n = (x_1, \dots, x_n)^T$. We first illustrate the Bayesian approach and draw connections with the traditional approach of Tikhonov's regularization when the integral in G is discretized. In this regard, let $x_i = (i-1)/n$, for $i = 1, \dots, n$. Letting $\boldsymbol{\theta} = (\theta(x_1), \dots, \theta(x_n))^T$ and \mathbf{K} be the $n \times n$ matrix with the (i, j) -th element $K(x_i, x_j)/n$, and $\boldsymbol{\epsilon}_n = (\epsilon_1, \dots, \epsilon_n)^T$, the discretized version of (2.3.2) can be represented as

$$\mathbf{y}_n = \mathbf{K}\boldsymbol{\theta} + \boldsymbol{\epsilon}_n. \quad (2.3.3)$$

We assume that $\boldsymbol{\epsilon}_n \sim N_n(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$, that is, an n -variate normal with mean $\mathbf{0}_n$, an n -dimensional vector with all components zero, and covariance $\sigma^2 \mathbf{I}_n$, where \mathbf{I}_n is the n -th order identity matrix.

2.3.1 Smooth prior on θ

To reflect the belief that the function θ is smooth, one may presume that

$$\theta(x_i) = \frac{\theta(x_{i-1}) + \theta(x_{i+1})}{2} + \tilde{\epsilon}_i, \quad (2.3.4)$$

where, for $i = 1, \dots, n$, $\tilde{\epsilon}_i \stackrel{iid}{\sim} N(0, \tilde{\sigma}^2)$. Thus, *a priori*, $\theta(x_i)$ is assumed to be an average of its nearest neighbors to quantify smoothness, with an additive random perturbation term. Letting

$$\mathbf{L} = \frac{1}{2} \begin{pmatrix} -1 & 2 & -1 & 0 & \cdots & \cdots \\ 0 & -1 & 2 & -1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 2 & -1 \end{pmatrix}, \quad (2.3.5)$$

and $\tilde{\boldsymbol{\epsilon}} = (\tilde{\epsilon}_1, \dots, \tilde{\epsilon}_n)^T$, it follows from (2.3.4) that

$$\mathbf{L}\boldsymbol{\theta} = \tilde{\boldsymbol{\epsilon}}, \quad (2.3.6)$$

Now, noting that the Laplacian of a twice-differentiable real-valued function f with independent arguments z_1, \dots, z_k is given by $\Delta f = \sum_{i=1}^k \frac{\partial^2 f}{\partial z_i^2}$, we have

$$\Delta\theta(x_j) \approx n^2(\mathbf{L}\boldsymbol{\theta})_j, \quad (2.3.7)$$

where $(\mathbf{L}\boldsymbol{\theta})_j$ is the j -th element of $\mathbf{L}\boldsymbol{\theta}$.

However, the rank of \mathbf{L} is $n - 1$, and boundary conditions on the Laplacian operator

is necessary to ensure positive definiteness of the operator. In our case, we assume that $\theta \equiv 0$ outside $[0, 1]$, so that we now assume $\theta(0) = \frac{\theta(x_1)}{2} + \tilde{\epsilon}_0$ and $\theta(x_n) = \frac{\theta(x_{n-1})}{2} + \tilde{\epsilon}_n$, where $\tilde{\epsilon}_0$ and $\tilde{\epsilon}_n$ are *iid* $N(0, \tilde{\sigma}^2)$. With this modification, the prior on $\boldsymbol{\theta}$ is given by

$$\pi(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2\tilde{\sigma}^2}\|\tilde{\mathbf{L}}\boldsymbol{\theta}\|^2\right), \quad (2.3.8)$$

where $\|\cdot\|$ is the Euclidean norm and

$$\tilde{\mathbf{L}} = \frac{1}{2} \begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & \cdots \\ -1 & 2 & -1 & 0 & \cdots & \cdots \\ 0 & -1 & 2 & -1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & \cdots & 0 & -1 & 2 \end{pmatrix}. \quad (2.3.9)$$

Rather than assuming zero boundary conditions, more generally one may assume that $\theta(0)$ and $\theta(x_n)$ are distributed as $N\left(0, \frac{\tilde{\sigma}^2}{\delta_0^2}\right)$ and $N\left(0, \frac{\tilde{\sigma}^2}{\delta_n^2}\right)$, respectively. The resulting modified matrix is then given by

$$\hat{\mathbf{L}} = \frac{1}{2} \begin{pmatrix} 2\delta_0 & 0 & 0 & 0 & \cdots & \cdots \\ -1 & 2 & -1 & 0 & \cdots & \cdots \\ 0 & -1 & 2 & -1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & \cdots & 0 & 0 & 2\delta_n \end{pmatrix}. \quad (2.3.10)$$

To choose δ_0 and δ_n , one may assume that

$$\text{Var} [\theta(0)] = \frac{\tilde{\sigma}^2}{\delta_0^2} = \text{Var} [\theta(x_n)] = \frac{\tilde{\sigma}^2}{\delta_n^2} = \text{Var} [\theta(x_{[n/2]})] = \tilde{\sigma}^2 \boldsymbol{\epsilon}_{[n/2]}^T (\hat{\mathbf{L}}^T \hat{\mathbf{L}})^{-1} \boldsymbol{\epsilon}_{[n/2]},$$

where $[n/2]$ is the largest integer not exceeding $n/2$, and $\boldsymbol{\epsilon}_{[n/2]}$ is the $[n/2]$ -th canonical basis vector in \mathbb{R}^{n+1} . It follows that

$$\delta_0^2 = \delta_n^2 = \frac{1}{\boldsymbol{\epsilon}_{[n/2]}^T (\hat{\mathbf{L}}^T \hat{\mathbf{L}})^{-1} \boldsymbol{\epsilon}_{[n/2]}}.$$

Since this requires solving a non-linear equation (since $\hat{\mathbf{L}}$ contains δ_0 and δ_n), for avoiding computational complexity one may simply employ the approximation

$$\delta_0^2 = \delta_n^2 = \frac{1}{\boldsymbol{\epsilon}_{[n/2]}^T (\tilde{\mathbf{L}}^T \tilde{\mathbf{L}})^{-1} \boldsymbol{\epsilon}_{[n/2]}},$$

where $\tilde{\mathbf{L}}$ is given by (2.3.9).

2.3.2 Non-smooth prior on θ

To begin with, let us assume that θ has several points of discontinuities on the grid of points $\{x_0, \dots, x_n\}$. To reflect this information in the prior, one may assume that $\theta(0) = 0$ and for $i = 1, \dots, n$, $\theta(x_i) = \theta(x_{i-1}) + \tilde{\epsilon}_i$, where, as before, $\tilde{\epsilon}_i$ are *iid* $N(0, \tilde{\sigma}^2)$.

Then, with

$$\mathbf{L}^* = \frac{1}{2} \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & \cdots & \cdots \\ -1 & 1 & 0 & 0 & \cdots & \cdots & \cdots \\ 0 & -1 & 1 & 0 & 0 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 1 & 0 & 0 \\ 0 & 0 & \cdots & 0 & -1 & 1 & 0 \end{pmatrix}, \quad (2.3.11)$$

the prior is given by

$$\pi(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2\tilde{\sigma}^2}\|\mathbf{L}^*\boldsymbol{\theta}\|^2\right). \quad (2.3.12)$$

One may also flexibly account for any particular big jump. For instance, if for some $\ell \in \{0, \dots, n\}$, the jump $\theta(x_\ell) - \theta(x_{\ell-1})$ is particularly large compared to the other jumps, then it can be assumed that $\theta(x_\ell) = \theta(x_{\ell-1}) + \epsilon_\ell^*$, with $\epsilon_\ell^* \sim N\left(0, \frac{\tilde{\sigma}^2}{\xi^2}\right)$, where $\xi < 1$. Letting \mathbf{D}_ℓ be the diagonal matrix with ξ^2 being the ℓ -th diagonal element and 1 being the other diagonal elements, the prior is then given by

$$\pi(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2\tilde{\sigma}^2}\|\mathbf{D}_\ell \mathbf{L}^* \boldsymbol{\theta}\|^2\right). \quad (2.3.13)$$

A more general prior can be envisaged where the number and location of the jump discontinuities are unknown. Then we may consider a diagonal matrix $\mathbf{D} = \text{diag}\{\xi_1, \dots, \xi_n\}$, so that conditionally on the hyperparameters ξ_1, \dots, ξ_n , the prior on $\boldsymbol{\theta}$ is given by

$$\pi(\boldsymbol{\theta}|\xi_1, \dots, \xi_n) \propto \exp\left(-\frac{1}{2\tilde{\sigma}^2}\|\mathbf{D} \mathbf{L}^* \boldsymbol{\theta}\|^2\right). \quad (2.3.14)$$

Prior on ξ_1, \dots, ξ_n may be considered to complete the specification. These may also be estimated by maximizing the marginal likelihood obtained by integrating out $\boldsymbol{\theta}$, which is known as the ML-II method; see Berger (1985). Calvetti and Somersalo (2007) also

advocate likelihood based methods.

2.3.3 Posterior distribution

For convenience, let us generically denote the matrices \mathbf{L} , $\tilde{\mathbf{L}}$, $\hat{\mathbf{L}}$, \mathbf{L}^* , $\mathbf{D}_\ell \mathbf{L}^*$, $\mathbf{D}\mathbf{L}^*$, by $\boldsymbol{\Gamma}^{-\frac{1}{2}}$. Then it can be easily verified that the posterior of θ admits the following generic form:

$$\pi(\boldsymbol{\theta} | \mathbf{y}_n, \mathbf{x}_n) \propto \exp \left\{ - \left[\frac{1}{2\sigma^2} \|\mathbf{y}_n - \mathbf{K}\boldsymbol{\theta}\|^2 + \frac{1}{2\tilde{\sigma}^2} \|\boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{\theta}\|^2 \right] \right\}. \quad (2.3.15)$$

Note that the exponent of the posterior is of the form of the Tikhonov functional, which we denote by $T(\boldsymbol{\theta})$. The maximizer of the posterior, commonly known as the *maximum a posteriori* (MAP) estimator, is given by

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg \max_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta} | \mathbf{y}_n, \mathbf{x}_n) = \arg \min_{\boldsymbol{\theta}} T(\boldsymbol{\theta}). \quad (2.3.16)$$

In other words, the deterministic solution to the inverse problem obtained by Tikhonov's regularization is nothing but the Bayesian MAP estimator in our context.

Writing $\mathbf{H} = \frac{1}{\sigma^2} \mathbf{K}^T \mathbf{K} + \frac{1}{\tilde{\sigma}^2} \boldsymbol{\Gamma}^{-1}$, which is the Hessian of the Tikhonov functional (regularized misfit), and writing $\|\cdot\|_{\mathbf{H}} = \|\mathbf{H}^{\frac{1}{2}} \cdot\|$, it is clear that (2.3.15) can be simplified to the Gaussian form, given by

$$\pi(\boldsymbol{\theta} | \mathbf{y}_n, \mathbf{x}_n) \propto \exp \left\{ - \left\| \boldsymbol{\theta} - \frac{1}{\sigma^2} \mathbf{H}^{-1} \mathbf{K}^{-1} \mathbf{y}_n \right\|_{\mathbf{H}}^2 \right\}. \quad (2.3.17)$$

It follows from (2.3.17) that the inverse of the Hessian of the regularized misfit is the posterior covariance itself. From the above posterior it also trivially follows that

$$\hat{\boldsymbol{\theta}}_{MAP} = \frac{1}{\sigma^2} \mathbf{H}^{-1} \mathbf{K}^{-1} \mathbf{y}_n = \frac{1}{\sigma^2} \left(\frac{1}{\sigma^2} \mathbf{K}^T \mathbf{K} + \frac{1}{\tilde{\sigma}^2} \boldsymbol{\Gamma}^{-1} \right)^{-1} \mathbf{K}^T \mathbf{Y}_n, \quad (2.3.18)$$

which coincides with the Tikhonov solution for linear inverse problems. The connection between the traditional deterministic Tikhonov regularization approach with Bayesian analysis continues to hold even if the likelihood is non-Gaussian.

2.3.4 Exploration of the smoothness conditions

For deeper investigation of the smoothness conditions, let us write

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg \min_{\boldsymbol{\theta}} T(\boldsymbol{\theta}) = \sigma^2 \left(\frac{1}{2} \|\mathbf{y}_n - \tilde{\mathbf{y}}_n\|^2 + \frac{1}{2} \varrho \|\tilde{\boldsymbol{\Gamma}}^{\frac{1}{2}} \boldsymbol{\theta}\|^2 \right), \quad (2.3.19)$$

where $\tilde{\mathbf{y}}_n = \mathbf{K}\boldsymbol{\theta}$, $\varrho = \sigma^2/\tilde{\sigma}^2$ and $\tilde{\boldsymbol{\Gamma}}^{\frac{1}{2}} = \boldsymbol{\Gamma}^{-\frac{1}{2}}$. Now, from (2.3.7) it follows that for the smooth priors with the zero boundary conditions, our Tikhonov functional discretizes

$$T_\infty(\boldsymbol{\theta}) = \frac{1}{2} \|\mathbf{y}_n - \tilde{\mathbf{y}}_n\|^2 + \frac{1}{2} \varrho \|\Delta\theta\|_{L^2(0,1)}^2, \quad (2.3.20)$$

where $\|\cdot\|_{L^2(0,1)}^2 = \int_0^1 (\cdot)^2 dt$.

On the other hand, for the non-smooth prior (2.3.12), rather than discretizing $\Delta\theta$, $\nabla\theta$, that is, the gradient of θ , is discretized. In other words, for non-smooth priors, our Tikhonov functional discretizes

$$T_\infty(\boldsymbol{\theta}) = \frac{1}{2} \|\mathbf{y}_n - \tilde{\mathbf{y}}_n\|^2 + \frac{1}{2} \varrho \|\nabla\theta\|_{L^2(0,1)}^2. \quad (2.3.21)$$

Hence, realizations of prior (2.3.12) is less smooth compared to those of our smooth priors. However, the realizations (2.3.12) must be continuous. The priors given by (2.3.13) and (2.3.14) also support continuous functions as long as the hyperparameters are bounded away from zero. These facts, although clear, can be rigorously justified by functional analysis arguments, in particular, using the Sobolev imbedding theorem (see, for example, [Arbogast and Bona \(2008\)](#)).

2.4 Links between Bayesian inverse problems based on Gaussian process prior and deterministic regularizations

In this section, based on Rasmussen and Williams (2006), we illustrate the connections between deterministic regularizations such as those obtained from differential operators as above, and Bayesian inverse problems based on the very popular Gaussian process prior on the unknown function. A key tool for investigating such relationship is the reproducing kernel Hilbert space (RKHS).

2.4.1 RKHS

We adopt the following definition of RKHS provided in Rasmussen and Williams (2006):

Definition 3 (RKHS) *Let \mathcal{H} be a Hilbert space of real functions θ defined on an index set \mathcal{X} . Then \mathcal{H} is called an RKHS endowed with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ (and norm $\|\theta\|_{\mathcal{H}} = \langle \theta, \theta \rangle_{\mathcal{H}}$) if there exists a function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ with the following properties:*

- (a) *for every x , $\mathcal{K}(\cdot, x) \in \mathcal{H}$, and*
- (b) *\mathcal{K} has the reproducing property $\langle \theta(\cdot), \mathcal{K}(\cdot, x) \rangle_{\mathcal{H}} = \theta(x)$.*

Observe that since $\mathcal{K}(\cdot, x), \mathcal{K}(\cdot, x') \in \mathcal{H}$, it follows that $\langle \mathcal{K}(\cdot, x), \mathcal{K}(\cdot, x') \rangle_{\mathcal{H}} = \mathcal{K}(x, x')$. The Moore-Aronszajn theorem asserts that the RKHS uniquely determines \mathcal{K} , and vice versa. Formally,

Theorem 4 (Aronszajn (1950)) . *Let \mathcal{X} be an index set. Then for every positive definite function $\mathcal{K}(\cdot, \cdot)$ on $\mathcal{X} \times \mathcal{X}$ there exists a unique RKHS, and vice versa.*

Here, by positive definite function $\mathcal{K}(\cdot, \cdot)$ on $\mathcal{X} \times \mathcal{X}$, we mean $\int \mathcal{K}(x, x') g(x)g(x') d\nu(x)d\nu(x') > 0$ for all non-zero functions $g \in L_2(\mathcal{X}, \nu)$, where $L_2(\mathcal{X}, \nu)$ denotes the space of functions square-integrable on \mathcal{X} with respect to the measure ν .

Indeed, the subspace \mathcal{H}_0 of \mathcal{H} spanned by the functions $\{\mathcal{K}(\cdot, \mathbf{x}_i); i = 1, 2, \dots\}$ is dense in \mathcal{H} in the sense that every function in \mathcal{H} is a pointwise limit of a Cauchy sequence from \mathcal{H}_0 .

To proceed, we require the concepts of eigenvalues and eigenfunctions associated with kernels. In the following section we provide a briefing on these.

2.4.2 Eigenvalues and eigenfunctions of kernels

We borrow the statements of the following definition of eigenvalue and eigenfunction, and the subsequent statement of Mercer's theorem from [Rasmussen and Williams \(2006\)](#).

Definition 5 *A function $\psi(\cdot)$ that obeys the integral equation*

$$\int_{\mathcal{X}} \mathcal{C}(x, x') \psi(x) d\nu(x) = \lambda \psi(x'), \quad (2.4.1)$$

is called an eigenfunction of the kernel \mathcal{C} with eigenvalue λ with respect to the measure ν .

We assume that the ordering is chosen such that $\lambda_1 \geq \lambda_2 \geq \dots$. The eigenfunctions are orthogonal with respect to ν and can be chosen to be normalized so that $\int_{\mathcal{X}} \psi_i(\mathbf{x}) \psi_j(\mathbf{x}) d\nu(x) = \delta_{ij}$, where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.

The following well-known theorem (see, for example, [König \(1986\)](#)) expresses the positive definite kernel \mathcal{C} in terms of its eigenvalues and eigenfunctions.

Theorem 6 (Mercer's theorem) *Let (\mathcal{X}, ν) be a finite measure space and $\mathcal{C} \in L_{\infty}(\mathcal{X}^2, \nu^2)$ be a positive definite kernel. By $L_{\infty}(\mathcal{X}^2, \nu^2)$ we mean the set of all measurable functions $\mathcal{C} : \mathcal{X}^2 \mapsto \mathbb{R}$ which are essentially bounded, that is, bounded up to a set of ν^2 -measure zero. For any function \mathcal{C} in this set, its essential supremum, given by $\inf \{C \geq 0 : |\mathcal{C}(x_1, x_2)| < C, \text{ for almost all } (x_1, x_2) \in \mathcal{X} \times \mathcal{X}\}$ serves as the norm $\|\mathcal{C}\|$.*

Let $\psi_j \in L_2(\mathcal{X}, \nu)$ be the normalized eigenfunctions of \mathcal{C} associated with the eigenvalues $\lambda_j(\mathcal{C}) > 0$. Then

- (a) the eigenvalues $\{\lambda_j(\mathcal{C})\}_{j=1}^{\infty}$ are absolutely summable.
- (b) $\mathcal{C}(x, x') = \sum_{j=1}^{\infty} \lambda_j(\mathcal{C}) \psi_j(x) \bar{\psi}_j(x')$ holds ν^2 -almost everywhere, where the series converges absolutely and uniformly ν^2 -almost everywhere. In the above, $\bar{\psi}_j$ denotes the complex conjugate of ψ_j .

It is important to note the difference between the eigenvalue $\lambda_j(\mathcal{C})$ associated with the kernel \mathcal{C} and $\lambda_j(\Sigma_n)$ where Σ_n denotes the $n \times n$ Gram matrix with (i, j) -th element $\mathcal{C}(x_i, x_j)$. Observe that (see Rasmussen and Williams (2006)):

$$\lambda_j(\mathcal{C}) \psi_j(x') = \int_{\mathcal{X}} \mathcal{C}(x, x') \psi_j(x) d\nu(x) \approx \frac{1}{n} \sum_{i=1}^n \mathcal{C}(x_i, x') \psi_j(x_i), \quad (2.4.2)$$

where, for $i = 1, \dots, n$, $x_i \sim \nu$, assuming that ν is a probability measure. Now substituting $x' = x_i$; $i = 1, \dots, n$ in (2.4.2) yields the following approximate eigen system for the matrix Σ_n :

$$\Sigma_n \mathbf{u}_j \approx n \lambda_j(\mathcal{C}) \mathbf{u}_j, \quad (2.4.3)$$

where the i -th component of \mathbf{u}_j is given by

$$u_{ij} = \frac{\psi_j(x_i)}{\sqrt{n}}. \quad (2.4.4)$$

Since ψ_j are normalized to have unit norm, it holds that

$$\mathbf{u}_j^T \mathbf{u}_j = \frac{1}{n} \sum_{i=1}^n \psi_j^2(x_i) \approx \int_{\mathcal{X}} \psi_j^2(x) d\nu(x) = 1. \quad (2.4.5)$$

From (2.4.5) it follows that

$$\lambda_j(\Sigma_n) \approx n \lambda_j(\mathcal{C}). \quad (2.4.6)$$

Indeed, Theorem 3.4 of Baker (1977) shows that $n^{-1} \lambda_j(\Sigma_n) \rightarrow \lambda_j(\mathcal{C})$, as $n \rightarrow \infty$.

For our purposes the main usefulness of the RKHS framework is that $\|\theta\|_{\mathcal{H}}^2$ can be

perceived as a generalization of $\boldsymbol{\theta}^T \mathbf{K}^{-1} \boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\theta(x_1), \dots, \theta(x_n))^T$ and $\mathbf{K} = (\mathcal{K}(x_i, x_j))_{i,j=1,\dots,n}$, is the $n \times n$ matrix with (i, j) -th element $\mathcal{K}(x_i, x_j)$.

2.4.3 Inner product

Consider a real positive semidefinite kernel $\mathcal{K}(x, x')$ with an eigenfunction expansion $\mathcal{K}(x, x') = \sum_{i=1}^N \lambda_i \phi_i(x) \phi_i(x')$ relative to a measure μ . Mercer's theorem ensures that the eigenfunctions are orthonormal with respect to μ , that is, we have $\int \phi_i(x) \phi_j(x) d\mu(x) = \delta_{ij}$. Consider a Hilbert space of linear combinations of the eigenfunctions, that is, $\theta(x) = \sum_{i=1}^N \theta_i \phi_i(x)$ with $\sum_{i=1}^N \frac{\theta_i^2}{\lambda_i} < \infty$. Then the inner product $\langle \theta_1, \theta_2 \rangle_{\mathcal{H}}$ between $\theta_1 = \sum_{i=1}^N \theta_{1i} \phi_i(x)$, and $\theta_2 = \sum_{i=1}^N \theta_{2i} \phi_i(x)$ is of the form

$$\langle \theta_1, \theta_2 \rangle_{\mathcal{H}} = \sum_{i=1}^N \frac{\theta_{1i} \theta_{2i}}{\lambda_i}. \quad (2.4.7)$$

This induces the norm $\|\cdot\|_{\mathcal{H}}$, where $\|\theta\|_{\mathcal{H}}^2 = \sum_{i=1}^N \frac{\theta_i^2}{\lambda_i}$. A smoothness condition on the space is immediately imposed by requiring the norm to be finite – the eigenvalues must decay sufficiently fast.

The Hilbert space defined above is a unique RKHS with respect to \mathcal{K} , in that it satisfies the following reproducing property:

$$\langle \theta, \mathcal{K}(\cdot, x) \rangle = \sum_{i=1}^N \frac{\theta_i \lambda_i \phi_i(x)}{\lambda_i} = \theta(x). \quad (2.4.8)$$

Further, the kernel satisfies the following:

$$\langle \mathcal{K}(x, \cdot), \mathcal{K}(x', \cdot) \rangle = \sum_{i=1}^N \frac{\lambda_i^2 \phi_i(x) \phi_i(x')}{\lambda_i} = \mathcal{K}(x, x'). \quad (2.4.9)$$

Now, with reference to (2.4.6), observe that the square norm $\|\theta\|_{\mathcal{H}}^2 = \sum_{i=1}^N \theta_i^2 / \lambda_i$ and the quadratic form $\boldsymbol{\theta}^T \mathbf{K} \boldsymbol{\theta}$ have the same form if the latter is expressed in terms of the

eigenvectors of \mathbf{K} , albeit the latter has n terms, while the square norm has N terms.

2.4.4 Regularization

The ill-posed-ness of inverse problems can be understood from the fact that for any given data set \mathbf{y}_n , all functions that pass through the data set minimize any given measure of discrepancy $\mathbb{D}(\mathbf{y}_n, \boldsymbol{\theta})$ between the data \mathbf{y}_n and $\boldsymbol{\theta}$. To combat this, one considers minimization of the following regularized functional:

$$R(\boldsymbol{\theta}) = \mathbb{D}(\mathbf{y}_n, \boldsymbol{\theta}) + \frac{\tau}{2} \|\boldsymbol{\theta}\|_{\mathcal{H}}^2, \quad (2.4.10)$$

where the second term, which is the regularizer, controls smoothness of the function and τ is the appropriate Lagrange multiplier.

The well-known representer theorem (see, for example, [Kimeldorf and Wahba \(1971\)](#), [O'Sullivan et al. \(1986\)](#), [Wahba \(1990\)](#), [Schölkopf and Smola \(2002\)](#)) guarantees that each minimizer $\boldsymbol{\theta} \in \mathcal{H}$ can be represented as $\boldsymbol{\theta}(x) = \sum_{i=1}^n c_i \mathcal{K}(x, x_i)$, where \mathcal{K} is the corresponding reproducing kernel. If $\mathbb{D}(\mathbf{y}_n, \boldsymbol{\theta})$ is convex, then there is a unique minimizer $\hat{\boldsymbol{\theta}}$.

2.4.5 Gaussian process modeling of the unknown function $\boldsymbol{\theta}$

For simplicity, let us consider the model

$$y_i = \boldsymbol{\theta}(x_i) + \epsilon_i, \quad (2.4.11)$$

for $i = 1, \dots, n$, where $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$, where we assume σ to be known for simplicity of illustration. Let $\boldsymbol{\theta}(x)$ be modeled by a Gaussian process with mean function $\mu(x)$ and covariance kernel \mathcal{K} associated with the RKHS. In other words, for any $x \in \mathcal{X}$, $E[\boldsymbol{\theta}(x)] = \mu(x)$ and for any $x_1, x_2 \in \mathcal{X}$, $Cov(\boldsymbol{\theta}(x_1), \boldsymbol{\theta}(x_2)) = \mathcal{K}(x_1, x_2)$.

Assuming for convenience that $\mu(x) = 0$ for all $x \in \mathcal{X}$, it follows that the posterior

distribution of $\theta(x^*)$ for any $x^* \in \mathcal{X}$ is given by

$$\pi(\theta(x^*)|\mathbf{y}_n, \mathbf{x}_n) \equiv N(\hat{\mu}(x^*), \hat{\sigma}^2(x^*)), \quad (2.4.12)$$

where, for any $x^* \in \mathcal{X}$,

$$\hat{\mu}(x^*) = \mathbf{s}^T(x^*) (\mathbf{K} + \sigma^2 \mathbb{I}_n)^{-1} \mathbf{y}_n; \quad (2.4.13)$$

$$\hat{\sigma}^2(x^*) = \mathcal{K}(x^*, x^*) - \mathbf{s}^T(x^*) (\mathbf{K} + \sigma^2 \mathbb{I}_n)^{-1} \mathbf{s}(x^*), \quad (2.4.14)$$

with $\mathbf{s}(x^*) = (\mathcal{K}(x^*, x_1), \dots, \mathcal{K}(x^*, x_n))^T$.

Observe that the posterior mean admits the following representation:

$$\hat{\mu}(x^*) = \sum_{i=1}^n \tilde{c}_i \mathcal{K}(x^*, x_i), \quad (2.4.15)$$

where \tilde{c}_i is the i -th element of $(\mathbf{K} + \sigma^2 \mathbb{I}_n)^{-1} \mathbf{y}_n$.

In other words, the posterior mean of the Gaussian process based model is consistent with the representer theorem.

2.5 Regularization using differential operators and connection with Gaussian process

For $x = (x_1, \dots, x_d)^T \in \mathbb{R}^d$, let

$$\|\mathcal{L}^m \theta\|^2 = \int \sum_{j_1+\dots+j_d=m} \left(\frac{\partial^m \theta(x)}{\partial x_1^{j_1} \cdots \partial x_d^{j_d}} \right)^2, \quad (2.5.1)$$

and

$$\|\mathcal{P}\theta\|^2 = \sum_{m=0}^M b_m \|\mathcal{L}^m \theta\|^2, \quad (2.5.2)$$

for some $M > 0$, where the co-efficients $b_m \geq 0$. In particular, we assume for our purpose that $b_0 > 0$. It is clear that $\|\mathcal{P}\theta\|^2$ is translation and rotation invariant. This norm penalizes θ in terms of its derivatives up to order M .

2.5.1 Relation to RKHS

It can be shown, using the fact that the complex exponentials $\exp(2\pi i s^T x)$ are eigen functions of the differential operator, that

$$\|\mathcal{P}\theta\|^2 = \int \sum_{m=0}^M b_m (4\pi^2 s^T s)^m |\tilde{\theta}(s)|^2 ds, \quad (2.5.3)$$

where $\tilde{\theta}(s)$ is the Fourier transform of $\theta(s)$. Comparison of (2.5.3) with (2.4.7) yields the power spectrum of the form $\left[\sum_{m=0}^M b_m (4\pi^2 s^T s)^m \right]^{-1}$ which yields the following kernel by Fourier inversion:

$$\mathcal{K}(x, x') = \mathcal{K}(x - x') = \int \frac{\exp(2\pi i s^T (x - x'))}{\sum_{m=0}^M b_m (4\pi^2 s^T s)^m} ds. \quad (2.5.4)$$

Calculus of variations can also be used to minimize $R(\theta)$ with respect to θ , which yields (using the Euler-Lagrange equation)

$$\theta(x) = \sum_{i=1}^n b_i \mathcal{G}(x - x_i), \quad (2.5.5)$$

with

$$\sum_{i=1}^m (-1)^m b_m \nabla^m \mathcal{G} = \delta_{x-x'}, \quad (2.5.6)$$

where \mathcal{G} is known as the Green's function. Using Fourier transform on (2.5.6) it can be shown that the Green's function is nothing but the kernel \mathcal{K} given by (2.5.4). Moreover, it follows from (2.5.6) that $\sum_{i=1}^m (-1)^m b_m \nabla^m$ and \mathcal{K} are inverses of each other.

Examples of kernels derived from differential operators are as follows. For $d =$

1, setting $b_0 = b^2$, $b_1 = 1$ and $b_m = 0$ for $m \geq 2$, one obtains $\mathcal{K}(x, x') = \mathcal{K}(x - x') = \frac{1}{2b} \exp(-b|x - x'|)$, which is the covariance of the Ornstein-Uhlenbeck process. For general d dimension, setting $b_m = b^{2m}/(m!2^m)$, yields $\mathcal{K}(x, x') = \mathcal{K}(x - x') = \frac{1}{(2\pi b^2)^{d/2}} \exp[-\frac{1}{2b^2}(x - x')^T(x - x')]$.

Considering a grid \mathbf{x}_n , note that

$$\|\mathcal{P}\theta\|^2 \approx \sum_{m=0}^M b_m (D_m \boldsymbol{\theta})^T (D_m \boldsymbol{\theta}) = \boldsymbol{\theta}^T \left(\sum_{m=0}^M D_m^T D_m \right) \boldsymbol{\theta}, \quad (2.5.7)$$

where D_m is a suitable finite-difference approximation of the differential operator. Note that such finite-difference approximation has been explored in Section 2.3, which we now investigate in a rigorous setting. Also, since (2.5.7) is quadratic in $\boldsymbol{\theta}$, assuming a prior for $\boldsymbol{\theta}$, the logarithm of which has this form, and further assuming that $\log[\mathbb{D}(\mathbf{y}_n, \boldsymbol{\theta})]$ is a log-likelihood quadratic in $\boldsymbol{\theta}$, a Gaussian posterior results.

2.5.2 Spline models and connection with Gaussian process

Let us consider the penalty function to be $\|\mathcal{L}^m \theta\|^2$. Then polynomials up to degree $m - 1$ are not penalized and so, are in the null space of the regularization operator. In this case, it can be shown that a minimizer of $R(\theta)$ is of the form

$$\theta(x) = \sum_{j=1}^k d_j \psi_j(x) + \sum_{i=1}^n c_i G(x, x_i), \quad (2.5.8)$$

where $\{\psi_1, \dots, \psi_k\}$ are polynomials that span the null space and the Green's function G is given by (see [Duchon \(1977\)](#), [Meinguet \(1979\)](#))

$$G(x, x') = G(x - x') = \begin{cases} c_{m,d} |x - x'|^{2m-d} \log|x - x'| & \text{if } 2m > d \text{ and } d \text{ even} \\ c_{m,d} |x - x'|^{2m-d} & \text{otherwise.} \end{cases}, \quad (2.5.9)$$

where $c_{m,D}$ are constants (see [Wahba \(1990\)](#) for the explicit form).

We now specialize the above arguments to the spline set-up. As before, let us consider the model $y_i = \theta(x_i) + \epsilon_i$, where, for $i = 1, \dots, n$, $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$. For simplicity, we consider the one-dimensional set-up, and consider the cubic spline smoothing problem that minimizes

$$R(\theta) = \sum_{i=1}^n (y_i - \theta(x_i))^2 + \tau \int_0^1 [\theta''(x)]^2 dx, \quad (2.5.10)$$

where $0 < x_1 < \dots < x_n < 1$. The solution to this minimization problem is given by

$$\theta(x) = \sum_{j=0}^1 d_j x^j + \sum_{i=1}^n c_i (x - x_i)_+^3, \quad (2.5.11)$$

where, for any x , $(x)_+ = x$ if $x > 0$ and zero otherwise.

Following Wahba (1978), let us consider

$$f(x) = \sum_{j=0}^1 \beta_j x^j + \theta(x), \quad (2.5.12)$$

where $\beta = (\beta_0, \beta_1)^T \sim N(\mathbf{0}, \sigma_\beta^2 \mathbb{I}_2)$, and θ is a zero mean Gaussian process with covariance

$$\sigma_\theta^2 \mathcal{K}(x, x') = \int_0^1 (x-u)_+ (x'-u)_+ du = \sigma_\theta^2 \left(\frac{|x-x'| v^2}{2} + \frac{v^3}{3} \right), \quad (2.5.13)$$

where $v = \min\{x, x'\}$.

Taking $\sigma_\beta^2 \rightarrow \infty$ makes the prior of β vague, so that penalty on the polynomial terms in the null space is effectively washed out. It follows that

$$E[\theta(x^*) | \mathbf{y}_n, \mathbf{x}_n] = \mathbf{h}(x^*)^T \hat{\beta} + \mathbf{s}(x^*)^T \hat{\mathbf{K}}^{-1} (\mathbf{y}_n - \mathbf{H}^T \hat{\beta}), \quad (2.5.14)$$

where, for any x , $\mathbf{h}(x) = (1, x)^T$, $\mathbf{H} = (\mathbf{h}(x_1), \dots, \mathbf{h}(x_n))$, $\hat{\mathbf{K}}$ is the covariance matrix corresponding to $\sigma_\theta^2 \mathcal{K}(x_i, x_j) + \sigma^2 \delta_{ij}$, and $\hat{\beta} = (\mathbf{H} \hat{\mathbf{K}}^{-1} \mathbf{H})^{-1} \mathbf{H} \hat{\mathbf{K}}^{-1} \mathbf{y}_n$.

Since the elements of $\mathbf{s}(x^*)$ are piecewise cubic polynomials, it is easy to see that the posterior mean (2.5.14) is also a piecewise cubic polynomial. It is also clear that (2.5.14) is a first order polynomial on $[0, x_1]$ and $[x_n, 1]$.

Connection with the ℓ -fold integrated Wiener process

Shepp (1966) considered the ℓ -fold integrated Wiener process, for $\ell = 0, 1, 2 \dots$, as follows:

$$W_\ell(x) = \int_0^1 \frac{(x-u)_+^\ell}{\ell!} Z(u) du, \quad (2.5.15)$$

where Z is a Gaussian white noise process with covariance $\delta(u - u')$. As a special case, note that W_0 is the standard Wiener process. In our case, note that

$$\mathcal{K}(x, x') = Cov(W_1(x), W_1(x')). \quad (2.5.16)$$

The above ideas can be easily extended to the case of the regularizer $\int [f^{(m)}(x)]^2 dx$, for $m \geq 1$ by replacing $(x-u)_+$ with $(x-u)_+^{m-1}/(m-1)!$ and letting $\mathbf{h}(x) = (1, x, \dots, x^{m-1})^T$.

2.6 The Bayesian approach to inverse problems in Hilbert spaces

We assume the following model

$$y = G(\theta) + \epsilon, \quad (2.6.1)$$

where y , θ and ϵ are in Banach or Hilbert spaces.

2.6.1 Bayes theorem for general inverse problems

We will consider the model stated by equation (2.6.1). Let \mathcal{Y} and Θ denote the sample spaces for y and θ , respectively. Let us first assume that both are separable Banach

spaces. Assume μ_0 to be the prior measure for θ . Assuming well-defined joint distribution for (y, θ) , let us denote the posterior of θ given y as μ_y . Let $\epsilon \sim Q_0$ where Q_0 is such that ϵ and θ are independent. Let us denote the conditional distribution of y given θ by Q_θ , obtained from a translation of Q_0 by $G(\theta)$. Assume that $Q_\theta \ll Q_0$. Thus, for some potential $\Phi : \Theta \times \mathcal{Y} \mapsto \mathbb{R}$,

$$\frac{dQ_\theta}{dQ_0} = \exp(-\Phi(\theta, y)). \quad (2.6.2)$$

Thus, for fixed θ , $\Phi(\theta, \cdot) : \mathcal{Y} \mapsto \mathbb{R}$ is measurable and $E_{Q_0}[\exp(-\Phi(\theta, y))] = 1$. Note that $-\Phi(\cdot, y)$ is nothing but the log-likelihood.

Let ν_0 denote the product measure

$$\nu_0(d\theta, dy) = \mu_0(d\theta)Q_0(dy), \quad (2.6.3)$$

and let us assume that Φ is ν_0 -measurable. Then $(\theta, y) \in \Theta \times \mathcal{Y}$ is distributed according to the measure $\nu(d\theta, dy) = \mu_0(d\theta)Q_\theta(dy)$. It then also follows that $\nu \ll \nu_0$, with

$$\frac{d\nu}{d\nu_0}(\theta, y) = \exp(-\Phi(\theta, y)). \quad (2.6.4)$$

Then we have the following statement of Bayes' theorem for general inverse problems:

Theorem 7 (Bayes theorem for general inverse problems) *Assume that $\Phi : \Theta \times \mathcal{Y} \mapsto \mathbb{R}$ is ν_0 -measurable and*

$$C = \int_{\Theta} \exp(-\Phi(\theta, y)) \mu_0(dy) > 0, \quad (2.6.5)$$

for Q_0 -almost surely all y . Then the posterior of θ given y , which we denote by μ^y , exists under ν . Also, $\mu^y \ll \mu_0$ and for all y ν_0 -almost surely,

$$\frac{d\mu^y}{d\mu_0}(\theta) = \frac{1}{C} \exp(-\Phi(\theta, y)). \quad (2.6.6)$$

Now assume that Θ and \mathcal{Y} are Hilbert spaces. Suppose $\epsilon \sim \mathbf{N}(0, \Gamma)$. Then the following theorem holds:

Theorem 8 (Vollmer (2013))

$$\frac{d\mu^y}{d\mu_0} \propto \exp\left(-\frac{1}{2}\|G(\theta)\|_\Gamma^2 + \langle y, G(\theta) \rangle_\Gamma\right), \quad (2.6.7)$$

where $\langle \cdot, \cdot \rangle_\Gamma = \langle \Gamma^{-1} \cdot, \cdot \rangle$, and $\|\cdot\|_\Gamma$ is the norm induced by $\langle \cdot, \cdot \rangle_\Gamma$.

For the model $y_i = \theta(x_i) + \epsilon_i$ for $i = 1, \dots, n$, with $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$, the posterior is of the form

$$\frac{d\mu^y}{d\mu_0} \propto \exp\left(-\sum_{i=1}^n \frac{(y_i - \theta(x_i))^2}{2\sigma^2}\right). \quad (2.6.8)$$

2.6.2 Connection with regularization methods

It is not immediately clear if the Bayesian approach in the Hilbert space setting has connection with the deterministic regularization methods, but Vollmer (2013) prove consistency of the posterior assuming certain stability results which are used to prove convergence of regularization methods; see Engl *et al.* (1996).

2.7 Conclusion

In this chapter, we have clarified the similarities and dissimilarities between the traditional inverse problems and the inverse regression problems. In particular, we have argued that only the latter class of problems qualify as authentic inverse problems in they have significantly different goals compared to the corresponding forward problems. Moreover, they include the traditional inverse problems on learning unknown functions as a special case, as exemplified by our palaeoclimate and Milky Way examples. We advocate the Bayesian paradigm for both classes of problems, not only because of its inherent flexibility, coherency and posterior uncertainty quantification, but also because the prior

acts as a natural penalty which is very important to regularize the so-called ill-posed inverse problems. The well-known Tikhonov regularizer is just a special case from this perspective.

3

Thesis Layout and Overview of Our Contributions

3.1 Thesis layout

Before touching upon our contributions briefly, let us first provide a layout of the thesis, which is also intended to provide a glimpse of the topics covered and the overall contributions.

We proceed towards establishing Bayesian covariate consistency in the inverse regression context by first establishing posterior asymptotic theory of unknown functions modeled by appropriate stochastic processes, including Gaussian processes, in the contexts of normal and double-exponential nonparametric regression (Chapter 4), followed by binary and Poisson nonparametric regression (Chapter 5). Relevant development of posterior convergence theories for unknown functions is necessary since, as already

pointed out in Chapter 1, the underlying inverse regression problem often involves unknown functions modeled nonparametrically by appropriate stochastic processes.

Judiciously making use of the results of Chapters 4 and 5, we establish Bayesian covariate consistency in the leave-one-out cross-validation setup in Chapter 6, introducing a specialized, data and (unknown) function dependent prior for the covariates. Bayesian consistency of the IRD approach is established in the same chapter, utilizing the results of Bayesian consistency of the inverse cross-validation posterior distributions.

In Chapter 7 we establish the asymptotic theory of Bayes factor for model and variable selection in the general setup, while the asymptotic theories of pseudo-Bayes factor for model and variable selection in the forward and inverse regression contexts are established in Chapter 8, showing that the final results are in agreement with those for Bayes factor.

We seek further improvement of Bayesian inverse model and variable selection through a novel Bayesian multiple testing procedure for such purpose, which we introduce and develop in Chapter 9.

Our efforts towards development of theoretical and methodological aspects of inverse regression problems culminate in an attempt to settle a very important applied problem. Indeed, in Chapter 10 we attempt to address the interesting question on global climate change, namely, if the future world will indeed experience the alarming global warming phenomenon as projected by the influential global climate models that has caused great concern among the scientists and policymakers all over the world. We show that the question can be posed as a Bayesian inverse regression problem, and our model and methodologies yield results that do not support future global warming.

Finally, we summarize the thesis and provide future directions in Chapter 11.

3.2 An overview of our contributions

In Chapter 4, we investigate posterior convergence in nonparametric regression models where the unknown regression function is modeled by some appropriate stochastic process.

In this regard, we consider two setups. The first setup is based on Gaussian processes, where the covariates are either random or non-random and the noise may be either normally or double-exponentially distributed. In the second setup, we assume that the underlying regression function is modeled by some reasonably smooth, but unspecified stochastic process satisfying reasonable conditions. The distribution of the noise is also left unspecified, but assumed to be thick-tailed. As in the previous studies regarding the same problems, we do not assume that the truth lies in the postulated parameter space, thus explicitly allowing the possibilities of misspecification. We exploit the general results of [Shalizi \(2009\)](#) for our purpose and establish not only posterior consistency, but also the rates at which the posterior probabilities converge, which turns out to be the Kullback-Leibler divergence rate. We also investigate the more familiar posterior convergence rates. Interestingly, we show that the posterior predictive distribution can accurately approximate the best possible predictive distribution in the sense that the Hellinger distance, as well as the total variation distance between the two distributions can tend to zero, in spite of misspecifications.

Then in Chapter 5, we investigate posterior convergence of nonparametric binary and Poisson regression under possible model misspecification, assuming general stochastic process prior with appropriate properties. Our model setup and objective for binary regression is similar to that of [Ghosal and Roy \(2006\)](#) where the authors have used the approach of entropy bound and exponentially consistent tests with the sieve method to achieve consistency with respect to their Gaussian process prior. In contrast, for both binary and Poisson regression, using general stochastic process prior, our approach involves verification of asymptotic equipartition property along with the method of sieve, which is a manoeuvre of the general results of [Shalizi \(2009\)](#), useful even for misspecified models. Moreover, we establish not only posterior consistency but also the rates at which the posterior probabilities converge, which again turn out to be the Kullback-Leibler divergence rate. As in Chapter 4, we also investigate the traditional posterior convergence

rates. As in Chapter 4, here also we show that the posterior predictive distribution can accurately approximate the best possible predictive distribution in the sense that the Hellinger distance, as well as the total variation distance between the two distributions, can tend to zero, in spite of misspecifications.

In Chapter 6 we consider Bayesian inference in inverse regression problems where the objective is to infer about unobserved covariates from observed responses and covariates. We establish posterior consistency of such unobserved covariates in Bayesian inverse regression problems under appropriate priors in a leave-one-out cross-validation setup. We relate this to posterior consistency of the IRD approach of [Bhattacharya \(2013\)](#) for assessing model adequacy. We illustrate our theory and methods with various examples of Bayesian inverse regression, along with adequate simulation experiments.

Although there is a significant literature on the asymptotic theory of Bayes factor, the set-ups considered are usually specialized and often involves independent and identically distributed data. Even in such specialized cases, mostly weak consistency results are available. In Chapter 7, for the first time ever, we derive the almost sure convergence theory of Bayes factor in the general set-up that includes even dependent data and misspecified models. Somewhat surprisingly, the key to the proof of such a general theory is a simple application of a result of [Shalizi \(2009\)](#) to a well-known identity satisfied by the Bayes factor.

In the Bayesian literature on model comparison, Bayes factors play the leading role. In the classical statistical literature, model selection criteria are often devised used cross-validation ideas. Amalgamating the ideas of Bayes factor and cross-validation [Geisser and Eddy \(1979\)](#) created the pseudo-Bayes factor. The usage of cross-validation inculcates several theoretical advantages, computational simplicity and numerical stability in Bayes factors as the marginal density of the entire dataset is replaced with products of cross-validation densities of individual data points. However, the popularity of pseudo-Bayes factors is still negligible in comparison with Bayes factors, with respect to both

theoretical investigations and practical applications. In Chapter 8, we establish almost sure exponential convergence of pseudo-Bayes factors for large samples under a general setup consisting of dependent data and model misspecifications. We particularly focus on general parametric and nonparametric regression setups in both forward and inverse contexts. Depending upon forward and inverse regression ideas, our asymptotic theory manifests itself in terms of almost sure exponential convergence of the pseudo-Bayes factor in terms of the Kullback-Leibler divergence rate or its integrated version, between the competing and the true models. Our asymptotic theory encompasses general model selection, variable selection and combinations of both. We illustrate our theoretical results with various examples, providing explicit calculations. We also supplement our asymptotic theory with simulation experiments in small sample situations of Poisson log regression and geometric logit and probit regression, additionally addressing the variable selection problem. We consider both linear and nonparametric regression modeled by Gaussian processes for our purposes. Our simulation results provide quite interesting insights into the usage of pseudo-Bayes factors in forward and inverse setups.

In Chapter 9, we propose a novel Bayesian multiple testing formulation for model and variable selection in inverse setups, judiciously embedding the idea of IRD in a mixture framework consisting of the competing models. We develop the theory and methods in the general context encompassing parametric and nonparametric competing models, dependent data, as well as misspecifications. Our investigation shows that asymptotically the multiple testing procedure almost surely selects the best possible inverse model that minimizes the minimum Kullback-Leibler divergence from the true model. We also show that the error rates, namely, versions of the false discovery rate and the false non-discovery rate converge to zero almost surely as the sample size goes to infinity. Asymptotic α -control of versions of the false discovery rate and its impact on the convergence of false non-discovery rate versions, are also investigated. With an aim to compare our multiple testing procedure with pseudo-Bayes factor, we consider

the same simulation experiments with the same datasets reported in Chapter 8. The experiments involve small sample based selection among inverse Poisson log regression and inverse geometric logit and probit regression, where the regressions are either linear or based on Gaussian processes. Additionally, variable selection is also considered. Our multiple testing results turn out to be very encouraging in the sense of selecting the best models in all the cases and convincingly outperforming the pseudo-Bayes factors.

Global warming, the phenomenon of increasing global average temperature in the recent decades, is receiving wide attention due to its very significant adverse effects on climate. Whether global warming will continue even in the future, is a question that is most important to investigate. In this regard, the so-called general circulation models (GCMs) have attempted to project the future climate, and nearly all of them exhibit alarming rates of global temperature rise in the future. Although global warming in the current time frame is undeniable, it is important to assess the validity of the future predictions of the GCMs. In Chapter 10, we attempt such a study using our recently-developed Bayesian multiple testing paradigm for model selection in inverse regression problems. An important premise in the context of our GCM forecast assessment hinges on the question that how probable the current global warming phenomenon is if the future predictions by the GCMs are correct. This is the inverse regression aspect since the future depends upon the current, but here we wish to learn about the present pretending it to be unknown, assuming that the future is known. Our multiple testing framework coherently compares the combination of inverse aspect with the forward, to yield the best model. The model we assume for the global temperature time series is based on Gaussian process emulation of the black box scenario, realistically treating the dynamic evolution of the time series as unknown. We apply our ideas to datasets available from the Intergovernmental Panel on Climate Change (IPCC) website. The best GCM models selected by our method under different assumptions on future climate change scenarios do not convincingly support the present global warming pattern when only the future

predictions are considered known. We also consider all the GCM models under given assumptions on any particular future climate change scenario as an ensemble, which we model as multivariate time series, based on multidimensional Gaussian processes. Our results in the multivariate cases emphatically demonstrate that if the future GCM predictions are believed to be true, then the current global warming phenomenon must be highly unlikely. In other words, the GCM predictions available at the IPCC website do not seem to adequately represent the future climate change. What is more, using our Gaussian process idea, we forecast the future temperature time series given the current one. Interestingly, our results do not support drastic future global warming predicted by almost all the GCM models. Indeed, we show that except the predictions of the best GCM model in the “Commitment” scenario, none other fall in the high density regions of our Bayesian forecasted time series.

4

Posterior Convergence of Gaussian and General Stochastic Process Regression Under Possible Misspecifications

4.1 Introduction

In statistics, either frequentist or Bayesian, nonparametric regression plays a very significant role. The frequentist nonparametric literature, however, is substantially larger than the Bayesian counterpart. Here we cite the books Schimek (2013), Härdle *et al.* (2012), Efromovich (2008), Takezawa (2006), Wu and Zhang (2006), Eubank (1999), Green and Silverman (1993) and Härdle (1990), among a large number of books on frequentist nonparametric regression. The Bayesian nonparametric literature, which is relatively young but flourishing in the recent times (see, for example, Ghosal and van

derVaart (2017), Müller *et al.* (2015), Dey *et al.* (2012), Hjort *et al.* (2010), Ghosh and Ramamoorthi (2003)), offers much broader scope for interesting and innovative research.

The importance of Gaussian processes in nonparametric statistical modeling, particularly in the Bayesian context, is undeniable. It is widely used in density estimation (Lenk (1988), Lenk (1991), Lenk (2003)), nonparametric regression (Rasmussen and Williams (2006)), spatial data modeling (Cressie (1993), Banerjee *et al.* (2014)), machine learning (Rasmussen and Williams (2006)), emulation of computer models (Santner *et al.* (2003)), to name a few areas. Although applications of Gaussian processes have received and continue to receive much attention, in the recent years there seems to be a growing interest among researchers in the theoretical properties of approaches based on Gaussian processes. Specifically, investigation of posterior convergence of Gaussian process based approaches has turned out to be an important undertaking. In this respect, contributions are made by Choi and Schervish (2007), van der Vaart and van Zanten (2008), van der Vaart and van Zanten (2009), van der Vaart and van Zanten (2011), Knapik *et al.* (2011), Vollmer (2013), Yang *et al.* (2018), Knapik and Salomond (2018). Choi and Schervish (2007) address posterior consistency in Gaussian process regression, while the others also attempt to provide the rates of posterior convergence. However, the rates are so far computed under the assumption that the error distribution is normal and the error variance is either known, or if unknown, can be given a prior, but on a compact support bounded away from zero.

General priors for the regression function or thick-tailed noise distributions seemed to have received less attention. The asymptotic theory for such frameworks is even rare, Choi (2009) being an important exception. As much as we are aware of, rates of convergence are not available for nonparametric regression with general stochastic process prior on the regression function and thick-tailed noise distributions. Another important issue which seems to have received less attention in the literature, is the case of misspecified models. We are not aware of any published asymptotic theory pertaining

to misspecifications in nonparametric regression, for either Gaussian or non-Gaussian processes with either normal or non-normal errors.

In this chapter, we consider both Gaussian and general stochastic process regression under the same setups as Choi and Schervish (2007) and Choi (2009), respectively, assuming that the covariates may be either random or non-random. For the Gaussian process setup we consider both normal and double-exponential distribution for the error, with unknown error variance. In the general context, we assume non-Gaussian noise with unknown scale parameter supported on the entire positive part of the real line. Based on the general theory of posterior convergence provided in Shalizi (2009), we establish posterior convergence theories for both the setups. We allow the case of misspecified models, that is, if the true regression function and the true error variance are not even supported by the prior. Our approach also enables us to show that the relevant posterior probabilities converge at the Kullback-Leibler (KL) divergence rate, and that the posterior convergence rate with respect to the KL-divergence is just slower than n^{-1} , n being the number of observations. We further show that even in the case of misspecification, the posterior predictive distribution can approximate the best possible predictive distribution adequately, in the sense that the Hellinger distance, as well as the total variation distance between the two distributions can tend to zero. In Section 4.1.1 we provide a brief overview and intuitive explanation of the main assumptions and results of Shalizi, which we exploit in this chapter. The details are provided in Section 4.A1. The results of Shalizi are based on seven assumptions, which we refer to as (S1) – (S7) throughout this thesis.

4.1.1 A briefing of the main results of Shalizi

Let $\mathbf{Y}_n = (Y_1, \dots, Y_n)^T$, and let $f_\theta(\mathbf{Y}_n)$ and $f_{\theta_0}(\mathbf{Y}_n)$ denote the observed and the true likelihoods respectively, under the given value of the parameter θ and the true parameter θ_0 . We assume that $\theta \in \Theta$, where Θ is the (often infinite-dimensional)

parameter space. However, we *do not* assume that $\theta_0 \in \Theta$, thus allowing misspecification. The key ingredient associated with Shalizi's approach to proving convergence of the posterior distribution of θ is to show that the asymptotic equipartition property holds. To elucidate, let us consider the following likelihood ratio:

$$R_n(\theta) = \frac{f_\theta(\mathbf{Y}_n)}{f_{\theta_0}(\mathbf{Y}_n)}.$$

Then, to say that for each $\theta \in \Theta$, the generalized or relative asymptotic equipartition property holds, we mean

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log R_n(\theta) = -h(\theta), \quad (4.1.1)$$

almost surely, where $h(\theta)$ is the KL-divergence rate given by

$$h(\theta) = \lim_{n \rightarrow \infty} \frac{1}{n} E_{\theta_0} \left(\log \frac{f_{\theta_0}(\mathbf{Y}_n)}{f_\theta(\mathbf{Y}_n)} \right),$$

provided that it exists (possibly being infinite), where E_{θ_0} denotes expectation with respect to the true model. Let

$$h(A) = \text{ess inf}_{\theta \in A} h(\theta);$$

$$J(\theta) = h(\theta) - h(\Theta);$$

$$J(A) = \text{ess inf}_{\theta \in A} J(\theta).$$

Thus, $h(A)$ can be roughly interpreted as the minimum KL-divergence between the postulated and the true model over the set A . If $h(\Theta) > 0$, this indicates model misspecification. However, as we shall show, model misspecification need not always imply that $h(\Theta) > 0$. For $A \subset \Theta$, $h(A) > h(\Theta)$, so that $J(A) > 0$.

As regards the prior, it is required to construct an appropriate sequence of sieves \mathcal{G}_n such that $\mathcal{G}_n \rightarrow \Theta$ and $\pi(\mathcal{G}_n^c) \leq \alpha \exp(-\beta n)$, for some $\alpha > 0$.

With the above notions, verification of (4.1.1) along with several other technical conditions ensure that for any $A \subseteq \Theta$ such that $\pi(A) > 0$,

$$\lim_{n \rightarrow \infty} \pi(A|\mathbf{Y}_n) = 0, \quad (4.1.2)$$

almost surely, provided that $h(A) > h(\Theta)$. Under mild assumptions, it also holds that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \pi(A|\mathbf{Y}_n) = -J(A), \quad (4.1.3)$$

almost surely, where $\pi(\cdot|\mathbf{Y}_n)$ denotes the posterior distribution of θ given \mathbf{Y}_n . With respect to (4.1.2) note that $h(A) > h(\Theta)$ implies positive KL-divergence in A , even if $h(\Theta) = 0$. In other words, A is the set in which the postulated model fails to capture the true model in terms of the KL-divergence. Hence, expectedly, the posterior probability of that set converges to zero. The result (4.1.3) asserts that the rate at which the posterior probability of A converges to zero is about $\exp(-nJ(A))$. From the above results it is clear that the posterior concentrates on sets of the form $N_\epsilon = \{\theta : h(\theta) \leq h(\Theta) + \epsilon\}$, for any $\epsilon > 0$.

As regards the rate of posterior convergence, let $N_{\epsilon_n} = \{\theta : h(\theta) \leq h(\Theta) + \epsilon_n\}$, where $\epsilon_n \rightarrow 0$ such that $n\epsilon_n \rightarrow \infty$. Then under an additional technical assumption it holds, almost surely, that

$$\lim_{n \rightarrow \infty} \pi(N_{\epsilon_n}|\mathbf{Y}_n) = 1. \quad (4.1.4)$$

Moreover, it was shown by Shalizi that the squares of the Hellinger and the total variation distances between the posterior predictive distribution and the best possible predictive distribution under the truth, are asymptotically almost surely bounded above by $h(\Theta)$ and $4h(\Theta)$, respectively. In other words, if $h(\Theta) = 0$, then this entails very accurate approximation of the true predictive distribution by the posterior predictive distribution.

The rest of this chapter is structured as follows. We treat the Gaussian process

regression with normal and double exponential errors in Section 4.2. Specifically, our assumptions regarding the model and discussion of the assumptions are presented in Section 4.2.1. In Section 4.2.2 we present our main results of posterior convergence, along with the summary of the verification of Shalizi's assumptions, for the Gaussian process setup. The complete details are provided in Sections 4.A2 and 4.A3. We deal with rate of convergence and model misspecification issue for Gaussian process regression in Sections 4.2.3 and 4.2.4, respectively.

The case of general stochastic process regression with thick tailed error distribution is taken up in Section 4.3. The assumptions with their discussion are provided in Section 4.3.1, the main posterior results are presented in Section 4.3.2, and Section 4.3.3 addresses the rate of convergence and model misspecification issue. Finally, we make concluding remarks in Section 4.4. The relevant details are provided in Section 4.A4.

4.2 The Gaussian process regression setup

As in Choi and Schervish (2007), we consider the following model:

$$y_i = \eta(x_i) + \epsilon_i; \quad i = 1, \dots, n; \quad (4.2.1)$$

$$\eta(\cdot) \sim GP(\mu(\cdot), c(\cdot, \cdot)); \quad (4.2.2)$$

$$\epsilon_i \stackrel{iid}{\sim} f(\cdot | \sigma^2); \quad (4.2.3)$$

$$\sigma \sim \pi_\sigma(\cdot). \quad (4.2.4)$$

In (4.2.2), $GP(\mu(\cdot), c(\cdot, \cdot))$ stands for Gaussian process with mean function $\mu(\cdot)$ and positive definite covariance function $cov(\eta(x_1), \eta(x_2)) = c(x_1, x_2)$, for any $x_1, x_2 \in \mathcal{X}$, where \mathcal{X} is the domain of η . In (4.2.3), $f(\cdot | \sigma^2)$ is some appropriate density with variance parameter σ^2 .

As in Choi and Schervish (2007) we assume two separate distributions for the errors ϵ_i , independent zero-mean normal with variance σ^2 which we denote by $N(0, \sigma^2)$ and

independent double exponential distribution with median 0 and scale parameter σ with density

$$f(\epsilon) = \frac{1}{2\sigma} \exp\left(-\frac{|\epsilon|}{\sigma}\right); \quad \epsilon \in \mathbb{R}.$$

We denote the double exponential distribution by $DE(0, \sigma)$.

In our case, let $\theta = (\eta, \sigma)$ be the infinite-dimensional parameter associated with our Gaussian process model and let $\theta_0 = (\eta_0, \sigma_0)$ be the true (infinite-dimensional) parameter. Let Θ denote the infinite-dimensional parameter space.

4.2.1 Assumptions and their discussions

Regarding the model and the prior, we make the following assumptions:

- (A1) \mathcal{X} is a compact, d -dimensional space, for some finite $d \geq 1$, equipped with a suitable metric.
- (A2) The functions η are continuous on \mathcal{X} and for such functions the limit

$$\eta'_j(x) = \frac{\partial \eta(x)}{\partial x_j} = \lim_{h \rightarrow 0} \frac{\eta(x + h\delta_j) - \eta(x)}{h} \quad (4.2.5)$$

exists for each $x \in \mathcal{X}$, and is continuous on \mathcal{X} , for $j = 1, \dots, d$. In the above, δ_j is the d -dimensional vector where the j -th element is 1 and all the other elements are zero. We denote the above class of functions by $\mathcal{C}'(\mathcal{X})$.

- (A3) We assume the following for the covariates x_i , accordingly as they are considered an observed random sample, or non-random.
 - (i) $\{x_i : i = 1, 2, \dots\}$ is an observed sample associated with an *iid* sequence associated with some probability measure Q , supported on \mathcal{X} , which is independent of $\{\epsilon_i : i = 1, 2, \dots\}$.
 - (ii) $\{x_i : i = 1, 2, \dots\}$ is an observed non-random sample. In this case, we consider a specific partition of the d -dimensional space \mathcal{X} into n subsets such that

each subset of the partition contains at least one $x \in \{x_i : i = 1, 2, \dots\}$ and has Lebesgue measure L/n , for some $L > 0$.

(A4) Regarding the prior for σ , we assume that for large enough n ,

$$\pi_\sigma \left(\exp(-(\beta n)^{1/4}) \leq \sigma \leq \exp((\beta n)^{1/4}) \right) \geq 1 - c_\sigma \exp(-\beta n),$$

for $c_\sigma > 0$ and $\beta > 2h(\Theta)$.

(A5) The true regression function η_0 satisfies $\|\eta_0\| \leq \kappa_0 < \infty$. We *do not* assume that $\eta_0 \in \mathcal{C}'(\mathcal{X})$. For random covariate X , we assume that $\eta_0(X)$ is measurable.

Discussion of the assumptions

The compactness assumption on \mathcal{X} in Assumption (A1) guarantees that continuous functions on \mathcal{X} have finite sup-norms. Here, by sup-norm of any function f on \mathcal{X} , we mean $\|f\| = \sup_{x \in \mathcal{X}} |f(x)|$. Hence, our Gaussian process prior on η , which gives probability one to continuously differentiable functions, also ensures that $\|\eta\| < \infty$, almost surely. Compact support of the functions is commonplace in the Gaussian process literature; see, for example, Cramer and Leadbetter (1967), Adler (1981), Adler and Taylor (2007), Choi and Schervish (2007). The metric on \mathcal{X} is necessary for partitioning \mathcal{X} in the case of non-random covariates.

Condition (A2) is required for constructing appropriate sieves for proving our posterior convergence results. In particular, this is required to ensure that η is Lipschitz continuous in the sieves. Since a function is Lipschitz if and only if its partial derivatives are bounded, this serves our purpose, as continuity of the partial derivatives of η guarantees boundedness in the compact domain \mathcal{X} . Conditions guaranteeing the above continuity and smoothness properties required by (A2) must also be reflected in the underlying Gaussian process prior for η . The relevant conditions can be found in Cramer and Leadbetter (1967), Adler (1981) and Adler and Taylor (2007), which we assume in our

case. In particular, these require adequate smoothness assumptions on the mean function $\mu(\cdot)$ and the covariance function $c(\cdot, \cdot)$ of the Gaussian process prior. It follows that $\eta'_j; j = 1, \dots, d$, are also Gaussian processes. It clearly holds that $\mu(\cdot)$ and its partial derivatives also have finite sup-norms.

As regards (A3) (i), thanks to the strong law of large numbers (SLLN), given any η in the complement of some null set with respect to the prior, and given any sequence $\{x_i : i = 1, 2, \dots\}$ this assumption ensures that for any $\nu > 0$, as $n \rightarrow \infty$,

$$\frac{1}{n} \sum_{i=1}^n |\eta(x_i) - \eta_0(x_i)|^\nu \rightarrow \int_{\mathcal{X}} |\eta(x) - \eta_0(x)|^\nu dQ(X) = E_X |\eta(X) - \eta_0(X)|^\nu \text{ (say)}, \quad (4.2.6)$$

almost surely, where Q is some probability measure supported on \mathcal{X} .

Condition (A3) (ii) ensures that $\frac{1}{n} \sum_{i=1}^n |\eta(x_i) - \eta_0(x_i)|^\nu$ is a particular Riemann sum and hence (4.2.6) holds with Q being the Lebesgue measure on \mathcal{X} . We continue to denote the limit in this case by $E_X [\eta(X) - \eta_0(X)]^\nu$.

In the light of (4.2.6), condition (A3) will play important role in establishing the equipartition property, for both Gaussian and double exponential errors. Another important role of this condition is to ensure consistency of the posterior predictive distribution, in spite of some misspecifications.

Condition (A4) ensures that the prior probabilities of the complements of the sieves are exponentially small. Such a requirement is common to most Bayesian asymptotic theories.

The essence of (A5) is to allow misspecification of the prior for η in a way that the true regression function is not even supported by the prior, even though it has finite sup-norm. In contrast, Choi and Schervish (2007) assumed that η_0 has continuous first-order partial derivatives. The assumption of measurability of $\eta_0(X)$ is a very mild technical condition.

Let $\Theta = \mathcal{C}'(\mathcal{X}) \times \mathbb{R}^+$ denote the infinite-dimensional parameter space for our Gaussian process model.

4.2.2 Posterior convergence of Gaussian process regression under normal and double exponential errors

In this section we provide a summary of our results leading to posterior convergence of Gaussian process regression when the errors are assumed to be either normal or double exponential. The details are provided in the supplement. The key results associated with the asymptotic equipartition property are provided in Lemma 9 and Theorem 10, the proofs of which are provided in the supplement in the context of detailed verification of Shalizi's assumptions.

Lemma 9 *Under the Gaussian process model and conditions (A1) and (A3), the KL-divergence rate $h(\theta)$ exists for $\theta \in \Theta$, and is given by*

$$h(\theta) = \log\left(\frac{\sigma}{\sigma_0}\right) - \frac{1}{2} + \frac{\sigma_0^2}{2\sigma^2} + \frac{1}{2\sigma^2} E_X [\eta(X) - \eta_0(X)]^2, \quad (4.2.7)$$

for the normal errors, and

$$h(\theta) = \log\left(\frac{\sigma}{\sigma_0}\right) - 1 + \frac{1}{\sigma} E_X |\eta(X) - \eta_0(X)| + \frac{\sigma_0}{\sigma} E_X \left[\exp\left(-\frac{|\eta(X) - \eta_0(X)|}{\sigma_0}\right) \right], \quad (4.2.8)$$

for the double exponential errors.

Theorem 10 *Under the Gaussian process model with normal and double exponential errors and conditions (A1) and (A3), the asymptotic equipartition property holds, and is given by*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log R_n(\theta) = -h(\theta), \text{ almost surely.}$$

The convergence is uniform on any compact subset of Θ .

Lemma 9 and Theorem 10 ensure that conditions (S1) – (S3) of Shalizi hold, and (S4)

holds since $h(\theta)$ is almost surely finite. We construct the sieves \mathcal{G}_n as

$$\mathcal{G}_n = \left\{ (\eta, \sigma) : \|\eta\| \leq \exp((\beta n)^{1/4}), \exp(-(\beta n)^{1/4}) \leq \sigma \leq \exp((\beta n)^{1/4}), \|\eta'_j\| \leq \exp((\beta n)^{1/4}); j = 1, \dots, d \right\}. \quad (4.2.9)$$

It follows that $\mathcal{G}_n \rightarrow \Theta$ as $n \rightarrow \infty$ and the properties of the Gaussian processes η , η' , together with (A4) ensure that $\pi(\mathcal{G}_n^c) \leq \alpha \exp(-\beta n)$, for some $\alpha > 0$. This result, continuity of $h(\theta)$, compactness of \mathcal{G}_n and the uniform convergence result of Theorem 10, together ensure (S5).

Now observe that the aim of assumption (S6) is to ensure that (see the proof of Lemma 7 of Shalizi (2009)) for every $\varepsilon > 0$ and for all n sufficiently large,

$$\frac{1}{n} \log \int_{\mathcal{G}_n} R_n(\theta) d\pi(\theta) \leq -h(\mathcal{G}_n) + \varepsilon, \text{ almost surely.}$$

Since $h(\mathcal{G}_n) \rightarrow h(\Theta)$ as $n \rightarrow \infty$, it is enough to verify that for every $\varepsilon > 0$ and for all n sufficiently large,

$$\frac{1}{n} \log \int_{\mathcal{G}_n} R_n(\theta) d\pi(\theta) \leq -h(\Theta) + \varepsilon, \text{ almost surely.} \quad (4.2.10)$$

In this regard, first observe that

$$\begin{aligned} \frac{1}{n} \log \int_{\mathcal{G}_n} R_n(\boldsymbol{\theta}) d\pi(\boldsymbol{\theta}) &\leq \frac{1}{n} \log \left[\sup_{\boldsymbol{\theta} \in \mathcal{G}_n} R_n(\boldsymbol{\theta}) \pi(\mathcal{G}_n) \right] \\ &= \frac{1}{n} \log \left[\sup_{\boldsymbol{\theta} \in \mathcal{G}_n} R_n(\boldsymbol{\theta}) \right] + \frac{1}{n} \log \pi(\mathcal{G}_n) \\ &= \sup_{\boldsymbol{\theta} \in \mathcal{G}_n} \frac{1}{n} \log R_n(\boldsymbol{\theta}) + \frac{1}{n} \log \pi(\mathcal{G}_n) \\ &\leq \frac{1}{n} \sup_{\boldsymbol{\theta} \in \mathcal{G}_n} \log R_n(\boldsymbol{\theta}), \end{aligned} \quad (4.2.11)$$

where the last inequality holds since $\frac{1}{n} \log \pi(\mathcal{G}_n) \leq 0$. Now, letting $\mathcal{S} = \{\boldsymbol{\theta} : h(\boldsymbol{\theta}) \leq \kappa\}$,

where $\kappa > h(\boldsymbol{\Theta})$ is as large as desired,

$$\begin{aligned} \sup_{\boldsymbol{\theta} \in \mathcal{G}_n} \frac{1}{n} \log R_n(\boldsymbol{\theta}) &= \sup_{\boldsymbol{\theta} \in (\mathcal{G}_n \cap \mathcal{S}) \cup (\mathcal{G}_n \cap \mathcal{S}^c)} \frac{1}{n} \log R_n(\boldsymbol{\theta}) \\ &\leq \max \left\{ \sup_{\boldsymbol{\theta} \in \mathcal{G}_n \cap \mathcal{S}} \frac{1}{n} \log R_n(\boldsymbol{\theta}), \sup_{\boldsymbol{\theta} \in \mathcal{S}^c} \frac{1}{n} \log R_n(\boldsymbol{\theta}) \right\}. \end{aligned} \quad (4.2.12)$$

In Sections 4.A2.5 and 4.A3.5 we have proved continuity of $h(\theta)$ for Gaussian and double exponential errors, respectively. Now observe that $\|\eta\| \leq \|\eta - \eta_0\| + \|\eta_0\|$, so that $\|\eta\| \rightarrow \infty$ implies $\|\eta - \eta_0\| \rightarrow \infty$ (since $\|\eta_0\| < \infty$). Hence, for each η , there exists a subset \mathcal{X}_η of \mathcal{X} depending upon η such that $Q(\mathcal{X}_\eta) > 0$ and $\sup_{x \in \mathcal{X}_\eta} |\eta(x) - \eta_0(x)| \rightarrow \infty$ as $\|\eta\| \rightarrow \infty$. It then follows that $E_{\mathbf{X}} |\eta(\mathbf{X}) - \eta_0(\mathbf{X})| \rightarrow \infty$ and $E_{\mathbf{X}} (\eta(\mathbf{X}) - \eta_0(\mathbf{X}))^2 \rightarrow \infty$ as $\|\eta\| \rightarrow \infty$. Hence observe that $\|\theta\| \rightarrow \infty$ if $\sigma \rightarrow \infty$ and $\|\eta\| \rightarrow \infty$, or if σ tends to zero or some non-negative constant and $\|\eta\| \rightarrow \infty$. In both the cases $h(\boldsymbol{\theta}) \rightarrow \infty$, for both Gaussian and double exponential errors. In other words, $h(\boldsymbol{\theta})$ is a continuous coercive function in this sense (see for example, [Lange \(2010\)](#) for concepts of coercive functions on finite-dimensional Euclidean spaces). Using similar principles as in the context of continuous coercive functions on finite-dimensional Euclidean spaces, it can be shown that \mathcal{S} is a closed and bounded set. Since for all η , η' is uniformly bounded on \mathcal{G}_n , it follows that equicontinuity of η holds on $\mathcal{G}_n \cap \mathcal{S}$. Thus, the sets $\mathcal{G}_n \cap \mathcal{S}$ are compact, for $n \geq 1$.

Now, since $\mathcal{G}_n \cap \mathcal{S}$ is compact for $n \geq 1$, by Theorem 10, for any $\epsilon > 0$, there exists $n_0 \geq 1$, such that for $n \geq n_0$, $\frac{1}{n} \log R_n(\boldsymbol{\theta}) \leq -h(\boldsymbol{\theta}) + \epsilon$, for all $\boldsymbol{\theta} \in \mathcal{G}_n \cap \mathcal{S}$. Then

$$\sup_{\boldsymbol{\theta} \in \mathcal{G}_n \cap \mathcal{S}} \frac{1}{n} \log R_n(\boldsymbol{\theta}) \leq -h(\mathcal{G}_n \cap \mathcal{S}) + \epsilon \leq -h(\boldsymbol{\Theta}) + \epsilon.$$

That is,

$$\sup_{\boldsymbol{\theta} \in \mathcal{G}_n \cap \mathcal{S}} \frac{1}{n} \log R_n(\boldsymbol{\theta}) \leq -h(\boldsymbol{\Theta}) \text{ almost surely, as } n \rightarrow \infty. \quad (4.2.13)$$

We now show that

$$\sup_{\theta \in \mathcal{S}^c} \frac{1}{n} \log R_n(\theta) \leq -h(\Theta) \text{ almost surely, as } n \rightarrow \infty. \quad (4.2.14)$$

First note that if $\sup_{\theta \in \mathcal{S}^c} \frac{1}{n} \log R_n(\theta) > -h(\Theta)$ infinitely often, then $\frac{1}{n} \log R_n(\theta) > -h(\Theta)$ for some $\theta \in \mathcal{S}^c$ infinitely often. But $\frac{1}{n} \log R_n(\theta) > -h(\Theta)$ if and only if $\frac{1}{n} \log R_n(\theta) + h(\theta) > h(\theta) - h(\Theta)$, for $\theta \in \mathcal{S}^c$. Hence, if we can show that

$$P \left(\left| \frac{1}{n} \log R_n(\theta) + h(\theta) \right| > \kappa - h(\Theta), \text{ for } \theta \in \mathcal{S}^c \text{ infinitely often} \right) = 0, \quad (4.2.15)$$

then (4.2.14) will be proved. We use the Borel-Cantelli lemma to prove (4.2.15). In other words, we prove in the supplement, in the context of verifying condition (S6) of Shalizi, that

Theorem 11 *For both normal and double exponential errors, under (A1)–(A5), it holds that*

$$\sum_{n=1}^{\infty} \int_{\mathcal{S}^c} P \left(\left| \frac{1}{n} \log R_n(\theta) + h(\theta) \right| > \kappa - h(\Theta) \right) d\pi(\theta) < \infty. \quad (4.2.16)$$

Since $h(\theta)$ is continuous, (S7) holds trivially. In other words, all the assumptions (S1)–(S7) are satisfied for Gaussian process regression, for both normal and double exponential errors. Formally, our results lead to the following theorem.

Theorem 12 *Assume the Gaussian process regression model where the errors are either normally or double-exponentially distributed. Then under the conditions (A1) – (A5), (4.1.2) holds. Also, for any measurable set A with $\pi(A) > 0$, if $\beta > 2h(A)$, where h is given by (4.2.7) for normal errors and (4.2.8) for double-exponential errors, or if $A \subset \cap_{k=n}^{\infty} \mathcal{G}_k$ for some n , where \mathcal{G}_k is given by (4.2.9), then (4.1.2) and (4.1.3) hold.*

4.2.3 Rate of convergence

For Shalizi's approach to the rate of convergence, it is first required to observe that for each measurable $A \subseteq \Theta$, for every $\delta > 0$, there exists a random natural number $\tau(A, \delta)$ such that $n^{-1} \log \int_A R_n(\theta) d\pi(\theta) \leq \delta + \limsup_n n^{-1} \log \int_A R_n(\theta) d\pi(\theta)$ for all $n > \tau(A, \delta)$, provided the lim sup is finite.

Shalizi considered the set $N_{\epsilon_n} = \{\theta : h(\theta) \leq h(\Theta) + \epsilon_n\}$, where $\epsilon_n \rightarrow 0$ and $n\epsilon_n \rightarrow \infty$, as $n \rightarrow \infty$, and proved the following result.

Theorem 13 *Under (S1)–(S7), if for each $\delta > 0$,*

$$\tau(\mathcal{G}_n \cap N_{\epsilon_n}^c, \delta) \leq n \quad (4.2.17)$$

eventually almost surely, then (4.1.4) holds almost surely.

To investigate the rate of convergence in our cases, we need to show that for any $\varepsilon > 0$ and all n sufficiently large,

$$\frac{1}{n} \log \int_{\mathcal{G}_n \cap N_{\epsilon_n}^c} R_n(\theta) d\pi(\theta) \leq -h(\mathcal{G}_n \cap N_{\epsilon_n}^c) + \varepsilon. \quad (4.2.18)$$

For $\epsilon_n \downarrow 0$ such that $n\epsilon_n \rightarrow 0$ as $n \rightarrow \infty$, it holds that $N_{\epsilon_n}^c \uparrow \Theta$. Since $\mathcal{G}_n \uparrow \Theta$ as well, $h(\mathcal{G}_n \cap N_{\epsilon_n}^c) \downarrow h(\Theta)$, since $h(\theta)$ is continuous in θ . Combining these arguments with (4.2.18) makes it clear that if we can show

$$\frac{1}{n} \log \int_{\mathcal{G}_n \cap N_{\epsilon_n}^c} R_n(\theta) d\pi(\theta) \leq -h(\Theta) + \varepsilon, \quad (4.2.19)$$

for any $\varepsilon > 0$ and all n sufficiently large, where $\epsilon_n \downarrow 0$ such that $n\epsilon_n \rightarrow 0$ as $n \rightarrow \infty$,

then that ϵ_n is the rate of convergence. Now, the same steps as (4.2.11) lead to

$$\begin{aligned} \frac{1}{n} \log \int_{\mathcal{G}_n \cap N_{\epsilon_n}^c} R_n(\theta) d\pi(\theta) &\leq \frac{1}{n} \log \left[\sup_{\theta \in \mathcal{G}_n} R_n(\theta) \pi(\mathcal{G}_n) \right] \\ &\leq \frac{1}{n} \sup_{\theta \in \mathcal{G}_n} \log R_n(\theta). \end{aligned} \quad (4.2.20)$$

In light of (4.2.20), (4.2.12), (4.2.13), (4.2.14) and (4.2.15) we only need to verify (4.2.16) to establish (4.2.19). As we have already verified (4.2.16) for both Gaussian and double exponential errors, (4.2.19) stands verified.

In other words, (4.2.17), and hence (4.1.4) hold for both the Gaussian process models with Gaussian and double exponential errors, so that their convergence rate is given by ϵ_n . In other words, the posterior rate of convergence with respect to KL-divergence is just slower than n^{-1} (just slower than $n^{-\frac{1}{2}}$ with respect to Hellinger distance), for both kinds of errors that we consider. Our result can be formally stated as the following theorem.

Theorem 14 *For Gaussian process regression with either normal or double exponential errors, under (A1)–(A5), (4.1.4) holds almost surely, for $\epsilon_n \downarrow 0$ such that $n\epsilon_n \rightarrow \infty$.*

4.2.4 Consequences of model misspecification

Suppose that the true function η_0 consists of countable number of discontinuities but has continuous first order partial derivatives at all other points. Then $\eta_0 \notin \mathcal{C}'(\mathcal{X})$, that is, η_0 is not in the parameter space. Now, assume that there exists some $\tilde{\eta} \in \mathcal{C}'(\mathcal{X})$ such that $\tilde{\eta}(x) = \eta_0(x)$ for all $x \in \mathcal{X}$ where η_0 is continuous.

Note that it is always possible to obtain discontinuous η_0 given $\tilde{\eta} \in \mathcal{C}'(\mathcal{X})$ by creating countable number of points of discontinuities in $\tilde{\eta}$ (for example, let $\tilde{\eta}(x) = x$ and $\eta_0(x) = x$ if $x \neq 1$ and $\eta_0(1) = 10$). On the other hand, there need not exist even continuous $\tilde{\eta}$ corresponding to any η_0 with countable number of discontinuities (for

instance, $\eta_0(x) = x/|x|$ when $x \neq 0$ and $\eta_0(0) = 0$, does not admit any continuous modification). However, in many cases such η_0 does exist.

Then, if the probability measure Q of (A3) is dominated by the Lebesgue measure, it follows from (4.2.7) and (4.2.8), that $h(\Theta) = 0$ for both the Gaussian and double exponential error models. In this case, the posterior of η concentrates around $\tilde{\eta}$, which is the same as η_0 except at the countable number of discontinuities of η_0 . If (η_0, σ_0) is such that $0 < h(\Theta) < \infty$, then the posterior concentrates around the minimizers of $h(\theta)$, provided such minimizers exist in Θ .

Now, following Shalizi, let us define the one-step-ahead predictive distribution of θ by $F_\theta^n \equiv F_\theta(Y_n | Y_1, \dots, Y_{n-1})$, with the convention that $n = 1$ gives the marginal distribution of the first observation. Similarly, let $P^n \equiv P^n(Y_n | Y_1, \dots, Y_{n-1})$, which is the best prediction one could make had P been known. The posterior predictive distribution is given by $F_\pi^n = \int_\Theta F_\theta^n d\pi(\theta | \mathbf{Y}_n)$. With the above definitions, Shalizi (2009) proved the following results:

Theorem 15 *Under assumptions (S1)–(S7), with probability 1,*

$$\limsup_{n \rightarrow \infty} \rho_H^2(P^n, F_\pi^n) \leq h(\Theta); \quad (4.2.21)$$

$$\limsup_{n \rightarrow \infty} \rho_{TV}^2(P^n, F_\pi^n) \leq 4h(\Theta), \quad (4.2.22)$$

where ρ_H and ρ_{TV} are Hellinger and total variation metrics, respectively.

Since, for both our Gaussian process models with normal and double exponential errors, $h(\Theta) = 0$ if η_0 consists of countable number of discontinuities, it follows from (4.2.21) and (4.2.22) that in spite of such misspecification, the posterior predictive distribution does a good job in learning the best possible predictive distribution in terms of the popular Hellinger and the total variation distance. We state our result formally as the following theorem.

Theorem 16 *In the Gaussian process regression problem with either normal or double exponential errors, assume that the true function η_0 consists of countable number of discontinuities but has continuous first order partial derivatives at all other points. Also assume that the probability measure Q of (A3) is dominated by the Lebesgue measure. Then under (A1) – (A5),*

$$\limsup_{n \rightarrow \infty} \rho_H(P^n, F_\pi^n) = 0;$$

$$\limsup_{n \rightarrow \infty} \rho_{TV}(P^n, F_\pi^n) = 0,$$

almost surely.

4.3 The general nonparametric regression setup

Following Choi (2009) we consider the following model:

$$y_i = \eta(x_i) + \epsilon_i; \quad i = 1, \dots, n; \tag{4.3.1}$$

$$\epsilon_i \stackrel{iid}{\sim} \frac{1}{\sigma} \phi\left(\frac{\epsilon_i}{\sigma}\right); \quad \sigma > 0; \tag{4.3.2}$$

$$\eta(\cdot) \sim \pi_\eta(\cdot); \tag{4.3.3}$$

$$\sigma \sim \pi_\sigma(\cdot). \tag{4.3.4}$$

In (4.3.2), we model the random errors $\epsilon_i; i = 1, \dots, n$ as *iid* samples from some density $\frac{1}{\sigma} \phi\left(\frac{\cdot}{\sigma}\right)$. In (4.3.3), π_η stands for any reasonable stochastic process prior, which may or may not be Gaussian, and in (4.3.4), π_σ is some appropriate prior on σ .

4.3.1 Additional assumptions and their discussions

Regarding the model and the prior, we make the following assumptions in addition to (A1) – (A5) presented in Section 4.2.1:

(A6) The prior on η is chosen such that for $\beta > 2h(\Theta)$,

$$\begin{aligned}\pi\left(\|\eta\| \leq \exp\left((\beta n)^{1/4}\right)\right) &\geq 1 - c_\eta \exp(-\beta n); \\ \pi\left(\|\eta'_j\| \leq \exp\left((\beta n)^{1/4}\right)\right) &\geq 1 - c_{\eta'_j} \exp(-\beta n), \text{ for } j = 1, \dots, d,\end{aligned}\quad (4.3.5)$$

where c_η and $c_{\eta'_j}$; $j = 1, \dots, d$, are positive constants.

(A7) $\phi(\cdot)$ is symmetric about zero; that is, for any $x \in \mathbb{R}$, $\phi(x) = \phi(|x|)$. Further, $\log \phi$ is L -Lipschitz, that is, there exists a $L > 0$ such that $|\log \phi(x_1) - \log \phi(x_2)| \leq L|x_1 - x_2|$, for any $x_1, x_2 \in \mathbb{R}$.

(A8) For $x \in \mathcal{X}$, let $g_{\eta, \sigma}(x) = E_{\theta_0} \left[\log \phi \left(\frac{y - \eta(x)}{\sigma} \right) \right] = \int_{-\infty}^{\infty} \log \phi \left(\frac{\sigma_0 z + \eta_0(x) - \eta(x)}{\sigma} \right) \phi(z) dz$. Then given (η, σ) , $U_i = \log \phi \left(\frac{y_i - \eta(x_i)}{\sigma} \right) - g_{\eta, \sigma}(x_i)$ are independent sub-exponential random variables satisfying for any $i = 1, \dots, n$,

$$E_{\theta_0} [\exp(\lambda U_i)] \leq \exp\left(\frac{\lambda^2 s_{\eta, \sigma}^2}{2}\right), \text{ for } |\lambda| \leq s_{\eta, \sigma}^{-1}, \quad (4.3.6)$$

where, for $c_1 > 0$, $c_2 > 0$,

$$s_{\eta, \sigma} = \frac{c_1 \|\eta - \eta_0\| + c_2}{\sigma}. \quad (4.3.7)$$

(A9) For $\sigma > 0$, $\int_{-\infty}^{\infty} |\log \phi(\frac{\sigma_0}{\sigma} z)| \phi(z) dz \leq \frac{c_3}{\sigma}$, where $c_3 > 0$. Also, $\int_{-\infty}^{\infty} |z| \phi(z) dz < \infty$.

(A10) (i) $E_X [g_{\eta, \sigma}(X)]$ is jointly continuous in (η, σ) ;
(ii) $E_X [g_{\eta, \sigma}(X)] \rightarrow \infty$ as $\|\theta\| = \|\eta\| + \sigma \rightarrow \infty$.

Discussion of the new assumptions

Condition (A6) ensures that the prior probabilities of the complements of the sieves are exponentially small. Such a requirement is common to most Bayesian asymptotic

theories. In particular, the first two inequalities are satisfied by Gaussian process priors even if $\exp((\beta n)^{1/4})$ is replaced by $\sqrt{\beta n}$.

Assumption (A7) is the same as that of Choi (2009), and holds in the case of double exponential errors, for instance.

Conditions (A8), (A9) and (A10) are reasonably mild conditions, and as shown in the supplement, are satisfied by double exponential errors.

As before, let $\Theta = \mathcal{C}'(\mathcal{X}) \times \mathbb{R}^+$ denote the infinite-dimensional parameter space for our model.

4.3.2 Posterior convergence

As before, we provide a summary of our results leading to posterior convergence in our general setup. The details are provided in the supplement.

Lemma 17 *Under our model assumptions and conditions (A1) and (A3), the KL-divergence rate $h(\theta)$ exists for $\theta \in \Theta$, and is given by*

$$h(\theta) = \log \left(\frac{\sigma}{\sigma_0} \right) + c - EX [g_{\eta, \sigma}(X)], \quad (4.3.8)$$

where $c = E_{\theta_0} \left[\log \phi \left(\frac{y_i - \eta_0(x_i)}{\sigma_0} \right) \right] = \int_{-\infty}^{\infty} [\log \phi(z)] \phi(z) dz$.

Theorem 18 *Under our model assumptions and conditions (A1) and (A3), the asymptotic equipartition property holds, and is given by*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log R_n(\theta) = -h(\theta), \text{ almost surely.}$$

The convergence is uniform on any compact subset of Θ .

Lemma 17 and Theorem 18 ensure that conditions (S1) – (S3) of Shalizi hold, and (S4) holds since $h(\theta)$ is almost surely finite. We construct the sieves \mathcal{G}_n as in (4.2.9). Hence,

as before, $\mathcal{G}_n \rightarrow \Theta$ as $n \rightarrow \infty$ and the assumptions on η, η' given by (A6), together with (A4) ensure that $\pi(\mathcal{G}_n^c) \leq \alpha \exp(-\beta n)$, for some $\alpha > 0$. This result, continuity of $h(\theta)$, compactness of \mathcal{G}_n and the uniform convergence result of Theorem 10, together ensure (S5).

As regards (S6), let us note that from the definition of $g_{\eta,\sigma}(x)$ and Lipschitz continuity of $\log \phi$, it follows that $E_X[g_{\eta,\sigma}(X)]$ is Lipschitz continuous in η . However, we still need to assume that $E_X[g_{\eta,\sigma}(X)]$ is jointly continuous in $\theta = (\eta, \sigma)$. Due to (A10) it follows that $h(\theta)$ is continuous in θ and $h(\theta) \rightarrow \infty$ as $\|\theta\| \rightarrow \infty$. In other words, $h(\theta)$ is a continuous coercive function as in the Gaussian process setup. Hence, as before, it is seen that $\mathcal{G}_n \cap \mathcal{S}$ is a compact set. With these observations, we then have the following result analogous to the Gaussian process case, the proof which is provided in the supplement.

Theorem 19 *In our setup, under (A1)–(A10), it holds that*

$$\sum_{n=1}^{\infty} \int_{\mathcal{S}^c} P\left(\left|\frac{1}{n} \log R_n(\theta) + h(\theta)\right| > \kappa - h(\Theta)\right) d\pi(\theta) < \infty.$$

Since $h(\theta)$ is continuous, (S7) holds trivially. Thus, all the assumptions (S1)–(S7) are satisfied, showing that Theorems 1 and 2 hold. Formally, our results lead to the following theorem.

Theorem 20 *Assume the hierarchical model given by (4.3.1), (4.3.2), (4.3.3) and (4.3.4). Then under the conditions (A1) – (A10), (4.1.2) holds. Also, for any measurable set A with $\pi(A) > 0$, if $\beta > 2h(A)$, where h is given by (4.3.8), or if $A \subset \cap_{k=n}^{\infty} \mathcal{G}_k$ for some n , where \mathcal{G}_k is given by (4.2.9), then (4.1.3) holds.*

4.3.3 Rate of convergence and consequences of model misspecification

For the general nonparametric model, the same result as Theorem 14 holds, under (A1)–(A10). Also, the same issues regarding model misspecification as detailed in Section

4.2.4 continues to be relevant in this setup. In other words, Theorem 16 holds under (A1) – (A10).

4.4 Conclusion

The fields of both theoretical and applied Bayesian nonparametric regression are dominated by Gaussian process priors and Gaussian noise. From the asymptotics perspective, even in the Gaussian setup, a comprehensive theory unifying posterior convergence for both random and non-random covariates along with the rate of convergence in the case of general priors for the unknown error variance, while also allowing for misspecification, seems to be very rare. Even more rare is the aforementioned investigations in the setting where a general stochastic process prior is on the unknown regression function is considered and the noise distribution is non-Gaussian and thick-tailed.

The approach of Shalizi allowed us to consider the asymptotic theory incorporating all the above issues, for both Gaussian and general stochastic process prior for the regression function. The approach, apart from enabling us to ensure consistency for both random and non-random covariates, allows us to compute the rate of convergence, while allowing misspecifications. Perhaps the most interesting result that we obtained is that even if the unknown regression function is misspecified, the posterior predictive distribution still captures the true predictive distribution asymptotically, for both Gaussian and general setups.

It seems that the most important condition among the assumptions of Shalizi is the asymptotic equipartition property. This directly establishes the KL property of the posterior which characterizes the posterior convergence, the rate of posterior convergence and misspecification. Interestingly, such a property that plays the key role, turned out to be relatively easy to establish in our context under reasonably mild conditions. On the other hand, in all the applications that we investigated so far, (S6) turned out to be the most difficult to verify. But the approach we devised to handle this condition and

the others, seem to be generally applicable for investigating posterior asymptotics in general Bayesian parametric and nonparametric problems.

Appendix

4.A1 Preliminaries for ensuring posterior consistency under general set-up

Following Shalizi (2009) we consider a probability space (Ω, \mathcal{F}, P) , and a sequence of random variables y_1, y_2, \dots , taking values in some measurable space (Ξ, \mathcal{Y}) , whose infinite-dimensional distribution is P . The natural filtration of this process is $\sigma(\mathbf{y}_n)$, the smallest σ -field with respect to which \mathbf{y}_n is measurable. In other words, the distribution P is an infinite-dimensional distribution since it is the joint distribution of infinitely many random variables corresponding to a valid stochastic process. As guaranteed by Kolmogorov's consistency result (see, for example, Billingsley (1995), Schervish (1995)), all finite-dimensional distributions associated with P can be obtained by marginalizing over the remaining (infinite number of) variables. The theoretical development requires no restrictive assumptions on P such as it being a product measure, Markovian, or exchangeable, thus paving the way for great generality.

We denote the distributions of processes adapted to $\sigma(\mathbf{y}_n)$ by F_θ , where θ is associated with a measurable space (Θ, \mathcal{T}) , and is generally infinite-dimensional. In other words, assuming that θ is the infinite-dimensional distribution of the stochastic process $\{Y_1, Y_2, \dots\}$, F_θ denotes the n -dimensional marginal distribution associated with θ ; n is suppressed for ease of notation. For parametric models, the probability measure θ corresponds to a probability density with respect to some dominating measure (such as Lebesgue or counting measure) and consists of finite number of parameters. For nonparametric models, θ is usually associated with an infinite number of parameters

and may not have a density with respect to σ -finite measures.

For the sake of convenience, we assume, as in Shalizi (2009), that P and all the F_θ are dominated by a common reference measure, with respective densities f_{θ_0} and f_θ . The usual assumptions that $P \in \Theta$ or even P lies in the support of the prior on Θ , are not required for Shalizi's result, rendering it very general indeed.

Given a prior π on θ , we assume that the posterior distributions $\pi(\cdot | \mathbf{X}_n)$ are dominated by a common measure for all $n \geq 1$; abusing notation, we denote the density at θ by $\pi(\theta | \mathbf{X}_n)$.

4.A1.1 Assumptions and theorems of Shalizi

(S1) Consider the following likelihood ratio:

$$R_n(\theta) = \frac{f_\theta(\mathbf{Y}_n)}{f_{\theta_0}(\mathbf{Y}_n)}. \quad (4.A1.1)$$

Assume that $R_n(\theta)$ is $\sigma(\mathbf{Y}_n) \times \mathcal{T}$ -measurable for all $n > 0$.

(S2) For every $\theta \in \Theta$, the KL-divergence rate

$$h(\theta) = \lim_{n \rightarrow \infty} \frac{1}{n} E \left(\log \frac{f_{\theta_0}(\mathbf{Y}_n)}{f_\theta(\mathbf{Y}_n)} \right). \quad (4.A1.2)$$

exists (possibly being infinite) and is \mathcal{T} -measurable. Note that in the *iid* set-up, $h(\theta)$ reduces to the KL-divergence between the true and the hypothesized model, so that $h(\theta)$ may be regarded as a generalized KL-divergence measure.

(S3) For each $\theta \in \Theta$, the generalized or relative asymptotic equipartition property holds, and so, almost surely,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log R_n(\theta) = -h(\theta).$$

Roughly, the terminology “asymptotic equipartition” refers to dividing up $\log [R_n(\theta)]$ into n factors for large n such that all the factors are asymptotically equal. Again,

considering the *iid* scenario helps clarify this point, as in this case each factor converges to the same Kullback-Leibler divergence between the true and the postulated model. With this understanding note that the purpose of condition (S3) is to ensure that relative to the true distribution, the likelihood of each θ decreases to zero exponentially fast, with rate being the KL-divergence rate.

(S4) Let $I = \{\theta : h(\theta) = \infty\}$. The prior π satisfies $\pi(I) < 1$.

Following the notation of Shalizi (2009), for $A \subseteq \Theta$, let

$$h(A) = \text{ess inf}_{\theta \in A} h(\theta); \quad (4.A1.3)$$

$$J(\theta) = h(\theta) - h(\Theta); \quad (4.A1.4)$$

$$J(A) = \text{ess inf}_{\theta \in A} J(\theta), \quad (4.A1.5)$$

where, for any function $g : \Theta \mapsto \mathbb{R}$, where \mathbb{R} is the real line,

$$\text{ess inf}_{\theta \in A} g(\theta) = \sup \{r \in \mathbb{R} : g(\theta) > r, \text{ for almost all } \theta \in A\},$$

is the essential infimum of g over the set A . Here ‘‘almost all’’ is with respect to the prior distribution. In words, essential infimum is the greatest lower bound which holds with prior probability one.

(S5) There exists a sequence of sets $\mathcal{G}_n \rightarrow \Theta$ as $n \rightarrow \infty$ such that:

(1)

$$\pi(\mathcal{G}_n) \geq 1 - \alpha \exp(-\beta n), \text{ for some } \alpha > 0, \beta > 2h(\Theta); \quad (4.A1.6)$$

(2) The convergence in (S3) is uniform in θ over $\mathcal{G}_n \setminus I$.

(3) $h(\mathcal{G}_n) \rightarrow h(\Theta)$, as $n \rightarrow \infty$.

The sets \mathcal{G}_n can be loosely interpreted as the sieves corresponding to the method of sieves (Geman and Hwang (1982)) such that the behaviour of the likelihood ratio and the posterior on the sets \mathcal{G}_n essentially carries over to Θ . This can be anticipated from the first and the second parts of the assumption; the second part ensuring in particular that the parts of Θ on which the log likelihood ratio may be ill-behaved have exponentially small prior probabilities. The third part is more of a technical condition that is useful in proving posterior convergence through the sets \mathcal{G}_n . For further details, see Shalizi (2009).

For each measurable $A \subseteq \Theta$, for every $\delta > 0$, there exists a random natural number $\tau(A, \delta)$ such that

$$n^{-1} \log \int_A R_n(\theta) \pi(\theta) d\theta \leq \delta + \limsup_{n \rightarrow \infty} n^{-1} \log \int_A R_n(\theta) \pi(\theta) d\theta, \quad (4.A1.7)$$

for all $n > \tau(A, \delta)$, provided $\limsup_{n \rightarrow \infty} n^{-1} \log \pi(\mathbb{I}_A R_n) < \infty$. Regarding this, the following assumption has been made by Shalizi:

- (S6) The sets \mathcal{G}_n of (S5) can be chosen such that for every $\delta > 0$, the inequality $n > \tau(\mathcal{G}_n, \delta)$ holds almost surely for all sufficiently large n .

To understand the essence of this assumption, note that for almost every data set $\{X_1, X_2, \dots\}$ there exists $\tau(\mathcal{G}_n, \delta)$ such that (4.A1.7) holds with A replaced by \mathcal{G}_n for all $n > \tau(\mathcal{G}_n, \delta)$. Since \mathcal{G}_n are sets with large enough prior probabilities, the assumption formalizes our expectation that $R_n(\theta)$ decays fast enough on \mathcal{G}_n . See Shalizi (2009) for more detailed explanation.

- (S7) The sets \mathcal{G}_n of (S5) and (S6) can be chosen such that for any set A with $\pi(A) > 0$,

$$h(\mathcal{G}_n \cap A) \rightarrow h(A), \quad (4.A1.8)$$

as $n \rightarrow \infty$.

4.A2. VERIFICATION OF THE ASSUMPTIONS OF SHALIZI FOR THE
GAUSSIAN PROCESS MODEL WITH NORMAL ERRORS

Under the above assumptions, Shalizi (2009) proved the following results.

Theorem 1 (Shalizi (2009)) Consider assumptions (S1)–(S7) and any set $A \in \mathcal{T}$ with $\pi(A) > 0$ and $h(A) > h(\Theta)$. Then,

$$\lim_{n \rightarrow \infty} \pi(A|\mathbf{Y}_n) = 0, \text{ almost surely.}$$

The rate of convergence of the log-posterior is given by the following result.

Theorem 2 (Shalizi (2009)) Consider assumptions (S1)–(S7) and any set $A \in \mathcal{T}$ with $\pi(A) > 0$. If $\beta > 2h(A)$, where β is given in (4.A1.6) under assumption (S5), or if $A \subset \cap_{k=n}^{\infty} \mathcal{G}_k$ for some n , then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \pi(A|\mathbf{Y}_n) = -J(A), \text{ almost surely.}$$

4.A2 Verification of the assumptions of Shalizi for the Gaussian process model with normal errors

4.A2.1 Verification of (S1)

note that

$$f_{\theta}(\mathbf{Y}_n) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \eta(x_i))^2 \right\}; \quad (4.A2.1)$$

$$f_{\theta_0}(\mathbf{Y}_n) = \frac{1}{(\sigma_0\sqrt{2\pi})^n} \exp \left\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^n (Y_i - \eta_0(x_i))^2 \right\}. \quad (4.A2.2)$$

The equations (4.A2.1) and (4.A2.2) yield, in our case,

$$\frac{1}{n} \log R_n(\theta) = \log \left(\frac{\sigma_0}{\sigma} \right) + \frac{1}{2\sigma_0^2} \times \frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i))^2 - \frac{1}{2\sigma^2} \times \frac{1}{n} \sum_{i=1}^n (y_i - \eta(x_i))^2. \quad (4.A2.3)$$

**4.A2. VERIFICATION OF THE ASSUMPTIONS OF SHALIZI FOR THE
GAUSSIAN PROCESS MODEL WITH NORMAL ERRORS**

We show that the right hand side of (4.A2.3), which we denote as $f(\mathbf{y}_n, \theta)$, is continuous in (\mathbf{y}_n, θ) , which is sufficient to confirm measurability of $R_n(\theta)$. Let $\|(\mathbf{y}_n, \theta)\| = \|\mathbf{y}_n\| + \|\theta\|$, where $\|\mathbf{y}_n\|$ is the Euclidean norm and $\|\theta\| = \|\eta\| + |\sigma|$, with $\|\eta\| = \sup_{x \in \mathcal{X}} |\eta(x)|$. Since \mathcal{X} is compact and η is almost surely continuous, it follows that $\|\eta\| < \infty$ almost surely.

Consider $\mathbf{y}_n = (y_1, y_2, \dots, y_n)^T$ and $\boldsymbol{\eta}_{0n} = (\eta_0(x_1), \dots, \eta_0(x_n))^T$. Then

$$\sum_{i=1}^n (y_i - \eta_0(x_i))^2 = \mathbf{y}_n^T \mathbf{y}_n - 2\mathbf{y}_n^T \boldsymbol{\eta}_{0n} + \boldsymbol{\eta}_{0n}^T \boldsymbol{\eta}_{0n} \quad (4.A2.4)$$

is clearly continuous in \mathbf{y}_n . Now note that

$$\frac{1}{n} \sum_{i=1}^n (y_i - \eta(x_i))^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i))^2 + \frac{1}{n} \sum_{i=1}^n (\eta(x_i) - \eta_0(x_i))^2 - \frac{2}{n} \sum_{i=1}^n (y_i - \eta_0(x_i))(\eta(x_i) - \eta_0(x_i)), \quad (4.A2.5)$$

where we have already proved continuity of the first term on the right hand side of (4.A2.5). To see continuity of $\frac{1}{n} \sum_{i=1}^n (\eta(x_i) - \eta_0(x_i))^2$ with respect to η , first consider any sequence $\{\eta_j : j = 1, 2, \dots\}$ satisfying $\|\eta_j - \tilde{\eta}\| \rightarrow 0$, as $j \rightarrow \infty$. Then

$$\begin{aligned} & \left| \frac{1}{n} \sum_{i=1}^n (\eta_j(x_i) - \eta_0(x_i))^2 - \frac{1}{n} \sum_{i=1}^n (\tilde{\eta}(x_i) - \eta_0(x_i))^2 \right| \\ & \leq \frac{1}{n} \sum_{i=1}^n |\eta_j(x_i) - \tilde{\eta}(x_i)| \times |(\eta_j(x_i) - \eta_0(x_i)) + (\tilde{\eta}(x_i) - \eta_0(x_i))| \\ & \leq \|\eta_j - \tilde{\eta}\| \times [\|\eta_j - \eta_0\| + \|\tilde{\eta} - \eta_0\|] \\ & \leq \|\eta_j - \tilde{\eta}\| \times [\|\eta_j - \tilde{\eta}\| + 2\|\tilde{\eta} - \eta_0\|] \\ & \rightarrow 0, \text{ as } j \rightarrow \infty. \end{aligned} \quad (4.A2.7)$$

This proves continuity of the second term of (4.A2.5).

For the third term of (4.A2.5) we now prove that for any $\tilde{\mathbf{y}} \in \mathbb{R}^n$, and for any sequence $\{\mathbf{y}_j : j = 1, 2, \dots\}$ (we denote the i -th component of \mathbf{y}_j as y_{ij}) such that $\|\mathbf{y}_j - \tilde{\mathbf{y}}\| \rightarrow 0$, as $j \rightarrow \infty$, and for any function $\tilde{\eta}$ associated with any sequence $\{\eta_j : j = 1, 2, \dots\}$ satisfying

**4.A2. VERIFICATION OF THE ASSUMPTIONS OF SHALIZI FOR THE
GAUSSIAN PROCESS MODEL WITH NORMAL ERRORS**

$\|\eta_j - \tilde{\eta}\| \rightarrow 0$, as $j \rightarrow \infty$, $\sum_{i=1}^n (y_{ij} - \eta_0(x_i))(\eta_j(x_i) - \eta_0(x_i)) \rightarrow \sum_{i=1}^n (\tilde{y}_i - \eta_0(x_i))(\tilde{\eta}(x_i) - \eta_0(x_i))$, as $j \rightarrow \infty$. Indeed, observe that

$$\begin{aligned} & \left| \sum_{i=1}^n (y_{ij} - \eta_0(x_i))(\eta_j(x_i) - \eta_0(x_i)) - \sum_{i=1}^n (\tilde{y}_i - \eta_0(x_i))(\tilde{\eta}(x_i) - \eta_0(x_i)) \right| \\ &= \left| \sum_{i=1}^n (y_{ij} - \tilde{y}_i)(\eta_j(x_i) - \tilde{\eta}(x_i)) + \sum_{i=1}^n (\tilde{y}_i - \eta_0(x_i))(\eta_j(x_i) - \tilde{\eta}(x_i)) \right. \\ &\quad \left. + \sum_{i=1}^n (y_{ij} - \tilde{y}_i)(\tilde{\eta}(x_i) - \eta_0(x_i)) \right| \\ &\leq n\|\mathbf{y}_j - \tilde{\mathbf{y}}\|\|\eta_j - \tilde{\eta}\| + n\|\tilde{\mathbf{y}} - \eta_0\|\|\eta_j - \tilde{\eta}\| + n\|\mathbf{y}_j - \tilde{\mathbf{y}}\|\|\tilde{\eta} - \eta_0\| \\ &\rightarrow 0, \text{ as } \|\mathbf{y}_j - \tilde{\mathbf{y}}\| \rightarrow 0 \text{ and } \|\eta_j - \tilde{\eta}\| \rightarrow 0, \text{ as } j \rightarrow \infty. \end{aligned}$$

Hence, $\sum_{i=1}^n (y_i - \eta_0(x_i))(\eta(x_i) - \eta_0(x_i))$ is continuous in \mathbf{y}_n and η . Continuity is clearly preserved if the above expression is divided by σ .

Also, the first term of $f(\mathbf{y}_n, \theta)$, given by $\log\left(\frac{\sigma_0}{\sigma}\right)$, is clearly continuous in σ . Thus, continuity of $f(\mathbf{y}_n, \theta)$ with respect to (\mathbf{y}_n, θ) is guaranteed, so that (S1) holds. Also observe that when the covariates are regarded as random, due to measurability of $\eta_0(X)$ assumed in (A4) and continuity of $\eta(x)$ in x .

4.A2.2 Verification of (S2) and proof of Lemma 9 for Gaussian errors

It follows from (4.A2.1) and (4.A2.2), that

$$\log \frac{f_{\theta_0}(\mathbf{y}_n)}{f_{\theta}(\mathbf{y}_n)} = n \log \left(\frac{\sigma}{\sigma_0} \right) - \frac{1}{2\sigma_0^2} \sum_{i=1}^n (y_i - \eta_0(x_i))^2 + \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \eta(x_i))^2, \quad (4.A2.8)$$

so that

$$\frac{1}{n} E_{\theta_0} \left(\log \frac{f_{\theta_0}(\mathbf{y}_n)}{f_{\theta}(\mathbf{y}_n)} \right) = \log \left(\frac{\sigma}{\sigma_0} \right) - \frac{1}{2} + \frac{\sigma_0^2}{2\sigma^2} + \frac{1}{2\sigma^2} \times \frac{1}{n} \sum_{i=1}^n (\eta(x_i) - \eta_0(x_i))^2. \quad (4.A2.9)$$

**4.A2. VERIFICATION OF THE ASSUMPTIONS OF SHALIZI FOR THE
GAUSSIAN PROCESS MODEL WITH NORMAL ERRORS**

By (A3), as $n \rightarrow \infty$,

$$\frac{1}{n} \sum_{i=1}^n (\eta(x_i) - \eta_0(x_i))^2 \rightarrow E_X [\eta(X) - \eta_0(X)]^2 = \int_{\mathcal{X}} [\eta(X) - \eta_0(X)]^2 dQ. \quad (4.A2.10)$$

Hence,

$$\frac{1}{n} E_{\theta_0} \left(\log \frac{f_{\theta_0}(\mathbf{y}_n)}{f_{\theta}(\mathbf{y}_n)} \right) \rightarrow \log \left(\frac{\sigma}{\sigma_0} \right) - \frac{1}{2} + \frac{\sigma_0^2}{2\sigma^2} + \frac{1}{2\sigma^2} E_X [\eta(X) - \eta_0(X)]^2, \text{ as } n \rightarrow \infty. \quad (4.A2.11)$$

We let

$$h(\theta) = \log \left(\frac{\sigma}{\sigma_0} \right) - \frac{1}{2} + \frac{\sigma_0^2}{2\sigma^2} + \frac{1}{2\sigma^2} E_X [\eta(X) - \eta_0(X)]^2.$$

4.A2.3 Verification of (S3) and proof of Theorem 10 for Gaussian errors

By SLLN, as $n \rightarrow \infty$,

$$\frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i))^2 \xrightarrow{a.s.} \sigma_0^2, \quad (4.A2.12)$$

where “ $\xrightarrow{a.s.}$ ” denotes convergence almost surely. Also,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (y_i - \eta(x_i))^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i))^2 + \frac{1}{n} \sum_{i=1}^n (\eta(x_i) - \eta_0(x_i))^2 \\ &\quad + \frac{2}{n} \sum_{i=1}^n (y_i - \eta_0(x_i)) (\eta_0(x_i) - \eta(x_i)). \end{aligned} \quad (4.A2.13)$$

By (4.A2.12) the first term on the right hand side of (4.A2.13) converges almost surely to σ_0^2 . The second term converges to $E_X [\eta(X) - \eta_0(X)]^2$ and the third term converges almost surely to zero by Kolmogorov's SLLN for independent random variables, noting that $y_i - \eta_0(x_i) = \epsilon_i$ are independent zero mean random variables and $\sum_{i=1}^{\infty} i^{-2} \text{Var}((y_i - \eta_0(x_i))(\eta_0(x_i) - \eta(x_i))) = \sigma_0^2 \sum_{i=1}^{\infty} i^{-2} (\eta_0(x_i) - \eta(x_i))^2 \leq \sigma_0^2 \|\eta - \eta_0\|^2$.

$\eta_0\|^2 \sum_{i=1}^{\infty} i^{-2} < \infty$. Hence, letting $n \rightarrow \infty$ in (4.A2.3), it follows that

$$\frac{1}{n} \log R_n(\theta) \xrightarrow{a.s.} \log \left(\frac{\sigma_0}{\sigma} \right) + \frac{1}{2} - \frac{\sigma_0^2}{2\sigma^2} - \frac{1}{2\sigma^2} E_X [\eta(X) - \eta_0(X)]^2 = -h(\theta). \quad (4.A2.14)$$

The above results of course remain the same if the covariates are assumed to be random.

4.A2.4 Verification of (S4)

Note that $h(\theta) \leq \log \left(\frac{\sigma}{\sigma_0} \right) - \frac{1}{2} + \frac{\sigma_0^2}{2\sigma^2} + \frac{\|\eta - \eta_0\|^2}{2\sigma^2}$, where $0 < \sigma < \infty$ and $0 < \|\eta - \eta_0\| < \infty$ with prior probability one. Hence, $h(\theta) < \infty$ with probability one, showing that (S4) holds.

4.A2.5 Verification of (S5)

Verification of (S5) (1)

Recall that

$$\mathcal{G}_n = \left\{ (\eta, \sigma) : \|\eta\| \leq \exp((\beta n)^{1/4}), \exp(-(\beta n)^{1/4}) \leq \sigma \leq \exp((\beta n)^{1/4}), \|\eta'_j\| \leq \exp((\beta n)^{1/4}); j = 1, \dots, d \right\}.$$

4.A2. VERIFICATION OF THE ASSUMPTIONS OF SHALIZI FOR THE
GAUSSIAN PROCESS MODEL WITH NORMAL ERRORS

Then $\mathcal{G}_n \rightarrow \Theta$, as $n \rightarrow \infty$. Now note that

$$\begin{aligned}
\pi(\mathcal{G}_n) &\geq \pi \left(\|\eta\| \leq \exp((\beta n)^{1/4}), \exp(-(\beta n)^{1/4}) \leq \sigma \leq \exp((\beta n)^{1/4}) \right) \\
&\quad - \pi \left(\left\{ \|\eta'_j\| \leq \exp((\beta n)^{1/4}); j = 1, \dots, d \right\}^c \right) \\
&= \pi \left(\|\eta\| \leq \exp((\beta n)^{1/4}), \exp(-(\beta n)^{1/4}) \leq \sigma \leq \exp((\beta n)^{1/4}) \right) \\
&\quad - \pi \left(\bigcup_{j=1}^d \left\{ \|\eta'_j\| > \exp((\beta n)^{1/4}) \right\} \right) \\
&\geq 1 - \pi \left(\|\eta\| > \exp((\beta n)^{1/4}) \right) - \pi \left(\left\{ \exp(-(\beta n)^{1/4}) \leq \sigma \leq \exp((\beta n)^{1/4}) \right\}^c \right) \\
&\quad - \sum_{j=1}^d \pi \left(\|\eta'_j\| > \exp((\beta n)^{1/4}) \right) \\
&\geq 1 - (c_\eta + c_\sigma + \sum_{j=1}^d c_{\eta'_j}) \exp(-\beta n), \tag{4.A2.15}
\end{aligned}$$

by the Borell-TIS inequality and (A5). In other words, (S5) (1) holds.

Verification of (S5) (2)

We now show that (S5) (2), namely, convergence in (S3) is uniform in θ over $\mathcal{G}_n \setminus I$ holds.

First note that $I = \emptyset$ in our case, so that $\mathcal{G}_n \setminus I = \mathcal{G}_n$.

To proceed further, we show that \mathcal{G}_n is compact. Note that $\mathcal{G}_n = \mathcal{G}_{n,\eta} \times \mathcal{G}_{n,\sigma}$, where

$$\mathcal{G}_{n,\eta} = \left\{ \eta : \|\eta\| \leq \exp((\beta n)^{1/4}), \|\eta'_j\| \leq \exp((\beta n)^{1/4}); j = 1, \dots, d \right\}$$

and

$$\mathcal{G}_{n,\sigma} = \left\{ \sigma : \exp(-(\beta n)^{1/4}) \leq \sigma \leq \exp((\beta n)^{1/4}) \right\}.$$

Since $\mathcal{G}_{n,\sigma}$ is compact and products of compact sets is compact, it is enough to prove compactness of $\mathcal{G}_{n,\eta}$. We use the Arzela-Ascoli lemma to prove that $\mathcal{G}_{n,\eta}$ is compact for each $n \geq 1$. In other words, $\mathcal{G}_{n,\eta}$ is compact if and only if it is closed, bounded and

**4.A2. VERIFICATION OF THE ASSUMPTIONS OF SHALIZI FOR THE
GAUSSIAN PROCESS MODEL WITH NORMAL ERRORS**

equicontinuous. By boundedness we mean $|\eta(x)| < M$ for each $x \in \mathcal{X}$ and for each $\eta \in \mathcal{G}_{n,\eta}$. Equicontinuity entails that for any $\epsilon > 0$, there exists $\delta > 0$ which depends only on ϵ such that $|\eta(x_1) - \eta(x_2)| < \epsilon$ whenever $\|x_1 - x_2\| < \delta$, for all $\eta \in \mathcal{G}_{n,\eta}$. Closedness and boundedness are obvious from the definition of $\mathcal{G}_{n,\eta}$. Equicontinuity follows from the fact that the elements of $\mathcal{G}_{n,\eta}$ are Lipschitz continuous thanks to boundedness of the partial derivatives. Thus, $\mathcal{G}_{n,\eta}$, and hence \mathcal{G}_n is compact.

Since \mathcal{G}_n is compact for all $n \geq 1$, uniform convergence as required will be proven if we can show that $\frac{1}{n} \log R_n(\theta) + h(\theta)$ is stochastically equicontinuous almost surely in $\theta \in \mathcal{G}$ for any $\mathcal{G} \in \{\mathcal{G}_n : n = 1, 2, \dots\}$ and $\frac{1}{n} \log R_n(\theta) + h(\theta) \rightarrow 0$, almost surely, for all $\theta \in \mathcal{G}$ (see Newey (1991), Billingsley (2013)) for the general theory of uniform convergence in compact sets under stochastic equicontinuity). Since, in the context of (S3) we have already shown almost sure pointwise convergence of $\frac{1}{n} \log R_n(\theta)$ to $-h(\theta)$, it is enough to verify stochastic equicontinuity of $\frac{1}{n} \log R_n(\theta) + h(\theta)$ in $\mathcal{G} \in \{\mathcal{G}_n : n = 1, 2, \dots\}$.

Stochastic equicontinuity usually follows easily if one can prove that the function concerned is almost surely Lipschitz continuous. Recall from (4.A2.3), (4.A2.4), (4.A2.5) and (4.A2.7) that if the term $\frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i))(\eta(x_i) - \eta_0(x_i))$ can be proved Lipschitz continuous in $\eta \in \mathcal{G}$, then $\frac{1}{n} \log R_n(\theta)$ is Lipschitz for $\eta \in \mathcal{G}$. Also, if $E_X [\eta(X) - \eta_0(X)]^2$ is Lipschitz in η , then it would follow from (4.2.7) that $h(\theta)$ is Lipschitz for $\eta \in \mathcal{G}$. Since sum of Lipschitz functions is Lipschitz, this would imply that $\frac{1}{n} \log R_n(\theta) + h(\theta)$ is Lipschitz in $\eta \in \mathcal{G}$. Since the first derivative of $\frac{1}{n} \log R_n(\theta) + h(\theta)$ with respect to σ is bounded (as σ is bounded in \mathcal{G}), it would then follow that $\frac{1}{n} \log R_n(\theta) + h(\theta)$ is Lipschitz for $\theta \in \mathcal{G}$. Hence, to see that $\frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i))(\eta(x_i) - \eta_0(x_i))$ is almost surely Lipschitz in $\eta \in \mathcal{G}$, note that for any $\eta_1, \eta_2 \in \mathcal{G}$,

$$\begin{aligned} & \left| \frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i))(\eta_1(x_i) - \eta_0(x_i)) - \frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i))(\eta_2(x_i) - \eta_0(x_i)) \right| \\ & \leq \|\eta_1 - \eta_2\| \times \frac{1}{n} \sum_{i=1}^n |y_i - \eta_0(x_i)|. \end{aligned}$$

**4.A2. VERIFICATION OF THE ASSUMPTIONS OF SHALIZI FOR THE
GAUSSIAN PROCESS MODEL WITH NORMAL ERRORS**

Hence, $\frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i))(\eta(x_i) - \eta_0(x_i))$ is Lipschitz in η and since $\frac{1}{n} \sum_{i=1}^n |y_i - \eta_0(x_i)| \rightarrow E_{\theta_0} |y_1 - \eta_0(x_1)| < \infty$ as $n \rightarrow \infty$, stochastic equicontinuity follows.

That $E_X [\eta(X) - \eta_0(X)]^2$ is also Lipschitz in \mathcal{G} can be seen from the fact that for $\eta_1, \eta_2 \in \mathcal{G}$,

$$\left| E_X [\eta_1(X) - \eta_0(X)]^2 - E_X [\eta_2(X) - \eta_0(X)]^2 \right| \leq \|\eta_1 - \eta_2\| \times [\|\eta_1\| + \|\eta_2\| + 2\|\eta_0\|],$$

where $\|\eta_0\| < \kappa_0$ by (A4) and for $j = 1, 2$, $\|\eta_j\| \leq \exp((\beta m)^{1/4})$, where $\mathcal{G} = \mathcal{G}_m$, for $m \geq 1$.

Verification of (S5) (3)

We now verify (S5) (3). For our purpose, let us show that $h(\theta)$ is continuous in θ . Continuity will easily follow if we can show that $E_X [\eta(X) - \eta_0(X)]^2$ is continuous in η . As before, let η_j be a sequence of functions converging to $\tilde{\eta}$ in the sense $\|\eta_j - \tilde{\eta}\| \rightarrow 0$ as $j \rightarrow \infty$. Then, since $\left| E_X [\eta_j(X) - \eta_0(X)]^2 - E_X [\tilde{\eta}(X) - \eta_0(X)]^2 \right| \leq \|\eta_j - \tilde{\eta}\| [\|\eta_j - \tilde{\eta}\| + 2\|\tilde{\eta} - \eta_0\|] \rightarrow 0$ as $j \rightarrow \infty$, continuity follows. Hence, continuity of $h(\theta)$, compactness of \mathcal{G}_n , along with its non-decreasing nature with respect to n implies that $h(\mathcal{G}_n) \rightarrow h(\Theta)$, as $n \rightarrow \infty$.

4.A2.6 Verification of (S6) and proof of Theorem 11 for Gaussian errors

Observe that

$$\begin{aligned} \frac{1}{n} \log R_n(\theta) + h(\theta) &= \left[\frac{1}{2\sigma_0^2} \times \frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i))^2 - \frac{1}{2} \right] + \left[\frac{1}{2\sigma^2} \times \frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i))^2 - \frac{\sigma_0^2}{2\sigma^2} \right] \\ &\quad + \left[\frac{1}{2\sigma^2} \times \frac{1}{n} \sum_{i=1}^n (\eta(x_i) - \eta_0(x_i))^2 - \frac{1}{2\sigma^2} E_X (\eta(X) - \eta_0(X))^2 \right] \\ &\quad + \left[\frac{1}{\sigma^2} \times \frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i)) (\eta(x_i) - \eta_0(x_i)) \right]. \end{aligned} \tag{4.A2.16}$$

Let $\kappa_1 = \kappa - h(\Theta)$. Then it follows from (4.A2.16) that for all $\theta \in \mathcal{G}$, we have

$$\begin{aligned} & P\left(\left|\frac{1}{n} \log R_n(\theta) + h(\theta)\right| > \kappa_1\right) \\ & \leq P\left(\left|\frac{1}{2\sigma_0^2} \times \frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i))^2 - \frac{1}{2}\right| > \frac{\kappa_1}{4}\right) + P\left(\left|\frac{1}{2\sigma^2} \times \frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i))^2 - \frac{\sigma_0^2}{2\sigma^2}\right| > \frac{\kappa_1}{4}\right) \\ & \quad + P\left(\left|\frac{1}{2\sigma^2} \times \frac{1}{n} \sum_{i=1}^n (\eta(x_i) - \eta_0(x_i))^2 - \frac{1}{2\sigma^2} E_X (\eta(X) - \eta_0(X))^2\right| > \frac{\kappa_1}{4}\right) \\ & \quad + P\left(\left|\frac{1}{\sigma^2} \times \frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i)) (\eta(x_i) - \eta_0(x_i))\right| > \frac{\kappa_1}{4}\right). \end{aligned} \quad (4.A2.17)$$

Note that $\sum_{i=1}^n \left(\frac{y_i - \eta_0(x_i)}{\sigma_0}\right)^2 = \mathbf{z}_n^T \mathbf{z}_n$, where $\mathbf{z}_n \sim N_n(\mathbf{0}_n, \mathbf{I}_n)$, the n -dimensional normal distribution with mean $\mathbf{0}_n = (0, 0, \dots, 0)^T$ and covariance matrix \mathbf{I}_n , the identity matrix. Using the Hanson-Wright inequality we bound the first term of the right hand side of (4.A2.17) as follows:

$$\begin{aligned} & P\left(\left|\frac{1}{2\sigma_0^2} \times \frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i))^2 - \frac{1}{2}\right| > \frac{\kappa_1}{4}\right) \\ & = P\left(|\mathbf{z}_n^T \mathbf{z}_n - n| > \frac{n\kappa_1}{2}\right) \\ & \leq 2 \exp\left(-n \min\left\{\frac{\kappa_1^2}{16c_0}, \frac{\kappa_1}{4c_0}\right\}\right), \end{aligned} \quad (4.A2.18)$$

where $c_0 > 0$ is a constant. It follows from (4.A2.18) that

$$\int_{\mathcal{S}^c} P\left(\left|\frac{1}{2\sigma_0^2} \times \frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i))^2 - \frac{1}{2}\right| > \frac{\kappa_1}{4}\right) d\pi(\theta) \leq 2 \exp\left(-n \min\left\{\frac{\kappa_1^2}{16c_0}, \frac{\kappa_1}{4c_0}\right\}\right). \quad (4.A2.19)$$

In almost the same way as in (4.A2.18), the second term of the right hand side of

4.A2. VERIFICATION OF THE ASSUMPTIONS OF SHALIZI FOR THE
GAUSSIAN PROCESS MODEL WITH NORMAL ERRORS

(4.A2.17) can be bounded as:

$$\begin{aligned}
 & P \left(\left| \frac{1}{2\sigma^2} \times \frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i))^2 - \frac{\sigma_0^2}{2\sigma^2} \right| > \frac{\kappa_1}{4} \right) \\
 & = P \left(|z_n^T z_n - n| > \frac{n\kappa_1\sigma^2}{2\sigma_0^2} \right) \\
 & \leq 2 \exp \left(-n \min \left\{ \frac{\kappa_1^2\sigma^4}{16c_0\sigma_0^4}, \frac{\kappa_1\sigma^2}{4c_0\sigma_0^2} \right\} \right). \tag{4.A2.20}
 \end{aligned}$$

Now

$$\begin{aligned}
 & \int_{\mathcal{S}^c} P \left(\left| \frac{1}{2\sigma^2} \times \frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i))^2 - \frac{\sigma_0^2}{2\sigma^2} \right| > \frac{\kappa_1}{4} \right) d\pi(\theta) \\
 & \leq \int_{\mathcal{G}_n} 2 \exp \left(-n \min \left\{ \frac{\kappa_1^2\sigma^4}{16c_0\sigma_0^4}, \frac{\kappa_1\sigma^2}{4c_0\sigma_0^2} \right\} \right) \pi(\sigma^2) d\sigma^2 \\
 & \quad + \int_{\mathcal{G}_n^c} 2 \exp \left(-n \min \left\{ \frac{\kappa_1^2\sigma^4}{16c_0\sigma_0^4}, \frac{\kappa_1\sigma^2}{4c_0\sigma_0^2} \right\} \right) \pi(\theta) d\theta \\
 & \leq \int_{\mathcal{G}_n} 2 \exp \left(-n \min \left\{ \frac{\kappa_1^2\sigma^4}{16c_0\sigma_0^4}, \frac{\kappa_1\sigma^2}{4c_0\sigma_0^2} \right\} \right) \pi(\sigma^2) d\sigma^2 + 2\pi(\mathcal{G}_n^c) \\
 & \leq \int_{\exp(-2(\beta n)^{1/4})}^{\exp(2(\beta n)^{1/4})} 2 \exp \left(-n \frac{\kappa_1^2\sigma^4}{16c_0\sigma_0^4} \right) \pi(\sigma^2) d\sigma^2 \\
 & \quad + \int_{\exp(-2(\beta n)^{1/4})}^{\exp(2(\beta n)^{1/4})} 2 \exp \left(-n \frac{\kappa_1\sigma^2}{4c_0\sigma_0^2} \right) \pi(\sigma^2) d\sigma^2 + 2\pi(\mathcal{G}_n^c) \\
 & = \int_{\exp(-2(\beta n)^{1/4})}^{\exp(2(\beta n)^{1/4})} 2 \exp \left(-n \frac{\kappa_1^2 u^{-2}}{16c_0\sigma_0^4} \right) \pi(u^{-1}) u^{-2} du \\
 & \quad + \int_{\exp(-2(\beta n)^{1/4})}^{\exp(2(\beta n)^{1/4})} 2 \exp \left(-n \frac{\kappa_1 u^{-1}}{4c_0\sigma_0^2} \right) \pi(u^{-1}) u^{-2} du + 2\pi(\mathcal{G}_n^c). \tag{4.A2.21}
 \end{aligned}$$

Let us first consider the first term of (4.A2.21). Note that the prior $\pi(u^{-1}) u^{-2}$ is such that large values of u receive small probabilities. Hence, if this prior is replaced by an appropriate function which has a thicker tail than the prior, then the resultant integral provides an upper bound for the first term of (4.A2.21). We consider a function $\tilde{\pi}(u)$ which is of mixture form depending upon n , that is, we let

4.A2. VERIFICATION OF THE ASSUMPTIONS OF SHALIZI FOR THE
GAUSSIAN PROCESS MODEL WITH NORMAL ERRORS

$\tilde{\pi}_n(u) = c_3 \sum_{r=1}^{M_n} \psi_{rn}^{\zeta_{rn}} \exp(-\psi_{rn} u^2) u^{2(\zeta_{rn}-1)}$, where $M_n \leq \exp((\beta n)^{1/4})$ is the number of mixture components, $c_3 > 0$, for $r = 1, \dots, M_n$, $\frac{1}{2} < \zeta_{rn} \leq c_4 n^q$, for $0 < q < 1/4$ and $n \geq 1$, where $c_4 > 0$, and $0 < \psi_1 \leq \psi_{rn} < c_5 < \infty$, for all r and n . In this case, with $C_1 = \frac{1}{16c_0\sigma_0^4}$,

$$\begin{aligned} & \int_{\exp(-2(\beta n)^{1/4})}^{\exp(2(\beta n)^{1/4})} \exp(-C_1 \kappa_1^2 n u^{-2}) \pi(u^{-1}) u^{-2} du \\ & \leq c_3 \sum_{r=1}^{M_n} \psi_{rn}^{\zeta_{rn}} \int_{\exp(-2(\beta n)^{1/4})}^{\exp(2(\beta n)^{1/4})} \exp[-(C_1 \kappa_1^2 n u^{-2} + \psi_1 u^2)] (u^2)^{\zeta_{rn}-1} du. \end{aligned} \quad (4.A2.22)$$

Now the r -th integrand of (4.A2.22) is maximized at $\tilde{u}_{rn}^2 = \frac{\zeta_{rn}-1+\sqrt{(\zeta_{rn}-1)^2+4C_1\psi_1\kappa_1^2n}}{2\psi_1}$, so that for sufficiently large n , $c_1 \kappa_1 \sqrt{\frac{n}{\psi_1}} \leq \tilde{u}_{rn}^2 \leq \tilde{c}_1 \kappa_1 \sqrt{\frac{n}{\psi_1}}$, for some positive constants c_1 and \tilde{c}_1 . Now, for sufficiently large n , we have $\frac{\tilde{u}_{rn}^2}{\log \tilde{u}_{rn}^2} \geq \frac{\zeta_{rn}-1}{\psi_1(1-c_2)}$, for $0 < c_2 < 1$. Hence, for sufficiently large n , $C_1 \kappa_1^2 n \tilde{u}_{rn}^{-2} + \psi_1 \tilde{u}_{rn}^2 - (\zeta_{rn} - 1) \log(\tilde{u}_{rn}^2) \geq c_2 \psi_1 \tilde{u}_{rn}^2 \geq C_2 \kappa_1 \sqrt{\psi_1 n}$ for some positive constant C_2 . From these and (4.A2.22) it follows that

$$\begin{aligned} & \int_{\exp(-2(\beta n)^{1/4})}^{\exp(2(\beta n)^{1/4})} 2 \exp\left(-n \frac{\kappa_1^2 u^{-2}}{16c_0\sigma_0^4}\right) \pi(u^{-1}) u^{-2} du \\ & = c_3 \sum_{r=1}^{M_n} \psi_{rn}^{\zeta_{rn}} \int_{\exp(-2(\beta n)^{1/4})}^{\exp(2(\beta n)^{1/4})} \exp[-(C_1 \kappa_1^2 n u^{-2} + \psi_1 u^2)] (u^2)^{\zeta_{rn}-1} du \\ & \leq c_3 M_n \exp\left[-\left(C_2 \kappa_1 \sqrt{n\psi_1} - 2(\beta n)^{1/4} - \tilde{c}_5 n^q\right)\right] \\ & \leq c_3 \exp\left[-\left(C_2 \kappa_1 \sqrt{n\psi_1} - 3(\beta n)^{1/4} - \tilde{c}_5 n^q\right)\right]. \end{aligned} \quad (4.A2.23)$$

for some constant \tilde{c}_5 . The negative of the exponent of (4.A2.23) is clearly positive for large n .

For the second term of (4.A2.21), we consider $\tilde{\pi}_n(u) = c_3 \sum_{r=1}^{M_n} \psi_{rn}^{\zeta_{rn}} \exp(-\psi_{rn} u) u^{(\zeta_{rn}-1)}$, with $M_n \leq \exp((\beta n)^{1/4})$ being the number of mixture components, $c_3 > 0$, for $r = 1, \dots, M_n$, $0 < \zeta_{rn} \leq c_4 n^q$, for $0 < q < 1/4$ and $n \geq 1$, where $c_4 > 0$, and $0 < \psi_1 \leq \psi_{rn} < c_5 < \infty$, for all r and n . Thus, the only difference here with the previous

4.A2. VERIFICATION OF THE ASSUMPTIONS OF SHALIZI FOR THE
GAUSSIAN PROCESS MODEL WITH NORMAL ERRORS

definition of $\tilde{\pi}_n(u)$ is that here $\zeta_{rn} > 0$ instead of $\zeta_{rn} > \frac{1}{2}$, which is due to the fact that here u^2 is replaced with u . In the same way as in (4.A2.23), it then follows that

$$\int_{\exp(-2(\beta n)^{1/4})}^{\exp(2(\beta n)^{1/4})} 2 \exp\left(-n \frac{\kappa_1 u^{-1}}{4c_0 \sigma_0^2}\right) \pi(u^{-1}) u^{-2} du \leq c_3 \exp\left[-\left(C_2 \sqrt{\kappa_1 n \psi_1} - 3(\beta n)^{1/4} - \tilde{c}_5 n^q\right)\right]. \quad (4.A2.24)$$

Again, the negative of the exponent of (4.A2.24) is clearly positive for large n .

For the third term, let us first consider the case of random covariates X . Here observe that by Hoeffding's inequality (Hoeffding (1963)),

$$\begin{aligned} P\left(\left|\frac{1}{2\sigma^2} \times \frac{1}{n} \sum_{i=1}^n (\eta(x_i) - \eta_0(x_i))^2 - \frac{1}{2\sigma^2} E_X (\eta(X) - \eta_0(X))^2\right| > \frac{\kappa_1}{4}\right) \\ \leq 2 \exp\left\{-\frac{nC\kappa_1^2\sigma^4}{\|\eta - \eta_0\|^2}\right\}, \end{aligned} \quad (4.A2.25)$$

where $C > 0$ is a constant. Note that $\|\eta - \eta_0\|$ is clearly the upper bound of $|\eta(\cdot) - \eta_0(\cdot)|$. Such an upper bound is finite since \mathcal{X} is compact, $\eta(\cdot)$ is continuous on \mathcal{X} , and $\|\eta_0\| < \infty$. The same inequality holds when the covariates are non-random; here we can view $(\eta(x_i) - \eta_0(x_i))^2; i = 1, \dots, n$, as a set of independent realizations from some independent stochastic process. It follows that

$$\begin{aligned} & \int_{\mathcal{S}^c} P\left(\left|\frac{1}{2\sigma^2} \times \frac{1}{n} \sum_{i=1}^n (\eta(x_i) - \eta_0(x_i))^2 - \frac{1}{2\sigma^2} E_X (\eta(X) - \eta_0(X))^2\right| > \frac{\kappa_1}{4}\right) d\pi(\theta) \\ & \leq 2 \int_{\mathcal{G}_n} \exp\left\{-\frac{nC\kappa_1^2\sigma^4}{\|\eta - \eta_0\|^2}\right\} d\pi(\theta) + \pi(\mathcal{G}_n^c) \\ & = 2 \int_{\|\eta\| \leq \exp((\beta n)^{1/4})} \left[\int_{\exp(-2(\beta n)^{1/4})}^{\exp(2(\beta n)^{1/4})} \exp\left(-\frac{nC\kappa_1^2 u^{-2}}{\|\eta - \eta_0\|^2}\right) \pi(u^{-1}) u^{-2} du \right] \pi(\|\eta\|) d\|\eta\| \\ & \quad + \pi(\mathcal{G}_n^c). \end{aligned} \quad (4.A2.26)$$

Replacing $\pi(u^{-1}) u^{-2}$ with $\tilde{\pi}_n(u) = c_3 \sum_{r=1}^{M_n} \psi_{rn}^{\zeta_{rn}} \exp(-\psi_{rn} u^2) u^{2(\zeta_{rn}-1)}$, where $M_n \leq \exp((\beta n)^{1/4})$ is the number of mixture components, $c_3 > 0$, for $r = 1, \dots, M_n$, $\frac{1}{2} < \zeta_{rn} \leq$

4.A2. VERIFICATION OF THE ASSUMPTIONS OF SHALIZI FOR THE
GAUSSIAN PROCESS MODEL WITH NORMAL ERRORS

$c_4 n^q$, for $0 < q < 1/4$ and $n \geq 1$, where $c_4 > 0$, and $0 < \psi_1 \leq \psi_{rn} < c_5 < \infty$, for all r and n , and using the same techniques as before, we obtain

$$\begin{aligned} & \int_{\exp(-2(\beta n)^{1/4})}^{\exp(2(\beta n)^{1/4})} \exp\left(-\frac{nC\kappa_1^2 u^{-2}}{\|\eta - \eta_0\|^2}\right) \pi(u^{-1}) u^{-2} du \\ & \leq c_3 \times \exp\left\{3(\beta n)^{1/4} + n^q \log c_5\right\} \times \exp\left\{-\frac{C_1\kappa_1\sqrt{\psi_1 n}}{(\|\eta\| + \|\eta_0\|)}\right\}, \end{aligned} \quad (4.A2.27)$$

for some constant $C_1 > 0$. Now, using the same techniques as before, we obtain

$$\begin{aligned} & \int_{\|\eta\| \leq \exp((\beta n)^{1/4})} \exp\left\{-\frac{C_1\kappa_1\sqrt{\psi_1 n}}{(\|\eta\| + \|\eta_0\|)}\right\} \pi(\|\eta\|) d\|\eta\| \\ & = \int_{v \leq \|\eta_0\| + \exp((\beta n)^{1/4})} \exp\left(-\frac{C_1\kappa_1\sqrt{\psi_1 n}}{v}\right) \pi(v - \|\eta_0\|) dv \\ & \leq c_3 \sum_{r=1}^{M_n} c_5^{n^q} \int_{v \leq \|\eta_0\| + \exp((\beta n)^{1/4})} \exp\left\{-\left(\frac{C_1\kappa_1\sqrt{\psi_1 n}}{v} + \psi_1 v - (\zeta_{rn} - 1) \log v\right)\right\} dv \end{aligned} \quad (4.A2.28)$$

$$\leq 2c_3 \exp\left\{-\left(C_2\sqrt{\kappa_1}n^{1/4} - 2(\beta n)^{1/4} - n^q \log c_5\right)\right\}, \quad (4.A2.29)$$

with $\pi(v - \|\eta_0\|)$ replaced with the mixture as before. Here $M_n \leq \exp((\beta n)^{1/4})$, $c_3 > 0$, for $r = 1, \dots, M_n$, $0 < \zeta_{rn} \leq c_4 n^q$, for $0 < q < 1/4$ and $n \geq 1$, where $c_4 > 0$, and $0 < \psi_1 \leq \psi_{rn} < c_5 < \infty$, for all r and n . Note that the negative of the exponent of the r -th term of (4.A2.27) is minimized for $\tilde{v}_{rn} = \frac{\zeta_{rn}-1+\sqrt{(\zeta_{rn}-1)^2+4\psi_1 C_1 \kappa_1 \sqrt{\psi_1 n}}}{2\psi_1}$, and for large n it holds that $\frac{\tilde{C}_1 \sqrt{\kappa_1} n^{1/4}}{2\psi_1} \leq \tilde{v}_{rn} \leq \frac{\tilde{C}_2 \sqrt{\kappa_1} n^{1/4}}{2\psi_1}$, for some positive constants \tilde{C}_1 and \tilde{C}_2 . Also, for large n , $\tilde{v}_{rn} \psi_1 (1 - c_2) \geq (\zeta_{rn} - 1) \log \tilde{v}_{rn}$, for $0 < c_2 < 1$. Hence (4.A2.29) follows from (4.A2.28) using $\frac{C_1 \kappa_1 \sqrt{\psi_1 n}}{\tilde{v}_{rn}} + \psi_1 \tilde{v}_{rn} - (\zeta_{rn} - 1) \log \tilde{v}_{rn} \geq \psi_1 \tilde{v}_{rn} - (\zeta_{rn} - 1) \log \tilde{v}_{rn} \geq c_2 \tilde{v}_{rn} \psi_1 \geq C_2 \sqrt{\kappa_1} n^{1/4}$, for some $C_2 > 0$.

4.A2. VERIFICATION OF THE ASSUMPTIONS OF SHALIZI FOR THE
GAUSSIAN PROCESS MODEL WITH NORMAL ERRORS

Combining (4.A2.26), (4.A2.27) and (4.A2.29), we obtain

$$\begin{aligned} & \int_{\mathcal{S}^c} P \left(\left| \frac{1}{2\sigma^2} \times \frac{1}{n} \sum_{i=1}^n (\eta(x_i) - \eta_0(x_i))^2 - \frac{1}{2\sigma^2} E_X (\eta(X) - \eta_0(X))^2 \right| > \frac{\kappa_1}{4} \right) d\pi(\theta) \\ & \leq C_1 \exp \left\{ - \left(C_2 \sqrt{\kappa_1} n^{1/4} - 5 (\beta n)^{1/4} - 2n^q \log c_5 \right) \right\} + \pi(\mathcal{G}_n^c), \end{aligned} \quad (4.A2.30)$$

where C_1 and C_2 are appropriate positive constants. Since κ_1 is as large as desired, it follows that (4.A2.29) is summable.

For the fourth term, note that

$$Z_n = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \eta_0(x_i)}{\sigma_0} \right) (\eta(x_i) - \eta_0(x_i)) \sim N \left(0, \frac{1}{n^2} \sum_{i=1}^n (\eta(x_i) - \eta_0(x_i))^2 \right).$$

Then since

$$\sum_{i=1}^n (\eta(x_i) - \eta_0(x_i))^2 \leq n \left(\sup_{x \in \mathcal{X}} |\eta(x) - \eta_0(x)| \right)^2 = n \|\eta - \eta_0\|^2,$$

$$\begin{aligned} & P \left(\left| \frac{1}{\sigma^2} \times \frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i)) (\eta(x_i) - \eta_0(x_i)) \right| > \frac{\kappa_1}{4} \right) = P \left(|Z_n| > \frac{\kappa_1 \sigma^2}{4\sigma_0} \right) \\ & \leq 2 \exp \left(- \frac{C n \kappa_1^2 \sigma^4}{\sigma_0^2 \|\eta - \eta_0\|^2} \right), \end{aligned} \quad (4.A2.31)$$

for some $C > 0$. Hence, in the same way as (4.A2.30), we obtain using (4.A2.31),

$$\begin{aligned} & \int_{\mathcal{S}^c} P \left(\left| \frac{1}{\sigma^2} \times \frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i)) (\eta(x_i) - \eta_0(x_i)) \right| > \frac{\kappa_1}{4} \right) d\pi(\theta) \\ & \leq \int_{\mathcal{G}_n^c} 2 \exp \left(- \frac{C n \kappa_1^2 \sigma^4}{\sigma_0^2 \|\eta - \eta_0\|^2} \right) d\pi(\theta) + 2\pi(\mathcal{G}_n^c) \\ & \leq C_1 \exp \left\{ - \left(C_2 \sqrt{\kappa_1} n^{1/4} - 5 (\beta n)^{1/4} - 2n^q \log c_5 \right) \right\} + \pi(\mathcal{G}_n^c), \end{aligned} \quad (4.A2.32)$$

for relevant positive constants C_1, C_2, c_5 .

Combining (4.A2.17), (4.A2.19), (4.A2.21), (4.A2.23), (4.A2.24), (4.A2.26), (4.A2.30), (4.A2.32), and noting that $\sum_{n=1}^{\infty} \pi(\mathcal{G}_n^c) < \sum_{n=1}^{\infty} \alpha \exp(-\beta n) < \infty$, we obtain

$$\int_{\mathcal{S}^c} P\left(\left|\frac{1}{n} \log R_n(\theta) + h(\theta)\right| > \kappa_1\right) d\pi(\theta) < \infty.$$

4.A2.7 Verification of (S7)

For any set A such that $\pi(A) > 0$, $\mathcal{G}_n \cap A \uparrow A$. It follows from this and continuity of h that $h(\mathcal{G}_n \cap A) \downarrow h(A)$ as $n \rightarrow \infty$, so that (S7) holds.

4.A3 Verification of Shalizi's conditions for Gaussian process regression with double exponential error distribution

4.A3.1 Verification of (S1)

In this case,

$$\frac{1}{n} \log R_n(\theta) = \log\left(\frac{\sigma_0}{\sigma}\right) + \frac{1}{\sigma_0} \times \frac{1}{n} \sum_{i=1}^n |y_i - \eta_0(x_i)| - \frac{1}{\sigma} \times \frac{1}{n} \sum_{i=1}^n |y_i - \eta(x_i)|. \quad (4.A3.1)$$

4.A3. VERIFICATION OF SHALIZI'S CONDITIONS FOR GAUSSIAN PROCESS
93 REGRESSION WITH DOUBLE EXPONENTIAL ERROR DISTRIBUTION

As before, note that

$$\begin{aligned}
& \left| \frac{1}{n} \sum_{i=1}^n |y_{1i} - \eta_0(x_i)| - \frac{1}{n} \sum_{i=1}^n |y_{2i} - \eta_0(x_i)| \right| \\
& \leq \frac{1}{n} \sum_{i=1}^n \left| |y_{1i} - \eta_0(x_i)| - |y_{2i} - \eta_0(x_i)| \right| \\
& \leq \frac{1}{n} \sum_{i=1}^n |y_{1i} - y_{2i}| \\
& \leq n^{-\frac{1}{2}} \sqrt{\sum_{i=1}^n (y_{1i} - y_{2i})^2} \\
& = n^{-\frac{1}{2}} \|\mathbf{y}_{1n} - \mathbf{y}_{2n}\|,
\end{aligned}$$

from which Lipschitz continuity follows. Similarly,

$$\begin{aligned}
& \left| \frac{1}{n} \sum_{i=1}^n |y_{1i} - \eta_1(x_i)| - \frac{1}{n} \sum_{i=1}^n |y_{2i} - \eta_2(x_i)| \right| \\
& \leq \frac{1}{n} \sum_{i=1}^n |y_{1i} - \eta_1(x_i) - y_{2i} + \eta_2(x_i)| \\
& \leq \frac{1}{n} \sum_{i=1}^n [|y_{1i} - y_{2i}| + |\eta_1(x_i) - \eta_2(x_i)|] \\
& \leq n^{-\frac{1}{2}} \|\mathbf{y}_1 - \mathbf{y}_2\| + \|\eta_1 - \eta_2\|,
\end{aligned} \tag{4.A3.2}$$

which implies continuity of $\frac{1}{n} \sum_{i=1}^n |y_i - \eta(x_i)|$ with respect to \mathbf{y} and η . In other words, (4.A2.14) is continuous and hence measurable, as before. Measurability, when the covariates are considered random, also follows as before, using measurability of $\eta_0(X)$ as assumed in (A4).

4.A3.2 Verification of (S2) and proof of Lemma 9 for double-exponential errors

Now note that if $\epsilon_i = y_i - \eta_0(x_i)$ has the double exponential density of the form

$$f(\epsilon) = \frac{1}{2\sigma} \exp\left(-\frac{|\epsilon|}{\sigma}\right); \quad \epsilon \in \mathbb{R}.$$

with σ replaced with σ_0 , then

$$E_{\theta_0} |y_i - \eta_0(x_i)| = \sigma_0; \quad (4.A3.3)$$

$$\begin{aligned} E_{\theta_0} |y_i - \eta(x_i)| &= E_{\theta_0} |(y_i - \eta_0(x_i)) + (\eta_0(x_i) - \eta(x_i))| \\ &= |\eta_0(x_i) - \eta(x_i)| + \sigma_0 \exp\left(-\frac{|\eta_0(x_i) - \eta(x_i)|}{\sigma_0}\right). \end{aligned} \quad (4.A3.4)$$

It follows from (4.A3.3), (4.A3.4) and (A3), that

$$\frac{1}{n} \sum_{i=1}^n E_{\theta_0} |y_i - \eta_0(x_i)| = \sigma_0; \quad (4.A3.5)$$

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n E_{\theta_0} |y_i - \eta(x_i)| &= \frac{1}{n} \sum_{i=1}^n \left[|\eta(x_i) - \eta_0(x_i)| + \sigma_0 \exp\left(-\frac{|\eta(x_i) - \eta_0(x_i)|}{\sigma_0}\right) \right] \\ &\rightarrow E_X |\eta(X) - \eta_0(X)| + \sigma_0 E_X \left[\exp\left(-\frac{|\eta(X) - \eta_0(X)|}{\sigma_0}\right) \right], \text{ as } n \rightarrow \infty. \end{aligned} \quad (4.A3.6)$$

Using (4.A3.5) and (4.A3.6) we see that as $n \rightarrow \infty$,

$$\begin{aligned} \frac{1}{n} E_{\theta_0} [\log R_n(\theta)] &= \log\left(\frac{\sigma_0}{\sigma}\right) + \frac{1}{\sigma_0} \times \frac{1}{n} \sum_{i=1}^n E_{\theta_0} |y_i - \eta_0(x_i)| - \frac{1}{\sigma} \times \frac{1}{n} \sum_{i=1}^n E_{\theta_0} |y_i - \eta(x_i)| \\ &\rightarrow \log\left(\frac{\sigma_0}{\sigma}\right) + 1 - \frac{1}{\sigma} E_X |\eta(X) - \eta_0(X)| - \frac{\sigma_0}{\sigma} E_X \left[\exp\left(-\frac{|\eta(X) - \eta_0(X)|}{\sigma_0}\right) \right], \\ &= -h(\theta), \end{aligned} \quad (4.A3.7)$$

where

$$h(\theta) = \log\left(\frac{\sigma}{\sigma_0}\right) - 1 + \frac{1}{\sigma} E_X |\eta(X) - \eta_0(X)| + \frac{\sigma_0}{\sigma} E_X \left[\exp\left(-\frac{|\eta(X) - \eta_0(X)|}{\sigma_0}\right) \right].$$

As in the case of Gaussian errors, the results remain the same if the covariates are assumed to be random.

4.A3.3 Verification of (S3) and proof of Theorem 10 for double exponential errors

We now show that for all $\theta \in \Theta$, $\lim_{n \rightarrow \infty} \frac{1}{n} \log R_n(\theta) = -h(\theta)$, almost surely. First note that

$$\begin{aligned} \left| \frac{1}{n} R_n(\theta) + h(\theta) \right| &\leq \left| \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \eta_0(x_i)|}{\sigma_0} - 1 \right| \\ &+ \left| \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \eta(x_i)|}{\sigma} - \frac{1}{\sigma} E_X |\eta(X) - \eta_0(X)| - \frac{\sigma_0}{\sigma} E_X \left[\exp\left(-\frac{|\eta(X) - \eta_0(X)|}{\sigma_0}\right) \right] \right|. \end{aligned} \tag{4.A3.8}$$

Since $\frac{|y_i - \eta_0(x_i)|}{\sigma_0}$ has the exponential distribution with mean one, the term $\left| \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \eta_0(x_i)|}{\sigma_0} - 1 \right| \rightarrow 0$ almost surely as $n \rightarrow \infty$ by the strong law of large numbers. That the term (4.A3.8) also tends to zero almost surely as $n \rightarrow \infty$ can be shown using the Borel-Cantelli lemma, using the inequality (4.A3.19), and replacing κ_1 in that inequality with any $\delta_1 > 0$. In other words, it holds that for all $\theta \in \Theta$, $\lim_{n \rightarrow \infty} \frac{1}{n} \log R_n(\theta) = -h(\theta)$, almost surely. Also, it follows from (4.A3.1), (4.A3.2), (4.2.8), Lipschitz continuity of $x \mapsto \exp(-|x|)$, boundedness of the first derivative with respect to σ , that $\frac{1}{n} \log R_n(\theta) + h(\theta)$ is Lipschitz on $\theta \in \mathcal{G}_n \setminus I = \mathcal{G}_n$, which is compact. As a result, it follows that $\frac{1}{n} \log R_n(\theta) + h(\theta)$ is stochastically equicontinuous in $\mathcal{G} \in \{\mathcal{G}_1, \mathcal{G}_2, \dots\}$. Hence, the convergence $\lim_{n \rightarrow \infty} \frac{1}{n} \log R_n(\theta) = -h(\theta)$ occurs uniformly for $\theta \in \mathcal{G}$, almost surely.

4.A3.4 Verification of (S4)

Note that $h(\theta) \leq \log\left(\frac{\sigma}{\sigma_0}\right) - 1 + \frac{\|\eta - \eta_0\| + \sigma_0}{\sigma}$. Now $0 < \|\eta - \eta_0\| < \infty$ and $0 < \sigma < \infty$ with prior probability one. Consequently, it follows that $h(\theta) < \infty$ with probability one, so that $I = \emptyset$ and hence, $\mathcal{G}_n \setminus I = \mathcal{G}_n$.

4.A3.5 Verification of (S5)

Verification of (S5) (1) and (S5) (2) remains the same as for Gaussian noise. (S5) (3) follows in the same way as for Gaussian noise if we can show that $h(\theta)$ is continuous in θ . To see that $h(\theta)$ is continuous in θ , again assume that $\eta_j \rightarrow \tilde{\eta}$ as $j \rightarrow \infty$ in the sense that $\|\eta_j - \tilde{\eta}\| \rightarrow 0$ as $j \rightarrow \infty$. Then $|E_X |\eta_j(X) - \eta_0(X)| - E_X |\tilde{\eta}(X) - \eta_0(X)|| \leq E_X |\eta_j(X) - \tilde{\eta}(X)| \leq \|\eta_j - \tilde{\eta}\| \rightarrow 0$ as $j \rightarrow \infty$. Also,

$$\begin{aligned} & \left| E_X \left[\exp \left(-\frac{|\eta_j(X) - \eta_0(X)|}{\sigma_0} \right) \right] - E_X \left[\exp \left(-\frac{|\tilde{\eta}(X) - \eta_0(X)|}{\sigma_0} \right) \right] \right| \\ & \leq E_X \left[\exp(-|\tilde{\eta}(X) - \eta_0(X)|) \times \left| \exp \left(-\frac{(|\eta_j(X) - \eta_0(X)| - |\tilde{\eta}(X) - \eta_0(X)|)}{\sigma_0} \right) - 1 \right| \right] \\ & \leq E_X \left[\exp(-|\tilde{\eta}(X) - \eta_0(X)|) \times \left| \exp \left(\frac{|\eta_j(X) - \tilde{\eta}(X)|}{\sigma_0} \right) - 1 \right| \right] \\ & \leq \left| \exp \left(\frac{\|\eta_j - \tilde{\eta}\|}{\sigma_0} \right) - 1 \right| \times E_X [\exp(-|\tilde{\eta}(X) - \eta_0(X)|)] \\ & \rightarrow 0, \text{ as } j \rightarrow \infty. \end{aligned}$$

Continuity of $h(\theta)$ hence follows easily.

4.A3.6 Verification of (S6) and proof of Theorem 11 for double exponential errors

It follows from (4.A3.8) that for all $\theta \in \Theta$, for $\kappa_1 = \kappa - h(\Theta)$, we have

$$\begin{aligned} P\left(\left|\frac{1}{n} \log R_n(\theta) + h(\theta)\right| > \kappa_1\right) &\leq P\left(\left|\frac{1}{\sigma_0} \times \frac{1}{n} \sum_{i=1}^n |y_i - \eta_0(x_i)| - 1\right| > \frac{\kappa_1}{2}\right) \\ &+ P\left(\left|\frac{1}{\sigma_0} \times \frac{1}{n} \sum_{i=1}^n |y_i - \eta(x_i)| - \frac{1}{\sigma_0} E_X |\eta(X) - \eta_0(X)|\right.\right. \\ &\quad \left.\left.- \frac{\sigma_0}{\sigma} E_X \left(\exp\left\{-\frac{|\eta(X) - \eta_0(X)|}{\sigma_0}\right\}\right)\right| > \frac{\kappa_1}{2}\right). \end{aligned} \quad (4.A3.9)$$

Since $\frac{|y_i - \eta_0(x_i)|}{\sigma_0}$ are exponential random variables with expectation one, it follows that $\frac{|y_i - \eta_0(x_i)|}{\sigma_0} - 1$ are zero-mean, independent sub-exponential random variables with some parameter $s > 0$. Hence, by Bernstein's inequality ([Uspensky \(1937\)](#), [Bennett \(1962\)](#), [Massart \(2003\)](#)),

$$P\left(\left|\frac{1}{\sigma_0} \times \frac{1}{n} \sum_{i=1}^n |y_i - \eta_0(x_i)| - 1\right| > \frac{\kappa_1}{2}\right) \leq 2 \exp\left(-\frac{n}{2} \min\left\{\frac{\kappa_1^2}{4s^2}, \frac{\kappa_1}{2s}\right\}\right).$$

Hence,

$$\int_{S^c} P\left(\left|\frac{1}{\sigma_0} \times \frac{1}{n} \sum_{i=1}^n |y_i - \eta_0(x_i)| - 1\right| > \frac{\kappa_1}{2}\right) \leq 2 \exp\left(-\frac{n}{2} \min\left\{\frac{\kappa_1^2}{4s^2}, \frac{\kappa_1}{2s}\right\}\right). \quad (4.A3.10)$$

Let $\bar{\varphi} = E_X |\eta(X) - \eta_0(X)| + \sigma_0 E_X \left(\exp\left\{-\frac{|\eta(X) - \eta_0(X)|}{\sigma_0}\right\}\right)$. Also, letting $\varphi(x) = |\eta(x) - \eta_0(x)| + \sigma_0 \left(\exp\left\{-\frac{|\eta(x) - \eta_0(x)|}{\sigma_0}\right\}\right)$, note that

$$\frac{1}{n} \sum_{i=1}^n \varphi(x_i) \rightarrow \bar{\varphi}, \text{ as } n \rightarrow \infty. \quad (4.A3.11)$$

With this, the second term of (4.A3.9) can be bounded as follows:

$$\begin{aligned}
 & P \left(\left| \frac{1}{\sigma} \times \frac{1}{n} \sum_{i=1}^n |y_i - \eta(x_i)| - \bar{\varphi} \right| > \frac{\kappa_1}{2} \right) \\
 &= P \left(\sigma^{-1} \left| \frac{1}{n} \sum_{i=1}^n \{|y_i - \eta(x_i)| - \varphi(x_i)\} + \frac{1}{n} \sum_{i=1}^n \varphi(x_i) - \bar{\varphi} \right| > \frac{\kappa_1}{2} \right) \\
 &\leq P \left(\sigma^{-1} \left| \frac{1}{n} \sum_{i=1}^n \{|y_i - \eta(x_i)| - \varphi(x_i)\} \right| > \frac{\kappa_1}{4} \right) + P \left(\sigma^{-1} \left| \frac{1}{n} \sum_{i=1}^n \varphi(x_i) - \bar{\varphi} \right| > \frac{\kappa_1}{4} \right).
 \end{aligned} \tag{4.A3.12}$$

In the case of random or non-random covariates X , again by Hoeffding's inequality,

$$P \left(\sigma^{-1} \left| \frac{1}{n} \sum_{i=1}^n \varphi(x_i) - \bar{\varphi} \right| > \frac{\kappa_1}{4} \right) \leq \exp \left\{ - \frac{nC\kappa_1^2\sigma^2}{(\|\eta\| + c_0)^2} \right\}, \tag{4.A3.13}$$

where $C > 0$ is a constant. Note that $(\|\eta\| + c_0)$, with $c_0 = \|\eta_0\| + \sigma_0$, is an upper bound of $|\varphi(\cdot)|$. Again, such an upper bound exists since \mathcal{X} is compact and $\eta(\cdot)$ is continuous on \mathcal{X} . Application of the same method as proving (4.A2.24) and (4.A2.30) yields

$$\begin{aligned}
 & \int_{\mathcal{S}^c} P \left(\sigma^{-1} \left| \frac{1}{n} \sum_{i=1}^n \varphi(x_i) - \bar{\varphi} \right| > \frac{\kappa_1}{4} \right) \pi(\theta) d\theta \\
 &\leq C_1 \exp \left\{ - \left(C_2 \sqrt{\kappa_1} n^{1/4} - 5(\beta n)^{1/4} - 2n^q \log c_5 \right) \right\} + \pi(\mathcal{G}_n^c),
 \end{aligned} \tag{4.A3.14}$$

where as before κ_1 is large enough to make the exponent of (4.A3.14) negative.

For the first term of (4.A3.12), let us first prove that $|y_i - \eta(x_i)| - \varphi(x_i)$ are sub-exponential random variables. Then we can apply Bernstein's inequality to directly bound the term. We need to show that $E_{\theta_0} [\exp \{t(|y_i - \eta(x_i)|) - \varphi(x_i)\}] \leq \exp \left(\frac{t^2 s^2}{2} \right)$ for $|t| \leq s^{-1}$, for some $s > 0$.

4.A3.7 Case 1: $t \geq 0$, $\eta(x_i) - \eta_0(x_i) > 0$

Direct calculation shows that

$$\begin{aligned} & E_{\theta_0} [\exp \{t(|y_i - \eta(x_i)|) - \varphi(x_i)\}] \\ &= \exp(-t\varphi(x_i)) \times \frac{\exp \{(\eta(x_i) - \eta_0(x_i))t\} - \exp \left(\frac{\eta_0(x_i) - \eta(x_i)}{\sigma_0} \right)}{1 - \sigma_0^2 t^2} \\ &\leq \frac{\exp \{t(\varphi(x_i) + \eta(x_i) - \eta_0(x_i))\}}{1 - \sigma_0^2 t^2} \\ &\leq \frac{\exp \{t(2\|\eta - \eta_0\| + \sigma_0)\}}{1 - \sigma_0^2 t^2}. \end{aligned} \quad (4.A3.15)$$

To show that (4.A3.15) is bounded above by $\exp(t^2 s^2 / 2)$, we need to show that

$$f(t) = \frac{t^2 s^2}{2} - 2(\|\eta - \eta_0\| + \sigma_0)t + \log(1 - \sigma_0^2 t^2) \geq 0. \quad (4.A3.16)$$

For $t > 0$, it is sufficient to show that

$$\frac{ts^2}{2} \geq 2(\|\eta - \eta_0\| + \sigma_0) - \frac{\log(1 - \sigma_0^2 t^2)}{t}. \quad (4.A3.17)$$

Now, $-\frac{\log(1 - \sigma_0^2 t^2)}{t} \rightarrow 0$, as $t \rightarrow 0$. Hence, for any $\epsilon > 0$, there exists $\delta(\epsilon) > 0$ such that $t \leq \delta(\epsilon)$ implies $-\frac{\log(1 - \sigma_0^2 t^2)}{t} < \epsilon$. Let $s \geq \frac{C_1\|\eta - \eta_0\| + C_2}{\delta(\epsilon)}$, where $C_1 > 0$ and $C_2 > 0$ are sufficiently large quantities. Hence, if $\delta(\epsilon)^2 \leq t \leq \delta(\epsilon)$, then (4.A3.17), and hence (4.A3.16), is satisfied. Now, $f(t)$ given by (4.A3.16) is continuous in t and $f(0) = 0$. Hence, (4.A3.16) holds even for $0 \leq t \leq \delta(\epsilon)^2$. In other words,

$$E_{\theta_0} [\exp \{t(|y_i - \eta(x_i)|) - \varphi\}] \leq \exp \left(\frac{t^2 s^2}{2} \right), \text{ for } 0 \leq t \leq s^{-1} \leq \frac{\delta(\epsilon)}{C_1\|\eta - \eta_0\| + C_2} \leq \delta(\epsilon). \quad (4.A3.18)$$

4.A3.8 Case 2: $t \geq 0, \eta(x_i) - \eta_0(x_i) < 0$

In this case,

$$\begin{aligned}
 & E_{\theta_0} [\exp \{t(|y_i - \eta(x_i)|) - \varphi(x_i)\}] \\
 &= \exp(-t\varphi(x_i)) \times \frac{\exp \{(\eta_0(x_i) - \eta(x_i))t\} + \sigma_0 t \exp \left(\frac{\eta(x_i) - \eta_0(x_i)}{\sigma_0} \right)}{1 - \sigma_0^2 t^2} \\
 &\leq \exp(t\varphi(x_i)) \times \frac{2 \exp \{(\eta_0(x_i) - \eta(x_i))t\}}{1 - \sigma_0^2 t^2} \\
 &\leq \frac{\exp \{t(\varphi(x_i) + (\eta_0(x_i) - \eta(x_i)))\}}{\frac{1 - \sigma_0^2 t^2}{2}} \\
 &\leq \frac{\exp \{t(2\|\eta - \eta_0\| + \sigma_0)\}}{\frac{1 - \sigma_0^2 t^2}{2}}.
 \end{aligned}$$

As in Section 4.A3.7 it can be seen that (4.A3.18) holds.

4.A3.9 Case 3: $t \leq 0, \eta(x_i) - \eta_0(x_i) > 0$

Here

$$\begin{aligned}
 & E_{\theta_0} [\exp \{t(|y_i - \eta(x_i)|) - \varphi(x_i)\}] \\
 &= \exp(-t\varphi(x_i)) \times \frac{\exp \{(\eta(x_i) - \eta_0(x_i))t\} - \sigma_0 |t| \exp \left(\frac{\eta_0(x_i) - \eta(x_i)}{\sigma_0} \right)}{1 - \sigma_0^2 t^2} \\
 &\leq \exp(-t\varphi(x_i)) \times \frac{1}{1 - \sigma_0^2 t^2} \\
 &\leq \frac{\exp \{-t(\|\eta - \eta_0\| + \sigma_0)\}}{1 - \sigma_0^2 t^2}.
 \end{aligned}$$

Here we need to have $|t| \left[\frac{|t|s^2}{2} - (\|\eta - \eta_0\| + \sigma_0) + \frac{\log(1 - \sigma_0^2 t^2)}{|t|} \right] > 0$. In the same way as before it follows that

$$E_{\theta_0} [\exp \{t(|y_i - \eta(x_i)|) - \varphi\}] \leq \exp \left(\frac{t^2 s^2}{2} \right), \text{ for } 0 \leq |t| \leq s^{-1} \leq \frac{\delta(\epsilon)}{C_1 \|\eta - \eta_0\| + C_2} \leq \delta(\epsilon).$$

4.A3.10 Case 4: $t \leq 0$, $\eta(x_i) - \eta_0(x_i) < 0$

In this case,

$$\begin{aligned} & E_{\theta_0} [\exp \{t (|y_i - \eta(x_i)|) - \varphi(x_i)\}] \\ &= \exp(-t\varphi(x_i)) \times \frac{\exp \{(\eta_0(x_i) - \eta(x_i))t\} - \sigma_0|t| \exp \left(\frac{\eta(x_i) - \eta_0(x_i)}{\sigma_0} \right)}{1 - \sigma_0^2 t^2} \\ &\leq \exp(-t\varphi(x_i)) \times \frac{1}{1 - \sigma_0^2 t^2} \\ &\leq \frac{\exp \{-t (\|\eta - \eta_0\| + \sigma_0)\}}{1 - \sigma_0^2 t^2}. \end{aligned}$$

Hence, (4.A3.19) holds.

Hence, for $i = 1, \dots, n$, $|y_i - \eta(x_i)| - E(|y_i - \eta(x_i)|)$ are zero-mean, independent sub-exponential random variables with parameter s . In particular, we can set $s = \frac{C_1 \|\eta - \eta_0\| + C_2}{\delta(\epsilon)}$.

Hence, by Bernstein's inequality,

$$\begin{aligned} & P \left(\sigma^{-1} \left| \frac{1}{n} \sum_{i=1}^n \{|y_i - \eta(x_i)| - \varphi(x_i)\} \right| > \frac{\kappa_1}{4} \right) \\ &\leq 2 \max \left\{ P \left(\frac{\sigma^{-1}}{n} \sum_{i=1}^n \{|y_i - \eta(x_i)| - \varphi(x_i)\} > \frac{\kappa_1}{4} \right), P \left(\frac{\sigma^{-1}}{n} \sum_{i=1}^n \{|y_i - \eta(x_i)| - \varphi(x_i)\} < -\frac{\kappa_1}{4} \right) \right\} \\ &\leq 2 \exp \left(-\frac{n}{2} \min \left\{ \frac{\kappa_1^2 \sigma^2}{16s^2}, \frac{\kappa_1 \sigma}{4s} \right\} \right) \\ &= 2 \exp \left(-\frac{n}{2} \min \left\{ \frac{\kappa_1^2 \delta(\epsilon)^2 \sigma^2}{16(C_1 \|\eta - \eta_0\| + C_2)^2}, \frac{\kappa_1 \delta(\epsilon) \sigma}{4(C_1 \|\eta - \eta_0\| + C_2)} \right\} \right). \end{aligned} \tag{4.A3.19}$$

Hence,

$$\begin{aligned} & \int_{\mathcal{S}^c} P \left(\sigma^{-1} \left| \frac{1}{n} \sum_{i=1}^n \{|y_i - \eta(x_i)| - \varphi(x_i)\} \right| > \frac{\kappa_1}{4} \right) d\pi(\theta) \\ & \leq \int_{\mathcal{G}_n} 2 \exp \left(-\frac{n\kappa_1^2 \delta(\epsilon)^2 \sigma^2}{32(C_1\|\eta - \eta_0\| + C_2)^2} \right) d\pi(\theta) + \int_{\mathcal{G}_n} 2 \exp \left(-\frac{n\kappa_1 \delta(\epsilon) \sigma}{8(C_1\|\eta - \eta_0\| + C_2)} \right) \\ & \quad (4.A3.20) \end{aligned}$$

$$+ 2\pi(\mathcal{G}_n^c). \quad (4.A3.21)$$

Applying the same techniques as proving (4.A2.30) we obtain

$$\int_{\mathcal{G}_n} 2 \exp \left(-\frac{n\kappa_1^2 \delta(\epsilon)^2 \sigma^2}{32(C_1\|\eta - \eta_0\| + C_2)^2} \right) d\pi(\theta) \leq C_1 \exp \left\{ - \left(C_2 \sqrt{\kappa_1} n^{1/4} - 5(\beta n)^{1/4} - 2n^q \log c_5 \right) \right\}, \quad (4.A3.22)$$

for appropriate positive constants C_1, C_2, c_5 .

For the second integral of (4.A3.20), observe that for appropriate positive constant c_0 and C ,

$$\begin{aligned} & \int_{\mathcal{G}_n} 2 \exp \left(-\frac{n\kappa_1 \delta(\epsilon) \sigma}{8(C_1\|\eta - \eta_0\| + C_2)} \right) \\ & \leq 2 \int_{\|\eta\| \leq \exp((\beta n)^{1/4})} \left[\int_{\exp(-2(\beta n)^{1/4})}^{\exp(-2(\beta n)^{1/4})} \exp \left(-\frac{C\kappa_1 n u^{-1}}{\|\eta\| + c_0} \right) 2\pi(u^{-2}) u^{-3} du \right] \pi(\|\eta\|) d\|\eta\| \\ & \quad (4.A3.23) \end{aligned}$$

Replacing $2\pi(u^{-2})u^{-3}$ with the mixture form as before with $0 < \zeta_{rn} < c_5 n^q$, where $0 < q < 1/4$, and the rest remaining the same as before, we obtain

$$\int_{\exp(-2(\beta n)^{1/4})}^{\exp(-2(\beta n)^{1/4})} \exp \left(-\frac{C\kappa_1 n u^{-1}}{\|\eta\| + c_0} \right) 2\pi(u^{-2}) u^{-3} du \leq \exp \left(-\frac{C_1 \kappa_1 \sqrt{n}}{\sqrt{\|\eta\|} + c_0} \right), \quad (4.A3.24)$$

for some appropriate positive constant C_1 .

Now we obtain

$$\begin{aligned}
 & \int_{\|\eta\| \leq \exp((\beta n)^{1/4})} \exp\left(-\frac{C_1 \kappa_1 \sqrt{n}}{\sqrt{\|\eta\| + c_0}}\right) \pi(\|\eta\|) d\|\eta\| \\
 &= \int_{0 \leq v \leq \sqrt{c_0 + \exp((\beta n)^{1/4})}} \exp\left(-\frac{C_2 \kappa_1 \sqrt{n}}{v}\right) \pi(v^2 - c_0) 2vdv \\
 &\leq \tilde{C}_1 \exp\left\{-\left(\tilde{C}_2 \sqrt{\kappa_1} n^{1/4} - \frac{9}{2} (\beta n)^{1/4} - 2n^q \log c_5\right)\right\}, \tag{4.A3.25}
 \end{aligned}$$

for appropriate positive constants \tilde{C}_1 and \tilde{C}_2 . From (4.A3.24) and (4.A3.25) it follows that (4.A3.25) is an upper bound for (4.A3.23). Combining this with (4.A3.19), (4.A3.20), (4.A3.21) and (4.A3.22), we obtain

$$\begin{aligned}
 & P\left(\sigma^{-1} \left| \frac{1}{n} \sum_{i=1}^n \{|y_i - \eta(x_i)| - \varphi(x_i)\} \right| > \frac{\kappa_1}{4}\right) \\
 &\leq C_1 \exp\left\{-\left(C_2 \sqrt{\kappa_1} n^{1/4} - 5(\beta n)^{1/4} - 2n^q \log c_5\right)\right\} \\
 &\quad + \tilde{C}_1 \exp\left\{-\left(\tilde{C}_2 \sqrt{\kappa_1} n^{1/4} - \frac{9}{2} (\beta n)^{1/4} - 2n^q \log c_5\right)\right\} + 2\pi(\mathcal{G}_n^c). \tag{4.A3.26}
 \end{aligned}$$

Gathering (4.A3.10), (4.A3.14) and (4.A3.26) we see that

$$\sum_{n=1}^{\infty} \int_{\mathcal{S}^c} P\left(\left|\frac{1}{n} \log R_n(\theta) + h(\theta)\right| > \delta\right) \pi(\theta) d\theta < \infty. \tag{4.A3.27}$$

4.A3.11 Verification of (S7)

Verification of (S7) is exactly the same as for Gaussian errors.

4.A4 Verification of the assumptions of Shalizi for the general stochastic process model

Note that

$$f_\theta(\mathbf{y}_n) = \frac{1}{\sigma^n} \prod_{i=1}^n \phi(y_i - \eta(x_i)); \quad (4.A4.1)$$

$$f_{\theta_0}(\mathbf{y}_n) = \frac{1}{\sigma_0^n} \prod_{i=1}^n \phi(y_i - \eta_0(x_i)). \quad (4.A4.2)$$

4.A4.1 Verification of (S1)

The equations (4.A4.1) and (4.A4.2) yield, in our case,

$$\frac{1}{n} \log R_n(\theta) = \log \left(\frac{\sigma_0}{\sigma} \right) + \frac{1}{n} \sum_{i=1}^n \log \phi \left(\frac{y_i - \eta_0(x_i)}{\sigma_0} \right) - \frac{1}{n} \sum_{i=1}^n \log \phi \left(\frac{y_i - \eta(x_i)}{\sigma} \right). \quad (4.A4.3)$$

We show that the right hand side of (4.A4.3), which we denote as $f(\mathbf{y}_n, \theta)$, is continuous in (\mathbf{y}_n, θ) , which is sufficient to confirm measurability of $R_n(\theta)$. Let $\|(\mathbf{y}_n, \theta)\| = \|\mathbf{y}_n\| + \|\theta\|$, where $\|\mathbf{y}_n\|$ is the Euclidean norm and $\|\theta\| = \|\eta\| + |\sigma|$, with $\|\eta\| = \sup_{x \in \mathcal{X}} |\eta(x)|$. Since \mathcal{X} is compact and η is almost surely continuous, it follows that $\|\eta\| < \infty$ almost surely.

Consider $\mathbf{y}_{1n} = (y_{11}, y_{12}, \dots, y_{1n})^T$, $\mathbf{y}_{2n} = (y_{21}, y_{22}, \dots, y_{2n})^T$, θ_1 and θ_2 . Using the Lipschitz condition of (A7), we obtain

$$\left| \frac{1}{n} \sum_{i=1}^n \log \phi \left(\frac{y_{1i} - \eta_0(x_i)}{\sigma_0} \right) - \frac{1}{n} \sum_{i=1}^n \log \phi \left(\frac{y_{2i} - \eta_0(x_i)}{\sigma_0} \right) \right| \quad (4.A4.4)$$

$$\leq \frac{L}{n\sigma_0} \sum_{i=1}^n |y_{1i} - y_{2i}| \leq \frac{L}{n\sigma_0} \|\mathbf{y}_{1n} - \mathbf{y}_{2n}\|. \quad (4.A4.5)$$

Hence, the term $\frac{1}{n} \sum_{i=1}^n \log \phi \left(\frac{y_i - \eta_0(x_i)}{\sigma_0} \right)$ is Lipschitz continuous.

To prove continuity of the term $\frac{1}{n} \sum_{i=1}^n \log \phi \left(\frac{y_i - \eta(x_i)}{\sigma} \right)$, we first recall from (A7) that $\log \phi(x) = \log \phi(|x|)$ is Lipschitz continuous in x . Hence, if we can show that for

**4.A4. VERIFICATION OF THE ASSUMPTIONS OF SHALIZI FOR THE
GENERAL STOCHASTIC PROCESS MODEL**

each $i = 1, \dots, n$, $\frac{y_i - \eta(x_i)}{\sigma}$ is continuous in (\mathbf{y}_n, θ) , then this would prove continuity of $\frac{1}{n} \sum_{i=1}^n \log \phi \left(\frac{y_i - \eta(x_i)}{\sigma} \right)$ since sum and composition of continuous functions are continuous. Now, $|y_{1i} - \eta_1(x_i) - (y_{2i} - \eta_2(x_i))| \leq |y_{1i} - y_{2i}| + |\eta_1(x_i) - \eta_2(x_i)| \leq \|\mathbf{y}_{1n} - \mathbf{y}_{2n}\| + \|\eta_1 - \eta_2\|$, showing continuity of $y_i - \eta(x_i)$. Division of this term by $\sigma (> 0)$, preserves continuity.

Hence, $f(\mathbf{y}_n, \theta)$ is continuous with respect to (\mathbf{y}_n, θ) , so that (S1) holds in our case.

4.A4.2 Verification of (S2) and proof of Lemma 17

It follows from (4.A4.1) and (4.A4.2), that

$$E_{\theta_0} \left[\frac{1}{n} \log \frac{f_{\theta_0}(\mathbf{y}_n)}{f_{\theta}(\mathbf{y}_n)} \right] = \log \left(\frac{\sigma}{\sigma_0} \right) + \frac{1}{n} \sum_{i=1}^n E_{\theta_0} \left[\log \phi \left(\frac{y_i - \eta_0(x_i)}{\sigma_0} \right) \right] - \frac{1}{n} \sum_{i=1}^n E_{\theta_0} \left[\log \phi \left(\frac{y_i - \eta(x_i)}{\sigma} \right) \right]. \quad (4.A4.6)$$

Now $E_{\theta_0} \left[\log \phi \left(\frac{y_i - \eta_0(x_i)}{\sigma_0} \right) \right] = \int_{-\infty}^{\infty} [\log \phi(z)] \phi(z) dz = c$ (say), so that for any $n \geq 1$,

$$\frac{1}{n} \sum_{i=1}^n E_{\theta_0} \left[\log \phi \left(\frac{y_i - \eta_0(x_i)}{\sigma_0} \right) \right] = c. \quad (4.A4.7)$$

Now for any $x \in \mathcal{X}$, let

$$g_{\eta, \sigma}(x) = E_{\theta_0} \left[\log \phi \left(\frac{y - \eta(x)}{\sigma} \right) \right] = \int_{-\infty}^{\infty} \log \phi \left(\frac{\sigma_0 z + \eta_0(x) - \eta(x)}{\sigma} \right) \phi(z) dz. \quad (4.A4.8)$$

Let us first investigate continuity of $g_{\eta, \sigma}(x)$ with respect to x . To this end, observe that for $x_1, x_2 \in \mathcal{X}$, the following hold thanks to Lipschitz continuity of $\log \phi$:

$$\begin{aligned} & |g_{\eta, \sigma}(x_1) - g_{\eta, \sigma}(x_2)| \\ & \leq \int_{-\infty}^{\infty} \left| \log \phi \left(\frac{\sigma_0 z + \eta_0(x_1) - \eta(x_1)}{\sigma} \right) - \log \phi \left(\frac{\sigma_0 z + \eta_0(x_2) - \eta(x_2)}{\sigma} \right) \right| \phi(z) dz \\ & = \frac{L}{\sigma} \int_{-\infty}^{\infty} |(\eta_0(x_1) - \eta_0(x_2)) - (\eta(x_1) - \eta(x_2))| \phi(z) dz \\ & \leq \frac{L}{\sigma} (|\eta_0(x_1) - \eta_0(x_2)| + |\eta(x_1) - \eta(x_2)|). \end{aligned} \quad (4.A4.9)$$

**4.A4. VERIFICATION OF THE ASSUMPTIONS OF SHALIZI FOR THE
GENERAL STOCHASTIC PROCESS MODEL**

In our model, $\eta(x)$ is continuous in x , but $\eta_0(x)$ need not be so. If $\eta_0(x)$ is allowed to be continuous, then by (4.A4.9), $g_{\eta,\sigma}(x)$ is continuous in x . If $\eta_0(x)$ has at most countably many discontinuities, then $g_{\eta,\sigma}(x)$ is continuous everywhere on \mathcal{X} except perhaps at a countable number of points. In both the cases, $g_{\eta,\sigma}(x)$ is Riemann integrable when the covariates are considered deterministic. In that case,

$$\frac{1}{n} \sum_{i=1}^n E_{\theta_0} \left[\log \phi \left(\frac{y_i - \eta(x_i)}{\sigma} \right) \right] = \frac{1}{n} \sum_{i=1}^n g_{\eta,\sigma}(x_i) \rightarrow \int_{\mathcal{X}} g_{\eta,\sigma}(x) dx, \text{ as } n \rightarrow \infty. \quad (4.A4.10)$$

If $\{x_i : i = 1, 2, \dots\}$ is considered to be an *iid* realization from Q , then by the ergodic theorem

$$\frac{1}{n} \sum_{i=1}^n E_{\theta_0} \left[\log \phi \left(\frac{y_i - \eta(x_i)}{\sigma} \right) \right] = \frac{1}{n} \sum_{i=1}^n g_{\eta,\sigma}(x_i) \rightarrow \int_{\mathcal{X}} g_{\eta,\sigma}(x) dQ, \text{ as } n \rightarrow \infty. \quad (4.A4.11)$$

We denote both $\int_{\mathcal{X}} g_{\eta,\sigma}(x) dx$ and $\int_{\mathcal{X}} g_{\eta,\sigma}(x) dQ$ by $E_X [g_{\eta,\sigma}(X)]$. Note that both the integrals exist thanks to continuity of $g_{\eta,\sigma}(x)$ and compactness of \mathcal{X} . Combining (4.A4.7), (4.A4.10) and (4.A4.11) we obtain

$$E_{\theta_0} \left[\frac{1}{n} \log \frac{f_{\theta_0}(\mathbf{y}_n)}{f_{\theta}(\mathbf{y}_n)} \right] \rightarrow h(\theta), \quad (4.A4.12)$$

where $h(\theta)$ is given by (4.3.8). In other words, (S2) holds.

4.A4.3 Verification of (S3) and proof of Theorem 18

For any $\delta > 0$, and for any $\theta \in \Theta$,

$$\begin{aligned} P\left(\left|\frac{1}{n} \log R_n(\theta) + h(\theta)\right| > \delta\right) \\ = P\left(\left|\frac{1}{n} \sum_{i=1}^n \log \phi\left(\frac{y_i - \eta(x_i)}{\sigma}\right) - \frac{1}{n} \sum_{i=1}^n \log \phi\left(\frac{y_i - \eta_0(x_i)}{\sigma_0}\right) + c - E_X[g_{\eta,\sigma}(X)]\right| > \delta\right) \\ \leq P\left(\left|\frac{1}{n} \sum_{i=1}^n \log \phi\left(\frac{y_i - \eta(x_i)}{\sigma}\right) - E_X[g_{\eta,\sigma}(X)]\right| > \frac{\delta}{2}\right) \end{aligned} \quad (4.A4.13)$$

$$+ P\left(\left|\frac{1}{n} \sum_{i=1}^n \log \phi\left(\frac{y_i - \eta_0(x_i)}{\sigma_0}\right) - c\right| > \frac{\delta}{2}\right). \quad (4.A4.14)$$

Let us focus attention on the probability given by (4.A4.13).

$$\begin{aligned} P\left(\left|\frac{1}{n} \sum_{i=1}^n \log \phi\left(\frac{y_i - \eta(x_i)}{\sigma}\right) - E_X[g_{\eta,\sigma}(X)]\right| > \frac{\delta}{2}\right) \\ = P\left(\left|\frac{1}{n} \sum_{i=1}^n \left[\log \phi\left(\frac{y_i - \eta(x_i)}{\sigma}\right) - g_{\eta,\sigma}(x_i)\right] + \left[\frac{1}{n} \sum_{i=1}^n g_{\eta,\sigma}(x_i) - E_X[g_{\eta,\sigma}(X)]\right]\right| > \frac{\delta}{2}\right) \\ \leq P\left(\left|\frac{1}{n} \sum_{i=1}^n \left[\log \phi\left(\frac{y_i - \eta(x_i)}{\sigma}\right) - g_{\eta,\sigma}(x_i)\right]\right| > \frac{\delta}{4}\right) \end{aligned} \quad (4.A4.15)$$

$$+ P\left(\left|\frac{1}{n} \sum_{i=1}^n g_{\eta,\sigma}(x_i) - E_X[g_{\eta,\sigma}(X)]\right| > \frac{\delta}{4}\right). \quad (4.A4.16)$$

Let us first deal with the probability given by (4.A4.15), with $U_i = \log \phi\left(\frac{y_i - \eta(x_i)}{\sigma}\right) - g_{\eta,\sigma}(x_i)$.

Due to (A8), we apply Bernstein's inequality to obtain

$$\begin{aligned} P\left(\left|\frac{1}{n} \sum_{i=1}^n U_i\right| > \frac{\delta}{4}\right) &\leq 2 \max \left\{ P\left(\frac{1}{n} \sum_{i=1}^n U_i > \frac{\delta}{4}\right), P\left(\frac{1}{n} \sum_{i=1}^n U_i < -\frac{\delta}{4}\right) \right\} \\ &\leq 2 \exp\left(-\frac{n}{2} \min\left\{\frac{\delta^2}{16s_{\eta,\sigma}^2}, \frac{\delta}{4s_{\eta,\sigma}}\right\}\right). \end{aligned} \quad (4.A4.17)$$

**4.A4. VERIFICATION OF THE ASSUMPTIONS OF SHALIZI FOR THE
GENERAL STOCHASTIC PROCESS MODEL**

Now note that the probability given by (4.A4.16) is the probability of a deterministic quantity with respect to \mathbf{y}_n and due to (4.A4.10) and (4.A4.11), is identically zero for large enough n . In the case of random covariates, using (A9) we obtain

$$\begin{aligned} |g_{\eta,\sigma}(x)| &\leq \int_{-\infty}^{\infty} \left| \log \phi \left(\frac{\sigma_0}{\sigma} z \right) \right| \phi(z) dz + \frac{L\|\eta - \eta_0\|}{\sigma} \\ &\leq \frac{c_3 + L\|\eta - \eta_0\|}{\sigma} = \tilde{c}_{\eta,\sigma} \text{ (say)}. \end{aligned} \quad (4.A4.18)$$

$g_{\eta,\sigma}(x_i)$ are independent, and satisfy (4.A4.18). Hence, Hoeffding's inequality yields

$$\begin{aligned} P \left(\left| \frac{1}{n} \sum_{i=1}^n g_{\eta,\sigma}(x_i) - E_X [g_{\eta,\sigma}(X)] \right| > \frac{\delta}{4} \right) \\ \leq \exp \left\{ -\frac{n^2 \delta^2}{144n\tilde{c}_{\eta,\sigma}^2} \right\} = \exp \left\{ -\frac{n\delta^2}{144\tilde{c}_{\eta,\sigma}^2} \right\}. \end{aligned} \quad (4.A4.19)$$

The probability given by (4.A4.14) can be bounded in the same way as (4.A4.17). Indeed, we have

$$P \left(\left| \frac{1}{n} \sum_{i=1}^n \log \phi \left(\frac{y_i - \eta_0(x_i)}{\sigma_0} \right) - c \right| > \frac{\delta}{2} \right) \leq 2 \exp \left(-\frac{n}{2} \min \left\{ \frac{\delta^2}{16s_{\eta_0,\sigma_0}^2}, \frac{\delta}{4s_{\eta_0,\sigma_0}} \right\} \right). \quad (4.A4.20)$$

Combining the above results, it is seen that for any $\delta > 0$, and for each $\theta \in \Theta$, there exists $a_\theta > 0$, depending on θ such that $P \left(\left| \frac{1}{n} \log R_n(\theta) + h(\theta) \right| > \delta \right) \leq 5 \exp \{-na_\theta\}$, which is summable. Hence, by the Borel-Cantelli lemma, $\frac{1}{n} \log R_n(\theta) \rightarrow -h(\theta)$, almost surely, as $n \rightarrow \infty$, for all $\theta \in \Theta$. Thus, (S3) holds.

4.A4.4 Verification of (S4)

Using (4.A4.18) it is easily seen that

$$h(\theta) \leq \left| \log \left(\frac{\sigma}{\sigma_0} \right) \right| + |c| + \int_{-\infty}^{\infty} \left| \log \phi \left(\frac{\sigma_0}{\sigma} z \right) \right| \phi(z) dz + \frac{\|\eta_0\| + \|\eta\|}{\sigma}. \quad (4.A4.21)$$

*4.A4. VERIFICATION OF THE ASSUMPTIONS OF SHALIZI FOR THE
GENERAL STOCHASTIC PROCESS MODEL*

Since almost surely with respect to the prior π_σ , $0 < \sigma < \infty$, and $\|\eta\| < \infty$ almost surely with respect to the prior of η , and since $\|\eta_0\| < \infty$, it follows from (4.A4.21), that $\pi(h(\theta) = \infty) = 0$, showing that (S4) holds.

4.A4.5 Verification of (S5)

Verification of (S5) (1)

Recall from (4.2.9) that

$$\mathcal{G}_n = \left\{ (\eta, \sigma) : \|\eta\| \leq \exp((\beta n)^{1/4}), \exp(-(\beta n)^{1/4}) \leq \sigma \leq \exp((\beta n)^{1/4}), \|\eta'_j\| \leq \exp((\beta n)^{1/4}); j = 1, \dots, d \right\}.$$

Then $\mathcal{G}_n \rightarrow \Theta$, as $n \rightarrow \infty$. Now note that

$$\begin{aligned} \pi(\mathcal{G}_n) &= \pi \left(\|\eta\| \leq \exp((\beta n)^{1/4}), \exp(-(\beta n)^{1/4}) \leq \sigma \leq \exp((\beta n)^{1/4}) \right) \\ &\quad - \pi \left(\left\{ \|\eta'_j\| \leq \exp((\beta n)^{1/4}); j = 1, \dots, d \right\}^c \right) \\ &= \pi \left(\|\eta\| \leq \exp((\beta n)^{1/4}), \exp(-(\beta n)^{1/4}) \leq \sigma \leq \exp((\beta n)^{1/4}) \right) \\ &\quad - \pi \left(\bigcup_{j=1}^d \left\{ \|\eta'_j\| > \exp((\beta n)^{1/4}) \right\} \right) \\ &\geq 1 - \pi \left(\|\eta\| > \exp((\beta n)^{1/4}) \right) - \pi \left(\left\{ \exp(-(\beta n)^{1/4}) \leq \sigma \leq \exp((\beta n)^{1/4}) \right\}^c \right) \\ &\quad - \sum_{j=1}^d \pi \left(\|\eta'_j\| > \exp((\beta n)^{1/4}) \right) \\ &\geq 1 - (c_\eta + c_\sigma + \sum_{j=1}^d c_{\eta'_j}) \exp(-\beta n), \end{aligned} \tag{4.A4.22}$$

by (A5) and (A6). In other words, (S5) (1) holds.

Verification of (S5) (2)

We now show that (S5) (2), namely, convergence in (S3) is uniform in θ over $\mathcal{G}_n \setminus I$ holds. In our case, by (S4), $h(\theta) < \infty$ with probability one, so that $I = \emptyset$ and $\mathcal{G}_n \setminus I = \mathcal{G}_n$. Since we have already proved in the context of (S3) that $\lim_{n \rightarrow \infty} \frac{1}{n} \log R_n(\theta) = -h(\theta)$, almost surely, for all $\theta \in \Theta$, (S5) (2) will be verified if we can further prove that \mathcal{G}_n is compact for each $n \geq 1$ and if $\frac{1}{n} \log R_n(\theta) + h(\theta)$ is Lipschitz in $\theta \in \mathcal{G}$, for any $\mathcal{G} \in \{\mathcal{G}_n : n = 1, 2, \dots\}$.

Compactness of \mathcal{G}_n , for all $n \geq 1$, follows as before. Hence uniform convergence as required will be proven if we can show that $\frac{1}{n} \log R_n(\theta) + h(\theta)$ is stochastically equicontinuous almost surely in $\theta \in \mathcal{G}$ for any $\mathcal{G} \in \{\mathcal{G}_n : n = 1, 2, \dots\}$ and $\frac{1}{n} \log R_n(\theta) + h(\theta) \rightarrow 0$, almost surely, for all $\theta \in \mathcal{G}$. As before, we rely on Lipschitz continuity. To see that $\frac{1}{n} \log R_n(\theta) + h(\theta)$ is Lipschitz in $\theta \in \mathcal{G}$, first observe that it follows from the arguments in Section 4.A4.1 that $\frac{1}{n} \log R_n(\theta)$ is Lipschitz in η , when the data are held constant. Moreover, the derivative with respect to σ is bounded since $\log \phi$ is Lipschitz and since σ is bounded in \mathcal{G} . In other words, $\frac{1}{n} \log R_n(\theta)$ is almost surely Lipschitz in θ . Thus, if we can show that $h(\theta)$ is also Lipschitz in θ , then this would prove that $\frac{1}{n} \log R_n(\theta) + h(\theta)$ is almost surely Lipschitz in θ . For our purpose, it is sufficient to show that $E_X [g_{\eta, \sigma}(X)]$ is Lipschitz in (η, σ) . Since for any $\eta_1, \eta_2, \sigma \in \Theta$,

$$\begin{aligned}
 & |E_X [g_{\eta_1, \sigma}(X)] - E_X [g_{\eta_2, \sigma}(X)]| \\
 & \leq E_X |g_{\eta_1, \sigma}(X) - g_{\eta_2, \sigma}(X)| \\
 & = E_X \left[\int_{-\infty}^{\infty} \left| \log \phi \left(\frac{\sigma_0 z + \eta_0(X) - \eta_1(X)}{\sigma} \right) - \log \phi \left(\frac{\sigma_0 z + \eta_0(X) - \eta_2(X)}{\sigma} \right) \right| \phi(z) dz \right] \\
 & \leq \frac{L}{\sigma} E_X \left[\int_{-\infty}^{\infty} |\eta_1(X) - \eta_2(X)| \phi(z) dz \right] \\
 & \leq \frac{L_2}{\sigma} \|\eta_1 - \eta_2\|,
 \end{aligned} \tag{4.A4.23}$$

for some $L_2 > 0$, $E_X [g_{\eta, \sigma}(X)]$ is Lipschitz in η . Now recall that under the assumption (A9), $\int_{-\infty}^{\infty} |z| \phi(z) dz < \infty$. With this, we now show that $E_X [g_{\eta, \sigma}(X)]$ has bounded first

**4.A4. VERIFICATION OF THE ASSUMPTIONS OF SHALIZI FOR THE
GENERAL STOCHASTIC PROCESS MODEL**

derivative with respect to σ in the interior of \mathcal{G} . Observe that

$$\begin{aligned}
 & |r|^{-1} |g_{\eta, \sigma+r}(x) - g_{\eta, \sigma}(x)| \\
 & \leq |r|^{-1} \left[\int_{-\infty}^{\infty} \left| \log \phi \left(\frac{\sigma_0 z + \eta_0(x) - \eta(x)}{\sigma + r} \right) - \log \phi \left(\frac{\sigma_0 z + \eta_0(x) - \eta(x)}{\sigma} \right) \right| \phi(z) dz \right] \\
 & \leq L \left[\int_{-\infty}^{\infty} \left(\frac{\sigma_0 |z| + \|\eta - \eta_0\|}{\sigma(\sigma + r)} \right) \phi(z) dz \right] \quad (\text{since } \log \phi \text{ is Lipschitz}) \\
 & \leq \frac{L}{\sigma(\sigma + r)} \left(\sigma_0 \int_{-\infty}^{\infty} |z| \phi(z) dz + \|\eta - \eta_0\| \right). \tag{4.A4.24}
 \end{aligned}$$

By (A9), $\int_{-\infty}^{\infty} |z| \phi(z) dz < \infty$, and $\sigma, \sigma + r$ (both in the interior of \mathcal{G}) are both upper and lower bounded in \mathcal{G} , the lower bound being strictly positive. Hence, (4.A4.24) is integrable with respect to (the distribution) of X , since \mathcal{X} is compact. Hence, by the dominated convergence theorem, differentiation with respect to σ can be performed inside the double integral associated with $E_X [g_{\eta, \sigma}(X)]$. Since $\log \phi$ has bounded first derivative as it is Lipschitz and since σ is lower bounded by a positive quantity in \mathcal{G} , it follows that $E_X [g_{\eta, \sigma}(X)]$ has bounded first derivative with respect to σ . Combined with the result that $E_X [g_{\eta, \sigma}(X)]$ is Lipschitz in η , this yields that $E_X [g_{\eta, \sigma}(X)]$ is Lipschitz in (η, σ) . In conjunction with the result that $\frac{1}{n} \log R_n(\theta)$ is almost surely Lipschitz in θ , it holds that $\frac{1}{n} \log R_n(\theta) + h(\theta)$ is almost surely Lipschitz in $\theta \in \mathcal{G}$. In other words, (S5) (2) stands verified.

Verification of (S5) (3)

To verify (S5) (3), note that continuity of $h(\theta)$, compactness of \mathcal{G}_n , along with its non-decreasing nature with respect to n implies that $h(\mathcal{G}_n) \rightarrow h(\Theta)$, as $n \rightarrow \infty$.

4.A4.6 Verification of (S6) and proof of Theorem 19

Let $\kappa_1 = \kappa - h(\Theta)$. Then it follows from (4.A4.14), (4.A4.15), (4.A4.16), (4.A4.17), (4.A4.19) and (4.A4.20), that

$$\begin{aligned} & \int_{\mathcal{S}^c} P \left(\left| \frac{1}{n} \log R_n(\theta) + h(\theta) \right| > \kappa_1 \right) d\pi(\theta) \\ & \leq \int_{\mathcal{S}^c} P \left(\left| \frac{1}{n} \sum_{i=1}^n \left[\log \phi \left(\frac{y_i - \eta(x_i)}{\sigma} \right) - g_{\eta, \sigma}(x_i) \right] \right| > \frac{\kappa_1}{4} \right) d\pi(\theta) \\ & \quad + \int_{\mathcal{S}^c} P \left(\left| \frac{1}{n} \sum_{i=1}^n g_{\eta, \sigma}(x_i) - E_X [g_{\eta, \sigma}(X)] \right| > \frac{\kappa_1}{4} \right) d\pi(\theta) \\ & \quad + \int_{\mathcal{S}^c} P \left(\left| \frac{1}{n} \sum_{i=1}^n \log \phi \left(\frac{y_i - \eta_0(x_i)}{\sigma_0} \right) - c \right| > \frac{\kappa_1}{2} \right) d\pi(\theta) \\ & \leq \int_{\mathcal{S}^c} 2 \exp \left(-\frac{n}{2} \min \left\{ \frac{\kappa_1^2}{16s_{\eta, \sigma}^2}, \frac{\kappa_1}{4s_{\eta, \sigma}} \right\} \right) d\pi(\theta) + \int_{\mathcal{S}^c} \exp \left(-\frac{Cn\kappa_1^2}{\tilde{c}_{\eta, \sigma}^2} \right) d\pi(\theta) \quad (4.A4.25) \end{aligned}$$

$$+ \int_{\mathcal{S}^c} 2 \exp \left(-\frac{n}{2} \min \left\{ \frac{\kappa_1^2}{16s_{\eta_0, \sigma_0}^2}, \frac{\kappa_1}{4s_{\eta_0, \sigma_0}} \right\} \right) d\pi(\theta), \quad (4.A4.26)$$

for some relevant positive constant C .

Now, in the same way as (4.A3.26) we obtain

$$\begin{aligned} & \int_{\mathcal{S}^c} 2 \exp \left(-\frac{n}{2} \min \left\{ \frac{\kappa_1^2}{16s_{\eta, \sigma}^2}, \frac{\kappa_1}{4s_{\eta, \sigma}} \right\} \right) d\pi(\theta) \\ & \leq \int_{\mathcal{G}_n} 2 \exp \left(-\frac{n}{2} \frac{\kappa_1^2}{16s_{\eta, \sigma}^2} \right) d\pi(\theta) + \int_{\mathcal{G}_n} 2 \exp \left(-\frac{n}{2} \frac{\kappa_1}{4s_{\eta, \sigma}} \right) d\pi(\theta) + 2\pi(\mathcal{G}_n^c) \\ & \leq C_1 \exp \left\{ - \left(C_2 \sqrt{\kappa_1} n^{1/4} - 5(\beta n)^{1/4} - 2n^q \log c_5 \right) \right\} \\ & \quad + \tilde{C}_1 \exp \left\{ - \left(\tilde{C}_2 \sqrt{\kappa_1} n^{1/4} - \frac{9}{2} (\beta n)^{1/4} - 2n^q \log c_5 \right) \right\} + 2\pi(\mathcal{G}_n^c), \quad (4.A4.27) \end{aligned}$$

for relevant positive constants $C_1, C_2, \tilde{C}_1, \tilde{C}_2, c_5$. In the same way,

$$\int_{\mathcal{S}^c} \exp \left(-\frac{Cn\kappa_1^2}{\tilde{c}_{\eta, \sigma}^2} \right) d\pi(\theta) \leq C_{11} \exp \left\{ - \left(C_{21} \sqrt{\kappa_1} n^{1/4} - 5(\beta n)^{1/4} - 2n^q \log c_{51} \right) \right\} + \pi(\mathcal{G}_n^c), \quad (4.A4.28)$$

for relevant positive constants C_{11}, C_{21}, c_{51} .

Since (4.A4.27) and (4.A4.28) are summable, and since (4.A4.26) is summable (as the integrand is independent of parameters), it follows from (4.A4.25) and (4.A4.26) that

$$\int_{\mathcal{S}^c} P \left(\left| \frac{1}{n} \log R_n(\theta) + h(\theta) \right| > \kappa_1 \right) d\pi(\theta) < \infty,$$

showing that (S6) holds.

4.A4.7 Verification of (S7)

For any set A such that $\pi(A) > 0$, $\mathcal{G}_n \cap A \uparrow A$. It follows from this and continuity of h that $h(\mathcal{G}_n \cap A) \downarrow h(A)$ as $n \rightarrow \infty$, so that (S7) holds.

5

Posterior Convergence of Nonparametric Binary and Poisson Regression Under Possible Misspecifications

5.1 Introduction

The situation for applicability of nonparametric regression is frequently encountered in many practical scenarios where no parametric model fits the data. In particular, nonparametric regression for binary dependent variables is very common for various branches of statistics like medical and spatial statistics, whereas nonparametric version of Poisson regression is being used recently in many non-trivial scenarios such as for analyzing the likelihood and severity of vehicle crashes ([Ye et al. \(2018\)](#)). Interestingly, despite vast applicability of both binary and Poisson regression, it seems that the available literature

on nonparametric Poisson regression is scarce in comparison to the available literature on nonparametric binary regression. The Bayesian approach to nonparametric binary regression problem has been accounted for in [Diaconis and Freedman \(1993\)](#). An account of posterior consistency for Gaussian process prior in nonparametric binary regression modeling can be found in [Ghosal and Roy \(2006\)](#), where the authors suggested that similar consistency results should hold for the nonparametric Poisson regression setup. Literature on consistency results for nonparametric Poisson regression is very limited. [Pillai *et al.* \(2007\)](#) have obtained consistency results for Poisson regression using an approach similar to that of [Ghosal and Roy \(2006\)](#) under certain assumptions, but so far without explicit specifications and details with respect to the prior. On the other hand, our approach will be based on the results of [Shalizi \(2009\)](#), which is much different from [Ghosal and Roy \(2006\)](#) and capable of handling model misspecification. In addition to facilitating investigation of the traditional posterior convergence rate, the approach of [Shalizi \(2009\)](#) also enables us to investigate the rate at which the posterior converges, which turns out to be the KL divergence rate.

In this chapter, we investigate posterior convergence of nonparametric binary and Poisson regression where the nonparametric regression is modeled as some suitable stochastic process. In the binary situation, we consider a similar setup as that of [Ghosal and Roy \(2006\)](#), where the authors have considered binary observations with response probability as an unknown smooth function of a set of covariates, which was modeled using Gaussian process. Here we will consider a binary response variable Y and a d -dimensional covariate X belonging to a compact set. The probability function is given by $p(x) = P(Y = 1|X = x)$ along with a prior for p induced by some appropriate stochastic process $\eta(x)$ with the relation $p(x) = H(\eta(x))$ for a known, non-decreasing and continuously differentiable cumulative distribution function $H(\cdot)$. We will establish a posterior convergence theory for nonparametric binary regression based on the general theory of posterior convergence of [Shalizi \(2009\)](#). Our theory also includes the case of

misspecified models, that is, if the true regression function is not even supported by the prior. This approach to Bayesian asymptotics also permits us to show that the relevant posterior probabilities converge at the KL divergence rate, and that the posterior convergence rate with respect to KL divergence is just slower than $\frac{1}{n}$, where n denotes the number of observations. We further show that even in the case of misspecification, the posterior predictive distribution can approximate the best possible predictive distribution adequately, in the sense that the Hellinger distance, as well as the total variation distance between the two distributions, can tend to zero.

For nonparametric Poisson regression, given x in the compact space of covariates, we model the mean function $\lambda(x)$ as $\lambda(x) = H(\eta(x))$, where H is a known, continuously differentiable function. Again, we investigate the general theory of posterior convergence, including misspecifications, rate of convergence of the posterior distribution and the usual posterior convergence rate, in Shalizi's framework.

The rest of this chapter is structured as follows. The basic premises for nonparametric binary and Poisson regression are provided in Sections 5.2 and 5.3, respectively. The required assumptions and their discussions are provided in Section 5.4. In Section 5.5, our main results on posterior convergence of binary and Poisson regression are provided, while Section 5.7 details the consequences of misspecifications. Concluding remarks are provided in Section 5.8.

The detailed proofs of verification of Shalizi's assumptions are provided in Appendices 5.A1 and 5.A2 for binary and Poisson regression setups, respectively.

5.2 Model setup and preliminaries of the binary regression

Let $\mathbf{Y}_n = (Y_1, Y_2, \dots, Y_n)^T$ be the vector of binary response random variables against the covariate vector $\mathbf{X}_n = (X_1, X_2, \dots, X_n)^T$. The corresponding realized values will be denoted by $\mathbf{y}_n = (y_1, y_2, \dots, y_n)^T$ and $\mathbf{x}_n = (x_1, x_2, \dots, x_n)^T$ respectively. Let the

11.5.2. MODEL SETUP AND PRELIMINARIES OF THE BINARY REGRESSION

model be specified as follows: for $i = 1, 2, \dots, n$,

$$Y_i|X_i \sim \text{Binomial}(1, p(X_i)) \quad (5.2.1)$$

$$p(x) = H(\eta(x)) \quad (5.2.2)$$

$$\eta(\cdot) \sim \pi_\eta, \quad (5.2.3)$$

where the link function H is a known, non-decreasing, continuously differentiable cumulative distribution function on the real line \mathbb{R} , π_η is the prior for some suitable stochastic process.

Note that the prior for p is induced by the prior for η . Our concern is to infer about the success probability function $p(x) = P(Y = 1|X = x)$ when the number of observations goes to infinity. We assume that for each $i \geq 1$, $x_i \in \mathcal{X}$, where $\mathcal{X} \subset \mathbb{R}^d$ is the compact space of covariates, with $d \geq 1$ being the dimension of the covariate space. We assume that d is finite and known. We also assume that the functions η have continuous first partial derivatives. We denote this class of functions by $\mathcal{C}'(\mathcal{X})$. We do not assume the truth η_0 in $\mathcal{C}'(\mathcal{X})$, allowing misspecification.

It is widely accepted to assume the function $H(\cdot)$ to be known as part of model assumption. For example, in logistic regression we choose the standard logistic cumulative distribution function as the link function, whereas in probit regression H is chosen to be the standard normal cumulative distribution function ϕ . More discussion on link function along with several other examples can be found in Choudhuri *et al.* (2007), Newton *et al.* (1996), Gelfand and Kuo (1991). A Bayesian method for estimation of p has been provided in Choudhuri *et al.* (2007). It has been shown in Ghosal and Roy (2006) that the sample paths of the Gaussian processes can well approximate a large class of functions and hence it is not essential to consider additional uncertainty in the link function H .

Let \mathfrak{C} be the counting measure on $\{0, 1\}$. Then according to the model assumption,

the conditional density of y given x with respect to \mathfrak{C} will be represented by the density function f as follows:

$$f(y|x) = p(x)^y (1 - p(x))^{1-y}. \quad (5.2.4)$$

The prior for f will be denoted by π . Let f_0 and p_0 denote the true density and true success probability, respectively. Then under the truth, the joint density is:

$$f_0(y|x) = p_0(x)^y (1 - p_0(x))^{1-y}. \quad (5.2.5)$$

One of the main objectives of this chapter is to show consistency of the posterior distribution of $\eta(\cdot)$ treated as a parameter arising from the parameter space $\Theta = \mathcal{C}'(\mathcal{X})$. Note that this would imply posterior consistency of $p(\cdot) = H(\eta(\cdot))$.

5.3 Model setup and preliminaries of Poisson regression

For Poisson regression, we let $\mathbf{Y}_n = (Y_1, Y_2, \dots, Y_n)^T$ be independent responses conditional on covariates $\mathbf{X}_n = (X_1, X_2, \dots, X_n)^T$, with realized values denoted by $\mathbf{y}_n = (y_1, y_2, \dots, y_n)^T$ and $\mathbf{x}_n = (x_1, x_2, \dots, x_n)^T$ as before. For each $i \geq 1$, $y_i \in \mathbb{N}$, where \mathbb{N} is the set of non-negative integers and as before we assume that $x_i \in \mathcal{X}$, where \mathcal{X} is a compact subset of the real line \mathbb{R}^d , where $d \geq 1$ is finite and known. The mean function is given by $\lambda(x) = H(\eta(x))$, where H is a known, non-negative continuously differentiable function on \mathbb{R} and $\eta \in \mathcal{C}'(\mathcal{X})$. Thus, here also the parameter space is $\Theta = \mathcal{C}'(\mathcal{X})$. The model is specified as follows: for $i = 1, 2, \dots, n$,

$$Y_i|X_i \sim \exp(-\lambda(X_i)) \frac{(\lambda(X_i))^y}{y!}; \quad (5.3.1)$$

$$\eta(\cdot) \sim \pi_\eta. \quad (5.3.2)$$

Now, let \mathfrak{C} be the counting measure on \mathbb{N} . According to the model assumption

for Poisson regression, the conditional density of y given x with respect to \mathfrak{C} will be represented by density function f as follows:

$$f(y|x) = \exp(-\lambda(x)) \frac{(\lambda(x))^y}{y!}. \quad (5.3.3)$$

The prior for f will be denoted by Π . We do not assume the truth η_0 to be in $\mathcal{C}'(\mathcal{X})$ as before, allowing misspecification. Let f_0 and λ_0 denote the true density and true mean function, respectively. Again, one of our main aims is to establish consistency of the posterior distribution of $\lambda(\cdot)$ through posterior consistency of $\eta(\cdot)$.

5.4 Assumptions and their discussions

We need to make some appropriate assumptions for establishing convergence of both the binary and Poisson regression models equipped with stochastic process prior. The latter also requires suitable assumptions. Many of the assumptions are similar to those taken in Chapter 4. Hence the purpose of such assumptions will be as discussed in Chapter 4, which we shall briefly touch upon here.

Assumption 1 \mathcal{X} is a compact, d -dimensional space, for some finite, known, $d \geq 1$, and is equipped with a suitable metric.

Assumption 2 Recall that in our notation, $\mathcal{C}'(\mathcal{X})$ denotes the class of continuously partially differentiable function on \mathcal{X} . In other words, the functions $\eta \in \mathcal{C}'(\mathcal{X})$ are continuous on \mathcal{X} and for such functions the limit

$$\eta'_j(x) = \frac{\partial \eta(x)}{\partial x_j} = \lim_{h \rightarrow 0} \frac{\eta(x + h\delta_j) - \eta(x)}{h} \quad (5.4.1)$$

exists for $j = 1, \dots, d$, for each $x \in \mathcal{X}$ and is continuous on \mathcal{X} . Here δ_j is the d -dimensional vector with the j -th element as 1 and all the other elements as zero.

Assumption 3 *The priors for η is chosen such that for $\beta > 2h(\Theta)$,*

$$\begin{aligned}\pi\left(\|\eta\| \leq \exp\left((\beta n)^{1/4}\right)\right) &\geq 1 - c_\eta \exp(-\beta n); \\ \pi\left(\|\eta'_j\| \leq \exp\left((\beta n)^{1/4}\right)\right) &\geq 1 - c_{\eta'_j} \exp(-\beta n), \text{ for } j = 1, \dots, d;\end{aligned}$$

where c_η and $c_{\eta'_j}$; $j = 1, \dots, d$, are positive constants. In the above, for any function $f : \mathcal{X} \mapsto \mathbb{R}$, $\|f\| = \sup_{x \in \mathcal{X}} |f(x)|$.

We treat the covariates as either random (observed or unobserved) or non-random (observed). Accordingly, in Assumption 4 below we provide conditions pertaining to these aspects.

Assumption 4 (i) *$\{x_i : i = 1, 2, \dots\}$ is an observed or unobserved sample associated with an iid sequence associated with some probability measure Q , supported on \mathcal{X} , which is independent of $\{y_i : i = 1, 2, \dots\}$*

(ii) *$\{x_i : i = 1, 2, \dots\}$ is an observed non-random sample. In this case, we consider a specific partition of the d -dimensional space \mathcal{X} into n subsets such that each subset of the partition contains at least one $x \in \{x_i : i = 1, 2, \dots\}$ and has Lebesgue measure $\frac{L}{n}$, for some $L > 0$.*

Assumption 5 *The true function η_0 is bounded in sup norm. In other words, the truth η_0 satisfies the following for some finite, positive constant κ_0 :*

$$\|\eta_0\| < \kappa_0 < \infty. \tag{5.4.2}$$

Observe that in general $\eta_0 \notin \mathcal{C}'(\mathcal{X})$. For random covariate X , we assume that $\eta_0(X)$ is measurable.

Assumption 6 For the binary regression model set up we assume a uniform positive lower bound κ_B for $\min\{p(\cdot), 1 - p(\cdot)\}$. In other words, for all $p \in \Theta$,

$$\inf\{\min(p(x), 1 - p(x)) : x \in \mathcal{X}\} \geq \kappa_B > 0.$$

Assumption 7 For the Poisson regression model set up we assume a uniform positive lower bound κ_P for $\lambda(\cdot)$. In other words, for all $\lambda \in \Lambda$,

$$\inf\{\lambda(x) : x \in \mathcal{X}\} \geq \kappa_P > 0.$$

5.4.1 Discussion of the assumptions

Assumption 1 is on compactness of \mathcal{X} , which guarantees that continuous functions on \mathcal{X} will have finite sup-norms.

Assumption 2 is as taken in Chapter 4 for the purpose of constructing appropriate sieves in order to show posterior convergence results. More precisely, Assumption 2 is required for to ensure that η is Lipschitz continuous in the sieves. Since a differentiable function is Lipschitz if and only if its partial derivatives are bounded, this serves our purpose, as continuity of the partial derivatives of η guarantees boundedness in the compact domain \mathcal{X} . In particular, if η is a Gaussian process, the conditions presented in Adler (1981), Adler and Taylor (2007), Cramer and Leadbetter (1967) guarantee the above continuity and smoothness properties required by Assumption 2. We refer to Chapter 4 for more discussion about this.

Assumption 3 is required for ensuring that the complements of the sieves have exponentially small probabilities. In particular, this assumption is satisfied if η is a Gaussian process, even if $\exp((\beta n)^{1/4})$ is replaced with $\sqrt{\beta n}$.

Assumption 4 is for the covariates x_i , accordingly as they are considered an observed random sample, unobserved random sample, or non-random. Note that thanks to the strong law of large numbers (SLLN), given any η in the complement of some null set

with respect to the prior, and given any sequence $\{x_i : i = 1, 2, \dots\}$ Assumption 4 (i) ensures that for any integrable function g , as $n \rightarrow \infty$,

$$\frac{1}{n} \sum_{i=1}^n g(x_i) \rightarrow \int_{\mathcal{X}} g(x) dQ(X) = E_X [g(X)] \text{ (say),} \quad (5.4.3)$$

almost surely, where Q is some probability measure supported on \mathcal{X} .

Assumption 4 (ii) ensures that $\frac{1}{n} \sum_{i=1}^n g(x_i)$ is a particular Riemann sum and hence (5.4.3) holds with Q being the Lebesgue measure on \mathcal{X} . We continue to denote the limit in this case by $E_X [g(X)]$.

Assumption 5 is equivalent to the Assumption(T) of Ghosal and Roy (2006). Assumption 5 actually implies that $p_0(x) = H(\eta_0(x))$ is bounded away from 0 and 1. For the Poisson regression model set up it follows that $\|\lambda_0\| < \infty$. It is to be noted that here we do not require to assume that $p_0 \in \Theta$ or $\lambda_0 \in \Lambda$, allowing model misspecifications.

Observe that, similar to Pillai *et al.* (2007) we need the parameter space for Poisson regression to be bounded away from zero (Assumption 7). As pointed out in Pillai *et al.* (2007), we cannot bypass this and as such this is not a mere pathway towards our proof. This is because, if almost all observations in a sample from a Poisson distribution are zero, then it impossible to extract the information about the (log) mean. Hence we must require at least some condition to make it bound away from zero. Similar argument is also applicable for binary regression, which is reflected in Assumption 6.

It is important to remark that Assumptions 6 and 7 are necessary only to validate Assumptions (S5) (3) (that is, the third part of Assumption (S5)) and (S6) of Shalizi, and unnecessary elsewhere. Although many of our proofs would be simpler if Assumptions 6 and 7 were used, we reserved these assumptions only to validate Assumptions (S5) (3) and (S6) of Shalizi.

For any $x \in \mathbb{R}$ and for any given set $A \subseteq R$, let $\mathbb{I}_A(x) = 1$ if $x \in A$ and 0 otherwise.

To achieve Assumptions 6 and 7, we set, for all $x \in \mathbb{R}$,

$$H(x) = \kappa_B \mathbb{I}_{\{G(x) \leq \kappa_B\}}(x) + G(x) \mathbb{I}_{\{\kappa_B < G(x) < 1 - \kappa_B\}}(x) + (1 - \kappa_B) \mathbb{I}_{\{G(x) \geq 1 - \kappa_B\}}(x), \quad (5.4.4)$$

for the binary case, where $0 < \kappa_B < 1/2$, and

$$H(x) = \kappa_P \mathbb{I}_{\{G(x) \leq \kappa_P\}}(x) + G(x) \mathbb{I}_{\{G(x) > \kappa_P\}}(x), \quad (5.4.5)$$

where $\kappa_P > 0$. In (5.4.4), G is a continuously differentiable distribution function on \mathbb{R} and in (5.4.5), G is a non-negative continuously differentiable function on \mathbb{R} .

5.5 Main results on posterior convergence

Here we will state a summary of our main results regarding posterior convergence of nonparametric binary regression and Poisson regression. The key results associated with the asymptotic equipartition property are provided in Theorems 3 – 6, proofs of which are provided in Appendix 5.A1 (for binary regression) and in Appendix 5.A2 (for Poisson regression).

Theorem 3 *Let Q and the counting measure \mathfrak{C} on $\{0, 1\}$ be the measures associated with the random variable X and the binary random variable Y respectively. Denote $E_{X,Y}(\cdot) = \int \int \cdot d\mathfrak{C} dQ$ and $E_X(\cdot) = \int \cdot dQ$. Then under the nonparametric binary regression model, under Assumption 4, the KL divergence rate $h_1(p)$ exists for $\eta \in \Theta$, and is given by*

$$h_1(p) = \left[E_X \left(p_0(X) \log \left\{ \frac{p_0(X)}{p(X)} \right\} \right) + E_X \left((1 - p_0(X)) \log \left\{ \frac{(1 - p_0(X))}{(1 - p(X))} \right\} \right) \right]. \quad (5.5.1)$$

Alternatively, $h_1(p)$ admits the following form:

$$h_1(p) = E_{X,Y} \left(f_0(X, Y) \log \left\{ \frac{f_0(X, Y)}{f(X, Y)} \right\} \right), \quad (5.5.2)$$

where f and f_0 are as defined in (5.2.4) and (5.2.5).

Theorem 4 Let Q and the counting measure \mathfrak{C} on \mathbb{N} be associated with the random variable X and the count random variable Y , respectively. Denote $E_{X,Y}(\cdot) = \int \int \cdot d\mathfrak{C} dQ$ and $E_X(\cdot) = \int \cdot dQ$. Then under the nonparametric Poisson regression model, under Assumption 4, the KL divergence rate $h_2(\lambda)$ exists for $\eta \in \Theta$, and is given by

$$h_2(\lambda) = \left[E_X (\lambda(X) - \lambda_0(X)) + E_X \left(\lambda_0(X) \log \left\{ \frac{\lambda_0(X)}{\lambda(X)} \right\} \right) \right]. \quad (5.5.3)$$

Theorem 5 Under the nonparametric binary regression model and Assumption 4, the asymptotic equipartition property holds, and is given by

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log [R_n(p)] = -h_1(p). \quad (5.5.4)$$

The convergence is uniform on any compact subset of Θ .

Theorem 6 Under the nonparametric Poisson regression model and Assumption 4, the asymptotic equipartition property holds, and is given by

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log [R_n(\lambda)] = -h_2(\lambda). \quad (5.5.5)$$

The convergence is uniform on any compact subset of Θ .

Theorems 3 and 5 for binary regression and Theorems 4 and 6 for Poisson regression ensure that conditions (S1) to (S3) of Shalizi (2009) hold, and (S4) holds for both binary and Poisson regression because of compactness of \mathcal{X} and continuity of H and η . The detailed proofs are presented in Appendix 5.A1.4 and Appendix 5.A2.4, respectively.

We construct the sieves \mathcal{G}_n for binary regression model set up as follows:

$$\mathcal{G}_n = \{\eta \in \mathcal{C}'(\mathcal{X}) : \|\eta\| \leq \exp((\beta n)^{1/4}), \|\eta'_j\| \leq \exp((\beta n)^{1/4}); j = 1, 2, \dots, d\}. \quad (5.5.6)$$

It follows that $\mathcal{G}_n \rightarrow \Theta$ as $n \rightarrow \infty$. We consider the same form (5.5.6) for the sieves associated with Poisson regression. That \mathcal{G}_n is compact, is already proved in Chapter 4. For notational convenience, we denote the sieves for Poisson regression by \mathbb{G}_n .

Assumption 3 ensures that for binary regression, $\Pi(\mathcal{G}_n^c) \leq \alpha \exp(-\beta n)$ for some $\alpha > 0$ and similarly $\Pi(\mathbb{G}_n^c) \leq \alpha \exp(-\beta n)$ for Poisson regression. Now, these results, continuity of $h(p)$, $h(\lambda)$ (the proofs of continuity of $h(p)$ and $h(\lambda)$ follows using the same techniques as in Appendices 5.A1.1 and 5.A2.1), compactness of \mathcal{G}_n , \mathbb{G}_n and the uniform convergence results of Theorems 5 and 6, together ensure (S5) for both the model setups.

Now, as pointed out in Chapter 4 we observe that the aim of assumption (S6) is to ensure that (see the proof of Lemma 7 of Shalizi (2009)) for every $\epsilon > 0$ and for all sufficiently large n ,

$$\frac{1}{n} \log \int_{\mathcal{G}_n} R_n(p) d\pi(\eta) \leq h_1(\mathcal{G}_n) + \epsilon, \text{ almost surely.} \quad (5.5.7)$$

As $h(\mathcal{G}_n) \rightarrow h_1(\Theta)$ as $n \rightarrow \infty$, it is enough to verify that for every $\epsilon > 0$ and for all n sufficiently large,

$$\frac{1}{n} \log \int_{\mathcal{G}_n} R_n(p) d\pi(\eta) \leq h(\Theta) + \epsilon, \text{ almost surely.} \quad (5.5.8)$$

First we observe that

$$\frac{1}{n} \log \int_{\mathcal{G}_n} R_n(p) d\pi(\eta) \leq \frac{1}{n} \sup_{\eta \in \mathcal{G}_n} \log R_n(p). \quad (5.5.9)$$

For large enough $\kappa > h_1(\Theta)$, consider $\mathcal{S} = \{\eta : h_1(p) \leq \kappa, \|\eta\| \leq M\}$, for $M > 0$.

Lemma 5.5.1 *For $M > 0$, $\mathcal{G}_n \cap \mathcal{S} = \mathcal{G}_n \cap \{\eta : h_1(p) \leq \kappa, \|\eta\| \leq M\}$ is a compact set.*

Proof. First recall that the proof of continuity of $h_1(p)$ in η follows easily using the same techniques as in Appendix 4.A2.1. Hence, it follows that \mathcal{S} is a closed and bounded set. The rest of the proof follows due to Lipschitz continuity of η on \mathcal{G}_n . ■

In a very similar manner, the following lemma also holds for Poisson model set up.

Lemma 5.5.2 $\mathbb{G}_n \cap \mathcal{S} = \mathbb{G}_n \cap \{\eta : h_2(\lambda) \leq \kappa, \|\eta\| \leq M\}$ is a compact set for $M > 0$.

Proof. Again, recall that continuity of $h_2(\lambda)$ in η can be shown using the same techniques as in Appendix 5.A2.1. The rest of the proof follows in the same way as that of Lemma 5.5.1. ■

Now observe that if κ_B of Assumption 6 is actually zero instead of positive, then as $\|\eta\| \rightarrow \infty$, $h_1(p) \rightarrow \infty$. Moreover, we have already shown continuity of $h_1(p)$ with respect to η . Hence, for sufficiently large M , $\|\eta\| > M$ implies $h_1(p) \geq \kappa$, provided that κ_B is sufficiently small. Hence, $\mathcal{S}^c = \{\eta : h_1(p) > \kappa\} \cup \{\eta : h_1(p) \leq \kappa, \|\eta\| > M\} = \{\eta : h_1(p) > \kappa\} \cup \{\eta : h_1(p) = \kappa, \|\eta\| > M\}$, for sufficiently small κ_B and sufficiently large M .

In the same way, for Poisson regression, if κ_P of Assumption 7 is actually zero instead of positive, then as $\|\eta\| \rightarrow \infty$, $h(\lambda) \rightarrow \infty$. This, along with continuity of $h_2(\lambda)$ with respect to η ensures that $h_2(\lambda) \geq \kappa$ when $\|\eta\| > M$, for sufficiently large M , if κ_P is small enough. Hence, for sufficiently small κ_P and sufficiently large M , $\mathcal{S}^c = \{\eta : h_2(\lambda) > \kappa\} \cup \{\eta : h_2(\lambda) \leq \kappa, \|\eta\| > M\} = \{\eta : h_2(\lambda) > \kappa\} \cup \{\eta : h_2(\lambda) = \kappa, \|\eta\| > M\}$, in the context of Poisson regression.

Using compactness of \mathcal{S} , in the same way as in Chapter 4 condition (S6) of Shalizi can be shown to be equivalent to (5.5.10) and (5.5.11) in Theorems 7 and 8 below, corresponding to binary and Poisson cases. In the supplement we show that these equivalent conditions are satisfied in our model setups.

Theorem 7 For the binary regression setup, (S6) is equivalent to the following, which

holds under Assumptions 1 – 6:

$$\sum_{n=1}^{\infty} \int_{\mathcal{S}^c} P \left(\left| \frac{1}{n} \log R_n(p) + h_1(p) \right| > \kappa - h_1(\Theta) \right) d\pi(\eta) < \infty. \quad (5.5.10)$$

Theorem 8 For the Poisson regression model set up, (S6) is equivalent to the following, which holds under Assumptions 1–5 and 7:

$$\sum_{n=1}^{\infty} \int_{\mathcal{S}^c} P \left(\left| \frac{1}{n} \log R_n(\lambda) + h_2(\lambda) \right| > \kappa - h_2(\Theta) \right) d\pi(\eta) < \infty. \quad (5.5.11)$$

Assumption (S7) of Shalizi also holds for both the model setups because of continuity of $h_1(p)$ and $h_2(\lambda)$. Hence, all the assumptions (S1)–(S7) are satisfied for binary and Poisson regression setups.

Overall, our results lead to the following theorems.

Theorem 9 Assume the nonparametric binary regression setup. Then under Assumptions 1–6, for $A \subseteq \Theta$ for which $\pi(A) > 0$ and $h_1(A) > h_1(\Theta)$,

$$\lim_{n \rightarrow \infty} \pi(A|\mathbf{Y}_n) = 0, \text{ almost surely.}$$

Also, for any measurable set A with $\pi(A) > 0$, if $\beta > 2h_1(A)$, where h_1 is given by (5.5.1), or if $A \subset \bigcap_{k=n}^{\infty} \mathcal{G}_k$ for some n , where \mathcal{G}_k is given by (5.5.6), then the following holds:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log [\pi(A|\mathbf{Y}_n)] = -J_1(A), \text{ almost surely.}$$

Theorem 10 Assume the nonparametric Poisson regression setup. Then under Assumptions 1–5 and 7, for $A \subseteq \Theta$ for which $\pi(A) > 0$ and $h_2(A) > h_2(\Theta)$,

$$\lim_{n \rightarrow \infty} \pi(A|\mathbf{Y}_n) = 0, \text{ almost surely.}$$

Also, for any measurable set A with $\pi(A) > 0$, if $\beta > 2h_2(A)$, where h_2 is given by

(5.5.3), or if $A \subset \bigcap_{k=n}^{\infty} \mathbb{G}_k$ for some n , where \mathbb{G}_k is of the same form as (5.5.6), then the following holds:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log [\pi(A|\mathbf{Y}_n)] = -J_2(A), \text{ almost surely.}$$

5.6 Rate of convergence

For Shalizi's approach to the rate of convergence, it is first required to observe that for each measurable $A \subseteq \Theta$, for every $\delta > 0$, there exists a random natural number $\tau(A, \delta)$ such that $n^{-1} \log \int_A R_n(\theta) d\pi(\theta) \leq \delta + \limsup_n n^{-1} \log \int_A R_n(\theta) d\pi(\theta)$ for all $n > \tau(A, \delta)$, provided the \limsup is finite.

Now consider a sequence of positive reals ϵ_n such that $\epsilon_n \rightarrow 0$ while $n\epsilon_n \rightarrow \infty$ as $n \rightarrow \infty$ and the set $N_{\epsilon_n} = \{\theta : h(\theta) \leq h(\Theta) + \epsilon_n\}$. Then the following result of Shalizi holds.

Theorem 11 (Shalizi (2009)) Assume (S1) to (S7) of Shalizi. If for each $\delta > 0$,

$$\tau(\mathcal{G}_n \cap N_{\epsilon_n}^c, \delta) \leq n$$

eventually almost surely, then almost surely the following holds:

$$\lim_{n \rightarrow \infty} (N_{\epsilon_n} | \mathbf{Y}_n) = 1.$$

In our contexts, let $N_{1,\epsilon_n} = \{\eta : h_1(p) \leq h_1(\Theta) + \epsilon_n\}$ and $N_{2,\epsilon_n} = \{\eta : h_2(\lambda) \leq h_2(\Theta) + \epsilon_n\}$. To investigate the rate of convergence in our cases, it follows from Chapter 4 that ϵ_n will be the rate of convergence for $\epsilon_n \rightarrow 0$, $n\epsilon_n \rightarrow \infty$ as $n \rightarrow \infty$, if we can show that the following hold:

$$\frac{1}{n} \log \int_{\mathcal{G}_n \cap N_{1,\epsilon_n}^c} R_n(p) d\pi(\eta) \leq -h_1(\Theta) + \epsilon, \quad (5.6.1)$$

and

$$\frac{1}{n} \log \int_{\mathbb{G}_n \cap N_{2,\epsilon_n}^c} R_n(\lambda) d\pi(\eta) \leq -h_2(\Theta) + \epsilon, \quad (5.6.2)$$

for any $\epsilon > 0$ and all n sufficiently large.

Following similar arguments of Chapter 4, we find that the posterior rate of convergence with respect to KL divergence is just slower than n^{-1} . To put it another way, it is just slower than $n^{-\frac{1}{2}}$ with respect to Hellinger distance for the model setups we consider. Our results can be formally stated in Theorem 12 for Binary regression and in Theorem 13 for Poisson regression.

Theorem 12 *For the nonparametric binary regression setup, under Assumptions 1–6, $\lim_{n \rightarrow \infty} (N_{\epsilon_n} | \mathbf{Y}_n) = 1$ holds almost surely.*

Theorem 13 *For the nonparametric Poisson regression setup, under Assumptions 1–5 and 7, $\lim_{n \rightarrow \infty} (N_{\epsilon_n} | \mathbf{Y}_n) = 1$ holds almost surely.*

5.7 Consequences of model misspecification

Suppose that the true function η_0 consists of countable number of discontinuities but has continuous first order partial derivatives at all other points. Then $\eta_0 \notin \mathcal{C}'(\mathcal{X})$. However, there exists some $\tilde{\eta} \in \mathcal{C}'(\mathcal{X})$ such that $\tilde{\eta}(x) = \eta_0(x)$ for all $x \in \mathcal{X}$ where η_0 is continuous. Similar situation is mentioned in Chapter 4. Observe that, if the probability measure Q of X is dominated by the Lebesgue measure, then from Theorems 3 and 4 we have $h_1(\Theta) = 0$ and $h_2(\Theta) = 0$. Then the posterior of η concentrates around $\tilde{\eta}$, which is the same as η_0 except at the countable number of discontinuities of η_0 . Corresponding $\tilde{p} = H(\tilde{\eta})$ and $\tilde{\lambda} = H(\tilde{\eta})$ will also differ from p_0 and λ_0 . If p_0 and λ_0 are such that $0 < h_1(\Theta) < \infty$ and $0 < h_2(\Theta) < \infty$ respectively, then the posteriors concentrate around the minimizers of $h_1(p)$ and $h_2(\lambda)$, provided such minimizers exist in Θ .

5.7.1 Consequences from the subjective Bayesian perspective

Bayesian posterior consistency has two apparently different viewpoints, namely, classical and subjective. Bayesian analysis starts with a prior knowledge, and updates the knowledge given the data, forming the posterior. It is of utmost importance to know whether the updated knowledge becomes more and more accurate and precise as data are collected indefinitely. This requirement is called consistency of the posterior distribution. From the classical Bayesian point of view we should believe in existence of a true model. On the contrary, if we look from the subjective Bayesian viewpoint, then we need not believe in true models. A subjective Bayesian thinks only in terms of the predictive distribution of future observations. But [Blackwell and Dubins \(1962\)](#), [Diaconis and Freedman \(1986\)](#) have shown that consistency is equivalent to inter subjective agreement, which means that two Bayesians will ultimately have very close posterior predictive distributions.

Shalizi considered the one-step-ahead predictive distribution of θ , given by $F_\theta^n \equiv F_\theta(Y_n|Y_1, \dots, Y_{n-1})$, with the convention that $n = 1$ gives the marginal distribution of the first observation. Accordingly, he also considered $P^n \equiv P^n(Y_n|Y_1, \dots, Y_{n-1})$, which is the best prediction one could make had P been known. The posterior predictive distribution is given by $F_\pi^n = \int_\Theta F_\theta^n d\pi(\theta|\mathbf{Y}_n)$. With the above definitions, the following results have been proved by Shalizi.

Theorem 14 ([Shalizi \(2009\)](#)) *Let ρ_H and ρ_{TV} be Hellinger and total variation metrics, respectively. Then with probability 1,*

$$\begin{aligned} \limsup_{n \rightarrow \infty} \rho_H^2(P^n, F_\pi^n) &\leq h(\Theta); \\ \limsup_{n \rightarrow \infty} \rho_{TV}^2(P^n, F_\pi^n) &\leq 4h(\Theta). \end{aligned}$$

In our nonparametric setup, $h_1(\Theta) = 0$ and $h_2(\Theta) = 0$ if η_0 consists of countable

number of discontinuities. Hence, from Theorem 14 it is clear that in spite of such misspecification, the posterior predictive distribution does a good job in learning the best possible predictive distribution in terms of the popular Hellinger and the total variation distance. We state our result formally as follows.

Theorem 15 *Consider the setups of nonparametric binary and Poisson regression. Assume that the truth function η_0 consists of countable number of discontinuities but has continuous first order partial derivatives at all other points. Then under Assumptions 1–6 (for binary regression) or under Assumptions 1–5 and 7 (for Poisson regression) the following hold:*

$$\begin{aligned}\limsup_{n \rightarrow \infty} \rho_H^2(P^n, F_\pi^n) &= 0; \\ \limsup_{n \rightarrow \infty} \rho_{TV}^2(P^n, F_\pi^n) &= 0.\end{aligned}$$

5.8 Conclusion

In this chapter we attempted to address posterior convergence of nonparametric binary and Poisson regression, along with the rate of convergence, while also allowing for misspecification, using the approach of Shalizi (2009). As in Chapter 4, we also have shown that, even in the case of misspecification, the posterior predictive distribution can be quite accurate asymptotically, which should be a point of interest from the subjective Bayesian viewpoint. The asymptotic equipartition property plays a central role even in binary and Poisson regression contexts. It is one of the crucial assumptions and yet relatively easy to establish under mild conditions. It actually brings forward the KL property of the posterior, which in turn characterizes the posterior convergence, and also the rate of posterior convergence and misspecification, as in the nonparametric regression with Gaussian and double-exponential errors dealt with in Chapter 4.

Appendix

5.A1 Verification of Assumptions (S1) to (S7) of Shalizi for binary regression

5.A1.1 Verification of (S1)

Observe that

$$f_p(\mathbf{Y}_n | \mathbf{X}_n) = \prod_{i=1}^n f(y_i | x_i) = \prod_{i=1}^n p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}, \quad (5.A1.1)$$

$$f_{p_0}(\mathbf{Y}_n | \mathbf{X}_n) = \prod_{i=1}^n f_0(y_i | x_i) = \prod_{i=1}^n p_0(x_i)^{y_i} (1 - p_0(x_i))^{1-y_i}. \quad (5.A1.2)$$

Therefore,

$$\frac{1}{n} \log R_n(p) = \frac{1}{n} \sum_{i=1}^n \left\{ \left(y_i \log \left(\frac{p(x_i)}{p_0(x_i)} \right) \right) + (1 - y_i) \log \left(\frac{1 - p(x_i)}{1 - p_0(x_i)} \right) \right\}. \quad (5.A1.3)$$

To show measurability of $R_n(p)$, first note that for any $a \in \mathbb{R}$,

$$\begin{aligned} & \left\{ (y_i, \eta) : y_i \log \left(\frac{p(x_i)}{p_0(x_i)} \right) + (1 - y_i) \log \left(\frac{1 - p(x_i)}{1 - p_0(x_i)} \right) < a \right\} \\ &= \left\{ \eta : \log \left(\frac{p(x_i)}{p_0(x_i)} \right) < a \right\} \cup \left\{ \eta : \log \left(\frac{1 - p(x_i)}{1 - p_0(x_i)} \right) < a \right\}. \end{aligned} \quad (5.A1.4)$$

Note that for given p , there exists $0 < \epsilon < 1/2$ such that $\epsilon < p(x) < 1 - \epsilon$, for all $x \in \mathcal{X}$.

Now consider a sequence $\tilde{\eta}_j$, $j = 1, 2, \dots$ such that $\|\tilde{\eta}_j - \eta\| \rightarrow 0$, as $j \rightarrow \infty$. Then, with $\tilde{p}_j(x) = H(\tilde{\eta}_j(x))$, note that there exists $j_0 \geq 1$ such that for $j \geq j_0$, $\epsilon < \tilde{p}_j(x) < 1 - \epsilon$,

for all $x \in \mathcal{X}$. Hence, using the inequality $1 - \frac{1}{x} \leq \log x \leq x - 1$ for $x > 0$, we obtain $\left| \log \left(\frac{\tilde{p}_j(x_i)}{p_0(x_i)} \right) \right| \leq C \|\tilde{p}_j - p\|$ and $\left| \log \left(\frac{1-\tilde{p}_j(x_i)}{1-p_0(x_i)} \right) \right| \leq C \|\tilde{p}_j - p\|$, for some $C > 0$, for all $x \in \mathcal{X}$. Hence, for $j \geq j_0$,

$$\left| \log \left(\frac{\tilde{p}_j(x_i)}{p_0(x_i)} \right) - \log \left(\frac{p(x_i)}{p_0(x_i)} \right) \right| = \left| \log \left(\frac{\tilde{p}_j(x_i)}{p(x_i)} \right) \right| \leq C \|\tilde{p}_j - p\|. \quad (5.A1.5)$$

Now, since H is continuously differentiable, using Taylor's series expansion up to the first order we obtain,

$$\begin{aligned} \|\tilde{p}_j - p\| &= \sup_{x \in \mathcal{X}} |H(\tilde{\eta}_j(x)) - H(\eta(x))| \\ &= \sup_{x \in \mathcal{X}} |H'(u(\tilde{\eta}_j(x), \eta(x)))| \|\tilde{\eta}_j - \eta\|, \end{aligned} \quad (5.A1.6)$$

where $u(\tilde{\eta}_j(x), \eta(x))$ lies between $\eta(x)$ and $\tilde{\eta}_j(x) - \eta(x)$. Since $\|\tilde{\eta}_j - \eta\| \rightarrow 0$, as $j \rightarrow \infty$, it follows from (5.A1.6) that $\|\tilde{p}_j - p\| \rightarrow 0$, as $j \rightarrow \infty$. This again implies, thanks to (5.A1.5), that $\left| \log \left(\frac{\tilde{p}_j(x_i)}{p_0(x_i)} \right) - \log \left(\frac{p(x_i)}{p_0(x_i)} \right) \right| \rightarrow 0$, as $j \rightarrow \infty$.

In other words, $\log \left(\frac{p(x_i)}{p_0(x_i)} \right)$ is continuous in η , and hence $\left\{ \eta : \log \left(\frac{p(x_i)}{p_0(x_i)} \right) < a \right\}$ of (5.A1.4) is measurable. Similarly, $\log \left(\frac{1-p(x_i)}{1-p_0(x_i)} \right)$ is also continuous in η , so that $\left\{ \eta : \log \left(\frac{1-p(x_i)}{1-p_0(x_i)} \right) < a \right\}$ is also measurable. Hence, the individual terms in (5.A1.3) are measurable. Since sums of measurable functions are measurable, it follows that $\log R_n(p)$, and hence $R_n(p)$, is measurable.

5.A1.2 Verification of (S2)

for every $\eta \in \Theta$, we need to show that the KL divergence rate

$$h_1(p) = \lim_{n \rightarrow \infty} \frac{1}{n} E_{p_0} \left[\log \left\{ \frac{f_{p_0}(\mathbf{Y}_n | \mathbf{X}_n)}{f_p(\mathbf{Y}_n | \mathbf{X}_n)} \right\} \right] = \lim_{n \rightarrow \infty} \frac{1}{n} E_{p_0} [-\log \{R_n(p)\}]$$

exists (possibly being infinite) and is \mathcal{T} -measurable.

Now,

$$\begin{aligned} \frac{1}{n} \log R_n(p) &= \frac{1}{n} \sum_{i=1}^n \{(y_i \log p(x_i)) + (1 - y_i) \log (1 - p(x_i))\} \\ &\quad - \frac{1}{n} \sum_{i=1}^n \{(y_i \log p_0(x_i)) + (1 - y_i) \log (1 - p_0(x_i))\}. \end{aligned} \quad (5.A1.7)$$

Therefore,

$$\begin{aligned} \frac{1}{n} E_{p_0} [-\log \{R_n(p)\}] &= \frac{1}{n} \sum_{i=1}^n \{(p_0(x_i) \log p_0(x_i)) + (1 - p_0(x_i)) \log (1 - p_0(x_i))\} \\ &\quad - \frac{1}{n} \sum_{i=1}^n \{(p_0(x_i) \log p(x_i)) + (1 - p_0(x_i)) \log (1 - p(x_i))\}. \end{aligned} \quad (5.A1.8)$$

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} E_{p_0} [-\log \{R_n(p)\}] &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \{(p_0(x_i) \log p_0(x_i)) + (1 - p_0(x_i)) \log (1 - p_0(x_i))\} \\ &\quad - \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \{(p_0(x_i) \log p(x_i)) + (1 - p_0(x_i)) \log (1 - p(x_i))\} \\ &= E_X \{(p_0(X) \log p_0(X)) + (1 - p_0(X)) \log (1 - p_0(X))\} \\ &\quad - E_X \{(p_0(X) \log p(X)) + (1 - p_0(X)) \log (1 - p(X))\}. \end{aligned} \quad (5.A1.9)$$

The last line follows from Assumption 4 and SLLN. Here $E_X(\cdot) = \int_{\mathcal{X}} \cdot dQ$.

Hence,

$$h_1(p) = \left[E_X \left(p_0(X) \log \left\{ \frac{p_0(X)}{p(X)} \right\} \right) + E_X \left((1 - p_0(X)) \log \left\{ \frac{(1 - p_0(X))}{(1 - p(X))} \right\} \right) \right]. \quad (5.A1.10)$$

It is easily seen that h_1 is continuous in η and hence measurable.

5.A1.3 Verification of (S3)

Here we need to verify the asymptotic equipartition, that is, almost surely with respect to f_{p_0} ,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log [R_n(p)] = -h_1(p) = \lim_{n \rightarrow \infty} \frac{1}{n} E_{p_0} \left[\log \left\{ \frac{f_p(\mathbf{Y}_n | \mathbf{X}_n)}{f_{p_0}(\mathbf{Y}_n | \mathbf{X}_n)} \right\} \right]. \quad (5.A1.11)$$

Observe that,

$$\begin{aligned} \frac{1}{n} \log R_n(p) &= \frac{1}{n} \sum_{i=1}^n \{(y_i \log p(x_i)) + (1 - y_i) \log (1 - p(x_i))\} \\ &\quad - \frac{1}{n} \sum_{i=1}^n \{(y_i \log p_0(x_i)) + (1 - y_i) \log (1 - p_0(x_i))\}. \end{aligned}$$

By rearranging the terms we get,

$$-\frac{1}{n} \log R_n(p) = \frac{1}{n} \sum_{i=1}^n \left\{ y_i \log \left(\frac{p_0(x_i)}{p(x_i)} \right) + (1 - y_i) \log \left(\frac{1 - p_0(x_i)}{1 - p(x_i)} \right) \right\}.$$

Using the inequality $1 - \frac{1}{x} \leq \log x \leq x - 1$ for $x > 0$, compactness of \mathcal{X} , and continuity of $p(x)$ in $x \in \mathcal{X}$ for given $\eta \in \Theta$, $\left| \log \left(\frac{p_0(x_i)}{p(x_i)} \right) \right| \leq C \|p - p_0\|$ and $\left| \log \left(\frac{1 - p_0(x_i)}{1 - p(x_i)} \right) \right| \leq$

$C\|p - p_0\|$, for some $C > 0$. Hence,

$$\begin{aligned}
 & \sum_{i=1}^{\infty} i^{-2} \operatorname{Var} \left[\left\{ y_i \log \left(\frac{p_0(x_i)}{p(x_i)} \right) + (1 - y_i) \log \left(\frac{1 - p_0(x_i)}{1 - p(x_i)} \right) \right\} \right] \\
 &= \sum_{i=1}^{\infty} i^{-2} p_0(x_i)(1 - p_0(x_i)) \\
 &\quad \times \left\{ \left[\log \left(\frac{p_0(x_i)}{p(x_i)} \right) \right]^2 + \left[\log \left(\frac{1 - p_0(x_i)}{1 - p(x_i)} \right) \right]^2 - 2 \log \left(\frac{p_0(x_i)}{p(x_i)} \right) \times \log \left(\frac{1 - p_0(x_i)}{1 - p(x_i)} \right) \right\} \\
 &\leq 4C^2 \|p - p_0\|^2 \sum_{i=1}^{\infty} i^{-2} \\
 &< \infty. \tag{5.A1.13}
 \end{aligned}$$

Observe that y_i are realizations of independent random variables. Hence by Kolmogorov's SLLN for independent random variables,

$$\begin{aligned}
 -\frac{1}{n} \log R_n(p) &= \frac{1}{n} \sum_{i=1}^n \left\{ y_i \log \left(\frac{p_0(x_i)}{p(x_i)} \right) + (1 - y_i) \log \left(\frac{1 - p_0(x_i)}{1 - p(x_i)} \right) \right\} \\
 &\rightarrow \left[E_X \left(p_0(X) \log \left\{ \frac{p_0(X)}{p(X)} \right\} \right) + E_X \left((1 - p_0(X)) \log \left\{ \frac{(1 - p_0(X))}{(1 - p(X))} \right\} \right) \right] = h_1(p),
 \end{aligned}$$

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

5.A1.4 Verification of (S4)

If $I = \{\eta : h_1(p) = \infty\}$ then we need to show $\pi(I) < 1$. Note that due to compactness of \mathcal{X} and continuity of H and η , given $\eta \in \Theta$, p is bounded away from 0 and 1, almost surely. Hence, $h_1(p) \leq \|p - p_0\| \times \left(\frac{1}{\inf_{x \in \mathcal{X}} p(x)} + \frac{1}{1 - \sup_{x \in \mathcal{X}} p(x)} \right) < \infty$, almost surely. In other words, (S4) holds. Indeed, note that under Assumption 6, $I = \emptyset$.

5.A1.5 Verification of (S5)

In our model, the parameter space is $\Theta = \mathcal{C}'(\mathcal{X})$. We need to show that there exists a sequence of sets $\mathcal{G}_n \rightarrow \Theta$ as $n \rightarrow \infty$ such that:

1. $h_1(\mathcal{G}_n) \rightarrow h_1(\Theta)$, as $n \rightarrow \infty$.
2. The inequality $\pi(\mathcal{G}_n) \geq 1 - \alpha \exp(-\beta n)$ holds for some $\alpha > 0, \beta > 2h(\Theta)$.
3. The convergence in (S3) is uniform in η over $\mathcal{G}_n \setminus I$.

Recall that in our case,

$$\mathcal{G}_n = \left\{ \eta \in \mathcal{C}'(\mathcal{X}) : \|\eta\| \leq \exp((\beta n)^{1/4}), \|\eta'_j\| \leq \exp((\beta n)^{1/4}); j = 1, 2, \dots, d \right\}.$$

so that $\mathcal{G}_n \rightarrow \Theta$ as $n \rightarrow \infty$.

Verification of (S5) (1)

We now verify that $h_1(\mathcal{G}_n) \rightarrow h_1(\Theta)$, as $n \rightarrow \infty$. Recall from our verification of (S2) that $h_1(p)$ is continuous in η . Hence, continuity of $h_1(p)$, compactness of \mathcal{G}_n along with its non-decreasing nature with respect to n implies that $h_1(\mathcal{G}_n) \rightarrow h_1(\Theta)$, as $n \rightarrow \infty$.

Verification of (S5) (2)

$$\begin{aligned}
\pi(\mathcal{G}_n) &= \pi \left(\|\eta\| \leq \exp((\beta n)^{1/4}) \right) \\
&\quad - \pi \left(\|\eta'_j\| \leq \exp((\beta n)^{1/4}); j = 1, 2, \dots, d \right) \\
&= \pi \left(\|\eta\| \leq \exp((\beta n)^{1/4}), \right) \\
&\quad - \pi \left(\bigcup_{j=1}^d \left\{ \|\eta'_j\| \leq \exp((\beta n)^{1/4}) \right\} \right) \\
&\geq 1 - \pi \left(\|\eta\| > \exp((\beta n)^{1/4}) \right) - \sum_{j=1}^d \pi \left(\|\eta'_j\| \leq \exp((\beta n)^{1/4}) \right) \\
&\geq 1 - \left(c_\eta + \sum_{j=1}^d c_{\eta'_j} \right) \exp(-\beta n).
\end{aligned}$$

where the last inequality follows from Assumption 3.

Verification of (S5) (3)

We need to show that uniform convergence in (S3) in η over $\mathcal{G}_n \setminus I$ holds, where $I = \{\eta : h(p) = \infty\}$ as in Section 5.A1.4. In our case, $I = \emptyset$ under Assumption 6. Hence, we need to show uniform convergence in (S3) in η over \mathcal{G}_n . We need to establish that \mathcal{G}_n is compact, but this has already been shown in Chapter 4. Indeed, recall that we proved compactness of \mathcal{G}_n for each $n \geq 1$ by showing that \mathcal{G}_n is closed, bounded and equicontinuous and then by using Arzela-Ascoli lemma to imply compactness. It should be noted that boundedness of the partial derivatives as in Assumption 1 is used to show Lipschitz continuity, hence equicontinuity.

Consider $\mathcal{G} \in \{\mathcal{G}_n : n = 1, 2, \dots\}$. Now, to show uniform convergence we only need to show the following:

- (i) $\frac{1}{n} \log R_n(p) + h_1(p)$ is stochastically equicontinuous almost surely in $\eta \in \mathcal{G}$,

(ii) $\frac{1}{n} \log R_n(p) + h_1(p) \rightarrow 0$ for all $\eta \in \mathcal{G}$ as $n \rightarrow \infty$.

We have already shown almost sure pointwise convergence of $n^{-1} \log R_n(p)$ to $-h_1(p)$ in Appendix 5.A1.3. Hence it is enough to verify stochastic equicontinuity of $\frac{1}{n} \log R_n(p) + h_1(p)$ in $\mathcal{G} \in \{\mathcal{G}_n : n = 1, 2, \dots\}$. Stochastic equicontinuity usually follows easily if one can prove that the function concerned is almost surely Lipschitz continuous. Observe that, if we can show that both $\frac{1}{n} \log R_n(p)$ and $h_1(p)$ are Lipschitz in η , then this would imply that $\frac{1}{n} \log R_n(p) + h_1(p)$ is Lipschitz (sum of Lipschitz functions is Lipschitz).

We now show that $\frac{1}{n} \log R_n(p)$ and $h_1(p)$ are both Lipschitz on \mathcal{G} . Note that,

$$\frac{1}{n} \log R_n(p) = \frac{1}{n} \sum_{i=1}^n \left\{ y_i \log \left(\frac{p(x_i)}{p_0(x_i)} \right) + (1 - y_i) \log \left(\frac{1 - p(x_i)}{1 - p_0(x_i)} \right) \right\}. \quad (5.A1.14)$$

Let p_1, p_2 correspond to $\eta_1, \eta_2 \in \Theta$. Since $0 < \kappa_B \leq p_1(x), p_2(x) \leq 1 - \kappa_B < 1$, for all $x \in \mathcal{X}$, there exists $C > 0$ such that $\left| \log \left(\frac{p_1(x)}{p_2(x)} \right) \right| \leq C \|p_1 - p_2\|$ and $\left| \log \left(\frac{1 - p_1(x)}{1 - p_2(x)} \right) \right| \leq C \|p_1 - p_2\|$, for $x \in \mathcal{X}$. Hence,

$$\begin{aligned} & \left| \frac{1}{n} \log R_n(p_1) - \frac{1}{n} \log R_n(p_2) \right| \\ &= \left| \frac{1}{n} \sum_{i=1}^n \left\{ y_i \log \left(\frac{p_1(x_i)}{p_2(x_i)} \right) + (1 - y_i) \log \left(\frac{1 - p_1(x_i)}{1 - p_2(x_i)} \right) \right\} \right| \\ &\leq 2C \|p_1 - p_2\|, \end{aligned}$$

showing Lipschitz continuity of $\frac{1}{n} \log R_n(p)$ with respect to p corresponding to $\eta \in \mathcal{G} = \mathcal{G}_m$. Since H is continuously differentiable, η and η' are bounded on \mathcal{G} , with the same bound for all η , it follows that p is Lipschitz on \mathcal{G} .

To see that $h_1(p)$ is also Lipschitz in $\mathcal{G} = \mathcal{G}_m$, it is enough to note that

$$\begin{aligned} |h_1(p_1) - h_1(p_2)| &= \left| E_X \left(p_0(X) \log \left(\frac{p_2(X)}{p_1(X)} \right) \right) + E_X \left((1 - p_0(X)) \log \left(\frac{1 - p_2(X)}{1 - p_1(X)} \right) \right) \right| \\ &\leq 2C\|p_1 - p_2\|, \end{aligned}$$

and the result follows since p is Lipschitz on \mathcal{G} .

5.A1.6 Verification of (S6)

We need to show:

$$\sum_{n=1}^{\infty} \int_{\mathcal{S}^c} P \left(\left| \frac{1}{n} \log R_n(p) + h_1(p) \right| > \kappa - h_1(\Theta) \right) d\pi(\eta) < \infty. \quad (5.A1.15)$$

Let us take $\kappa_1 = \kappa - h(\Theta)$. Observe that,

$$\begin{aligned} &\frac{1}{n} \log R_n(p) + h_1(p) \\ &= \frac{1}{n} \sum_{i=1}^n \left\{ y_i \log \left(\frac{p(x_i)}{p_0(x_i)} \right) + (1 - y_i) \log \left(\frac{1 - p(x_i)}{1 - p_0(x_i)} \right) \right\} \\ &\quad + \left[E_X \left(p_0(X) \log \left\{ \frac{p_0(X)}{p(X)} \right\} \right) + E_X \left((1 - p_0(X)) \log \left\{ \frac{(1 - p_0(X))}{(1 - p(X))} \right\} \right) \right] \\ &= \frac{1}{n} \sum_{i=1}^n \left\{ y_i \log \left(\frac{p(x_i)}{p_0(x_i)} \right) - E_X \left(p_0(X) \log \left\{ \frac{p(X)}{p_0(X)} \right\} \right) \right\} \\ &\quad + \frac{1}{n} \sum_{i=1}^n \left\{ (1 - y_i) \log \left(\frac{1 - p(x_i)}{1 - p_0(x_i)} \right) - E_X \left((1 - p_0(X)) \log \left\{ \frac{(1 - p(X))}{(1 - p_0(X))} \right\} \right) \right\}. \end{aligned}$$

5.A1. VERIFICATION OF ASSUMPTIONS (S1) TO (S7) OF SHALIZI FOR
BINARY REGRESSION

It follows that:

$$P \left(\left| \frac{1}{n} \log R_n(p) + h_1(p) \right| > \kappa_1 \right) \quad (5.A1.16)$$

$$\leq P \left(\left| \frac{1}{n} \sum_{i=1}^n \left\{ y_i \log \left(\frac{p(x_i)}{p_0(x_i)} \right) - E_X \left(p_0(X) \log \left\{ \frac{p(X)}{p_0(X)} \right\} \right) \right\} \right| > \frac{\kappa_1}{2} \right) \quad (5.A1.17)$$

$$+ P \left(\left| \frac{1}{n} \sum_{i=1}^n \left\{ (1 - y_i) \log \left(\frac{1 - p(x_i)}{1 - p_0(x_i)} \right) - E_X \left((1 - p_0(X)) \log \left\{ \frac{(1 - p(X))}{(1 - p_0(X))} \right\} \right) \right\} \right| > \frac{\kappa_1}{2} \right). \quad (5.A1.18)$$

Since y_i are binary, it follows using the inequalities $1 - \frac{1}{x} \leq \log x \leq x - 1$, for $x > 0$ and Assumptions 5 and 6, that the random variables $V_i = y_i \log \left(\frac{p(x_i)}{p_0(x_i)} \right)$ and $W_i = y_i \log \left(\frac{1 - p(x_i)}{1 - p_0(x_i)} \right)$ are absolutely bounded by $C \|p - p_0\|$, for some $C > 0$. We shall apply Hoeffding's inequality (Hoeffding (1963)) separately on the two terms of (5.A1.18) involving V_i and W_i .

Note that for $\eta \in \mathcal{G}_n$,

$$\begin{aligned} & P \left(\left| \frac{1}{n} \sum_{i=1}^n \left\{ y_i \log \left(\frac{p(x_i)}{p_0(x_i)} \right) - E_X \left(p_0(X) \log \left\{ \frac{p(X)}{p_0(X)} \right\} \right) \right\} \right| > \frac{\kappa_1}{2} \right) \\ & \leq P \left(\left| \frac{1}{n} \sum_{i=1}^n \left\{ y_i \log \left(\frac{p(x_i)}{p_0(x_i)} \right) - p_0(x_i) \log \left(\frac{p(x_i)}{p_0(x_i)} \right) \right\} \right| > \frac{\kappa_1}{4} \right) \\ & \quad + P \left(\left| \frac{1}{n} \sum_{i=1}^n \left\{ p_0(x_i) \log \left(\frac{p(x_i)}{p_0(x_i)} \right) - E_X \left(p_0(X) \log \left\{ \frac{p(X)}{p_0(X)} \right\} \right) \right\} \right| > \frac{\kappa_1}{4} \right) \\ & \leq 4 \exp \left\{ - \frac{n \kappa_1^2}{8C^2 \|p - p_0\|^2} \right\} \leq 4 \exp \left\{ - \frac{n \kappa_1^2}{8C^2 L^2 \|\eta - \eta_0\|^2} \right\}, \end{aligned} \quad (5.A1.19)$$

where $L > 0$ is the Lipschitz constant associated with H . Here it is important to note that for $\eta \in \mathcal{G}_n$, $H(\eta)$ is Lipschitz in η thanks to continuous differentiability of H , and boundedness of η and η' by the same constant on \mathcal{G}_n . Also note that (5.A1.19) holds irrespective of x_i ; $i = 1, \dots, n$ being random or non-random (see also Chapter 4).

Similarly, for $\eta \in \mathcal{G}_n$,

$$\begin{aligned} & P \left(\left| \frac{1}{n} \sum_{i=1}^n \left\{ (1 - y_i) \log \left(\frac{1 - p(x_i)}{1 - p_0(x_i)} \right) - E_X \left((1 - p_0(X)) \log \left\{ \frac{(1 - p(X))}{(1 - p_0(X))} \right\} \right) \right\} \right| > \frac{\kappa_1}{2} \right) \\ & \leq 4 \exp \left\{ - \frac{n\kappa_1^2}{8C^2L^2\|\eta - \eta_0\|^2} \right\}. \end{aligned} \quad (5.A1.20)$$

Now,

$$\begin{aligned} & \sum_{n=1}^{\infty} \int_{\mathcal{S}^c} P \left(\left| \frac{1}{n} \sum_{i=1}^n \left\{ y_i \log \left(\frac{p(x_i)}{p_0(x_i)} \right) - E_X \left(p_0(X) \log \left\{ \frac{p(X)}{p_0(X)} \right\} \right) \right\} \right| > \frac{\kappa_1}{2} \right) d\pi(p) \\ & \leq \sum_{n=1}^{\infty} \int_{\mathcal{G}_n} 4 \exp \left\{ - \frac{n\kappa_1^2}{8C^2L^2\|\eta - \eta_0\|^2} \right\} d\pi(\eta) + \sum_{n=1}^{\infty} \pi(\mathcal{G}_n^c), \end{aligned} \quad (5.A1.21)$$

and

$$\begin{aligned} & \sum_{n=1}^{\infty} \int_{\mathcal{S}^c} P \left(\left| \frac{1}{n} \sum_{i=1}^n \left\{ (1 - y_i) \log \left(\frac{1 - p(x_i)}{1 - p_0(x_i)} \right) - E_X \left((1 - p_0(X)) \log \left\{ \frac{1 - p(X)}{1 - p_0(X)} \right\} \right) \right\} \right| > \frac{\kappa_1}{2} \right) d\pi(p) \\ & \leq \sum_{n=1}^{\infty} \int_{\mathcal{G}_n} 4 \exp \left\{ - \frac{n\kappa_1^2}{8C^2L^2\|\eta - \eta_0\|^2} \right\} d\pi(\eta) + \sum_{n=1}^{\infty} \pi(\mathcal{G}_n^c). \end{aligned} \quad (5.A1.22)$$

Then proceeding in the same way as in the corresponding situation in Chapter 4, and noting that $\sum_{n=1}^{\infty} \pi(\mathcal{G}_n^c) < \infty$, we obtain (5.A1.15).

Hence (S6) holds.

Remark 5.A1.1 It is important to clarify the role of Assumption 6 here. For instance, let $H(\eta(x)) = \frac{\exp(\eta(x))}{1+\exp(\eta(x))}$, and let $\|\eta\| \leq \sqrt{\beta n}$ on \mathcal{G}_n , for simplicity of exposition. Then with our bounding method using the inequality $\log x \geq 1 - 1/x$ for $x > 0$, and noting that both $-\eta(x) \leq \|\eta\| \leq \sqrt{\beta n}$ and $\eta(x) \leq \|\eta\| \leq \sqrt{\beta n}$, we have $\log \left(\frac{p(x)}{p_0(x)} \right) \geq -\frac{\|p-p_0\|}{p(x)} \geq -2 \exp(2\sqrt{\beta n}) \|p - p_0\|$. Using $\log x \leq x - 1$ for $x > 0$, we obtain $\log \left(\frac{p(x)}{p_0(x)} \right) \leq 2 \exp(2\sqrt{\beta n}) \|p - p_0\|$. Thus, $\left| \log \left(\frac{p(x)}{p_0(x)} \right) \right| \leq 2 \exp(2\sqrt{\beta n}) \|p - p_0\|$. It would then follow that the exponent of the Hoeffding inequality is $o(1)$. This would fail to ensure

summability of the corresponding terms involving V_i . Thus, we need to ensure that $p(x)$ is bounded away from 0. Similarly, the infinite sum associated with W_i would not be finite unless $1 - p(x)$ is bounded away from 0.

5.A1.7 Verification of (S7)

This verification follows from the fact that $h_1(p)$ is continuous. Indeed, for any set A with $\pi(A) > 0$, $\mathcal{G}_n \cap A \uparrow A$. It follows from continuity of h_1 that $h_1(\mathcal{G}_n \cap A) \downarrow h_1(A)$ as $n \rightarrow \infty$ and hence (S7) holds.

5.A2 Verification of Assumptions (S1) to (S7) of Shalizi for Poisson regression

5.A2.1 Verification of (S1)

Observe that

$$f_\lambda(\mathbf{Y}_n | \mathbf{X}_n) = \prod_{i=1}^n f(y_i | x_i) = \prod_{i=1}^n \exp(-\lambda(x_i)) \frac{(\lambda(x_i))^{y_i}}{y_i!},$$

$$f_{\lambda_0}(\mathbf{Y}_n | \mathbf{X}_n) = \prod_{i=1}^n f_0(y_i | x_i) = \prod_{i=1}^n \exp(-\lambda_0(x_i)) \frac{(\lambda_0(x_i))^{y_i}}{y_i!}.$$

Therefore,

$$R_n(\lambda) = \exp \left(- \sum_{i=1}^n [\lambda(x_i) - \lambda_0(x_i)] \right) \prod_{i=1}^n \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right)^{y_i} \quad (5.A2.1)$$

and

$$\frac{1}{n} \log R_n(\lambda) = \left(-\frac{1}{n} \sum_{i=1}^n [\lambda(x_i) - \lambda_0(x_i)] \right) + \frac{1}{n} \sum_{i=1}^n y_i \log \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right). \quad (5.A2.2)$$

Note that for any $a \in \mathbb{R}$, $\left\{ (y_i, \eta) : y_i \log \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right) < a \right\} = \bigcup_{r=1}^{\infty} \left\{ \eta : r \log \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right) < a \right\}$.

Let $\tilde{\eta}_j$; $j = 1, 2, \dots$ be such that $\|\eta_j - \eta\| \rightarrow 0$, as $j \rightarrow \infty$. Then, letting $\tilde{\lambda}_j(x) = H(\tilde{\eta}_j(x))$, for all $x \in \mathcal{X}$, it follows, since $0 < C_1 \leq \lambda(x) \leq C_2 < \infty$ on \mathcal{X} , that there exists $j_0 \geq 1$ such that for $j \geq j_0$, $0 < C_1 \leq \tilde{\lambda}_j(x) \leq C_2 < \infty$. Hence, using the inequalities $1 - \frac{1}{x} \leq \log x \leq x - 1$ for $x > 0$, we obtain $\left| \log \left(\frac{\tilde{\lambda}_j(x_i)}{\lambda(x_i)} \right) \right| \leq C \|\tilde{\lambda}_j - \lambda\|$, for some $C > 0$, for $j \geq j_0 \geq 1$. It follows that

$$\left| r \log \left(\frac{\tilde{\lambda}_j(x_i)}{\lambda_0(x_i)} \right) - r \log \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right) \right| = r \left| \log \left(\frac{\tilde{\lambda}_j(x_i)}{\lambda(x_i)} \right) \right| \leq rC \|\tilde{\lambda}_j - \lambda\| \rightarrow 0,$$

in the same way as in the binary regression, using Taylor's series expansion up to the first order. Hence, $r \log \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right)$ is continuous in η , ensuring measurability of $\left\{ \eta : r \log \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right) < a \right\}$, and hence of $\left\{ (y_i, \eta) : y_i \log \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right) < a \right\}$. It follows that $\frac{1}{n} \sum_{i=1}^n y_i \log \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right)$ is measurable.

Also, continuity of $\lambda(x_i) - \lambda_0(x_i)$ with respect to η ensures measurability of $-\frac{1}{n} \sum_{i=1}^n [\lambda(x_i) - \lambda_0(x_i)]$. Thus, $\frac{1}{n} \log R_n(\lambda)$, and hence $R_n(\lambda)$, is measurable.

5.A2.2 Verification of (S2)

For every $\eta \in \Theta$, we need to show that the KL divergence rate

$$h_2(\lambda) = \lim_{n \rightarrow \infty} \frac{1}{n} E_{\lambda_0} \left[\log \left\{ \frac{f_{\lambda_0}(\mathbf{Y}_n | \mathbf{X}_n)}{f_{\lambda}(\mathbf{Y}_n | \mathbf{X}_n)} \right\} \right] = \lim_{n \rightarrow \infty} \frac{1}{n} E_{\lambda_0} [-\log \{R_n(\lambda)\}]$$

exists (possibly being infinite) and is \mathcal{T} -measurable.

Now,

$$\frac{1}{n} \log R_n(\lambda) = \left(-\frac{1}{n} \sum_{i=1}^n [\lambda(x_i) - \lambda_0(x_i)] \right) + \frac{1}{n} \sum_{i=1}^n y_i \log \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right).$$

Therefore,

$$\frac{1}{n} E_{\lambda_0} [-\log \{R_n(\lambda)\}] = \left(\frac{1}{n} \sum_{i=1}^n [\lambda(x_i) - \lambda_0(x_i)] \right) + \frac{1}{n} \sum_{i=1}^n \lambda_0(x_i) \log \left(\frac{\lambda_0(x_i)}{\lambda(x_i)} \right).$$

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} E_{\lambda_0} [-\log \{R_n(\lambda)\}] &= \lim_{n \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n [\lambda(x_i) - \lambda_0(x_i)] \right) + \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \lambda_0(x_i) \log \left(\frac{\lambda_0(x_i)}{\lambda(x_i)} \right) \\ &= E_X [\lambda(X) - \lambda_0(X)] + E_X \left[\lambda_0(X) \log \left(\frac{\lambda_0(X)}{\lambda(X)} \right) \right]. \end{aligned}$$

The last line holds due to Assumption 4 and SLLN. Here $E_X(\cdot) = \int_{\mathcal{X}} \cdot dQ$. In other words,

$$h_2(\lambda) = E_X [\lambda(X) - \lambda_0(X)] + E_X \left[\lambda_0(X) \log \left(\frac{\lambda_0(X)}{\lambda(X)} \right) \right]. \quad (5.A2.3)$$

It is easily seen that h_2 is continuous in η , and hence measurable.

5.A2.3 Verification of (S3)

Here we need to verify the asymptotic equipartition property, that is, almost surely with respect to the true model f_{λ_0} ,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log [R_n(\lambda)] = -h_2(\lambda) = \lim_{n \rightarrow \infty} \frac{1}{n} E_{\lambda_0} \left[\log \left\{ \frac{f_{\lambda}(\mathbf{Y}_n | \mathbf{X}_n)}{f_{\lambda_0}(\mathbf{Y}_n | \mathbf{X}_n)} \right\} \right]. \quad (5.A2.4)$$

Now,

$$-\frac{1}{n} \log R_n(\lambda) = \frac{1}{n} \sum_{i=1}^n \left\{ [\lambda(x_i) - \lambda_0(x_i)] + y_i \log \left(\frac{\lambda_0(x_i)}{\lambda(x_i)} \right) \right\}.$$

As before, for given λ , there exists $C > 0$ such that $\left| \log \left(\frac{\lambda_0(x_i)}{\lambda(x_i)} \right) \right| \leq C \|\lambda - \lambda_0\|$. Hence,

$$\begin{aligned}
 & \sum_{i=1}^{\infty} i^{-2} \operatorname{Var} \left[\left\{ [\lambda(x_i) - \lambda_0(x_i)] + y_i \log \left(\frac{\lambda_0(x_i)}{\lambda(x_i)} \right) \right\} \right] \\
 &= \sum_{i=1}^{\infty} i^{-2} \lambda_0(x_i) \left[\log \left(\frac{\lambda_0(x_i)}{\lambda(x_i)} \right) \right]^2 \\
 &\leq C^2 H(\kappa_0) (\|\lambda - \lambda_0\|)^2 \sum_{i=1}^{\infty} i^{-2} \\
 &< \infty. \tag{5.A2.5}
 \end{aligned}$$

Observe that y_i are observations from independent random variables. Hence from Kolmogorov's SLLN for independent random variables and from Assumption 4, (5.A2.4) holds as $n \rightarrow \infty$.

5.A2.4 Verification of (S4)

If $I = \{\eta : h_2(\lambda) = \infty\}$ then we need to show $\pi(I) < 1$. But this holds in almost the same way as for binary regression. In other words, (S4) holds for Poisson regression.

5.A2.5 Verification of (S5)

We need to verify that

1. $h_2(\mathbb{G}_n) \rightarrow h_2(\Theta)$, as $n \rightarrow \infty$;
2. The inequality $\pi(\mathbb{G}_n) \geq 1 - \alpha \exp(-\beta n)$ holds for some $\alpha > 0, \beta > 2h_2(\Theta)$;
3. The convergence in (S3) is uniform over $\mathbb{G}_n \setminus I$.

Verification of (S5) (1)

We now need to verify that $h_2(\mathbb{G}_n) \rightarrow h_2(\Theta)$ as $n \rightarrow \infty$. But this holds in the same way as for binary regression.

Verification of (S5) (2)

Again, this holds in the same way as for binary regression.

Verification of (S5) (3)

Using the same arguments as in the binary regression case, here we only need to show that $\frac{1}{n} \log(R_n(\lambda))$ and $h_2(\lambda)$ are both Lipschitz.

Recall that

$$\frac{1}{n} \log R_n(\lambda) = \frac{1}{n} \sum_{i=1}^n \left\{ [\lambda_0(x_i) - \lambda(x_i)] + y_i \log \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right) \right\}.$$

For any $\eta_1, \eta_2 \in \mathbb{G}$, there exists $C > 0$ such that $\left| \log \left(\frac{\lambda_1(x)}{\lambda_2(x)} \right) \right| \leq C \|\lambda_1 - \lambda_2\|$, for all $x \in \mathcal{X}$, where $\lambda_1 = H(\eta_1)$ and $\lambda_2 = H(\eta_2)$. Hence,

$$\left| \frac{1}{n} \log R_n(\lambda_1) - \frac{1}{n} \log R_n(\lambda_2) \right| \leq \|\lambda_1 - \lambda_2\| \left(1 + C \times \frac{1}{n} \sum_{i=1}^n y_i \right).$$

Thus, $\frac{1}{n} \log R_n(\lambda)$ is almost surely Lipschitz with respect to λ . Since, by Kolmogorov's SLLN for independent variables, $\frac{1}{n} \sum_{i=1}^n y_i \xrightarrow{a.s.} E_X(\lambda_0(X)) < \infty$, as $n \rightarrow \infty$, and since $\lambda = H(\eta)$ is Lipschitz in $\eta \in \mathbb{G}_n$ in the same way as in binary regression, the desired stochastic equicontinuity follows. Lipschitz continuity of $h_2(\lambda)$ in \mathbb{G}_n follows using similar techniques.

5.A2.6 Verification of (S6)

Since

$$\begin{aligned}
 & \sum_{n=1}^{\infty} \int_{\mathcal{S}^c} P \left(\left| \frac{1}{n} \log R_n(\lambda) + h_2(\lambda) \right| > \kappa - h_2(\Theta) \right) d\pi(\eta) \\
 & \leq \sum_{n=1}^{\infty} \int_{\mathbb{G}_n} P \left(\left| \frac{1}{n} \log R_n(\lambda) + h_2(\lambda) \right| > \kappa - h_2(\Theta) \right) d\pi(\eta) \\
 & \quad + \sum_{n=1}^{\infty} \int_{\mathbb{G}_n^c} P \left(\left| \frac{1}{n} \log R_n(\lambda) + h_2(\lambda) \right| > \kappa - h_2(\Theta) \right) d\pi(\eta) \\
 & \leq \sum_{n=1}^{\infty} \int_{\mathbb{G}_n} P \left(\left| \frac{1}{n} \log R_n(\lambda) + h_2(\lambda) \right| > \kappa - h_2(\Theta) \right) d\pi(\eta) + \sum_{n=1}^{\infty} \pi(\mathbb{G}_n^c), \quad (5.A2.6)
 \end{aligned}$$

and the second term of (5.A2.6) is finite, it is enough to show that the first term of (5.A2.6) is finite.

Let us take $\kappa_1 = \kappa - h_2(\Theta)$. Observe that for $\eta \in \mathbb{G}_n$,

$$\begin{aligned}
 & P \left(\left| \frac{1}{n} \log R_n(\lambda) + h_2(\lambda) \right| > \kappa_1 \right) \\
 & \leq P \left(\left| \frac{1}{n} \sum_{i=1}^n \left[\lambda_0(x_i) \log \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right) - E_X \left(\lambda_0(X) \log \left(\frac{\lambda(X)}{\lambda_0(X)} \right) \right) \right] \right| > \frac{\kappa_1}{3} \right) \quad (5.A2.7)
 \end{aligned}$$

$$+ P \left(\left| \frac{1}{n} \sum_{i=1}^n [(\lambda_0(x_i) - \lambda(x_i)) - E_X(\lambda_0(X) - \lambda(X))] \right| > \frac{\kappa_1}{3} \right) \quad (5.A2.8)$$

$$+ P \left(\left| \frac{1}{n} \sum_{i=1}^n \left[y_i \log \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right) - \lambda_0(x_i) \log \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right) \right] \right| > \frac{\kappa_1}{3} \right). \quad (5.A2.9)$$

Using Hoeffding's inequality and Lipschitz continuity of H in \mathbb{G}_n as in binary regression, we find that (5.A2.7) and (5.A2.8) are bounded above by $2 \exp \left(-\frac{C_1 n \kappa_1^2}{\|\eta - \eta_0\|^2} \right)$, and $\exp \left(-\frac{C_2 n \kappa_1^2}{\|\eta - \eta_0\|^2} \right)$, for some $C_1 > 0$ and $C_2 > 0$. These bounds hold even if the covariates are non-random.

To bound (5.A2.9), we shall first show that the summands are sub-exponential, and then shall apply Bernstein's inequality (see, for example, Uspensky (1937), Bennett

(1962), Massart (2003)). Direct calculation yields

$$\begin{aligned} & E \left[\exp \left\{ t \left(y_i \log \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right) - \lambda_0(x_i) \log \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right) \right) \right\} \right] \\ &= \exp \left[-t \lambda_0(x_i) \log \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right) \right] \times \exp \left[\lambda_0(x_i) \left\{ \exp \left(t \log \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right) \right) - 1 \right\} \right]. \end{aligned} \quad (5.A2.10)$$

The first factor of (5.A2.10) has the following upper bound:

$$\exp \left[-t \lambda_0(x_i) \log \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right) \right] \leq \exp (c_\lambda \|\lambda\| \times |t|). \quad (5.A2.11)$$

A bound for the second factor of (5.A2.10) is given as follows:

$$\begin{aligned} & \exp \left[\lambda_0(x_i) \left\{ \exp \left(t \log \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right) \right) - 1 \right\} \right] \\ & \leq \exp \left[\|\lambda_0\| \left(\exp \left(\frac{t \|\lambda - \lambda_0\|}{\kappa_P} \right) - 1 \right) \right] \\ & \leq \exp [\|\lambda_0\| (c_\lambda |t| + c_\lambda^2 t^2)], \end{aligned} \quad (5.A2.12)$$

for $|t| \leq c_\lambda^{-1}$, where $c_\lambda = C \|\lambda - \lambda_0\|$, for some $C > 0$.

Combining (5.A2.10), (5.A2.11) and (5.A2.12) we see that (5.A2.10) is bounded above by $\exp (c_\lambda^2 t^2)$ provided that

$$c_\lambda |t| \geq 2 / (\|\lambda_0\|^{-1} - 1) \geq 2 / (\kappa_P^{-1} - 1). \quad (5.A2.13)$$

The rightmost bound of (5.A2.13) is close to zero if κ_P is chosen sufficiently small. Now consider the function $g(t) = \exp (c_\lambda^2 t^2) - f(t)$, where $f(t)$ is given by (5.A2.10). Since $g(t)$ is continuous in t and $g(0) = 0$ and $g(t) > 0$ on $2 / (\kappa_P^{-1} - 1) \leq |t| \leq c_\lambda^{-1}$, it follows that on the sufficiently small interval $0 \leq |t| \leq 2 / (\kappa_P^{-1} - 1)$, $g(t) > 0$. In other words, (5.A2.10) is bounded above by $\exp (c_\lambda^2 t^2)$ for $0 \leq |t| \leq c_\lambda^{-1}$. Thus,

$z_i = y_i \log \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right) - \lambda_0(x_i) \log \left(\frac{\lambda(x_i)}{\lambda_0(x_i)} \right)$ are independent sub-exponential variables with parameter c_λ .

Bernstein's inequality, in conjunction with Lipschitz continuity of H on \mathbb{G}_n then ensures that (5.A2.9) is bounded above by $2 \exp \left[-\frac{n}{2} \min \left\{ \frac{C_1 \kappa_1^2}{\|\eta - \eta_0\|^2}, \frac{C_2 \kappa_1}{\|\eta - \eta_0\|} \right\} \right]$, for positive constants C_1 and C_2 .

The rest of the proof of finiteness of (5.A2.6) follows in the same (indeed, simpler) way as in Chapter 4. Hence (S6) holds.

Remark 5.A2.1 *Arguments similar to that of Remark 5.A1.1 shows that it is essential to have λ bounded away from zero.*

5.A2.7 Verification of (S7)

This verification follows from the fact that $h_2(\lambda)$ is continuous, similar to binary regression.

6

Posterior Consistency of Bayesian Inverse Regression and Inverse Reference Distributions

6.1 Introduction

As already pointed out in Chapter 1, the literature on goodness-of-fit for inverse regression models is non-existent, except for the IRD approach of [Bhattacharya \(2013\)](#), the basic premise with the LOO-CV setup and the key idea of which are discussed in Section 1.3.2. In this chapter we develop the asymptotic theory of IRD; in particular, we establish consistency of the IRD approach in the sense that with probability tending to one as the sample size tends to infinity, the approach declares the goodness-of-fit of the correct, data-generating model as satisfactory and the wrong models as unsatisfactory.

Our asymptotic theory of IRD relies on consistency of the LOO-CV posteriors associated with the covariates, and it has been shown in Section 1.3.1 that such consistency does not hold for general priors for the covariates, considered to be unknown for the sake of Bayesian cross-validation. In this chapter we introduce a specialized class of priors that depend upon the data as well as on the unknown model parameters, using which we establish consistency of the LOO-CV posteriors associated with the covariates.

Note that the LOO-CV posteriors, as well as the specialized prior for the unknown covariates, may involve unknown functions, modeled nonparametrically by appropriate stochastic processes, posterior consistency of which is required for our asymptotic theory of the LOO-CV posteriors of the unknown covariates, and hence of the IRD approach. In this regard, our posterior convergence results with respect to Gaussian and other general stochastic processes under different model setups like normal, double exponential, binary and Poisson, provide the necessary technical support. Note that unknown functions embedded in the inverse LOO-CV of the unknown covariates and the IRD approach also vindicate that inverse regression problems contain the traditional inverse problems as special cases, as already pointed out in Chapter 1.

Not only do we establish asymptotic results, we conduct adequate simulation experiments that uphold our methods and asymptotic investigations. In particular, we demonstrate consistency of the LOO-CV posteriors of the unknown covariates with our specialized prior using simulation studies under both parametric and nonparametric setups, which would in turn induce consistency of the respective IRD strategies.

The rest of this chapter is structured as follows. The general premise of our inverse regression model, LOO-CV and the IRD approach are described in Section 6.2. General consistency issues of the same are discussed in Section 6.3. We propose an appropriate prior for \tilde{x}_i and investigate its properties in Section 6.4, and in Section 6.5 prove consistency of the LOO-CV posteriors under reasonably mild conditions. Relating consistency of the LOO-CV posteriors, we prove consistency of the IRD approach in

Section 6.6. In Section 6.7 we provide a discussion on the issues and applicability of our asymptotic theory in various inverse regression contexts and in Section 6.8, we illustrate our asymptotic theory with simulation studies. Finally, we make concluding remarks in Section 6.9.

6.2 Preliminaries and the general setup

We consider experiment with n covariate observations x_1, x_2, \dots, x_n along with responses $\{y_{ij} : 1 \leq i \leq n, 1 \leq j \leq m\}$. In other words, the experiment considered here will allow us to have m samples of responses $\{y_{i1}, y_{i2}, \dots, y_{im}\}$ against covariate observations x_i , for $i = 1, 2, \dots, n$. Both x_i and y_{ij} are allowed to be multidimensional. In this chapter, we consider the large sample scenario where both $m, n \rightarrow \infty$.

For $i = 1, \dots, n$ and $j = 1, \dots, m$, consider the following general model setup: conditionally on x_i and θ ,

$$y_{ij} \sim f_\theta(x_i), \quad (6.2.1)$$

independently. In (6.2.1), f_θ is a known distribution depending upon (a set of) parameters $\theta \in \Theta$, where Θ is the parameter space, which may be infinite-dimensional. For the sake of generality, we shall consider $\theta = (\eta, \xi)$, where η is a function of the covariates, which we more explicitly denote as $\eta(x)$, where $x \in \mathcal{X}$, \mathcal{X} being the space of covariates. The part ξ of θ will be assumed to consist of other parameters, such as the unknown error variance.

6.2.1 Examples of the above model setup

- (i) $y_{ij} \sim \text{Poisson}(\theta x_i)$, where $\theta > 0$ and $x_i > 0$ for all i .
- (ii) $y_{ij} \sim \text{Bernoulli}(p_i)$, where $p_i = H(\eta(x_i))$, where H is some appropriate link function and η is some function with known or unknown form. For known, suitably

parameterized form, the model is parametric. If the form of η is unknown, one may model it by a Gaussian process, assuming adequate smoothness of the function.

- (iii) $y_{ij} \sim \text{Poisson}(\lambda_i)$, where $\lambda_i = H(\eta(x_i))$, where H is some appropriate link function and η is some function with known (parametric) or unknown (nonparametric) form. Again, in case of unknown form of η , the Gaussian process can be used as a suitable model under sufficient smoothness assumptions.
- (iv) $y_{ij} = \eta(x_i) + \epsilon_{ij}$, where η is a parametric or nonparametric function and ϵ_{ij} are *iid* Gaussian errors. In particular, $\eta(x_i)$ may be a linear regression function, that is, $\eta(x_i) = \beta'x_i$, where β is a vector of unknown parameters. Non-linear forms of η are also permitted. Also, η may be a reasonably smooth function of unknown form, modeled by some appropriate Gaussian process.

6.2.2 The Bayesian inverse LOO-CV setup and the IRD approach

In the Bayesian inverse LOO-CV setup, for $i \geq 1$, we successively leave out x_i from the data set, and attempt to predict the same using the rest of the dataset, in the form of the posterior $\pi(\tilde{x}_i | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm})$, where $\mathbf{Y}_{nm} = \{y_{ij} : i = 1, \dots, n; j = 1, \dots, m\}$, $\mathbf{X}_n = \{x_i : i = 1, \dots, n\}$ and $\mathbf{X}_{n,-i} = \mathbf{X}_n \setminus \{x_i\}$, and \tilde{x}_i is the random quantity corresponding to the left out x_i .

In this chapter, we are interested in proving that $\pi(\tilde{x}_i \in U_i^c | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}) \rightarrow 0$ almost surely as $m, n \rightarrow \infty$, where U_i is any neighborhood of x_i . Here, for any set A , A^c denotes the complement of A .

Note that the i -th LOO-CV posterior is given by

$$\pi(\tilde{x}_i | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}) = \int_{\Theta} \pi(\tilde{x}_i | \theta, \mathbf{y}_i) d\pi(\theta | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}). \quad (6.2.2)$$

In the IRD approach, we consider the distribution of any suitable statistic $T(\tilde{\mathbf{X}}_n)$, where the distribution of $\tilde{\mathbf{X}}_n = \{\tilde{x}_1, \dots, \tilde{x}_n\}$ is induced by the respective LOO-CV

posteriors of the form (6.2.2). The distribution of $T(\tilde{\mathbf{X}}_n)$ is referred to as the IRD in Bhattacharya (2013). Now consider the observed statistic $T(\mathbf{X}_n)$. In a nutshell, if $T(\mathbf{X}_n)$ falls within the desired $100(1 - \alpha)\%$ ($0 < \alpha < 1$) of the IRD, then the model is said to fit the data; otherwise, the model does not fit the data. Typical examples of $T(\mathbf{X}_n)$, which turned out to be useful in the palaeoclimate modeling context are (see Mukhopadhyay and Bhattacharya (2013)) are:

$$T_1(\mathbf{X}_n) = \sum_{i=1}^n \frac{(x_i - E_\pi(\tilde{x}_i))^2}{V_\pi(\tilde{x}_i)} \quad (6.2.3)$$

$$T_2(\mathbf{X}_n) = \sum_{i=1}^n \frac{|x_i - E_\pi(\tilde{x}_i)|}{\sqrt{V_\pi(\tilde{x}_i)}} \quad (6.2.4)$$

$$T_3(\mathbf{X}_n) = x_i \quad (6.2.5)$$

To obtain $T(\tilde{\mathbf{X}}_n)$ corresponding to $T(\mathbf{X}_n)$ above, we only need to replace x_i with \tilde{x}_i in (6.2.3) – (6.2.5). In the above, E_π and V_π denote the expectation and the variance, respectively, with respect to the LOO-CV posteriors. The statistic $T_3(\tilde{\mathbf{X}}_n)$ is \tilde{x}_i itself, so that the posterior of $T_3(\tilde{\mathbf{X}}_n)$ is nothing but the i -th LOO-CV posterior. Such a statistic can be important when there is particular interest in x_i , for instance, if one suspects outlyingness of x_i . An example of such an issue is considered in Bhattacharya and Haslett (2007).

6.3 Discussion regarding consistency of the LOO-CV and the IRD approach

The question now arises if the IRD approach is at all consistent. That is, whether by increasing n and m , the distribution of $T(\tilde{\mathbf{x}})$ will increasingly concentrate around $T(\mathbf{x})$. A sufficient condition for this to hold is consistency of the i -th LOO-CV posterior at x_i , for $i \geq 1$. From (6.2.2) it is clear that consistency of $\pi(\theta | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm})$ at the truth θ_0 is required for this purpose, but even if θ in $\pi(\tilde{x}_i | \theta, \mathbf{y}_i)$ is replaced with θ_0 , consistency of

(6.2.2) at x_i does not hold for arbitrary priors on \tilde{x}_i , and for fixed $m \geq 1$. This has been demonstrated in Chapter 1. with the help of a simple Poisson regression with mean θx_i , where both θ and x_i are positive quantities. Special priors on \tilde{x}_i is needed, along with the setup with $m \rightarrow \infty$, to achieve desired consistency of the LOO-CV posterior of \tilde{x}_i at x_i . In Section 6.4 we propose such an appropriate prior form and establish some requisite properties of the prior and $\pi(\tilde{x}_i|\theta, \mathbf{y}_i)$. With such prior and with conditions that ensure consistency of $\pi(\theta|\mathbf{X}_{n,-i}, \mathbf{Y}_{nm})$ at θ_0 , we establish consistency of the LOO-CV posteriors in Section 6.5.

Indeed, in the setups that we consider, for any $m \geq 1$, $\pi(\theta|\mathbf{X}_n, \mathbf{Y}_{nm})$ is consistent at the true value θ_0 . That is, for any neighbourhood V of θ_0 , for given $m \geq 1$, $\pi(\theta \in V|\mathbf{X}_n, \mathbf{Y}_{nm}) \rightarrow 1$ almost surely, as $n \rightarrow \infty$. Assuming complete separable metric space Θ , this is again equivalent to weak convergence of $\pi(\theta|\mathbf{X}_n, \mathbf{Y}_{nm})$ to δ_{θ_0} , as $n \rightarrow \infty$, for $m \geq 1$, for almost all data sequences (see, for example, Ghosh and Ramamoorthi (2003), Ghosal and van der Vaart (2017)).

In our situations, we assume that the conditions of Shalizi (2009) hold for $m \geq 1$, which would ensure consistency of $\pi(\theta|\mathbf{X}_n, \mathbf{Y}_{nm})$ is consistent at the true value θ_0 . The advantages of Shalizi's results include great generality of the model and prior including dependent setups, and reasonably easy to verify conditions. The results crucially hinge on verification of the asymptotic equipartition property. In Section 6.3.1 we show that Shalizi's result leads to weak convergence of the posterior of θ to the point mass at θ_0 , which will play an useful role in our proof of consistency of the LOO-CV posteriors.

6.3.1 Weak convergence of Shalizi's result

From (4.1.2) it follows that for any $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} \pi(\mathbb{N}_\epsilon^c | \mathbf{Y}_n) = 0, \quad (6.3.1)$$

where $\mathbb{N}_\epsilon = \{\theta : h(\theta) \leq h(\Theta) + \epsilon\}$. In our case, we shall not consider misspecification, as we are interested in ensuring posterior consistency. Thus, we have $h(\Theta) = 0$ in our context. Now observe that $h(\theta)$ given by (4.A1.2) is not a proper KL-divergence between two distributions. Thus the question arises if (6.3.1) suffices for posterior consistency, and hence weak convergence of the posterior to δ_{θ_0} . Lemma 16 below settles this question in the affirmative.

Lemma 16 *Given any neighborhood U of θ_0 , the set \mathbb{N}_ϵ is contained in U for sufficiently small ϵ .*

Proof. It is sufficient to prove that $h(\theta) > 0$ if and only if $\theta \neq \theta_0$. Note that $E_{\theta_0} \left(\log \frac{f_{\theta_0}(\mathbf{Y}_n)}{f_\theta(\mathbf{Y}_n)} \right)$ is a proper KL-divergence and hence is non-decreasing with n (see van Erven and Harremoës (2014)). Hence if $\theta \neq \theta_0$, then there exists $\varepsilon > 0$ such that $E_{\theta_0} \left(\log \frac{f_{\theta_0}(\mathbf{Y}_n)}{f_\theta(\mathbf{Y}_n)} \right) > \varepsilon$ for all $n \geq 1$. Hence, $h(\theta)$ given by (4.A1.2) is larger than ε if $\theta \neq \theta_0$. Of course, if $h(\theta) > 0$, we must have $\theta \neq \theta_0$, since otherwise, $E_{\theta_0} \left(\log \frac{f_{\theta_0}(\mathbf{Y}_n)}{f_\theta(\mathbf{Y}_n)} \right) = 0$ for all n , which would imply $h(\theta) = 0$. This proves the lemma. ■

It follows from Lemma 16 that for any neighborhood U of θ_0 , $\pi(U|\mathbf{Y}_n) \rightarrow 1$, almost surely, as $n \rightarrow \infty$. Thus, $\pi(\cdot|\mathbf{Y}_n) \xrightarrow{w} \delta_{\theta_0}(\cdot)$, almost surely, as $n \rightarrow \infty$, where “ \xrightarrow{w} ” denotes weak convergence.

6.4 Prior for \tilde{x}_i

We consider the following prior for \tilde{x}_i : given θ ,

$$\tilde{x}_i \sim \text{Uniform}(B_{im}(\theta)), \quad (6.4.1)$$

where

$$B_{im}(\theta) = \left(\left\{ x : H(\eta(x)) \in \left[\bar{y}_i - \frac{cs_i}{\sqrt{m}}, \bar{y}_i + \frac{cs_i}{\sqrt{m}} \right] \right\} \right), \quad (6.4.2)$$

for some appropriate transformation H . In (6.4.2), $\bar{y}_i = \frac{1}{m} \sum_{j=1}^m y_{ij}$ and $s_i^2 = \frac{1}{m-1} \sum_{j=1}^m (y_{ij} - \bar{y}_i)^2$, and $c \geq 1$ is some constant. We denote this prior by $\pi(\tilde{x}_i|\eta)$. Lemma 17 shows that the density or any probability associated with $\pi(\tilde{x}_i|\eta)$ is continuous with respect to η .

6.4.1 Illustrations

- (i) $y_{ij} \sim \text{Poisson}(\theta x_i)$, where $\theta > 0$ and $x_i > 0$ for all i . Here, under the prior $\pi(\tilde{x}_i|\theta)$, \tilde{x}_i has uniform distribution on the set $B_{im}(\theta) = \left\{ x > 0 : \frac{\bar{y}_i - \frac{cs_i}{\sqrt{m}}}{\theta} \leq x \leq \frac{\bar{y}_i + \frac{cs_i}{\sqrt{m}}}{\theta} \right\}$.
- (ii) $y_{ij} \sim \text{Poisson}(\lambda_i)$, where $\lambda_i = \lambda(x_i)$, with $\lambda(x) = H(\eta(x))$. Here H is a known, one-to-one, continuously differentiable function and $\eta(\cdot)$ is an unknown function modeled by Gaussian process. Here, the prior for \tilde{x}_i is the uniform distribution on $B_{im}(\eta) = \left\{ x : \eta(x) \in H^{-1} \left\{ \left[\bar{y}_i - \frac{cs_i}{\sqrt{m}}, \bar{y}_i + \frac{cs_i}{\sqrt{m}} \right] \right\} \right\}$.
- (iii) $y_{ij} \sim \text{Bernoulli}(p_i)$, where $p_i = \lambda(x_i)$, with $\lambda(x) = H(\eta(x))$. Here H is a known, increasing, continuously differentiable, cumulative distribution function and $\eta(\cdot)$ is an unknown function modeled by some appropriate Gaussian process. Here, the prior for \tilde{x}_i is the uniform distribution on $B_{im}(\eta) = \left\{ x : \eta(x) \in H^{-1} \left\{ \left[\bar{y}_i - \frac{cs_i}{\sqrt{m}}, \bar{y}_i + \frac{cs_i}{\sqrt{m}} \right] \right\} \right\}$.
- (iv) $y_{ij} = \eta(x_i) + \epsilon_{ij}$, where $\eta(\cdot)$ is an unknown function modeled by some appropriate Gaussian process, and ϵ_{ij} are *iid* zero-mean Gaussian noise with variance σ^2 . Here, the prior for \tilde{x}_i is the uniform distribution on $B_{im}(\eta) = \left\{ x : \eta(x) \in \left[\bar{y}_i - \frac{cs_i}{\sqrt{m}}, \bar{y}_i + \frac{cs_i}{\sqrt{m}} \right] \right\}$. If $\eta(x_i) = \alpha + \beta x_i$, then the prior for \tilde{x}_i is the uniform distribution on $[a, b]$, where $a = \min \left\{ \frac{\bar{y}_i - \frac{cs_i}{\sqrt{m}} - \alpha}{\beta}, \frac{\bar{y}_i + \frac{cs_i}{\sqrt{m}} - \alpha}{\beta} \right\}$ and $b = \max \left\{ \frac{\bar{y}_i - \frac{cs_i}{\sqrt{m}} - \alpha}{\beta}, \frac{\bar{y}_i + \frac{cs_i}{\sqrt{m}} - \alpha}{\beta} \right\}$.

6.4.2 Some properties of the prior

Our proposed prior for \tilde{x}_i possesses several useful properties necessary for our asymptotic theory. These are formally provided in the lemmas below.

Lemma 17 *The prior density $\pi(\tilde{x}_i|\eta)$ or any probability associated with $\pi(\tilde{x}_i|\eta)$ is continuous with respect to η .*

Proof. Let $\{\eta_k : k = 1, 2, \dots\}$ be a sequence of functions such that $\|\eta_k - \eta\| \rightarrow 0$, as $k \rightarrow \infty$, where $\|\cdot\|$ denotes the sup norm. It then follows that for any set A ,

$$\{x : \eta_k(x) \in A\} \cap B_{im}(\eta_k) \rightarrow \{x : \eta(x) \in A\} \cap B_{im}(\eta), \text{ as } k \rightarrow \infty.$$

Hence, as $k \rightarrow \infty$,

$$Leb(\{x : \eta_k(x) \in A\} \cap B_{im}(\eta_k)) \rightarrow Leb(\{x : \eta(x) \in A\} \cap B_{im}(\eta)),$$

where, for any set A , $Leb(A)$ denotes the Lebesgue measure of A . This proves the lemma. \blacksquare

If the density of \mathbf{y}_i given \tilde{x}_i and θ , which we denote by $f(\mathbf{y}_i|\theta, \tilde{x}_i)$, is continuous in θ and Θ is bounded then it would follow from Lemma 17 and the dominated convergence theorem that $\pi(\tilde{x}_i|\theta, \mathbf{y}_i)$ and its associated probabilities are also continuous in θ . Below we formally present the result as Lemma 18.

Lemma 18 *If $f(\mathbf{y}_i|\theta, \tilde{x}_i)$ is continuous in θ and Θ is bounded, then the density $\pi(\tilde{x}_i|\theta, \mathbf{y}_i)$ or any probability associated with $\pi(\tilde{x}_i|\theta, \mathbf{y}_i)$ is continuous with respect to θ .*

However, we usually can not assume a compact parameter space. For example, such compactness assumption is invalid for Gaussian process priors for θ . But in most situations, continuity of the density of $\pi(\tilde{x}_i|\theta, \mathbf{y}_i)$ and its associated probabilities with respect to θ hold even without the compactness assumption, provided $f(\mathbf{y}_i|\theta, \tilde{x}_i)$ is continuous in θ . We thus make the following realistic assumption:

Assumption 8 $\pi(\tilde{x}_i|\theta, \mathbf{y}_i)$ is continuous in θ .

The following result holds due to Assumption 8 and Scheffe's theorem (see, for example, Schervish (1995)).

Lemma 19 *If Assumption 8 holds, then any probability associated with $\pi(\tilde{x}_i|\theta, \mathbf{y}_i)$ is continuous in θ .*

6.5 Consistency of the LOO-CV posteriors

For consistency of the LOO-CV posteriors given by (6.2.2), we first need to ensure weak convergence of $\pi(\theta|\mathbf{X}_{n,-i}, \mathbf{Y}_{nm})$ almost surely to δ_{θ_0} , as $n \rightarrow \infty$, for $m \geq 1$. This holds if and only if $\pi(\theta|\mathbf{X}_n, \mathbf{Y}_{nm})$ is consistent at θ_0 . This can be seen by noting that the i -th factor of $\log R_n(\theta)$, obtained by integrating out \tilde{x}_i , does not play any role in (4.1.1) and (4.A1.2), so that these limits remain the same as in the case of $\pi(\theta|\mathbf{X}_n, \mathbf{Y}_{nm})$. The other conditions of Shalizi also remain the same for both the posteriors $\pi(\theta|\mathbf{X}_n, \mathbf{Y}_{nm})$ and $\pi(\theta|\mathbf{X}_{n,-i}, \mathbf{Y}_{nm})$.

Hence, assuming that conditions (S1)–(S7) of Shalizi are verified for $\pi(\theta|\mathbf{X}_n, \mathbf{Y}_{nm})$, for fixed m , it follows that $\pi(\cdot|\mathbf{X}_{n,-i}, \mathbf{Y}_{nm}) \xrightarrow{w} \delta_{\theta_0}(\cdot)$, almost surely, as $n \rightarrow \infty$.

For any neighborhood U_i of x_i , note that the probability $\pi(\tilde{x}_i \in U_i^c|\theta, \mathbf{y}_i)$ is continuous in θ due to Lemma 19. Moreover, since it is a probability, it is bounded. Hence, by the Portmanteau theorem, using (6.2.2) and consistency of $\pi(\theta|\mathbf{X}_{n,-i}, \mathbf{Y}_{nm})$ it holds almost surely that

$$\begin{aligned} \pi(\tilde{x}_i \in U_i^c|\mathbf{X}_{n,-i}, \mathbf{Y}_{nm}) &= \int_{\Theta} \pi(\tilde{x}_i \in U_i^c|\theta, \mathbf{y}_i) d\pi(\theta|\mathbf{X}_{n,-i}, \mathbf{Y}_{nm}) \\ &\xrightarrow{a.s.} \pi(\tilde{x}_i \in U_i^c|\theta_0, \mathbf{y}_i), \text{ as } n \rightarrow \infty, \text{ for any } m \geq 1. \end{aligned} \quad (6.5.1)$$

We formalize this result as the following theorem.

Theorem 20 *Assume conditions (S1)–(S7) of Shalizi. Then for $i \geq 1$, under the prior (6.4.1) and Assumption 8, (6.5.1) holds almost surely, for any $m \geq 1$, for any neighborhood U_i of x_i .*

Let us now make the following extra assumptions:

Assumption 9 $f(\mathbf{y}_i|\theta_0, \tilde{x}_i)$ is continuous in \tilde{x}_i .

Assumption 10 η_0 is a one-to-one function.

With these assumptions, we have the following result.

Theorem 21 Under the prior (6.4.1) and Assumptions 9 and 10, for any neighborhood U_i of x_i , for any $i \geq 1$,

$$\pi(\tilde{x}_i \in U_i^c | \theta_0, \mathbf{y}_i) \xrightarrow{a.s.} 0, \text{ as } m \rightarrow \infty. \quad (6.5.2)$$

Proof. Note that

$$\pi(\tilde{x}_i \in U_i^c | \theta_0, \mathbf{y}_i) = \frac{\int_{U_i^c} \pi(\tilde{x}_i | \theta_0) f(\mathbf{y}_i | \theta_0, \tilde{x}_i) d\tilde{x}_i}{\int_{U_i^c} \pi(\tilde{x}_i | \theta_0) f(\mathbf{y}_i | \theta_0, \tilde{x}_i) d\tilde{x}_i + \int_{U_i} \pi(\tilde{x}_i | \theta_0) f(\mathbf{y}_i | \theta_0, \tilde{x}_i) d\tilde{x}_i}. \quad (6.5.3)$$

Let us consider $\int_{U_i^c} \pi(\tilde{x}_i | \theta_0) f(\mathbf{y}_i | \theta_0, \tilde{x}_i) d\tilde{x}_i$ of (6.5.3). Since the support of \tilde{x}_i is compact, Assumption 9 ensures that $f(\mathbf{y}_i | \theta_0, \tilde{x}_i)$ is bounded. Hence,

$$\int_{U_i^c} \pi(\tilde{x}_i | \theta_0) f(\mathbf{y}_i | \theta_0, \tilde{x}_i) d\tilde{x}_i \leq K \int_{U_i^c} \pi(\tilde{x}_i | \theta_0) d\tilde{x}_i = K \pi(\tilde{x}_i \in U_i^c | \theta_0), \quad (6.5.4)$$

for some positive constant K . Now note that $\pi(\tilde{x}_i \in U_i^c | \theta_0) = \pi(\tilde{x}_i \in U_i^c \cap B_{im}(\theta_0) | \theta_0)$, and Assumption 10 ensures that $B_{im}(\theta_0) \rightarrow \{x_i\}$ almost surely, as $m \rightarrow \infty$, for all $i \geq 1$. It follows that there exists $m_0 \geq 1$ such that $U_i^c \cap B_{im}(\theta_0) = \emptyset$, for $m \geq m_0$. Hence, $\pi(\tilde{x}_i \in U_i^c \cap B_{im}(\theta_0) | \theta_0) \rightarrow 0$, as $m \rightarrow \infty$. This implies, in conjunction with (6.5.4) and (6.5.3), that (6.5.2) holds.

■

Combining Theorems 20 and 21 yields the following main result.

Theorem 22 Assume conditions (S1)–(S7) of Shalizi. Then with the prior (6.4.1),

under further Assumptions 8 – 10, for $i \geq 1$,

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \pi(\tilde{x}_i \in U_i^c | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}) = 0, \text{ almost surely,} \quad (6.5.5)$$

for any neighborhood U_i of x_i .

6.6 Consistency of the IRD approach

Due to practical usefulness, we consider consistency of IRD associated with (6.2.3) – (6.2.5). Among these, the IRD associated with T_3 is just the i -th LOO-CV posterior, which is consistent by Theorem 22. For T_1 and T_2 , we consider slight modification by dividing the right hand sides of (6.2.3) and (6.2.4) by n , and adding some small quantity $\varepsilon > 0$ to $V_\pi(\tilde{x}_i)$. These adjustments are not significant for practical applications, but seems to be necessary for our asymptotic theory. With these, we provide the consistency result and its for the IRD corresponding to T_1 ; that corresponding to T_2 would follow in the same way.

Theorem 23 *Assume conditions (S1)–(S7) of Shalizi, and the prior (6.4.1). Also let Assumptions 8 – 10 hold, for $i \geq 1$, Define for some $\varepsilon > 0$, the following:*

$$T_1(\tilde{\mathbf{X}}_n) = \frac{1}{n} \sum_{i=1}^n \frac{(\tilde{x}_i - E_\pi(\tilde{x}_i))^2}{V_\pi(\tilde{x}_i) + \varepsilon}$$

and

$$T_1(\mathbf{X}_n) = \frac{1}{n} \sum_{i=1}^n \frac{(x_i - E_\pi(\tilde{x}_i))^2}{V_\pi(\tilde{x}_i) + \varepsilon}.$$

Then

$$\left| T_1(\tilde{\mathbf{X}}_n) - T_1(\mathbf{X}_n) \right| \xrightarrow{P} 0, \text{ as } m \rightarrow \infty, n \rightarrow \infty, \text{ almost surely.} \quad (6.6.1)$$

In the above, “ \xrightarrow{P} ” denotes convergence in probability.

Proof. The assumptions of this theorem ensures consistency of the LOO-CV posteriors due to Theorem 22. This again is equivalent to almost sure weak convergence of the i -th cross-validation posterior to $\delta_{\{x_i\}}$, for $i \geq 1$. This is again equivalent to convergence in (cross-validation posterior) distribution of \tilde{x}_i , to the degenerate quantity x_i , almost surely. Due to degeneracy, this is again equivalent to convergence in probability, almost surely.

For notational clarity we denote \tilde{x}_i by \tilde{x}_i^{nm} , whose LOO-CV posterior is $\pi(\cdot | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm})$. Let also $\tilde{\mathbf{X}}^{nm} = \{\tilde{x}_1^{nm}, \dots, \tilde{x}_n^{nm}\}$, so that we now denote $T_1(\tilde{\mathbf{X}}_n)$ by $T_1(\tilde{\mathbf{X}}^{nm})$. It follows from the above arguments that for $i \geq 1$,

$$\tilde{x}_i^{nm} \xrightarrow{P} x_i, \text{ almost surely, as } m \rightarrow \infty, n \rightarrow \infty. \quad (6.6.2)$$

Now consider $T_1(\tilde{\mathbf{X}}^{nm}) - T_1(\mathbf{X}_n)$, which is an average of n terms, the i -th term being

$$z_i^{nm} = \frac{(\tilde{x}_i^{nm} - E_\pi(\tilde{x}_i^{nm}))^2 - (x_i - E_\pi(\tilde{x}_i^{nm}))^2}{V_\pi(\tilde{x}_i^{nm}) + \varepsilon}. \quad (6.6.3)$$

Due to bounded support of \tilde{x}_i^{nm} and (6.6.2), uniform integrability entails $E_\pi(\tilde{x}_i) \rightarrow x_i$ and $V_\pi(\tilde{x}_i) \rightarrow 0$, almost surely. The latter two results ensure, along with (6.6.2), that for $i \geq 1$,

$$z_i^{nm} \xrightarrow{P} 0, \text{ as } m \rightarrow \infty, n \rightarrow \infty, \text{ almost surely.} \quad (6.6.4)$$

Now note that if z_i^{nm} were non-random, then $z_i^{nm} \rightarrow 0$, as $m \rightarrow \infty, n \rightarrow \infty$, would imply $\frac{1}{n} \sum_{i=1}^n z_i^{nm} \rightarrow 0$ as $m \rightarrow \infty, n \rightarrow \infty$. Hence, by Theorem 7.15 of Schervish (1995) (page 398), it follows that

$$T_1(\tilde{\mathbf{X}}^{nm}) - T_1(\mathbf{X}_n) \xrightarrow{P} 0, \text{ as } m \rightarrow \infty, n \rightarrow \infty, \text{ almost surely.}$$

In other words, (6.6.1) holds. ■

6.7 Discussion of the applicability of our asymptotic results in the inverse regression contexts

From the development of the asymptotic results it is clear that there are two separate aspects that ensures consistency of the LOO-CV posteriors. The first is consistency of the posterior of the parameter(s) θ , and then consistency of $\pi(\tilde{x}_i|\theta, \mathbf{y}_i)$. Once consistency of the posterior of θ is ensured, our prior for \tilde{x}_i then guarantees consistency of the posterior of \tilde{x}_i at x_i . To verify consistency of the posterior of θ , we referred to the general conditions of Shalizi because of their wide applicability, including dependent setups, and relatively easy verifiability of the conditions. Indeed, the seven conditions of Shalizi have been verified in the contexts of general stochastic process (including Gaussian process) regression (Chapter 4) with both Gaussian and double exponential errors, binary and Poisson regression involving general stochastic process (including Gaussian process) and known link functions (Chapter 5). Moreover, for finite-dimensional parametric problems, the conditions are much simpler to verify. Thus, the examples provided in Section 6.4.1 are relevant in this context, and the LOO-CV posteriors, and hence the IRD, are consistent. Furthermore, Chandra and Bhattacharya (2020a) and Chandra and Bhattacharya (2020b) establish the conditions of Shalizi in an autoregressive regression context, even for the so-called “large p , small n ” paradigm. In such cases, our asymptotic results for the LOO-CV posteriors and the IRD, will hold.

There is one minor point to touch upon regarding our requirement for ensuring consistency. In all the aforementioned works regarding verification of Shalizi’s conditions, $m = 1$ was considered. For our asymptotic theory, we first require consistency of θ as $n \rightarrow \infty$, for fixed $m \geq 1$, and then take the limit as $m \rightarrow \infty$. This is of course satisfied if consistency holds for $m = 1$, as for more information about θ brought in for larger values of m , consistency automatically continues to hold. Indeed, for fixed $m \geq 1$, the limit as $n \rightarrow \infty$ does not depend upon m , as the posterior of θ converges weakly to the

point mass at θ_0 , almost surely. Thus, it is always sufficient to verify consistency of the posterior of θ for $m = 1$.

6.8 Simulation studies

6.8.1 Poisson parametric regression

Let us first consider the case where $y_{ij} \sim \text{Poisson}(\theta x_i)$, as briefed in Section 6.4.1 (i). Here we investigate consistency of the posterior of \tilde{x}_i . We generate the data by simulating $\theta \sim \text{Uniform}(0, 2)$, $x_i \sim \text{Uniform}(0, 2)$, $i = 1, \dots, n$, and then by generating $y_{ij} \sim \text{Poisson}(\theta x_i)$, for $j = 1, \dots, m$ and $i = 1, \dots, n$. We set $\pi(\theta) = 1; \theta > 0$, for the prior for θ .

Since numerical integration turned out to be unstable, we resort to Gibbs sampling from the posterior, noting that the full conditional distributions of θ and \tilde{x}_i are of the forms

$$\begin{aligned} [\theta|\tilde{x}_i, \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}] &\propto \theta^{\sum_{i=1}^n \sum_{j=1}^m y_{ij}} \exp\left\{-m\theta\left(\tilde{x}_i + \sum_{j \neq i} x_j\right)\right\} I_{\left[\frac{\max\{0, \bar{y}_i - cs_i/\sqrt{m}\}}{\tilde{x}_i}, \frac{\bar{y}_i + cs_i/\sqrt{m}}{\tilde{x}_i}\right]}(\theta); \\ [\tilde{x}_i|\theta, \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}] &\propto \tilde{x}_i^{m\bar{y}_i} \exp(-m\theta\tilde{x}_i) I_{\left[\frac{\max\{0, \bar{y}_i - cs_i/\sqrt{m}\}}{\theta}, \frac{\bar{y}_i + cs_i/\sqrt{m}}{\theta}\right]}(\tilde{x}_i). \end{aligned}$$

It follows that $[\theta|\tilde{x}_i, \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}]$ has the gamma distribution with shape parameter $\sum_{i=1}^n \sum_{j=1}^m y_{ij} + 1$ and rate parameter $m\left(\tilde{x}_i + \sum_{j \neq i} x_j\right)$, truncated on $\left[\frac{\max\{0, \bar{y}_i - cs_i/\sqrt{m}\}}{\tilde{x}_i}, \frac{\bar{y}_i + cs_i/\sqrt{m}}{\tilde{x}_i}\right]$. Similarly, $[\tilde{x}_i|\theta, \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}]$ has the gamma distribution with shape parameter $m\bar{y}_i + 1$ and rate parameter $m\tilde{x}_i$, truncated on $\left[\frac{\max\{0, \bar{y}_i - cs_i/\sqrt{m}\}}{\theta}, \frac{\bar{y}_i + cs_i/\sqrt{m}}{\theta}\right]$.

For our investigation, we set $i = 1$. That is, without loss of generality, we address consistency of the posterior of \tilde{x}_1 via simulation study. As for the choice of c , we set $c = 20$. This choice ensured that the full conditional distributions have reasonably large support, for given values of n and m . We run our Gibbs sampler for 11000 iterations, and discard the first 1000 iterations as burn-in.

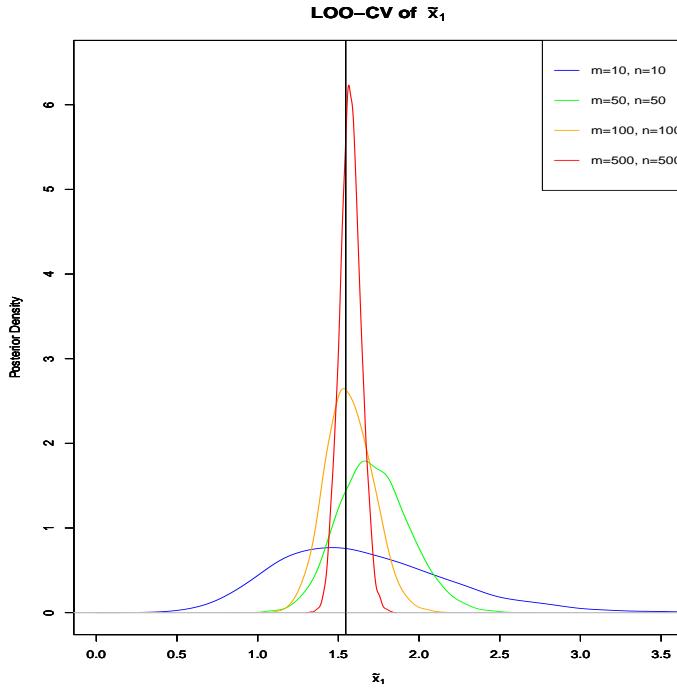


Figure 6.8.1: Demonstration of posterior consistency in inverse parametric Poisson regression. The vertical line denotes the true value.

Figure 6.8.1 displays the posterior densities of \tilde{x}_1 for different values of m and n ; here, for convenience of presentation, we have set $m = n$. The vertical line denotes the true value x_1 . The diagram vividly depicts that the LOO-CV posterior of \tilde{x}_1 concentrates more and more around x_1 as n and m increase.

6.8.2 Poisson nonparametric regression

We now consider the case where $y_{ij} \sim \text{Poisson}(\lambda(x_i))$, where $\lambda(x) = H(\eta(x))$, as briefed in Section 6.4.1 (ii). In particular, we let $H(\cdot) = \exp(\cdot)$ and $\eta(\cdot)$ be a Gaussian process with mean function $\mu(x) = \alpha + \beta x$ and covariance $\text{Cov}(\eta(x_1), \eta(x_2)) = \sigma^2 \exp\{-(x_1 - x_2)^2\}$, where σ is unknown. We assume that the true data-generating distribution is $y_{ij} \sim \text{Poisson}(\lambda(x_i))$, with $\lambda(x) = \exp(\alpha_0 + \beta_0(x))$. We generate the data

by simulating $\alpha_0 \sim Uniform(-1, 1)$, $\beta_0 \sim Uniform(-1, 1)$ and $x_i \sim Uniform(-1, 1)$; $i = 1, \dots, n$, and then finally simulating $y_{ij} \sim Poisson(\lambda(x_i))$; $j = 1, \dots, m$, $i = 1, \dots, n$.

For our convenience, we reparameterize σ^2 as $\exp(\omega)$, where $-\infty < \omega < \infty$. For the prior on the parameters, we set $\pi(\alpha, \beta, \omega) = 1$, for $-\infty < \alpha, \beta, \omega < \infty$. Now note that the prior for \tilde{x}_i , which is uniform on $B_{im}(\eta) = \left\{x : \eta(x) \in H^{-1}\left\{\left[\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}}, \bar{y}_i + \frac{c_2 s_i}{\sqrt{m}}\right]\right\}\right\}$, does not have a closed form, since the form of $\eta(x)$ is unknown. However, if m is large, the interval $H^{-1}\left\{\left[\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}}, \bar{y}_i + \frac{c_2 s_i}{\sqrt{m}}\right]\right\}$ is small, and $\eta(x)$ falling in this small interval can be reasonably well-approximated by a straight line. Hence, we set $\eta(x) = \mu(x) = \alpha + \beta x$, for $\eta(x)$ falling in this interval. In our case, it follows that $[\tilde{x}_i|\eta] \sim Uniform(a, b)$, where

$$a = \min \left\{ \beta^{-1} \left(\log \left(\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}} \right) - \alpha \right), \beta^{-1} \left(\log \left(\bar{y}_i + \frac{c_2 s_i}{\sqrt{m}} \right) - \alpha \right) \right\}$$

and

$$b = \max \left\{ \beta^{-1} \left(\log \left(\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}} \right) - \alpha \right), \beta^{-1} \left(\log \left(\bar{y}_i + \frac{c_2 s_i}{\sqrt{m}} \right) - \alpha \right) \right\}.$$

We set $c_1 = 1$ and $c_2 = 100$, for ensuring positive value of $\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}}$ (so that logarithm of this quantity is well-defined) and a reasonably large support of the prior for \tilde{x}_i . As before, we set $i = 1$, for our purpose, thus focussing on posterior consistency of \tilde{x}_1 only.

In this example, both numerical integration and Gibbs sampling are infeasible. Hence, we resort to Transformation based Markov Chain Monte Carlo (TMCMC) ([Dutta and Bhattacharya \(2014\)](#)) for simulating from the posterior. In particular, we use the additive transformation and update all the unknowns simultaneously, in a single block. More specifically, at each iteration $t = 1, 2, \dots$, we first generate $\epsilon \sim N(0, 1)$, a standard normal variable. Then, letting $(\tilde{x}_1^{(t)}, \alpha^{(t)}, \beta^{(t)}, \omega^{(t)}, \eta^{(t)}(x_2), \dots, \eta^{(t)}(x_n))$ denote the values of the unknowns at the t -th iteration, at the $(t+1)$ -th iteration we set $\alpha = \alpha^{(t)} \pm 0.5\epsilon$, $\beta = \beta^{(t)} \pm 0.5\epsilon$, $\omega = \omega^{(t)} \pm 0.05\epsilon$, and $\eta(x_k) = \eta^{(t)}(x_k) \pm 0.00005\epsilon$; $k = 2, \dots, n$. For updating \tilde{x}_1 we set $\tilde{x}_1 = \tilde{x}_1^{(t)} \pm a_{nm}\epsilon$, where we let the scale a_{nm} depend upon n and m . Specifically, since for our experiments we set $n = m = 10, 50, 100, 100$, we choose

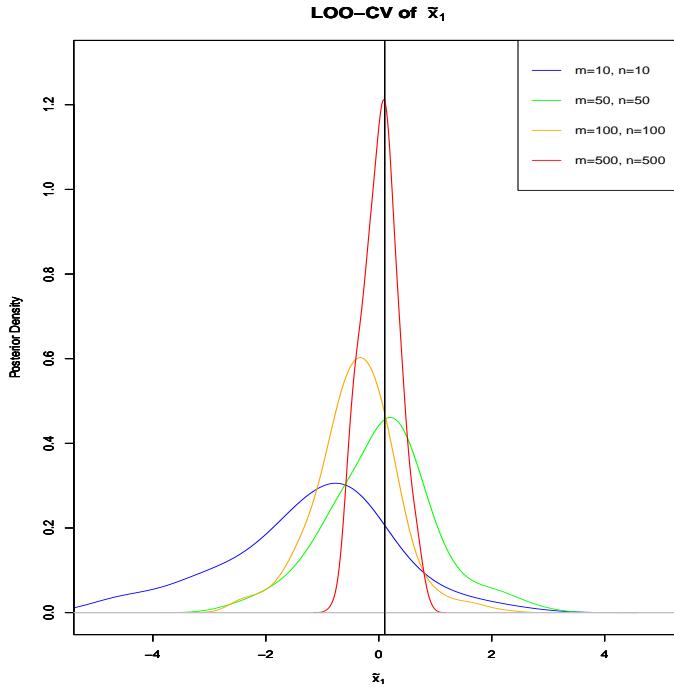


Figure 6.8.2: Demonstration of posterior consistency in inverse nonparametric Poisson regression. The vertical line denotes the true value.

$a_{nm} = 0.8, 0.65, 0.65, 0.45$, respectively, for such values of n and m . We accept these proposed values with an appropriate acceptance probability (see Dutta and Bhattacharya (2014) for details), provided the prior conditions are satisfied. This strategy has yielded reasonable mixing properties of the additive TMCMC algorithm, for all values of n and m chosen. We run our additive TMCMC algorithm for 11000 iterations, discarding the first 1000 iterations as burn-in.

Figure 6.8.2 shows the posterior densities of \tilde{x}_1 for this nonparametric inverse regression problem for different values of n and m . Again, it is clearly evident that the posterior concentrates more and more around the true value x_1 , as n and m are increased.

6.9 Conclusion

In this chapter, we have proposed a prior for \tilde{x}_i that seems to be natural for ensuring consistency of the LOO-CV posteriors, and hence of the IRD approach. Crucially, we need m observations corresponding to each x_i , and m is taken to infinity for the asymptotic theory. Note that for $m = 1$, or for any finite m , consistency of the LOO-CV posterior of \tilde{x}_i not achievable, even though consistency of the corresponding posterior of θ is attainable for any $m \geq 1$. This issue sets apart the problem of LOO-CV consistency from the usual parameter consistency.

An interesting issue is that, for forward Bayesian problems, the posterior predictive distribution of the i -th response y_i does not tend to point mass at y_i , even if the corresponding posterior of θ is consistent at θ_0 . The reason is that the distribution of y_i given θ and x_i is specified as per the modeled likelihood, and does not admit any prior construction as in the inverse setup. Since the modeled response variable is always associated with positive variability, even under the true model, the posterior predictive distribution of y_i always has positive variance, and hence, can not be consistent at y_i . From this perspective, even in forward problems, it perhaps makes sense to consider the IRD approach for model validation. Indeed, our simulation studies demonstrate the effectiveness of the IRD approach to model validation compared to the forward approach.

As a final remark, we mention that for our prior on \tilde{x}_i we required independence among $\{y_{i1}, \dots, y_{im}\}$, for the strong law of large numbers to hold for \bar{y}_i and s_i^2 . However, independence is not strictly necessary, as the ergodic theorem can often be utilized for ensuring limits in the strong sense.

7

A Short Note on Almost Sure Convergence of Bayes Factors in the General Setup

7.1 Introduction

Bayes factors are well-established in the Bayesian literature for the purpose of model comparison. Briefly, given data $\mathbf{Y}_n = \{Y_1, Y_2, \dots, Y_n\}$, where n is the sample size, consider the problem of comparing any two models \mathcal{M}_1 and \mathcal{M}_2 associated with parameter spaces Θ_1 and Θ_2 , respectively. For $i = 1, 2$, let the likelihoods, priors and the marginal densities for the two models be $L_n(\theta_i|\mathcal{M}_i) = f_{\theta_i}(\mathbf{Y}_n|\mathcal{M}_i)$, $\pi(\theta_i|\mathcal{M}_i)$ and $m(\mathbf{Y}_n|\mathcal{M}_i) = \int_{\Theta_i} L_n(\theta_i|\mathcal{M}_i)\pi(d\theta_i|\mathcal{M}_i)$, respectively. Then the Bayes factor of model \mathcal{M}_1 against \mathcal{M}_2

is given by

$$B_n^{(12)} = \frac{m(\mathbf{Y}_n | \mathcal{M}_1)}{m(\mathbf{Y}_n | \mathcal{M}_2)}. \quad (7.1.1)$$

Thus, $B_n^{(12)}$ can be interpreted as the quantification of the evidence of model \mathcal{M}_1 against model \mathcal{M}_2 , given data \mathbf{Y}_n . A comprehensive account of Bayes factors is provided in [Kass and Raftery \(1995\)](#).

The asymptotic study of Bayes factor involves investigation of limiting properties of $B_n^{(12)}$ as n goes to infinity. In particular, it is essential to guarantee the consistency property that $B_n^{(12)}$ goes to infinity almost surely when \mathcal{M}_1 is the better model and to zero almost surely when \mathcal{M}_2 is the better model. It is also important to obtain the rate of convergence of the Bayes factor. In the case of independent and identically distributed (*iid*) data, a relevant result is provided in [Walker \(2004\)](#) and [Walker et al. \(2004\)](#). Such strong “almost sure” convergence results are rare however, even when the data are independent but not identically distributed. Recently, [Maitra and Bhattacharya \(2016a\)](#) obtained a strong, general result when the data are independent but not identically distributed and applied it to time-varying covariate and drift function selection in the context of systems of stochastic differential equations (see also [Maitra and Bhattacharya \(2016b\)](#) for further application of Bayes factor asymptotics in stochastic differential equations). The other existing works on Bayes factor asymptotics are problem specific and even in such particular set-ups strong consistency results are seldom available (but see, for example, [Dawid \(1992\)](#), [Kundu and Dunson \(2014\)](#), [Choi and Rousseau \(2015\)](#)). For a comprehensive review of Bayes factor consistency, see [Chib and Kuffner \(2016\)](#).

We are interested in more general frameworks where the data may be dependent and where the possible models are perhaps all misspecified. We are not aware of any existing work on Bayes factor asymptotics in this direction. However, posterior convergence has been addressed by [Shalizi \(2009\)](#), and indeed, Theorem 2 of [Shalizi \(2009\)](#) combined with a well-known identity satisfied by Bayes factors, holds the key to an elegant almost sure convergence result for the Bayes factor. The result depends explicitly on the average

Kullback-Leibler divergence between the competing and the true models, even in such a general set-up. Here it is important to emphasize that although Chib and Kuffner (2016) is essentially a review paper, the authors demonstrate for the first time with a specific example of nested models that the identity satisfied by the Bayes factor may be exploited to prove weak consistency of the latter, and provide general discussion regarding “in probability” Bayes factor convergence assuming that the identity is satisfied by the Bayes factor.

The rest of this chapter is structured as follows. In Section 7.2, based on Shalizi (2009) we describe the general setting for our Bayes factor investigation, and provide the result of Shalizi (2009) on which our main result on Bayes factor hinges. In Section 7.3 we provide our results on Bayes factor convergence. We make concluding remarks in Section 6.9. Additional details are provided in the Appendix.

7.2 The general setup for model comparison using Bayes factors

Here we assume the same setup detailed in Chapter 4.A1. As in Shalizi (2009), we assume that P and all the F_θ are dominated by a common measure with densities p and f_θ , respectively. In Shalizi (2009) and in our case, the assumption that $P \in \Theta$ is not required so that all possible models are allowed to be misspecified.

Given a prior π on θ , we assume that the posterior distributions $\pi(\cdot|\mathbf{Y}_n)$ are dominated by a common measure for all $n \geq 1$; abusing notation, we denote the density at θ by $\pi(\theta|\mathbf{Y}_n)$.

Let $L_n(\theta) = f_\theta(\mathbf{Y}_n)$ be the likelihood and $p_n = p(\mathbf{Y}_n)$ be the marginal density of \mathbf{Y}_n under the true model P . Below we furnish Theorem 2 of Shalizi (2009) which shall play the key role for our purpose of deriving almost sure convergence of Bayes factors.

Theorem 24 (Theorem 2 of Shalizi (2009)) Consider assumptions (S1)–(S6). Then for all θ such that $\pi(\theta) > 0$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log [\pi(\theta | \mathbf{Y}_n)] = -J(\theta), \quad (7.2.1)$$

almost surely with respect to the true model P , where $J(\theta)$ is given by (4.A1.4).

7.3 Convergence of Bayes factors

For the model comparison problem using Bayes factors, we now assume the likelihoods and the priors of all the competing models satisfy (S1)–(S6), in addition to satisfying that P and all the F_θ for $\theta \in \Theta_1 \cup \Theta_2$ have densities with respect to a common σ -finite measure. We also assume that for $i = 1, 2$, the posterior $\pi(\cdot | \mathbf{Y}_n, \mathcal{M}_i)$ associated with model \mathcal{M}_i is dominated by the prior $\pi(\cdot | \mathcal{M}_i)$, which is again absolutely continuous with respect to some appropriate σ -finite measure. These latter assumptions ensure that up to the normalizing constant, the posterior density associated with \mathcal{M}_i is factorizable into the prior density times the likelihood. Indeed, for any $\theta_i \in \Theta_i$,

$$\log [m(\mathbf{Y}_n | \mathcal{M}_i)] = \log [L_n(\theta_i | \mathcal{M}_i)] + \log [\pi(\theta_i | \mathcal{M}_i)] - \log [\pi(\theta_i | \mathbf{Y}_n, \mathcal{M}_i)]. \quad (7.3.1)$$

Hence, the logarithm of the Bayes factor is given, for any $\theta_1 \in \Theta_1$ and $\theta_2 \in \Theta_2$, by (see, for example, Chib (1995), Chib and Kuffner (2016))

$$\log [B_n^{(12)}] = \log \left[\frac{L_n(\theta_1 | \mathcal{M}_1)}{L_n(\theta_2 | \mathcal{M}_2)} \right] + \log \left[\frac{\pi(\theta_1 | \mathcal{M}_1)}{\pi(\theta_2 | \mathcal{M}_2)} \right] - \log \left[\frac{\pi(\theta_1 | \mathbf{Y}_n, \mathcal{M}_1)}{\pi(\theta_2 | \mathbf{Y}_n, \mathcal{M}_2)} \right],$$

so that

$$\begin{aligned} \frac{1}{n} \log [B_n^{(12)}] &= \frac{1}{n} \log [R_n(\theta_1|\mathcal{M}_1)] - \frac{1}{n} \log [R_n(\theta_2|\mathcal{M}_2)] \\ &\quad + \frac{1}{n} \log [\pi(\theta_1|\mathcal{M}_1)] - \frac{1}{n} \log [\pi(\theta_2|\mathcal{M}_2)] \\ &\quad - \frac{1}{n} \log [\pi(\theta_1|\mathbf{Y}_n, \mathcal{M}_1)] + \frac{1}{n} \log [\pi(\theta_2|\mathbf{Y}_n, \mathcal{M}_2)], \end{aligned} \quad (7.3.2)$$

where, for $i = 1, 2$, $R_n(\theta_i|\mathcal{M}_i) = \frac{L_n(\theta_i|\mathcal{M}_i)}{p_n}$.

Now let $J_i(\theta_i) = h_i(\theta_i) - h_i(\Theta_i)$, where $h_i(\theta_i)$ is defined as in (4.A1.2) with $L_n(\theta)$ replaced with $L_n(\theta_i|\mathcal{M}_i)$, and $h_i(A) = \text{ess inf}_{\theta_i \in A_i} h_i(\theta_i)$, for any $A_i \subseteq \Theta_i$. Assumption (S3) then yields

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log [R_n(\theta_i|\mathcal{M}_i)] = -h_i(\theta_i), \quad (7.3.3)$$

almost surely, and assuming that both the models and their associated priors satisfy assumptions (S1)–(S6), it follows using Theorem 24 that for $i = 1, 2$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log [\pi(\theta_i|\mathbf{Y}_n, \mathcal{M}_i)] = -J_i(\theta_i), \quad (7.3.4)$$

almost surely.

Assuming that for $i = 1, 2$, $\pi(\theta_i|\mathcal{M}_i) > 0$ for all $\theta_i \in \Theta_i$, note that $\frac{1}{n} \log [\pi(\theta_i|\mathcal{M}_i)] \rightarrow 0$ as $n \rightarrow \infty$, so that it follows using (7.3.2), (7.3.3) and (7.3.4), that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log [B_n^{(12)}] = -[h_1(\Theta_1) - h_2(\Theta_2)], \quad (7.3.5)$$

almost surely with respect to P . We formalize this main result in the form of the following theorem:

Theorem 25 (Bayes factor convergence) *Assume that for $i = 1, 2$, the competing models \mathcal{M}_i satisfy assumptions (S1)–(S6), with parameter spaces Θ_i , in addition to satisfying that P and all the F_θ for $\theta \in \Theta_1 \cup \Theta_2$ have densities with respect to a common*

σ -finite measure. We also assume that the posterior associated with \mathcal{M}_i is dominated by the prior, which is again absolutely continuous with respect to some appropriate σ -finite measure, and that the priors satisfy $\pi(\theta_i | \mathcal{M}_i) > 0$ for all $\theta_i \in \Theta_i$. Then (7.3.5) holds almost surely with respect to the true infinite-dimensional probability measure P .

Theorem 25 provides an elegant convergence result for Bayes factors, explicitly in terms of differences between average Kullback-Leibler divergences between the competing and the true models. That such a result holds in the general set-up that includes even dependent data and misspecified models, is very encouraging. Indeed, we are not aware of any such result in the general set-up, although in the *iid* situation Walker (2004) and Walker *et al.* (2004) prove strong convergence of Bayes factor in terms of Kullback-Leibler divergences, taking misspecification into account. Theorem 25 readily leads to the following corollaries.

Corollary 26 (Consistency of Bayes factor) *Without loss of generality, let \mathcal{M}_1 be the correct model and \mathcal{M}_2 be incorrect. Then $L_n(\theta_1 | \mathcal{M}_1) = p_n$ for all $\theta_1 \in \Theta_1$, so that $h_1(\theta_1) = 0$ for all $\theta_1 \in \Theta_1$, implying that $h_1(\Theta_1) = 0$. On the other hand, $h_2(\Theta_2) > 0$, so that by Theorem 25, $\lim_{n \rightarrow \infty} \frac{1}{n} \log [B_n^{(12)}] = h_2(\Theta_2)$. In other words, $B_n^{(12)} \rightarrow \infty$ exponentially fast, confirming consistency of the Bayes factor. If \mathcal{M}_1 is not necessarily the correct model but is a better model than \mathcal{M}_2 in the sense that its average Kullback-Leibler divergence $h_1(\Theta_1)$ is smaller than $h_2(\Theta_2)$, then again $B_n^{(12)} \rightarrow \infty$ exponentially fast, guaranteeing consistency.*

Corollary 27 (Selection among a countable class of models) *Theorem 25 and Corollary 26 make it explicit that if the class of competing models is countable and contains the true model, it is selected by the Bayes factor, otherwise Bayes factor selects the model for which the average Kullback-Leibler divergence from the true model is minimized among the (countable) class of misspecified models, provided that the infimum is attained by some model.*

Corollary 28 (The case when two or more models are asymptotically correct)

For simplicity let us consider two models \mathcal{M}_1 and \mathcal{M}_2 as before with parameter spaces Θ_1 and Θ_2 respectively. From Theorem 25 it follows that $\frac{1}{n} \log [B_n^{(12)}] \rightarrow 0$ almost surely if and only if $h_1(\Theta_1) = h_2(\Theta_2)$, that is, if and only if both the models are asymptotically correct in the average Kullback-Leibler sense. Note that the zero limit of $\frac{1}{n} \log [B_n^{(12)}]$ is the only logical limit here since any non-zero limit would lead the Bayes factor to lend infinitely more support to one model compared to the other even though both the competing models are correct asymptotically. The situation of zero limit of $\frac{1}{n} \log [B_n^{(12)}]$ may arise in the case of comparisons between nested models or when testing parametric versus nonparametric models. In these cases even though both the competing models are correct asymptotically, one may be a much larger model. For reasons of parsimony it then makes sense to choose the model with smaller dimensionality. If both the models are infinite-dimensional, for example, when comparing two sets of basis functions, then model combination seems to be the right step.

In the Appendix we illustrate Theorem 25 with an example with autoregressive models of the first order ($AR(1)$ models) comparing (asymptotically) stationary versus nonstationary models when the true model is (asymptotically) stationary. We show that asymptotically the Bayes factor heavily favours the (asymptotically) stationary model.

In Corollary 28, we have referred to comparisons with nonparametric models. However, recall that the results of Shalizi require the true model P and all the postulated models F_θ to have densities with respect to a common dominating measure, and also the posteriors $\pi(\cdot | \mathbf{Y}_n)$ to be dominated by a common reference measure for all $n \geq 1$. These conditions are typically satisfied by parametric models, but not necessarily by nonparametric models. Indeed, in the case of the traditional nonparametric Bayesian analysis using the Dirichlet process prior, there is no parametric form of the likelihood as there is no density of the data \mathbf{Y}_n under this nonparametric set-up. Also, the prior is not dominated by any σ -finite measure, and so does not have any density. In other words, not all

nonparametric models lead to posteriors that can be factorized as proportional to prior times likelihood, as our Bayes factor treatment requires. However, as we clarify in the Appendix with a series of various examples of nonparametric Bayesian set-ups, in general the aforementioned factorization of the posterior holds in Bayesian nonparametrics and the domination requirements of Shalizi also hold in general. However, we emphasize that we did not yet verify assumptions (S1)–(S6) for all the examples, as we reserve this task for our future paper to be communicated elsewhere.

7.4 Conclusion

In this chapter, we have obtained an elegant almost sure convergence result for Bayes factors in the general set-up where the data may be dependent and where all possible models are allowed to be misspecified. To our knowledge, this is a first-time effort in this direction. Interestingly, in spite of the importance of the result, it follows rather trivially from Shalizi’s result on posterior consistency applied to the identity satisfied by Bayes factors. We assert that although similar results can be shown to hold in simpler set-ups (see Walker (2004) and Walker *et al.* (2004) for the *iid* set-up and Maitra and Bhattacharya (2016a) for the independent and non-identical set-up) and perhaps under specific models, our contribution is a proof of a strong convergence result under a very general set-up that has not been considered before.

The generality of our result will enable Bayes factor based asymptotic comparisons of various models in various set-ups, for example, k -th order Markov models, hidden Markov models, spatial Markov random field models, models based on dependent systems of stochastic differential equations, parametric versus nonparametric models in the dependent data setting (Ghosal *et al.* (2008) consider the *iid* set-up and study “in-probability” convergence of Bayes factor comparing specific finite and infinite-dimensional models). dependent versus independent model set-ups, to name only a few. Moreover, even in the *iid* data contexts, the existing Bayes factor asymptotic results for the specific

problems are usually not directly based on Kullback-Leibler divergence. Since our result directly make use of Kullback-Leibler divergence in any set-up, it is much more appealing from this perspective compared to the existing results.

In our future endeavors, we shall explore the effectiveness of our result in various specific set-ups, along with comparisons with existing results whenever applicable.

Appendix

7.A1 Illustration of our result on Bayes factor with competing $AR(1)$ models

Let the true model P stand for the following $AR(1)$ model:

$$y_t = \rho_0 y_{t-1} + \epsilon_t, \quad t = 1, 2, \dots, \tag{7.A1.1}$$

where $y_0 \equiv 0$, $|\rho_0| < 1$ and $\epsilon_t \stackrel{iid}{\sim} N(0, \sigma_0^2)$, for $t = 1, 2, \dots$. We assume the competing models \mathcal{M}_1 and \mathcal{M}_2 to be of the same form as (7.A1.1) but with the true parameter ρ_0 replaced with the unknown parameters ρ_1 and ρ_2 , respectively, such that $|\rho_1| < 1$ and $\rho_2 \in (-1, 1)^c$, where $(-1, 1)^c$ denotes complement of $(-1, 1)$. For model \mathcal{M}_i ; $i = 1, 2$, we assume that $y_0 \equiv 0$ and $\epsilon_t \stackrel{iid}{\sim} N(0, \sigma_i^2)$; $t = 1, 2, \dots$. For simplicity of illustration we assume for the time being that σ_1 and σ_2 are known, that is, $\sigma_1 = \sigma_2 = \sigma_0$, but see Section 7.A3 where we allow σ_1 and σ_2 to be unknown. Thus, we are interested in comparing (asymptotically) stationary and nonstationary $AR(1)$ models where the true $AR(1)$ model is (asymptotically) stationary. Note that $\Theta_1 = (-1, 1)$ and $\Theta_2 = (-1, 1)^c$. We consider priors $\pi(\cdot | \mathcal{M}_i)$; $i = 1, 2$, both of which have densities with respect to the Lebesgue measure. Let us first verify assumptions (S1)–(S6) with respect to \mathcal{M}_1 . All the probabilities and expectations below are with respect to the true model P . Notationally, in this time series context we denote the sample size by the more natural notation T rather than n .

7.A1.1 Verification of (S1) for \mathcal{M}_1

Note that

$$\log R_T(\rho_1) = \left(\frac{\rho_0 - \rho_1}{\sigma_0^2} \right) \left[\left(\sum_{t=1}^T y_{t-1}^2 \right) \left(\frac{\rho_0 + \rho_1}{2} \right) - \sum_{t=1}^T y_t y_{t-1} \right]. \quad (7.A1.2)$$

Thanks to continuity it is clear that $R_T(\rho_1)$ is $\mathcal{F}_T \times \mathcal{T}$ measurable. In other words, (S1) holds.

7.A1.2 Verification of (S2) for \mathcal{M}_1

It is easy to verify that under the true model P the autocovariance function is given by

$$\text{Cov}(y_{t+h}, y_t) \sim \frac{\sigma_0^2 \rho_0^h}{1 - \rho_0^2}; \quad h \geq 0, \quad (7.A1.3)$$

where for any two sequences $\{a_t\}_{t=1}^\infty$ and $\{b_t\}_{t=1}^\infty$, $a_t \sim b_t$ stands for $a_t/b_t \rightarrow 1$ as $t \rightarrow \infty$.

This leads to

$$\begin{aligned} E[\log R_T(\rho_1)] &= - \left(\frac{\rho_1 - \rho_0}{\sigma_0^2} \right) \left[\left(\sum_{t=1}^T E(y_{t-1}^2) \right) \left(\frac{\rho_1 + \rho_0}{2} \right) - \sum_{t=1}^T E(y_t y_{t-1}) \right] \\ &\sim -(\rho_1 - \rho_0) \left[\frac{(T-1)(\rho_1 + \rho_0)}{2(1 - \rho_0^2)} - \frac{(T-1)\rho_0}{(1 - \rho_0^2)} \right], \end{aligned}$$

so that

$$\frac{E[\log R_T(\rho_1)]}{T} \rightarrow -\frac{(\rho_1 - \rho_0)^2}{2(1 - \rho_0^2)}, \quad \text{as } T \rightarrow \infty.$$

In other words, (S2) holds, with

$$h_1(\rho_1) = \frac{(\rho_1 - \rho_0)^2}{2(1 - \rho_0^2)}. \quad (7.A1.4)$$

7.A1.3 Verification of (S3) for \mathcal{M}_1

Rather than proving pointwise almost sure convergence of $\frac{\log R_T(\rho_1)}{T}$ to $-h_1(\rho_1)$, we prove the stronger result of almost sure uniform convergence in our example. Indeed, note that

$$\begin{aligned} & \sup_{|\rho_1|<1} \left| \frac{\log R_T(\rho_1)}{T} + h_1(\rho_1) \right| \\ &= \sup_{|\rho_1|<1} \left| \frac{\rho_1 - \rho_0}{\sigma_0^2} \right| \times \left| \left(\frac{\sum_{t=1}^T y_{t-1}^2}{T} \right) \left(\frac{\rho_1 + \rho_0}{2} \right) - \frac{\sum_{t=1}^T y_t y_{t-1}}{T} - \frac{\sigma_0^2 (\rho_1 - \rho_0)}{2(1 - \rho_0^2)} \right| \\ &\leq \sup_{|\rho_1|\leq 1} \left| \frac{\rho_1 - \rho_0}{\sigma_0^2} \right| \times \left| \left(\frac{\sum_{t=1}^T y_{t-1}^2}{T} \right) \left(\frac{\rho_1 + \rho_0}{2} \right) - \frac{\sum_{t=1}^T y_t y_{t-1}}{T} - \frac{\sigma_0^2 (\rho_1 - \rho_0)}{2(1 - \rho_0^2)} \right| \\ &= \left| \frac{\hat{\rho}_1 - \rho_0}{\sigma_0^2} \right| \times \left| \left(\frac{\sum_{t=1}^T y_{t-1}^2}{T} \right) \left(\frac{\hat{\rho}_1 + \rho_0}{2} \right) - \frac{\sum_{t=1}^T y_t y_{t-1}}{T} - \frac{\sigma_0^2 (\hat{\rho}_1 - \rho_0)}{2(1 - \rho_0^2)} \right| \quad (7.A1.5) \\ &\leq \kappa \left| \left(\frac{\sum_{t=1}^T y_{t-1}^2}{T} \right) \left(\frac{\hat{\rho}_1 + \rho_0}{2} \right) - \frac{\sum_{t=1}^T y_t y_{t-1}}{T} - \frac{\sigma_0^2 (\hat{\rho}_1 - \rho_0)}{2(1 - \rho_0^2)} \right|, \quad (7.A1.6) \end{aligned}$$

where step (7.A1.5) follows due to compactness of $[-1, 1]$; here $\hat{\rho}_1 \in [-1, 1]$ depends upon the data. In (7.A1.6), κ is a finite positive constant greater than the bounded positive quantity $\left| \frac{\hat{\rho}_1 - \rho_0}{\sigma_0^2} \right|$.

Now observe that under P , the Markov chain $\{y_t : t = 1, 2, \dots\}$ is not only an asymptotically stationary process but is also irreducible and aperiodic (for definitions, see, for example, Meyn and Tweedie (1993) and Robert and Casella (2004)). The latter two properties are easy to see because the chain can travel from any value in the real line to any set with positive Lebesgue measure in just one step with positive probability. Thus, the ergodic theorem holds, so that as $T \rightarrow \infty$,

$$\frac{\sum_{t=1}^T y_{t-1}^2}{T} \rightarrow \frac{\sigma_0^2}{1 - \rho_0^2}, \quad (7.A1.7)$$

7.A1. ILLUSTRATION OF OUR RESULT ON BAYES FACTOR WITH
COMPETING AR(1) MODELS

almost surely with respect to P . To deal with $\frac{\sum_{t=1}^T y_t y_{t-1}}{T}$, note that under P ,

$$y_t y_{t-1} = \rho_0 y_{t-1}^2 + \epsilon_t y_{t-1}, \quad (7.A1.8)$$

and that $\{\epsilon_t y_{t-1} : t = 2, 3, \dots\}$ is also an asymptotically stationary, irreducible and aperiodic Markov chain. Hence, applying ergodic theorem to the latter Markov chain, we obtain, using independence of ϵ_t and y_{t-1} for all $t \geq 2$,

$$\frac{\sum_{t=1}^T \epsilon_t y_{t-1}}{T} \rightarrow 0, \quad (7.A1.9)$$

as $T \rightarrow \infty$, almost surely with respect to P . It follows by combining (7.A1.7), (7.A1.8) and (7.A1.9) that

$$\frac{\sum_{t=1}^T y_t y_{t-1}}{T} \rightarrow \frac{\sigma_0^2 \rho_0}{1 - \rho_0^2}, \quad (7.A1.10)$$

as $T \rightarrow \infty$, almost surely with respect to P . Applying (7.A1.7) and (7.A1.10) to (7.A1.6) yields

$$\begin{aligned} & \left| \left(\frac{\sum_{t=1}^T y_{t-1}^2}{T} \right) \left(\frac{\hat{\rho}_1 + \rho_0}{2} \right) - \frac{\sum_{t=1}^T y_t y_{t-1}}{T} - \frac{\sigma_0^2 (\hat{\rho}_1 - \rho_0)}{2(1 - \rho_0^2)} \right| \\ &= \left| \left(\frac{\sum_{t=1}^T y_{t-1}^2}{T} - \frac{\sigma_0^2}{1 - \rho_0^2} \right) \left(\frac{\hat{\rho}_1 + \rho_0}{2} \right) - \left(\frac{\sum_{t=1}^T y_t y_{t-1}}{T} - \frac{\sigma_0^2 \rho_0}{1 - \rho_0^2} \right) \right| \\ &\leq \left| \left(\frac{\hat{\rho}_1 + \rho_0}{2} \right) \right| \times \left| \frac{\sum_{t=1}^T y_{t-1}^2}{T} - \frac{\sigma_0^2}{1 - \rho_0^2} \right| + \left| \frac{\sum_{t=1}^T y_t y_{t-1}}{T} - \frac{\sigma_0^2 \rho_0}{1 - \rho_0^2} \right| \\ &\rightarrow 0, \end{aligned} \quad (7.A1.11)$$

as $T \rightarrow \infty$, almost surely with respect to P . In other words, (S3) holds and the convergence is uniform.

7.A1.4 Verification of (S4) for \mathcal{M}_1

In our example, (S4) holds trivially since $h_1(\rho_1) = \frac{(\rho_1 - \rho_0)^2}{2(1 - \rho_0^2)}$, and $|\rho| < 1$ almost surely. Specifically, $\pi(I|\mathcal{M}_1) = 0$.

7.A1.5 Verification of (S5) for \mathcal{M}_1

First note that $h_1(\Theta_1) = \operatorname{ess\ inf}_{\rho_1 \in \Theta_1} h_1(\rho_1) = \operatorname{ess\ inf}_{\rho_1 \in \Theta_1} \frac{(\rho_1 - \rho_0)^2}{2(1 - \rho_0^2)} = 0$. Next, let $\mathcal{G}_T = \Theta_1$, for $T > 0$. Then (S5) (1) and (S5) (2) hold trivially. Validation of (S5) (3) is exactly the same as our proof of uniform convergence of $\frac{\log R_T(\cdot)}{T}$ to $h_1(\cdot)$, provided in Section 7.A1.3. Hence, (S5) is satisfied.

7.A1.6 Verification of (S6) for \mathcal{M}_1

Under (S1) – (S3), which we have already verified, it holds that (see equation (18) of Shalizi (2009)) for any fixed \mathcal{G} of the sequence \mathcal{G}_T , for any $\epsilon > 0$ and for sufficiently large T ,

$$\frac{1}{T} \log \int_{\mathcal{G}} R_T(\rho_1) \pi(\rho_1 | \mathcal{M}_1) d\rho_1 \leq -h_1(\mathcal{G}) + \epsilon + \frac{1}{T} \log \pi(\mathcal{G} | \mathcal{M}_1). \quad (7.A1.12)$$

It follows that $\tau(\mathcal{G}_T, \delta)$ is almost surely finite for all T and δ . We now argue that for sufficiently large T , $\tau(\mathcal{G}_T, \delta) > T$ only finitely often with probability one. By equation (41) of Shalizi (2009),

$$\sum_{T=1}^{\infty} P(\tau(\mathcal{G}_T, \delta) > T) \leq \sum_{T=1}^{\infty} \sum_{m=T+1}^{\infty} P\left(\frac{1}{m} \log \int_{\mathcal{G}_T} R_m(\rho_1) \pi(\rho_1 | \mathcal{M}_1) d\rho_1 > \delta - h_1(\mathcal{G}_T)\right). \quad (7.A1.13)$$

7.A1. ILLUSTRATION OF OUR RESULT ON BAYES FACTOR WITH
COMPETING AR(1) MODELS

Since $\frac{1}{m} \log \int_{\mathcal{G}_T} R_m(\rho_1) \pi(\rho_1 | \mathcal{M}_1) d\rho_1 = \frac{1}{m} \log \int_{|\rho_1| \leq 1} R_m(\rho_1) \pi(\rho_1 | \mathcal{M}_1) d\rho_1$, by the mean value theorem for integrals,

$$\frac{1}{m} \log \int_{\mathcal{G}_T} R_m(\rho_1) \pi(\rho_1 | \mathcal{M}_1) d\rho_1 = \frac{1}{m} \log [R_m(\hat{\rho}_T) \pi(\Theta_1 | \mathcal{M}_1)] = \frac{1}{m} \log [R_m(\hat{\rho}_T)], \quad (7.A1.14)$$

for $\hat{\rho}_T \in [-1, 1]$ depending upon the data.

Since $h_1(\mathcal{G}_T) = h_1((-1, 1)) = 0$, and $h_1(\hat{\rho}_T) \geq 0$, it follows from

$$\frac{1}{m} \log \int_{\mathcal{G}_T} R_m(\rho_1) \pi(\rho_1 | \mathcal{M}_1) d\rho_1 > \delta - h_1(\mathcal{G}_T)$$

and (7.A1.14) that

$$\frac{1}{m} \log R_m(\hat{\rho}_T) + h_1(\hat{\rho}_T) > \delta + h_1(\hat{\rho}_T) > \delta.$$

Thus, it follows from (7.A1.13), (7.A1.6) and (7.A1.8), that

$$\begin{aligned} & \sum_{T=1}^{\infty} P(\tau(\mathcal{G}_T, \delta) > T) \\ & \leq \sum_{T=1}^{\infty} \sum_{m=T+1}^{\infty} P\left(\left|\frac{1}{m} \log R_m(\hat{\rho}_T) + h_1(\hat{\rho}_T)\right| > \delta\right) \\ & \leq \sum_{T=1}^{\infty} \sum_{m=T+1}^{\infty} P\left(\left|\left(\frac{\sum_{t=1}^m y_{t-1}^2}{m}\right)\left(\frac{\hat{\rho}_T - \rho_0}{2}\right) - \frac{\sum_{t=2}^m \epsilon_t y_{t-1}}{m} - \frac{\sigma_0^2 (\hat{\rho}_T - \rho_0)}{2(1 - \rho_0^2)}\right| > \frac{\delta}{\kappa}\right) \\ & \leq \sum_{T=1}^{\infty} \sum_{m=T+1}^{\infty} P\left(\left|\left(\frac{\sum_{t=1}^m y_{t-1}^2}{m}\right)\left(\frac{\hat{\rho}_T - \rho_0}{2}\right) - \frac{\sigma_0^2 (\hat{\rho}_T - \rho_0)}{2(1 - \rho_0^2)}\right| > \frac{\delta}{2\kappa}\right) \quad (7.A1.15) \end{aligned}$$

$$+ \sum_{T=1}^{\infty} \sum_{m=T+1}^{\infty} P\left(\left|\frac{\sum_{t=2}^m \epsilon_t y_{t-1}}{m}\right| > \frac{\delta}{2\kappa}\right). \quad (7.A1.16)$$

We first show that (7.A1.15) is convergent. To simplify arguments, we first approximate

7.A1. ILLUSTRATION OF OUR RESULT ON BAYES FACTOR WITH
COMPETING AR(1) MODELS

$y_t = \sum_{k=1}^t \rho_0^{t-k} \epsilon_k$ by

$$\tilde{y}_t = \sum_{k=t-t_0}^t \rho_0^{t-k} \epsilon_k \quad (7.A1.17)$$

in the “in probability” sense. In \tilde{y}_t , t_0 is such that, for any given $\varepsilon > 0$, for $t > t_0$,

$$\max \left\{ E |\epsilon_1| \times \frac{\rho_0^{t_0+1}}{1-\rho_0}, \frac{\sigma_0^2 \rho_0^{2(t_0+1)}}{1-\rho_0^2} \right\} < \varepsilon. \quad (7.A1.18)$$

Since \tilde{y}_t consists of only $t_0 + 1$ terms for any $t > t_0$, it is easier to handle compared to y_t , whose number of terms increases with t . Importantly, \tilde{y}_t and \tilde{y}_{t+t_0+k} are independent, for any $k \geq 1$. This property, which is not possessed by y_t , will be instrumental for making most of the terms zero associated with multinomial expansions required in our proceeding.

For the “in probability” fact, note that

$$E |y_t - \tilde{y}_t| \leq E |\epsilon_1| \sum_{k=1}^{t-t_0-1} \rho_0^{t-k} = E |\epsilon_1| \times \frac{\rho_0^{t_0+1} (1 - \rho_0^{t-t_0-1})}{1 - \rho_0} < \varepsilon, \quad (7.A1.19)$$

and

$$E |y_t - \tilde{y}_t|^2 = \sigma_0^2 \sum_{k=1}^{t-t_0-1} \rho_0^{2(t-k)} = \frac{\sigma_0^2 \rho_0^{2(t_0+1)} (1 - \rho_0^{2(t-t_0-1)})}{1 - \rho_0^2} < \varepsilon, \quad (7.A1.20)$$

due to (7.A1.18). Since $\varepsilon > 0$ is arbitrary, it follows that

$$|y_t - \tilde{y}_t| \xrightarrow{P} 0, \text{ as } t \rightarrow \infty, \quad (7.A1.21)$$

where “ \xrightarrow{P} ” indicates “in probability” convergence. Now, $|y_t^2 - \tilde{y}_t^2| = |y_t + \tilde{y}_t| \times |y_t - \tilde{y}_t|$, where y_t is an irreducible, aperiodic Markov chain with mean zero Gaussian asymptotic stationary distribution with variance $\sigma_0^2 / (1 - \rho_0^2)$, and \tilde{y}_t is also asymptotically Gaussian with mean zero and variance $\sigma_0^2 (1 - \rho_0^{2(t_0+1)}) / (1 - \rho_0^2)$. Hence, $|y_t + \tilde{y}_t|$ converges

7.A1. ILLUSTRATION OF OUR RESULT ON BAYES FACTOR WITH
COMPETING AR(1) MODELS

in probability to a finite random variable, and because of (7.A1.21), it follows from the above representation that

$$|y_t^2 - \tilde{y}_t^2| \xrightarrow{P} 0, \text{ as } t \rightarrow \infty. \quad (7.A1.22)$$

It then follows from the representation

$$\left| \frac{\sum_{t=1}^T y_t^2}{T} - \frac{\sum_{t=1}^T \tilde{y}_t^2}{T} \right| \leq \frac{\sum_{t=1}^T |y_t^2 - \tilde{y}_t^2|}{T},$$

(7.A1.22), and Theorem 7.15 of Schervish (1995) that

$$\left| \frac{\sum_{t=1}^m y_t^2}{m} - \frac{\sum_{t=1}^m \tilde{y}_t^2}{m} \right| \xrightarrow{P} 0, \text{ as } m \rightarrow \infty. \quad (7.A1.23)$$

Now note that for any finite integer $p \geq 1$,

$$\sup_{m \geq 1} E \left(\frac{\sum_{t=1}^m y_t^2}{m} - \frac{\sum_{t=1}^m \tilde{y}_t^2}{m} \right)^p \leq 2^{p-1} \sup_{m \geq 1} E \left(\frac{\sum_{t=1}^m y_t^2}{m} \right)^p + 2^{p-1} \sup_{m \geq 1} E \left(\frac{\sum_{t=1}^m \tilde{y}_t^2}{m} \right)^p. \quad (7.A1.24)$$

Noting that the multinomial expansion $(a_1 + a_2 + \dots + a_m)^p = \sum_{b_1+b_2+\dots+b_m=p} \prod_{j=1}^m a_j^{b_j}$ (where b_1, \dots, b_m are non-negative integers) consists of $\binom{m+p-1}{p}$ terms, it follows using asymptotic stationarity of y_t and \tilde{y}_t that both the expectations on the right hand side of (7.A1.24) are of the order $O(1)$, as $m \rightarrow \infty$. Also, since for any finite m , the expectations are finite, it follows that the right hand side of (7.A1.24) is finite, from which uniform integrability, and hence

$$E \left| \frac{\sum_{t=1}^m y_t^2}{m} - \frac{\sum_{t=1}^m \tilde{y}_t^2}{m} \right|^p \rightarrow 0, \text{ as } m \rightarrow \infty, \quad (7.A1.25)$$

follows for integers $p \geq 1$. Hence, using binomial expansion, the Cauchy-Schwartz

7.A1. ILLUSTRATION OF OUR RESULT ON BAYES FACTOR WITH
COMPETING AR(1) MODELS

inequality and (7.A1.25), it follows that

$$\begin{aligned} & E \left| \frac{\sum_{t=1}^m y_t^2}{m} \right|^p - E \left| \frac{\sum_{t=1}^m \tilde{y}_t^2}{m} \right|^p \\ &= E \left| \left(\frac{\sum_{t=1}^m y_t^2}{m} - \frac{\sum_{t=1}^m \tilde{y}_t^2}{m} \right) + \frac{\sum_{t=1}^m \tilde{y}_t^2}{m} \right|^p - E \left| \frac{\sum_{t=1}^m \tilde{y}_t^2}{m} \right|^p \\ &\leq \sum_{k=0}^{p-1} \binom{p}{k} \left\{ E \left(\left| \frac{\sum_{t=1}^m y_t^2}{m} - \frac{\sum_{t=1}^m \tilde{y}_t^2}{m} \right|^{2(p-k)} \right) \right\}^{1/2} \times \left\{ E \left(\left| \frac{\sum_{t=1}^m \tilde{y}_t^2}{m} \right|^{2k} \right) \right\}^{1/2}, \end{aligned}$$

so that

$$\begin{aligned} \frac{E \left| \frac{\sum_{t=1}^m y_t^2}{m} \right|^p}{E \left| \frac{\sum_{t=1}^m \tilde{y}_t^2}{m} \right|^p} - 1 &\leq \sum_{k=0}^{p-1} \binom{p}{k} \left\{ E \left(\left| \frac{\sum_{t=1}^m y_t^2}{m} - \frac{\sum_{t=1}^m \tilde{y}_t^2}{m} \right|^{2(p-k)} \right) \right\}^{1/2} \times \left\{ \frac{E \left(\left| \frac{\sum_{t=1}^m \tilde{y}_t^2}{m} \right|^{2k} \right)}{E \left| \frac{\sum_{t=1}^m \tilde{y}_t^2}{m} \right|^{2p}} \right\}^{1/2} \\ &\rightarrow 0, \text{ as } m \rightarrow \infty \text{ (due to (7.A1.25))}. \end{aligned} \tag{7.A1.26}$$

In other words, for $p \geq 1$,

$$E \left| \frac{\sum_{t=1}^m y_t^2}{m} \right|^p \sim E \left| \frac{\sum_{t=1}^m \tilde{y}_t^2}{m} \right|^p, \text{ as } m \rightarrow \infty. \tag{7.A1.27}$$

Hence, while applying Markov's inequality to the probability terms of the series (7.A1.15), we can replace the moments associated with y_t with those associated with \tilde{y}_t , for $m > T_0$, where T_0 is sufficiently large.

Now observe that

$$\begin{aligned} & P \left(\left| \left(\frac{\sum_{t=1}^m y_{t-1}^2}{m} \right) \left(\frac{\hat{\rho}_T - \rho_0}{2} \right) - \frac{\sigma_0^2 (\hat{\rho}_T - \rho_0)}{2(1 - \rho_0^2)} \right| > \frac{\delta}{2\kappa} \right) \\ &\leq P \left(\left| \left(\frac{\sum_{t=1}^m [y_{t-1}^2 - E(y_{t-1}^2)]}{m} \right) \left(\frac{\hat{\rho}_T - \rho_0}{2} \right) \right| > \frac{\delta}{4\kappa} \right) \end{aligned} \tag{7.A1.28}$$

$$+ P \left(\left| \frac{\hat{\rho}_T - \rho_0}{2} \right| \times \left| \frac{\sum_{t=1}^m E(y_t^2)}{m} - \frac{\sigma_0^2}{1 - \rho_0^2} \right| > \frac{\delta}{4\kappa} \right). \tag{7.A1.29}$$

7.A1. ILLUSTRATION OF OUR RESULT ON BAYES FACTOR WITH
COMPETING AR(1) MODELS

For $m > T_0$, where T_0 is sufficiently large, $\left| \frac{\hat{\rho}_T - \rho_0}{2} \right| \times \left| \frac{\sum_{t=1}^m E(y_t^2)}{m} - \frac{\sigma_0^2}{1-\rho_0^2} \right| < \frac{\delta}{4\kappa}$, so that the (7.A1.29) is exactly zero for $m > T_0$. Using Markov's inequality for (7.A1.29) where $m > T_0$ and replacing y_t with \tilde{y}_t in the right hand side of Markov's inequality using (7.A1.27) we obtain

$$\begin{aligned} P \left(\left| \left(\frac{\sum_{t=1}^m [y_{t-1}^2 - E(y_{t-1}^2)]}{m} \right) \left(\frac{\hat{\rho}_T - \rho_0}{2} \right) \right| > \frac{\delta}{4\kappa} \right) \\ < C \left(\frac{4\kappa}{\delta} \right)^5 \left(\frac{\hat{\rho}_T - \rho_0}{2} \right)^5 E \left(\frac{\sum_{t=1}^m [\tilde{y}_{t-1}^2 - E(\tilde{y}_{t-1}^2)]}{m} \right)^5, \end{aligned} \quad (7.A1.30)$$

where C is a positive constant. Now, $(\sum_{t=1}^m [\tilde{y}_{t-1}^2 - E(\tilde{y}_{t-1}^2)])^5$ admits the multinomial expansion of the form $(a_1 + a_2 + \dots + a_m)^5 = \sum_{b_1+b_2+\dots+b_m=5} \prod_{t=1}^m a_t^{b_t}$, where $a_t = [\tilde{y}_{t-1}^2 - E(\tilde{y}_{t-1}^2)]$ and b_1, \dots, b_m are non-negative integers. Observe that for any $t \geq 1$, a_t and a_{t+k} are independent for any $k \geq 1$, which enables factorization of $E(\prod_{t=1}^m a_t^{b_t})$ into products of expectations of the independent terms. Since $E(a_t) = 0$ for $t = 2, \dots, m$, the expected product term becomes zero whenever it consists of at least one term of the form $E(a_t)$, for any $t = 2, \dots, m$.

For the sake of convenience, let $m = (s+1)(t_0+1)$, where $s (\geq 1)$ is an integer. Let $A_l = \{a_t : t = (l-1)t_0 + 1, \dots, l(t_0+1)\}$, for $l = 1, \dots, (s+1)$. Then A_l and A_{l+2+r} are independent sets for any integer $l \geq 1$ and any integer $r \geq 0$.

When at least one $b_t = 1$, the following argument gives an upper bound on the number of ways $E(\prod_{t=1}^m a_t^{b_t})$ can be non-zero. Consider selecting 5 sets, say, $\{A_l, A_{l+1}, A_{l+2}, A_{l+3}, A_{l+4}\}$ from $\{A_1, \dots, A_{s+1}\}$, for some $l \geq 1$. Let $B_l = \{b_t : t = (l-1)t_0 + 1, \dots, l(t_0+1)\}$ for $l = 1, \dots, (s+1)$, and consider setting one element of each of B_{l+r} ; $r = 0, \dots, 4$, to be 1 and the rest of the b_t 's to be zero. Then the number of such cases, namely, $O((s+1))$ (since t_0 is a constant), provides an upper bound on the number of possible ways $E(\prod_{t=1}^m a_t^{b_t})$ can be non-zero when at least one $b_t = 1$.

Further cases of non-zero $E(\prod_{t=1}^m a_t^{b_t})$ can occur when one of the b_t 's is 5 and the

7.A1. ILLUSTRATION OF OUR RESULT ON BAYES FACTOR WITH
COMPETING AR(1) MODELS

rest are zeros, and when one of the b_t is 3, another is 2, and the rest are zeros, so that there are $m + m(m - 1) = m^2$ cases with respect to such choices.

Hence, in all there are $O(m^2)$ possible cases when $E\left(\prod_{t=1}^m a_t^{b_t}\right)$ is non-zero, and in the remaining cases $E\left(\prod_{t=1}^m a_t^{b_t}\right) = 0$. In other words,

$$\left(\frac{4\kappa}{\delta}\right)^5 \left(\frac{\hat{\rho}_T - \rho_0}{2}\right)^5 E\left(\frac{\sum_{t=1}^m [\tilde{y}_{t-1}^2 - E(\tilde{y}_{t-1}^2)]}{m}\right)^5 = O(m^{-3}), \quad (7.A1.31)$$

since $\hat{\rho}_T \in [-1, 1]$.

Now, (7.A1.15) converges if and only if

$$\sum_{T=T_0}^{\infty} \sum_{m=T+1}^{\infty} P\left(\left|\left(\frac{\sum_{t=1}^m y_{t-1}^2}{m}\right)\left(\frac{\hat{\rho}_T - \rho_0}{2}\right) - \frac{\sigma_0^2(\hat{\rho}_T - \rho_0)}{2(1 - \rho_0^2)}\right| > \frac{\delta}{\kappa}\right) < \infty, \quad (7.A1.32)$$

for sufficiently large T_0 . Due to (7.A1.28), (7.A1.29) (which is exactly zero for $m > T_0$), (7.A1.30) and (7.A1.31), we see that (7.A1.32) is dominated by some finite positive constant times the series

$$\begin{aligned} \sum_{T=T_0}^{\infty} \sum_{m=T+1}^{\infty} \frac{1}{m^3} &= \frac{1}{(T_0+1)^3} + \frac{1}{(T_0+2)^3} + \frac{1}{(T_0+3)^3} + \dots \\ &\quad + \frac{1}{(T_0+2)^3} + \frac{1}{(T_0+3)^3} + \dots \\ &\quad + \frac{1}{(T_0+3)^3} + \dots \\ &\quad + \dots \\ &\quad \vdots \\ &= \sum_{k=1}^{\infty} \frac{k}{(T_0+k)^3}. \end{aligned} \quad (7.A1.33)$$

The series (7.A1.33) is convergent since it is bounded above by $\sum_{k=1}^{\infty} \frac{(T_0+k)}{(T_0+k)^3} \leq \sum_{k=1}^{\infty} \frac{1}{k^2} <$

∞ .

Similar (and simpler) arguments and using the result

$$\left| \frac{\sum_{t=1}^m \epsilon_t y_{t-1}}{m} - \frac{\sum_{t=1}^m \epsilon_t \tilde{y}_{t-1}}{m} \right| \leq \frac{\sum_{t=1}^m |\epsilon_t| |y_{t-1} - \tilde{y}_{t-1}|}{m} \xrightarrow{P} 0, \text{ as } m \rightarrow \infty,$$

shows that the series (7.A1.16) also converges. Hence, (S6) stands verified.

Thus, (S1)–(S6) holds for \mathcal{M}_1 .

7.A2 Verification of Shalizi's conditions for model \mathcal{M}_2

We now verify the same set of conditions for \mathcal{M}_2 . As in \mathcal{M}_1 , (S1) and (S2) easily hold; here $h_2(\rho_2) = \frac{(\rho_2 - \rho_0)^2}{2(1 - \rho_0^2)}$ is of the same form as h_1 . With respect to (S3) we verify pointwise convergence as required, rather than uniform convergence as in \mathcal{M}_1 . Using (7.A1.7), (7.A1.8), (7.A1.9) and (7.A1.10), it is easily seen that $\frac{\log R_T(\rho_2)}{T} + h_2(\rho_2) \rightarrow 0$ almost surely, for all $\rho_2 \in \Theta_2$. As in \mathcal{M}_1 , it is clear that $\pi(I|\mathcal{M}_2) = 0$ so that (S4) holds.

As regards (S5), note that

$$h_2(\Theta_2) = \min \left\{ \frac{(1 - \rho_0)^2}{2(1 - \rho_0^2)}, \frac{(1 + \rho_0)^2}{2(1 - \rho_0^2)} \right\}. \quad (7.A2.1)$$

Now, in contrast with \mathcal{M}_1 , here let $\mathcal{G}_T = \left\{ \rho_2 \in \Theta_2 : |\rho_2| \leq \beta^{\frac{1}{q_1}} T^{\frac{1}{q_1}} \right\}$, where $q_1 > 5$. This q_1 is the power associated with the Markov inequality of the form similar to (7.A1.30) required in verification of (S6) for model \mathcal{M}_2 . It is easily seen that $\mathcal{G}_T \rightarrow \Theta_2$ and $h_2(\mathcal{G}_T) \rightarrow h_2(\Theta_2)$, as $T \rightarrow \infty$, so that (S5) (1) holds. To see that (S5) (2) is satisfied, note that by Markov's inequality, $\pi(\mathcal{G}_T) > 1 - E(\exp(\alpha|\rho_2|^{q_1}) \exp(-\alpha\beta T))$, where $\alpha (> 0)$ is such that $E(\exp(\alpha|\rho_2|^{q_1})) < \infty$. We choose β so large that $\alpha\beta > h_2(\Theta_2)$.

Since \mathcal{G}_T is compact for all $T \geq 1$, uniform convergence as required will be proven if we can show that $\frac{1}{T} \log R_T(\rho_2) + h_2(\rho_2)$ is stochastically equicontinuous almost surely in $\rho_2 \in \mathcal{G}$ for any $\mathcal{G} \in \{\mathcal{G}_T : T = 1, 2, \dots\}$ and $\frac{1}{T} \log R_T(\rho_2) + h_2(\rho_2) \rightarrow 0$, almost surely,

for all $\rho_2 \in \mathcal{G}$. Since we have already verified pointwise convergence of the above for all $\rho_2 \in \Theta_2$ while verifying (S4), it remains to prove stochastic equicontinuity of $\frac{1}{T} \log R_T(\cdot) + h_2(\cdot)$. Stochastic equicontinuity usually follows easily if one can prove that the function concerned is almost surely Lipschitz continuous. In our case, for any $\rho_2^{(1)}, \rho_2^{(2)} \in \mathcal{G}$,

$$\begin{aligned} & \left| \frac{1}{T} \log R_T(\rho_2^{(1)}) + h_2(\rho_2^{(1)}) - \frac{1}{T} \log R_T(\rho_2^{(2)}) - h_2(\rho_2^{(2)}) \right| \\ &= \left| \rho_2^{(1)} - \rho_2^{(2)} \right| \times \left| \left(\frac{\sum_{t=1}^T y_{t-1}^2}{T} \right) \left(\frac{\rho_2^{(1)} + \rho_2^{(2)}}{2\sigma_0^2} \right) - \frac{\sum_{t=1}^T y_t y_{t-1}}{T} \times \frac{1}{\sigma_0^2} - \frac{\sigma_0^2}{2(1 - \rho_0^2)} \right|. \end{aligned} \quad (7.A2.2)$$

By (7.A1.7) and (7.A1.10), $\frac{\sum_{t=1}^T y_{t-1}^2}{T}$ and $\frac{\sum_{t=1}^T y_t y_{t-1}}{T}$ converge almost surely to $\sigma_0^2/(1 - \rho_0^2)$ and $\sigma_0^2 \rho_0/(1 - \rho_0^2)$, respectively, while $\rho_2^{(1)} + \rho_2^{(2)}$ is bounded since $\rho_2^{(1)}, \rho_2^{(2)} \in \mathcal{G}$. It follows that $\frac{1}{T} \log R_T(\rho_2) + h_2(\rho_2)$ is stochastically equicontinuous. Hence, uniform convergence as required by (S5) (3), follows. That is, (S5) is satisfied for \mathcal{M}_2 .

We now verify (S6). First note that almost sure finiteness of $\tau(\mathcal{G}_T, \delta)$ is guaranteed in the same way as in model \mathcal{M}_1 . We hence need to verify that $T \geq \tau(\mathcal{G}_T, \delta)$ almost surely, for all sufficiently large T and for all $\delta > 0$. Again by equation (41) of Shalizi (2009),

$$\sum_{T=1}^{\infty} P(\tau(\mathcal{G}_T, \delta) > T) \leq \sum_{T=1}^{\infty} \sum_{m=T+1}^{\infty} P \left(\frac{1}{m} \log \int_{\mathcal{G}_T} R_m(\rho_2) \pi(\rho_2 | \mathcal{M}_2) d\rho_2 > \delta - h_2(\mathcal{G}_T) \right). \quad (7.A2.3)$$

Now since $R_n(\rho_2)$ is also continuous in ρ_2 , by the mean value theorem for integrals it holds that

$$\frac{1}{m} \log \int_{\mathcal{G}_T} R_m(\rho_2) d\pi(\rho_2) = \frac{1}{m} \log [R_m(\tilde{\rho}_2) \pi(\mathcal{G}_T)] = \frac{1}{m} \log R_m(\tilde{\rho}_2) + \frac{1}{m} \log \pi(\mathcal{G}_T), \quad (7.A2.4)$$

for $\tilde{\rho}_2 \in \mathcal{G}_T$, perhaps depending upon the data. Thus, $\frac{1}{m} \log \int_{\mathcal{G}_T} R_m(\rho_2) d\pi(\rho_2) >$

$\delta - h_2(\mathcal{G}_T)$ implies, since $h_2(\tilde{\rho}_2) \geq h_2(\mathcal{G}_T)$, that $\frac{1}{m} \log R_m(\tilde{\rho}_2) + \frac{1}{m} \log \pi(\mathcal{G}_T) + h_2(\tilde{\rho}_2) > \delta$, so that $\frac{1}{m} \log R_m(\tilde{\rho}_2) + h_2(\tilde{\rho}_2) > \delta$, as $\delta - \frac{1}{m} \log \pi(\mathcal{G}_T) > \delta$. Again this implies that $|\frac{1}{m} \log R_m(\tilde{\rho}_2) + h_2(\tilde{\rho}_2)| > \delta$, from which it finally follows that $\sup_{\theta \in \mathcal{G}_T} |\frac{1}{m} \log R_m(\rho_2) + h_2(\rho_2)| > \delta$. Hence,

$$P \left(\frac{1}{m} \log \int_{\mathcal{G}_T} R_m(\rho_2) d\pi(\rho_2) > \delta - h_2(\mathcal{G}_T) \right) \leq P \left(\sup_{\rho_2 \in \mathcal{G}_T} \left| \frac{1}{m} \log R_m(\rho_2) + h_2(\rho_2) \right| > \delta \right). \quad (7.A2.5)$$

Since \mathcal{G}_T is compact, there exist finite number of open sets \mathcal{O}_{iT} ; $i = 1, \dots, p_T$, with $p_T (\geq 1)$ finite for each $T \geq 1$, such that $\mathcal{G}_T \subseteq \cup_{i=1}^{p_T} \mathcal{O}_{iT}$. Here, for $i = 1, \dots, p_T$, $\mathcal{O}_{iT} = \{\rho_2 : |\rho_2 - c_{iT}| \leq r/2\}$, where $c_{iT} \in \mathcal{G}_T$ and $r > 0$ is such that by stochastic equicontinuity, $|\rho_2^{(1)} - \rho_2^{(2)}| < r$ implies

$$\left| \frac{1}{m} \log R_m(\rho_2^{(1)}) + h_2(\rho_2^{(1)}) - \frac{1}{m} \log R_m(\rho_2^{(2)}) - h_2(\rho_2^{(2)}) \right| \leq \delta/2, \quad (7.A2.6)$$

for sufficiently large m , almost surely. Indeed, observe that if $\rho_2^{(1)}, \rho_2^{(2)} \in \mathcal{O}_{iT}$ for any $i = 1, \dots, p_T$, then $|\rho_2^{(1)} - \rho_2^{(2)}| \leq |\rho_2^{(1)} - c_{iT}| + |\rho_2^{(2)} - c_{iT}| < r$. With these, it then follows that

$$\begin{aligned} & P \left(\sup_{\rho_2 \in \mathcal{G}_T} \left| \frac{1}{m} \log R_m(\rho_2) + h_2(\rho_2) \right| > \delta \right) \\ &= 1 - P \left(\sup_{\theta \in \mathcal{G}_T} \left| \frac{1}{m} \log R_m(\rho_2) + h_2(\rho_2) \right| \leq \delta \right) \\ &= 1 - P \left(\sup_{\theta \in \mathcal{O}_{iT}} \left| \frac{1}{m} \log R_m(\rho_2) + h_2(\rho_2) \right| \leq \delta, i = 1, \dots, p_T \right). \end{aligned} \quad (7.A2.7)$$

Now for any $\rho_{iT}^* \in \mathcal{O}_{iT}$,

$$\left| \frac{1}{m} \log R_m(\rho_{iT}^*) + h_2(\rho_{iT}^*) \right| \leq \delta/2, \quad (7.A2.8)$$

for sufficiently large m due to pointwise convergence of $|\frac{1}{m} \log R_m(\rho_2) + h_2(\rho_2)|$ to zero

for all $\rho_2 \in \Theta_2$ as we verified in the context of (S3) (indeed, due to uniform convergence to zero over $\mathcal{G}_T \setminus I$, as we verified in the context of (S5) (2)). Then for any $\rho_2 \in \mathcal{O}_{iT}$,

$$\begin{aligned} \left| \frac{1}{m} \log R_m(\rho_2) + h_2(\rho_2) \right| &\leq \left| \frac{1}{m} \log R_m(\rho_2) + h_2(\rho_2) - \frac{1}{m} \log R_m(\rho_{iT}^*) - h_2(\rho_{iT}^*) \right| \\ &\quad + \left| \frac{1}{m} \log R_m(\rho_{iT}^*) + h_2(\rho_{iT}^*) \right|. \end{aligned} \quad (7.A2.9)$$

By (7.A2.6) and (7.A2.8) respectively, the first and second terms of the right hand side of (7.A2.9) are less than $\delta/2$, for sufficiently large m . Hence, for sufficiently large m ,

$$\sup_{\rho_2 \in \mathcal{O}_{iT}} \left| \frac{1}{m} \log R_m(\rho_2) + h_2(\rho_2) \right| \leq \delta.$$

It then follows from (7.A2.7) that

$$\begin{aligned} P \left(\sup_{\theta \in \mathcal{G}_T} \left| \frac{1}{m} \log R_m(\rho_2) + h_2(\rho_2) \right| > \delta \right) \\ \leq 1 - P \left(\left| \frac{1}{m} \log R_m(\rho_{iT}^*) + h_2(\rho_{iT}^*) \right| \leq \delta, i = 1, \dots, p_T \right) \\ \leq \sum_{i=1}^{p_T} P \left(\left| \frac{1}{m} \log R_m(\rho_{iT}^*) + h_2(\rho_{iT}^*) \right| > \delta \right). \end{aligned} \quad (7.A2.10)$$

Combining (7.A2.10) with (7.A2.5) and (7.A2.3) we observe that it is now required to prove finiteness of the following sum:

$$\sum_{T=1}^{\infty} \sum_{m=T+1}^{\infty} \sum_{i=1}^{p_T} P \left(\left| \frac{1}{m} \log R_m(\rho_{iT}^*) + h_2(\rho_{iT}^*) \right| > \delta \right). \quad (7.A2.11)$$

Using the same ideas as in Section 7.A1.6 for verification of (S6) for \mathcal{M}_1 , but the right hand side of the Markov's inequality (7.A1.30) raised to an appropriate power $q_1 (> 5)$ instead of 5, we find that $P \left(\left| \frac{1}{m} \log R_m(\rho_{iT}^*) + h_2(\rho_{iT}^*) \right| > \delta \right)$ is bounded above by an expression of the form $a_{iT} m^{-q_2}$, for some $a_{iT} > 0$ depending on ρ_{iT}^* , where $q_2 > 4$.

But since $\rho_{iT}^* \in \mathcal{O}_{iT}$ with center c_{iT} satisfying $1 < |c_{iT}|^{q_1} < \beta T$ for all i , it is easy to see that $a_{iT}m^{-q_2} = O(Tm^{-q_2})$. Hence, $\sum_{i=1}^{p_T} P\left(\left|\frac{1}{m} \log R_m(\rho_{iT}^*) + h_2(\rho_{iT}^*)\right| > \delta\right) = O(p_T T m^{-q_2})$. Now p_T is the number of open balls with radius $r/2$ required to cover \mathcal{G}_T . By Lemma 1 of Lorentz (1966), $\frac{\beta T}{r} \leq p_T \leq \frac{6\beta T}{r}$. Hence, $\sum_{i=1}^{p_T} P\left(\left|\frac{1}{m} \log R_m(\rho_{iT}^*) + h_2(\rho_{iT}^*)\right| > \delta\right) = O(T^2 m^{-q_2})$. We then have, for sufficiently large T_0 ,

$$\sum_{T=T_0}^{\infty} \sum_{m=T+1}^{\infty} \sum_{i=1}^{p_T} P\left(\left|\frac{1}{m} \log R_m(\rho_{iT}^*) + h_2(\rho_{iT}^*)\right| > \delta\right) < \sum_{k=1}^{\infty} \frac{k}{(T_0 + k)^{q_2-2}} < \frac{1}{k^{q_2-3}} < \infty, \quad (7.A2.12)$$

since $q_2 > 4$. In other words, (S6) holds for model \mathcal{M}_2 .

Hence, Theorem 25, so that

$$\lim_{T \rightarrow \infty} \frac{1}{T} \log[B_T^{(12)}] = h_2(\Theta_2), \quad (7.A2.13)$$

that is, the Bayes factor heavily favours the (asymptotically) stationary model \mathcal{M}_1 over the nonstationary model \mathcal{M}_2 . Since the true model P is (asymptotically) stationary, this result is very encouraging.

7.A3 Convergence of Bayes factor when ρ_1, ρ_2, σ_1 and σ_2 are all unknown

When apart from unknown ρ_1 and ρ_2 , the error variances σ_1^2 and σ_2^2 associated with models \mathcal{M}_1 and \mathcal{M}_2 are also unknown, we consider the parameter spaces $\Theta_1 = \{(\rho_1, \sigma_1) : |\rho_1| < 1, \sigma_1 > 0\}$ and $\Theta_2 = \{(\rho_2, \sigma_2) : \rho_2 \in (-1, 1)^c, \sigma_2 > 0\}$ associated with models \mathcal{M}_1 and \mathcal{M}_2 , respectively. For $i = 1, 2$, we assume joint priors $\pi(\rho_i, \sigma_i | \mathcal{M}_i)$, having densities on Θ_i , with respect to the Lebesgue measure. It can be easily seen that

in this case, for $i = 1, 2$,

$$h_i(\rho_i, \sigma_i) = \frac{1}{2(1 - \rho_0^2)} \left[\left(\rho_0 - \frac{\sigma_0 \rho_i}{\sigma_i} \right)^2 + \frac{\sigma_0^2}{\sigma_i^2} - (1 - \rho_0^2) \log \frac{\sigma_0^2}{\sigma_i^2} - 1 \right]. \quad (7.A3.1)$$

Since $(1 - \rho_0^2) \log \frac{\sigma_0^2}{\sigma_i^2} + 1 \leq \log \frac{\sigma_0^2}{\sigma_i^2} + 1 \leq \frac{\sigma_0^2}{\sigma_i^2}$, (7.A3.1) is non-negative. Also, as in the case with $\sigma_1 = \sigma_2 = \sigma_0$, it holds that $h_1(\Theta_1) = 0$ and $h_2(\Theta_2) = \min \left\{ \frac{(1-\rho_0)^2}{2(1-\rho_0^2)}, \frac{(1+\rho_0)^2}{2(1-\rho_0^2)} \right\}$. Further, note that $\pi(I|\mathcal{M}_i) = 0$, for $i = 1, 2$. Thus, conditions (S1)–(S4) are easily seen to hold for both the competing models.

We now verify the remaining conditions for the models. As regards \mathcal{G}_T , here we set

$$\mathcal{G}_T = \left\{ (\rho_1, \sigma_1) : |\rho_1| < 1, \frac{1}{T^{1/2q_1}} \leq \sigma_1 \leq \beta T \right\}$$

for model \mathcal{M}_1 where $\beta > h_1(\Theta_1) = 0$, and for model \mathcal{M}_2 we set

$$\mathcal{G}_T = \left\{ (\rho_2 \in \Theta_2, \sigma_2 \geq 0) : |\rho_2| \leq \beta T, \frac{1}{T^{1/2q_1}} \leq \sigma_2 \leq \beta T \right\}.$$

Note that there exists $T_0 \geq 1$ such that $\frac{1}{T^{1/2q_1}} \leq \sigma_0 \leq \beta T$ for $T \geq T_0$. Hence, $h_1(\mathcal{G}_T) = h_1(\Theta_1) = 0$ and $h_2(\mathcal{G}_T) = h_2(\Theta_2) = \min \left\{ \frac{(1-\rho_0)^2}{2(1-\rho_0^2)}, \frac{(1+\rho_0)^2}{2(1-\rho_0^2)} \right\}$, for $T \geq T_0$. Hence, (S5) (1) holds for both \mathcal{M}_1 and \mathcal{M}_2 . Letting E_1 denote the expectation with respect to $\pi(\cdot|\mathcal{M}_1)$, now observe that for $\alpha > 0$ such that $E_1 [\exp(\alpha \sigma_1)] < \infty$ and

$$E_1 \left[\exp \left(\frac{\alpha\beta}{\sigma_1^{2q_1}} \right) \right] < \infty,$$

$$\begin{aligned} \pi(\mathcal{G}_T | \mathcal{M}_1) &= \pi \left(\frac{1}{T^{1/2q_1}} \leq \sigma_1 \leq \beta T \right) = \pi(\sigma_1 \leq \beta T) - \pi \left(\sigma_1 \leq \frac{1}{T^{1/2q_1}} \right) \\ &> 1 - E_1 [\exp(\alpha\sigma_1)] \exp(-\alpha\beta T) - \pi \left(\frac{1}{\sigma_1} \geq T^{\frac{1}{2q_1}} \right) \\ &> 1 - E_1 [\exp(\alpha\sigma_1)] \exp(-\alpha\beta T) - E_1 \left[\exp \left(\frac{\alpha\beta}{\sigma_1^{2q_1}} \right) \right] \exp(-\alpha\beta T) \\ &= 1 - \left(E_1 [\exp(\alpha\sigma_1)] + E_1 \left[\exp \left(\frac{\alpha\beta}{\sigma_1^{2q_1}} \right) \right] \right) \exp(-\alpha\beta T), \end{aligned} \quad (7.A3.2)$$

so that (S5) (2) holds for \mathcal{M}_1 , since $\alpha\beta > 0 = h_1(\Theta_1)$.

For \mathcal{M}_2 , denoting by E_2 the expectation with respect to $\pi(\cdot | \mathcal{M}_2)$, and assuming the existence of $\alpha > 0$ such that $E_2 [\exp(\alpha\sigma_2)] < \infty$, $E_2 \left[\exp \left(\frac{\alpha\beta}{\sigma_2^{2q_1}} \right) \right] < \infty$ and $E_2 [\exp(\alpha|\rho_2|)] < \infty$, note that

$$\pi(\mathcal{G}_T | \mathcal{M}_2) = \pi \left(\frac{1}{T^{1/2q_1}} \leq \sigma_2 \leq \beta T | \mathcal{M}_2 \right) - \pi \left(|\rho_2| > \beta T, \frac{1}{T^{1/2q_1}} \leq \sigma_2 \leq \beta T | \mathcal{M}_2 \right),$$

where

$$\pi \left(\frac{1}{T^{1/2q_1}} \leq \sigma_2 \leq \beta T | \mathcal{M}_2 \right) > 1 - \left(E_2 [\exp(\alpha\sigma_2)] + E_2 \left[\exp \left(\frac{\alpha\beta}{\sigma_2^{2q_1}} \right) \right] \right) \exp(-\alpha\beta T),$$

in the same way as (7.A3.2), and

$$\pi \left(|\rho_2| > \beta T, \frac{1}{T^{1/2q_1}} \leq \sigma_2 \leq \beta T | \mathcal{M}_2 \right) \leq \pi(|\rho_2| > \beta T | \mathcal{M}_2) < E_2 [\exp(\alpha|\rho_2|)] \exp(-\alpha\beta T),$$

by Markov's inequality. It follows that

$$\pi(\mathcal{G}_T | \mathcal{M}_2) > 1 - \left(E_2 [\exp(\alpha\sigma_2)] + E_2 \left[\exp \left(\frac{\alpha\beta}{\sigma_2^{2q_1}} \right) \right] + E_2 (\exp(\alpha|\rho_2|)) \right) \exp(-\alpha\beta T),$$

that is, (S5) (2) holds for \mathcal{M}_2 , with β large enough such that $\alpha\beta > h_2(\Theta_2)$. For both \mathcal{M}_1 and \mathcal{M}_2 , (S5) (3) can be seen to hold in almost the same way as in Section 7.A2 using compactness of $\mathcal{G} \in \{\mathcal{G}_k; k \geq 1\}$, and stochastic equicontinuity utilizing the assumption that σ_2 is bounded away from zero in \mathcal{G} . Indeed, (S6) for model \mathcal{M}_1 can be verified in almost the same way as in Section 7.A2. Here we note that the number of open balls with radius $r/2$ required to cover \mathcal{G}_T for \mathcal{M}_1 still remains of the order T as $|\rho_1|$ is bounded above by the constant 1 in \mathcal{G}_T . But since σ^2 is unknown, an extra factor of the order T would emerge after raising the right hand side of the Markov's inequality (7.A1.30) to the power q_1 , which is actually the lower bound of σ^{2q_1} , where $\sigma \in \mathcal{G}_T$ features in the aforementioned Markov inequality. Hence, $\sum_{i=1}^{p_T} P(|\frac{1}{m} \log R_m(\theta_{iT}^*) + h_1(\theta_{iT}^*)| > \delta) = O(T^2 m^{-q_2})$, where $q_2 > 4$, and $\theta_{iT}^* = (\rho_{iT}^*, \sigma_{iT}^*) \in \mathcal{O}_{iT}$. In exactly the same way as in (7.A2.12) it then follows that for sufficiently large T_0 , $\sum_{T=T_0}^{\infty} \sum_{m=T+1}^{\infty} \sum_{i=1}^{p_T} P(|\frac{1}{m} \log R_m(\theta_{iT}^*) + h_1(\theta_{iT}^*)| > \delta) < \infty$. Hence, (S6) holds for model \mathcal{M}_1 .

For model \mathcal{M}_2 , p_T is of the order T^2 , instead of T in the previous case. Note that here we need q_1 to be larger than in the previous case such that now $q_2 > 5$. Consequently, in the same way as before, (S6) can be seen to hold for model \mathcal{M}_2 .

Hence, Theorem 25 is applicable to this situation and the result remains the same as (7.A2.13).

7.A4 A first look at the applicability of our Bayes factor result to some infinite-dimensional models

7.A4.1 Traditional Dirichlet process model: undominated case

Theorem 25 requires the unnormalized posterior to admit factorization as the prior times the likelihood. It is well-known that for the original nonparametric models associated with the Dirichlet process prior (Ferguson (1973)) such factorization is not possible, since

there is no parametric form of the likelihood. In other words, if $[Y_1, \dots, Y_T | F] \stackrel{iid}{\sim} F$, where $F \sim DP(\alpha F_0)$, where $DP(\alpha F_0)$ stands for Dirichlet process with base measure F_0 and precision parameter α , then the likelihood associated with the data Y_1, \dots, Y_T does not have a parametric form, and although the posterior $\pi(F | \mathbf{Y}_T)$ is well-defined, it is not dominated by any σ -finite measure (see, for example, Proposition 7.7 of [Orbanz \(2014\)](#)), and hence does not have a density. This of course prevents factorization of the posterior of F as the prior times likelihood. Moreover, recall that [Shalizi \(2009\)](#) also assumes the existence of a common reference measure for the posteriors $\pi(\cdot | \mathbf{Y}_T)$, for all T , which does not hold here. Indeed, such an assumption is valid in the usual dominated case of Bayes theorem where the aforementioned factorization is possible; in such (usually parametric) cases, the prior is the natural common dominating measure (see [Schervish \(1995\)](#), for example).

7.A4.2 Dirichlet process mixture model: dominated case

Since Dirichlet process supports discrete distributions with probability one, the modeling style described in Section 7.A4.1 is inappropriate if the data \mathbf{Y}_T arises from some continuous distribution. Hence, for such data it is usual in Bayesian nonparametrics based on the Dirichlet process prior to consider the following mixture model (see, for example, [Ghosh and Ramamoorthi \(2003\)](#)):

$$[Y_1, \dots, Y_T | F] \stackrel{iid}{\sim} \int f(\cdot | \xi) dF(\xi), \quad (7.A4.1)$$

where $f(\cdot | \xi)$ is some standard continuous density, usually Gaussian, given $\xi \sim F$, where $F \sim DP(\alpha F_0)$. By Sethuraman's construction ([Sethuraman \(1994\)](#)), $F(\cdot) = \sum_{i=1}^{\infty} p_i \delta_{\xi_i}(\cdot)$, with probability one, where, for $i = 1, 2, \dots$, $\xi_i \stackrel{iid}{\sim} F_0$, and for any ξ , $\delta_{\xi}(\cdot)$ denotes the point mass on ξ . Also, for $i = 1, 2, \dots$, $p_i = V_i \prod_{j < i} (1 - V_j)$, where $V_i \stackrel{iid}{\sim} Beta(\alpha, 1)$. It is easy to verify that $\sum_{i=1}^{\infty} p_i = 1$, almost surely. Application of

Sethuraman's construction in (7.A4.1) yields the equivalent infinite mixture representation

$$[Y_1, \dots, Y_T | \theta] \stackrel{iid}{\sim} \sum_{i=1}^{\infty} p_i f(\cdot | \xi_i), \quad (7.A4.2)$$

where $\theta = (\xi_1, \xi_2, \dots, V_1, V_2, \dots)$ is the infinite-dimensional parameter. The prior on θ is already specified by the *iid* F_0 and $Beta(\alpha, 1)$ distributions, and is the infinite product probability measure associated with these *iid* distributions, so that each factor of the product of the probability measures is dominated by the Lebesgue measure. In this case, the posterior of θ admits the representation

$$\pi(\theta | \mathbf{Y}_T) \propto \pi(\theta) \prod_{t=1}^T \left[\sum_{i=1}^{\infty} p_i f(Y_t | \xi_i) \right], \quad (7.A4.3)$$

and hence the representation of Bayes factor in terms of the prior and the likelihood holds in this case, as required by Theorem 25. Moreover, the posterior $\pi(\cdot | \mathbf{Y}_T)$ is absolutely continuous with respect to $\pi(\cdot)$ for all T , as assumed by Shalizi (2009).

7.A4.3 Polya urn based mixture obtained by integrating out random F : dominated case but \mathcal{T} changes with T

Assume that for $t = 1, \dots, T$, $[Y_t | \phi_t] \sim f(\cdot | \phi_t)$, independently, and $\phi_1, \dots, \phi_T \stackrel{iid}{\sim} F$, where $F \sim DP(\alpha F_0)$. This is equivalent to the Dirichlet process mixture model (7.A4.1), but if F is integrated out, then the joint distribution of ϕ_1, \dots, ϕ_T is given by the Polya urn scheme, that is, $\phi_1 \sim F_0$, and for $t = 2, \dots, T$, $[\phi_t | \phi_1, \dots, \phi_{t-1}] \sim \frac{\alpha F_0}{\alpha+t-1} + \frac{\sum_{j=1}^{t-1} \delta_{\phi_j}}{\alpha+t-1}$ (see, for example, Ferguson (1973), Escobar and West (1995)). The joint prior distribution of ϕ_1, \dots, ϕ_T has a density with respect to a measure composed of Lebesgue measures in lower dimensions; see Lemma 1.99 of Schervish (1995) for the exact forms of the density and the dominating measure. Hence, in this case the posterior of ϕ_1, \dots, ϕ_T is proportional to the prior times the likelihood, where the likelihood is given by $\prod_{t=1}^T f(Y_t | \phi_t)$, and the posterior is dominated by the prior probability measure. Hence,

a countably infinite convex combination of the prior probability measures dominates the posterior of ϕ_1, \dots, ϕ_T for all T , as required for the results of Shalizi (2009) to hold. However, Shalizi (2009) assumes that the σ -field \mathcal{T} associated with the parameter space Θ does not change with T , which does not hold in this case.

7.A4.4 Polya urn based finite mixture: dominated case and \mathcal{T} remains fixed

Bhattacharya (2008) (see also Mukhopadhyay *et al.* (2011), Mukhopadhyay *et al.* (2012)) introduce the following finite mixture model based on Dirichlet process:

$$Y_1, \dots, Y_T \stackrel{iid}{\sim} \frac{1}{M} \sum_{i=1}^M f(\cdot | \phi_i); \quad (7.A4.4)$$

$$\phi_1, \dots, \phi_M \stackrel{iid}{\sim} F; \quad (7.A4.5)$$

$$F \sim DP(\alpha F_0), \quad (7.A4.6)$$

where $f(\cdot | \phi)$ is any standard density as before, given parameter(s) ϕ , and $M (> 1)$ is some fixed integer. Integrating out F yields the following Polya urn scheme for the joint distribution of ϕ_1, \dots, ϕ_M : $\phi_1 \sim F_0$, and for $t = 2, \dots, M$, $[\phi_t | \phi_1, \dots, \phi_{t-1}] \sim \frac{\alpha F_0}{\alpha+t-1} + \frac{\sum_{j=1}^{t-1} \delta_{\phi_j}}{\alpha+t-1}$. Here $\theta = (\phi_1, \dots, \phi_M)$, which is of fixed, finite size, even though the problem is induced by the nonparametric Dirichlet process prior. Also clearly the σ -field \mathcal{T} associated with the parameter space Θ does not change with T . Thus, in this set-up, not only is the posterior written in terms of product of the prior and the likelihood, but is dominated by the Polya urn based prior of θ , for all sample sizes T .

7.A4.5 Nonparametric Bayesian using the Polya tree prior: dominated case

Lavine (1992), Lavine (1994) proposed the Polya tree prior for the random probability measure F as an alternative to the Dirichlet process prior. Briefly, one starts with a

partition $\pi_1 = \{B_0, B_1\}$ of the sample space Ω , so that $\Omega = B_0 \cup B_1$. This procedure is then continued with $B_0 = B_{00} \cup B_{01}$, $B_1 = B_{10} \cup B_{11}$, etc. At level m , the partition is then $\pi_m = \{B_\epsilon : \epsilon = \epsilon_1 \dots \epsilon_m\}$, where ϵ are all binary sequences of length m . Let $\Pi = \{\pi_m : m = 1, 2, \dots\}$, and $\mathcal{A} = \{\alpha_\epsilon\}$ be a sequence of non-negative numbers, one for each partitioning subset. Now, if $Y_{\epsilon 0} = F(B_{\epsilon 0} | B_\epsilon) \sim Beta(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$ independently with respect to the ϵ 's, then F is said to have the Polya tree prior $PT(\Pi, \mathcal{A})$.

It can be shown that if $\alpha_\epsilon \propto m^{-1/2}$, the Polya tree prior reduces to the Dirichlet process prior, confirming that the latter is a special case of the Polya tree prior. However, the most important property of the Polya tree prior is that with appropriate choices of the α_ϵ , F can be made absolutely continuous with respect to the Lebesgue measure. Specifically, if $\alpha_\epsilon \propto m^2$, for the m -th level subset, then F is dominated by the Lebesgue measure almost surely. Hence, if $[Y_1, \dots, Y_T | F] \sim F$ and $F \sim PT(\Pi, \mathcal{A})$, with $\alpha_\epsilon \propto m^2$, then the likelihood is available almost surely. Here we may set $\theta = \{Y_{\epsilon 0} : \epsilon = \epsilon_1 \dots \epsilon_m, m = 1, 2, \dots\}$, which has the infinite product prior measure. The posterior of F given \mathbf{Y}_T , which is also a Polya tree process, is dominated by $\pi(\theta)$ for all $T > 0$. Similar issues hold for the extended Polya tree prior, namely, the optional Polya tree prior proposed by Wong and Ma (2010).

7.A4.6 Bayesian density estimation using the generalized lognormal process prior: dominated case

Lenk (1988) model the unknown density $f(x)$ with respect to measure λ as

$$f(x) = \frac{W(x)}{\int_{\mathcal{X}} W(s)d\lambda(s)}, \quad (7.A4.7)$$

where W is a generalized lognormal process over \mathcal{X} . The generalized lognormal process

has distribution Λ_ζ given by (see [Lenk \(1988\)](#))

$$\Lambda_\zeta(A) = \frac{E \left[(\int_{\mathcal{X}} W d\lambda)^\zeta \mathbb{I}_A \right]}{E \left[(\int_{\mathcal{X}} W d\lambda)^\zeta \right]}, \quad (7.A4.8)$$

where $-\infty < \zeta < \infty$ and the expectations are taken with respect to the usual lognormal process, that is, with respect to $W = \exp(Z)$, where Z is a Gaussian process. In (7.A4.8), \mathbb{I}_A is the indicator of the set A , where A belongs to the Borel σ -field associated with the space of functions from \mathcal{X} to $(0, \infty)$. The properties and moments of the lognormal process are provided in [Lenk \(1988\)](#).

In this formulation, the likelihood with respect to *iid* data Y_1, \dots, Y_T is defined via (7.A4.7). The prior distribution, as well as the posterior distribution of $\Theta = W$ for all $T \geq 1$, are absolutely continuous with respect to the distribution of the lognormal process $W = \exp(Z)$, where Z is a Gaussian process.

7.A4.7 Bayesian regression using Gaussian process: dominated case

Consider the Bayesian nonparametric regression setups embedded in normal, double-exponential, binary and Poisson models, as considered in Chapters 4 and 5. Let the unknown regression function $\eta(\cdot)$ be modeled by Gaussian process with mean function $\mu(\cdot)$ on \mathcal{X} and covariance function $\sigma^2 c(\cdot, \cdot)$ on $\mathcal{X} \times \mathcal{X}$, where \mathcal{X} is the space of covariates. From Mercer's theorem (see, for example, [Rasmussen and Williams \(2006\)](#)) it follows that the Gaussian process $\eta(\cdot)$ admits the representation below almost surely:

$$\eta(\cdot) = \mu(\cdot) + \sum_{i=1}^{\infty} \sqrt{\lambda_i} \psi_i(\cdot) e_i, \quad (7.A4.9)$$

where ψ_i and λ_i are the normalized eigenfunctions and eigenvalues of the positive definite function $c(\cdot, \cdot)$; and, for $i = 1, 2, \dots$, $e_i \stackrel{iid}{\sim} N(0, 1)$. The above representation for Gaussian processes is popularly known as the Karhunen-Loëve expansion (see, for example, [Ash](#)

and Gardner (1975)).

Hence, both the likelihood and the prior can be parameterized in terms of $\psi_i(\cdot)$; $i = 1, 2, \dots$ and $\boldsymbol{\epsilon} = \{e_i; i = 1, 2, \dots\}$, the latter being unknown and having the infinite product prior distribution such that $e_i \stackrel{iid}{\sim} N(0, 1)$; $i = 1, 2, \dots$. Letting $\theta = (\boldsymbol{\epsilon}, \vartheta)$, where ϑ stands for other finite-dimensional model parameters including σ^2 with probability measure φ , say, observe that the posterior distribution of θ , is clearly dominated by this infinite product prior measure times φ , for almost all datasets.

8

Convergence of Pseudo-Bayes Factors in Forward and Inverse Regression Problems

8.1 Introduction

The Bayesian statistical literature on model selection is rich in its collection of innovative methodologies, among which the most principled method of comparing different competing models seems to be offered by Bayes factors (BFs), through the ratio of the posterior and prior odds associated with the models under comparison, which reduces to the ratio of the marginal densities of the data under the two models. In Chapter 7 we established the almost sure convergence theory of BF in the general setup that includes even dependent data and misspecified models. The result depends explicitly on

the average KL-divergence between the competing and the true models. Thus, BFs have sound theoretical properties as well, which make them very useful for model comparison in general.

However, BFs are known to have several limitations. First, if the prior for the model parameter θ_j is improper, then the marginal density $m(\cdot|\mathcal{M}_j)$ is also improper and hence $m(\mathbf{Y}_n|\mathcal{M}_j)$ does not admit any sensible interpretation. Second, BFs suffer from the Jeffreys-Lindley-Bartlett paradox (see [Jeffreys \(1939\)](#), [Lindley \(1957\)](#), [Bartlett \(1957\)](#), [Robert \(1993\)](#), [Villa and Walker \(2015\)](#) for details and general discussions on the paradox). Furthermore, a drawback of BFs in practical applications is that the marginal density of the data \mathbf{Y}_n is usually quite challenging to compute accurately, even with sophisticated simulation techniques based on importance sampling, bridge sampling and path sampling (see, for example, [Meng and Wong \(1996\)](#), [Gelman and Meng \(1998\)](#); see also [Gronau *et al.* \(2017\)](#) for a relatively recent tutorial and many relevant references), particularly when the posterior is far from normal and when the dimension of the parameter space is large. Moreover, the marginal density is usually extremely close to zero if n is even moderately large. This causes numerical instability in computation of the BF.

The problems of BFs regarding improper prior, Jeffreys-Lindley-Bartlett paradox, and general computational difficulties associated with the marginal density can be simultaneously alleviated if the marginal density $m(\mathbf{Y}_n|\mathcal{M}_j)$ for model \mathcal{M}_j is replaced with the product of leave-one-out cross-validation posteriors $\prod_{i=1}^n \pi(y_i|\mathbf{Y}_{n,-i}, \mathcal{M}_j)$, where $\mathbf{Y}_{n,-i} = \mathbf{Y}_n \setminus \{y_i\} = \{y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n\}$, and

$$\pi(y_i|\mathbf{Y}_{n,-i}, \mathcal{M}_j) = \int_{\Theta_j} f(y_i|\theta_j, y_1, \dots, y_{i-1}, \mathcal{M}_j) d\pi(\theta_j|\mathbf{Y}_{n,-i}, \mathcal{M}_j) \quad (8.1.1)$$

is the i -th leave-one-out cross-validation posterior density evaluated at y_i . In the above equation (8.1.1), $f(y_i|\theta_j, y_1, \dots, y_{i-1}, \mathcal{M}_j)$ is the density of y_i given model parameters θ_j and y_1, \dots, y_{i-1} ; $\pi(\theta_j|\mathbf{Y}_{n,-i}, \mathcal{M}_j)$ is the posterior distribution of θ_j given $\mathbf{Y}_{n,-i}$.

Viewing $\prod_{i=1}^n \pi(y_i|\mathbf{Y}_{n,-i}, \mathcal{M}_j)$ as the surrogate for $m(\mathbf{Y}_n|\mathcal{M}_j)$, it seems reasonable to replace $BF^{(n)}(\mathcal{M}_1, \mathcal{M}_2)$ with the corresponding pseudo-Bayes factor (PBF) given by

$$PBF^{(n)}(\mathcal{M}_1, \mathcal{M}_2) = \frac{\prod_{i=1}^n \pi(y_i|\mathbf{Y}_{n,-i}, \mathcal{M}_1)}{\prod_{i=1}^n \pi(y_i|\mathbf{Y}_{n,-i}, \mathcal{M}_2)}. \quad (8.1.2)$$

In the case of independent observations, the above formula and the terminology “pseudo-Bayes factor” seem to be first proposed by [Geisser and Eddy \(1979\)](#). Their motivation for PBF did not seem to arise as providing solutions to the problems of BFs, however, but rather the urge to exploit the concept of cross-validation in Bayesian model selection, which had been proved to be indispensable for constructing model selection criteria in the classical statistical paradigm. Below we argue how this cross-validation idea helps solve the aforementioned problems of BFs.

First note that the posterior $\pi(\theta_j|\mathbf{Y}_{n,-i}, \mathcal{M}_j)$ is usually proper even for improper prior for θ_j if n is sufficiently large. Thus, $\pi(y_i|\mathbf{Y}_{n,-i}, \mathcal{M}_j)$ given by (8.1.1) is usually well-defined even for improper priors, unlike $m(\mathbf{Y}_n|\mathcal{M}_j)$. So, even though BF is ill-defined for improper priors, PBF is usually still well-defined.

Second, a clear theoretical advantage of PBF over BF is that PBF is immune to the problem of Jeffreys-Lindley-Bartlett paradox (see [Gelfand and Dey \(1994\)](#) for example), while BF is certainly not.

Finally, PBF enjoys significant computational advantages over BF. Note that straightforward Monte Carlo averages of $f(y_i|\theta_j, y_1, \dots, y_{i-1}, \mathcal{M}_j)$ over realizations of θ obtained from $\pi(\theta|\mathbf{Y}_{n,-i}, \mathcal{M}_j)$ by simulation techniques is sufficient to ensure good estimates of the cross-validation posterior density $\pi(y_i|\mathbf{Y}_{n,-i}, \mathcal{M}_j)$. Since $\pi(y_i|\mathbf{Y}_{n,-i}, \mathcal{M}_j)$ is the density of y_i individually, the estimate is also numerically stable compared to estimates of $m(\mathbf{Y}_n|\mathcal{M}_j)$. Hence, the sum of logarithms of the estimates of $\pi(y_i|\mathbf{Y}_{n,-i}, \mathcal{M}_j)$, for $i = 1, \dots, n$, results in quite accurate and stable estimates of $\log[\prod_{i=1}^n \pi(y_i|\mathbf{Y}_{n,-i}, \mathcal{M}_j)]$. In other words, PBF is far simpler to compute accurately than BF and is numerically far more stable and reliable.

In spite of the advantages of PBF over BF, it seems to be largely ignored in the statistical literature, both theoretically and application-wise. Some asymptotic theory of PBF has been attempted by [Gelfand and Dey \(1994\)](#) using independent observations, Laplace approximations and some essentially ad-hoc simplifying approximations and arguments. Application of PBF has been considered in [Bhattacharya \(2008\)](#) for demonstrating the superiority of his new Bayesian nonparametric Dirichlet process model over the traditional Dirichlet process mixture model. But apart from these works we are not aware of any other significant research involving PBF.

In this chapter, we establish the asymptotic theory for PBF in the general setup consisting of dependent observations, model misspecifications as well as covariates; inclusion of covariates also validates our asymptotic theory in the variable selection framework. Judiciously exploiting the posterior convergence treatise of [Shalizi \(2009\)](#) we prove almost sure exponential convergence of PBF in favour of the true model, the convergence explicitly depending upon the KL-divergence rate from the true model. For any two models different from the true model, we prove almost sure exponential convergence of PBF in favour of the better model, where the convergence depends explicitly upon the difference between KL-divergence rates from the true model. Thus, our PBF convergence results agree with the BF convergence results established in Chapter 7.

An important aspect of our PBF research involves establishing its convergence properties for inverse regression problems, and even if one of the two competing models involve inverse regression and the other forward regression. Recall that, crucially, Bayesian inverse regression problems require priors on the covariate values to be predicted. In this chapter, we consider two setups of inverse regression and establish almost sure exponential convergence of PBF in general inverse regression for both the setups. These include situations where one of the competing models involve forward regression and the other is associated with inverse regression.

We illustrate our asymptotic results with various theoretical examples in both forward and inverse regression contexts, including forward and inverse variable selection problems. We also follow up our theoretical investigations with simulation experiments in small samples involving Poisson and geometric forward and inverse regression models with relevant link functions and both linear regression and nonparametric regression, the latter modeled by Gaussian processes. We also illustrate variable selection in the aforementioned setups with two different covariates. The results that we obtain are quite encouraging and illuminating, providing useful insights into the behaviour of PBF for forward and inverse parametric and nonparametric regression.

The roadmap for the rest of this chapter is as follows. We begin our progress by discussing and formalizing the relevant aspects of forward and inverse regression problems and the associated pseudo-Bayes factors in Section 8.2. Convergence of PBF in the forward regression context is established in Section 8.3, while in Sections 8.4 and 8.5 we establish convergence of PBF in the two setups related to inverse regression. In Sections 8.6 and 8.7 we provide theoretical illustrations of PBF convergence in forward and inverse setups, respectively, with various examples including variable selection. Details of our simulation experiments with small samples involving Poisson and geometric linear and Gaussian process regression for relevant link functions, under both forward and inverse setups, are reported in Section 8.8, which also includes experiments on variable selection. Finally, we summarize our contributions and provide future directions in Section 8.9.

8.2 Preliminaries and general setup for forward and inverse regression problems

Let us first consider the forward regression setup.

8.2.1 Forward regression problem

For $i = 1, \dots, n$, let observed response y_i be related to observed covariate x_i through

$$y_1 \sim f(\cdot|\theta, x_1) \text{ and } y_i \sim f(\cdot|\theta, x_i, \mathbf{Y}^{(i-1)}) \text{ for } i = 2, \dots, n, \quad (8.2.1)$$

where for $i = 2, \dots, n$, $\mathbf{Y}^{(i)} = \{y_1, \dots, y_i\}$ and $f(\cdot|\theta, x_1)$, $f(\cdot|\theta, x_i, \mathbf{Y}^{(i-1)})$ are known densities depending upon (a set of) parameters $\theta \in \Theta$, where Θ is the parameter space, which may be infinite-dimensional. For the sake of generality, we shall consider $\theta = (\eta, \xi)$, where η is a function of the covariates, which we more explicitly denote as $\eta(x)$. The covariate $x \in \mathcal{X}$, \mathcal{X} being the space of covariates. The part ξ of θ will be assumed to consist of other parameters, such as the unknown error variance. For Bayesian forward regression problems, some prior needs to be assigned on the parameter space Θ . For notational convenience, we shall denote $f(\cdot|\theta, x_1)$ by $f(\cdot|\theta, x_1, \mathbf{Y}^{(0)})$, so that we can represent (8.2.1) more conveniently as

$$y_i \sim f(\cdot|\theta, x_i, \mathbf{Y}^{(i-1)}) \text{ for } i = 1, \dots, n. \quad (8.2.2)$$

8.2.2 Forward pseudo-Bayes factor

Letting $\mathbf{Y}_n = \{y_i : i = 1, \dots, n\}$, $\mathbf{X}_n = \{x_i : i = 1, \dots, n\}$, $\mathbf{Y}_{n,-i} = \mathbf{Y}_n \setminus \{y_i\}$ and $\mathbf{X}_{n,-i} = \mathbf{X}_n \setminus \{x_i\}$, let $\pi(y_i | \mathbf{Y}_{n,-i}, \mathbf{X}_n, \mathcal{M})$ denote the posterior density at y_i , given data $\mathbf{Y}_{n,-i}$, \mathbf{X}_n and model \mathcal{M} . Let the density of y_i given θ and x_i under model \mathcal{M} be denoted by $f(y_i | \theta, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M})$. Then note that

$$\pi(y_i | \mathbf{Y}_{n,-i}, \mathbf{X}_n, \mathcal{M}) = \int_{\Theta} f(y_i | \theta, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}) d\pi(\theta | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}), \quad (8.2.3)$$

where

$$\pi(\theta | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}) \propto \pi(\theta) \prod_{j \neq i; j=1}^n f(y_j | \theta, x_j, \mathbf{Y}^{(j-1)}, \mathcal{M}). \quad (8.2.4)$$

For any two models \mathcal{M}_1 and \mathcal{M}_2 , the forward pseudo Bayes factor (FPBF) of \mathcal{M}_1 against \mathcal{M}_2 based on the cross-validation posteriors of the form (8.2.3) is defined as follows:

$$FPBF^{(n)}(\mathcal{M}_1, \mathcal{M}_2) = \frac{\prod_{i=1}^n \pi(y_i | \mathbf{Y}_{n,-i}, \mathbf{X}_n, \mathcal{M}_1)}{\prod_{i=1}^n \pi(y_i | \mathbf{Y}_{n,-i}, \mathbf{X}_n, \mathcal{M}_2)}, \quad (8.2.5)$$

and we are interested in studying the limit $\lim_{n \rightarrow \infty} \frac{1}{n} \log FPBF^{(n)}(\mathcal{M}_1, \mathcal{M}_2)$ for almost all data sequences.

8.2.3 Inverse regression problem: first setup

In inverse regression, the basic premise remains the same as in forward regression detailed in Section 8.2.1. In other words, the distribution $f(\cdot | \theta, x_i, \mathbf{Y}^{(i-1)})$, parameter θ , the parameter and the covariate space remain the same as in the forward regression setup. However, unlike in Bayesian forward regression problems where a prior needs to be assigned only to the unknown parameter θ , a prior is also required for \tilde{x} , the unknown covariate observation associated with known response \tilde{y} , say. Given the entire dataset and \tilde{y} , the problem in inverse regression is to predict \tilde{x} . Hence, in the Bayesian inverse setup, a prior on \tilde{x} is necessary. Given model \mathcal{M} and the corresponding parameters θ , we denote such prior by $\pi(\tilde{x} | \theta, \mathcal{M})$. For Bayesian cross-validation in inverse problems it is pertinent to successively leave out $(y_i, x_i); i = 1, \dots, n$, and compute the posterior predictive distribution $\pi(\tilde{x}_i | \mathbf{Y}_n, \mathbf{X}_{n,-i})$, from y_i and the rest of the data $(\mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i})$ (see [Bhattacharya and Haslett \(2007\)](#)). But these posteriors are not useful for Bayes or pseudo-Bayes factors even for inverse regression setups. The reason is that the Bayes factor for inverse regression is still the ratio of posterior odds and prior odds associated with the competing models, which as usual translates to the ratio of the marginal densities of the data under the two competing models. The marginal densities depend upon the prior for (θ, \tilde{x}) , however, under the competing models. The pseudo-Bayes factor for inverse models is then the ratio of products of the cross-validation posteriors of y_i , where θ and \tilde{x}_i are marginalized out. Details of such inverse cross-validation posteriors

and the definition of pseudo-Bayes factors for inverse regression are given below.

Inverse pseudo-Bayes factor in this setup

In the inverse regression setup, first note that

$$\begin{aligned}
 & \pi(\tilde{x}_i, \theta | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}) \\
 &= \frac{\pi(\tilde{x}_i, \theta | \mathcal{M}) \prod_{j \neq i; j=1}^n f(y_j | \theta, x_j, \mathbf{Y}^{(j-1)}, \mathcal{M})}{\int_{\mathcal{X}} \int_{\Theta} d\pi(u, \psi) \prod_{j \neq i; j=1}^n f(y_j | \psi, x_j, \mathbf{Y}^{(j-1)}, \mathcal{M})} \\
 &= \frac{\pi(\tilde{x}_i | \theta, \mathcal{M}) \pi(\theta | \mathcal{M}) \prod_{j \neq i; j=1}^n f(y_j | \theta, x_j, \mathbf{Y}^{(j-1)}, \mathcal{M})}{\int_{\mathcal{X}} \int_{\Theta} d\pi(u | \psi, \mathcal{M}) d\pi(\psi | \mathcal{M}) \prod_{j \neq i; j=1}^n f(y_j | \psi, x_j, \mathbf{Y}^{(j-1)}, \mathcal{M})} \\
 &= \frac{\pi(\tilde{x}_i | \theta, \mathcal{M}) \pi(\theta | \mathcal{M}) \prod_{j \neq i; j=1}^n f(y_j | \theta, x_j, \mathbf{Y}^{(j-1)}, \mathcal{M})}{\int_{\Theta} d\pi(\psi | \mathcal{M}) \prod_{j \neq i; j=1}^n f(y_j | \psi, x_j, \mathbf{Y}^{(j-1)}, \mathcal{M})} = \pi(\tilde{x}_i | \theta, \mathcal{M}) \pi(\theta | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}). \tag{8.2.6}
 \end{aligned}$$

Using (8.2.6) we obtain

$$\begin{aligned}
 \pi(y_i | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}) &= \int_{\mathcal{X}} \int_{\Theta} f(y_i | \theta, \tilde{x}_i, \mathbf{Y}^{(i-1)}, \mathcal{M}) d\pi(\tilde{x}_i, \theta | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}), \\
 &= \int_{\Theta} g(\mathbf{Y}^{(i)}, \theta, \mathcal{M}) d\pi(\theta | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}),
 \end{aligned} \tag{8.2.7}$$

where

$$g(\mathbf{Y}^{(i)}, \theta, \mathcal{M}) = \int_{\mathcal{X}} f(y_i | \theta, \tilde{x}_i, \mathbf{Y}^{(i-1)}, \mathcal{M}) d\pi(\tilde{x}_i | \theta, \mathcal{M}), \tag{8.2.8}$$

and $\pi(\theta | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M})$ is the same as (8.2.4). For any two models \mathcal{M}_1 and \mathcal{M}_2 , the inverse pseudo Bayes factor (IPBF) of \mathcal{M}_1 against \mathcal{M}_2 based on cross-validation posteriors of the form (8.2.7) is given by

$$IPBF^{(n)}(\mathcal{M}_1, \mathcal{M}_2) = \frac{\prod_{i=1}^n \pi(y_i | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}_1)}{\prod_{i=1}^n \pi(y_i | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}_2)}, \tag{8.2.9}$$

and our goal is to investigate $\lim_{n \rightarrow \infty} \frac{1}{n} \log IPBF^{(n)}(\mathcal{M}_1, \mathcal{M}_2)$ for almost all data sequences.

8.2.4 Inverse regression problem: second setup

In the inverse regression context, we consider another setup under which we established consistency of the inverse cross-validation posteriors of \tilde{x}_i in Chapter 6. Here we consider experiments with covariate observations x_1, x_2, \dots, x_n along with responses $\mathbf{Y}_{nm} = \{y_{ij} : i = 1, \dots, n, j = 1, \dots, m\}$. In other words, the experiment considered here will allow us to have m samples of responses $\mathbf{y}_i = \{y_{i1}, y_{i2}, \dots, y_{im}\}$ against each covariate observation x_i , for $i = 1, 2, \dots, n$. Again, both x_i and y_{ij} are allowed to be multidimensional. Let $\mathbf{Y}_{nm,-i} = \mathbf{Y}_{nm} \setminus \{\mathbf{y}_i\}$.

For $i = 1, \dots, n$ consider the following general model setup: conditionally on θ , x_i and $\mathbf{Y}_j^{(i-1)} = \{y_{1j}, \dots, y_{i-1,j}\}$,

$$y_{ij} \sim f(\cdot | \theta, x_i, \mathbf{Y}_j^{(i-1)}) ; \quad j = 1, \dots, m, \quad (8.2.10)$$

independently, where $f(\cdot | \theta, x_1, \mathbf{Y}^{(0)}) = f(\cdot | \theta, x_1)$ as before.

We consider the prior for x_i to be of the same form as in Chapter 6.4, whose illustrations and properties are provided in Chapters 6.4.1 and 6.4.2, respectively.

Inverse pseudo-Bayes factor in this setup

For any two models \mathcal{M}_1 and \mathcal{M}_2 we define inverse pseudo-Bayes factor for model \mathcal{M}_1 against model \mathcal{M}_2 , for any $k \geq 1$, as

$$IPBF^{(n,m,k)}(\mathcal{M}_1, \mathcal{M}_2) = \frac{\prod_{i=1}^n \pi(y_{ik} | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M}_1)}{\prod_{i=1}^n \pi(y_{ik} | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M}_2)} \quad (8.2.11)$$

and study the limit $\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{1}{n} \log IPBF^{(n,m,k)}(\mathcal{M}_1, \mathcal{M}_2)$ for almost all data sequences. Note that since $\{y_{ik}; k \geq 1\}$ are distributed independently as $f(\cdot | \theta, x_i, \mathbf{Y}_k^{(i-1)})$ given any θ and x_i , it would follow that if the limit exists, it must be the same for all $k \geq 1$.

Suppose that the true data-generating parameter θ_0 is not contained in Θ , the parameter space considered. This is a case of misspecification that we must incorporate

in our convergence theory of PBF. Our PBF asymptotics draws on posterior convergence theory for (possibly infinite-dimensional) parameters that also allows misspecification. In this regard, the approach presented in Shalizi (2009) seems to be very appropriate. Before proceeding further, we first provide a brief overview of this approach, which we conveniently exploit for our purpose.

In what follows, we denote almost sure convergence by “ $\xrightarrow{a.s.}$ ”, almost sure equality by “ $\xlongequal{a.s.}$ ” and weak convergence by “ \xrightarrow{w} ”.

8.3 Convergence of PBF in forward problems

Let \mathcal{M}_0 denote the true model which is also associated with parameter $\theta \in \Theta_0$, where Θ_0 is a parameter space containing the true parameter θ_0 . Then the following result holds.

Theorem 29 *Assume conditions (S1)–(S7) of Shalizi, and let the infimum of $h(\theta)$ over Θ be attained at $\tilde{\theta} \in \Theta$, where $\tilde{\theta} \neq \theta_0$. Also assume that Θ and Θ_0 are complete separable metric spaces and that for $i \geq 1$, $f(y_i|\theta, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M})$ and $f(y_i|\theta, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_0)$ are bounded and continuous in θ . Then,*

$$\frac{1}{n} \log FPBF^{(n)}(\mathcal{M}, \mathcal{M}_0) = \frac{1}{n} \log \left[\frac{\prod_{i=1}^n \pi(y_i|\mathbf{Y}_{n,-i}, \mathbf{X}_n, \mathcal{M})}{\prod_{i=1}^n \pi(y_i|\mathbf{Y}_{n,-i}, \mathbf{X}_n, \mathcal{M}_0)} \right] \xrightarrow{a.s.} -h(\tilde{\theta}), \text{ as } n \rightarrow \infty, \quad (8.3.1)$$

where, for any θ ,

$$h(\theta) = \lim_{n \rightarrow \infty} \frac{1}{n} E_{\theta_0} \left\{ \sum_{i=1}^n \log \left[\frac{f(y_i|\theta_0, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_0)}{f(y_i|\theta, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M})} \right] \right\}. \quad (8.3.2)$$

Proof. By the hypotheses, (4.1.2) holds, from which it follows that for any $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} \pi(\mathbb{N}_\epsilon^c | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}) = 0, \quad (8.3.3)$$

where $\mathbb{N}_\epsilon = \{\theta : h(\theta) \leq h(\Theta) + \epsilon\}$.

Now, by hypothesis, the infimum of $h(\theta)$ over Θ is attained at $\tilde{\theta} \in \Theta$, where $\tilde{\theta} \neq \theta_0$. Then by (8.3.3), the posterior of θ given $\mathbf{Y}_{n,-i}$ and $\mathbf{X}_{n,-i}$, given by (8.2.4), concentrates around $\tilde{\theta}$, the minimizer of the limiting KL-divergence rate from the true distribution. Formally, given any neighborhood U of $\tilde{\theta}$, the set \mathbb{N}_ϵ is contained in U for sufficiently small ϵ . It follows that for any neighborhood U of $\tilde{\theta}$, $\pi(U|\mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}) \rightarrow 1$, almost surely, as $n \rightarrow \infty$. Since Θ is a complete, separable metric space, it follows that (see, for example, Ghosh and Ramamoorthi (2003), Ghosal and van der Vaart (2017))

$$\pi(\cdot|\mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}) \xrightarrow{w} \delta_{\tilde{\theta}}(\cdot), \text{ almost surely, as } n \rightarrow \infty. \quad (8.3.4)$$

Then, due to (8.3.4) and the Portmanteau theorem, as $f(y_i|\theta, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M})$ is bounded and continuous in θ , it holds using (8.2.3), that

$$\pi(y_i|\mathbf{Y}_{n,-i}, \mathbf{X}_n, \mathcal{M}) \xrightarrow{a.s.} f(y_i|\tilde{\theta}, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}), \text{ as } n \rightarrow \infty. \quad (8.3.5)$$

Now, due to (8.3.5),

$$\frac{1}{n} \sum_{i=1}^n \log \pi(y_i|\mathbf{Y}_{n,-i}, \mathbf{X}_n, \mathcal{M}) \xrightarrow{a.s.} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log f(y_i|\tilde{\theta}, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}), \text{ as } n \rightarrow \infty. \quad (8.3.6)$$

Also, essentially the same arguments leading to (8.3.5) yield

$$\pi(y_i|\mathbf{Y}_{n,-i}, \mathbf{X}_n, \mathcal{M}_0) \xrightarrow{a.s.} f(y_i|\theta_0, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_0), \text{ as } n \rightarrow \infty,$$

which ensures

$$\frac{1}{n} \sum_{i=1}^n \log \pi(y_i|\mathbf{Y}_{n,-i}, \mathbf{X}_n, \mathcal{M}_0) \xrightarrow{a.s.} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log f(y_i|\theta_0, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_0), \text{ as } n \rightarrow \infty. \quad (8.3.7)$$

From (8.3.6) and (8.3.7) we obtain

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log FPBF^{(n)}(\mathcal{M}, \mathcal{M}_0) \stackrel{a.s.}{=} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log \left[\frac{f(y_i | \tilde{\theta}, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M})}{f(y_i | \tilde{\theta}_0, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_0)} \right] \stackrel{a.s.}{=} -h(\tilde{\theta}), \quad (8.3.8)$$

where the rightmost step of (8.3.8), given by (8.3.2), follows due to (4.1.1). Hence, the result is proved. ■

For postulated model \mathcal{M}_j , let the KL-divergence rate h in (4.A1.2) be denoted by h_j , for $j \geq 1$.

Theorem 30 *For models \mathcal{M}_0 , \mathcal{M}_1 and \mathcal{M}_2 with complete separable parameter spaces Θ_0 , Θ_1 and Θ_2 , assume conditions (S1)–(S7) of Shalizi, and for $j = 1, 2$, let the infimum of $h_j(\theta)$ over Θ_j be attained at $\tilde{\theta}_j \in \Theta_j$, where $\tilde{\theta}_j \neq \theta_0$. Also assume that for $i \geq 1$, $f(y_i | \theta, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_j)$; $j = 1, 2$, and $f(y_i | \theta, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_0)$ are bounded and continuous in θ . Then,*

$$\frac{1}{n} \log FPBF^{(n)}(\mathcal{M}_1, \mathcal{M}_2) = \frac{1}{n} \log \left[\frac{\prod_{i=1}^n \pi(y_i | \mathbf{Y}_{n,-i}, \mathbf{X}_n, \mathcal{M}_1)}{\prod_{i=1}^n \pi(y_i | \mathbf{Y}_{n,-i}, \mathbf{X}_n, \mathcal{M}_2)} \right] \xrightarrow{a.s.} -[h(\tilde{\theta}_1) - h(\tilde{\theta}_2)], \text{ as } n \rightarrow \infty, \quad (8.3.9)$$

where, for $j = 1, 2$, and for any θ ,

$$h_j(\theta) = \lim_{n \rightarrow \infty} \frac{1}{n} E_{\theta_0} \left\{ \sum_{i=1}^n \log \left[\frac{f(y_i | \theta_0, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_0)}{f(y_i | \theta, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_j)} \right] \right\}. \quad (8.3.10)$$

Proof. The proof follows by noting that

$$\frac{1}{n} \log FPBF^{(n)}(\mathcal{M}_1, \mathcal{M}_2) = \frac{1}{n} \log FPBF^{(n)}(\mathcal{M}_1, \mathcal{M}_0) - \frac{1}{n} \log FPBF^{(n)}(\mathcal{M}_2, \mathcal{M}_0),$$

and then using (8.3.1) for $\frac{1}{n} \log FPBF^{(n)}(\mathcal{M}_1, \mathcal{M}_0)$ and $\frac{1}{n} \log FPBF^{(n)}(\mathcal{M}_2, \mathcal{M}_0)$. ■

8.4 Convergence results for PBF in inverse regression: first setup

Theorem 31 Assume conditions (S1)–(S7) of Shalizi, and let the infimum of $h(\theta)$ over Θ be attained at $\tilde{\theta} \in \Theta$, where $\tilde{\theta} \neq \theta_0$. Also assume that Θ and Θ_0 are complete separable metric spaces and that for $i \geq 1$, $g(\mathbf{Y}^{(i)}, \theta, \mathcal{M})$ and $g(\mathbf{Y}^{(i)}, \theta, \mathcal{M}_0)$ are bounded and continuous in θ . Then,

$$\frac{1}{n} \log IPBF^{(n)}(\mathcal{M}, \mathcal{M}_0) = \frac{1}{n} \log \left[\frac{\prod_{i=1}^n \pi(y_i | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M})}{\prod_{i=1}^n \pi(y_i | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}_0)} \right] \xrightarrow{a.s.} -h^*(\tilde{\theta}), \text{ as } n \rightarrow \infty, \quad (8.4.1)$$

where, for any θ ,

$$h^*(\theta) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log \left[\frac{g(\mathbf{Y}^{(i)}, \theta_0, \mathcal{M}_0)}{g(\mathbf{Y}^{(i)}, \theta, \mathcal{M})} \right],$$

provided that the limit exists.

Proof. Since $\pi(\cdot | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M})$ remains the same as in Theorem 29, it follows as before that

$$\pi(\cdot | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}) \xrightarrow{w} \delta_{\tilde{\theta}}(\cdot), \text{ almost surely, as } n \rightarrow \infty.$$

Then, since $g(y_i, \theta, \mathcal{M})$ is bounded and continuous in θ , the above ensures in conjunction with the Portmanteau theorem using (8.2.7), that

$$\pi(y_i | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}) \xrightarrow{a.s.} g(\mathbf{Y}^{(i)}, \tilde{\theta}, \mathcal{M}), \text{ as } n \rightarrow \infty. \quad (8.4.2)$$

Hence,

$$\frac{1}{n} \sum_{i=1}^n \log \pi(y_i | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}) \xrightarrow{a.s.} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log g(\mathbf{Y}^{(i)}, \tilde{\theta}, \mathcal{M}), \text{ as } n \rightarrow \infty. \quad (8.4.3)$$

Similarly,

$$\frac{1}{n} \sum_{i=1}^n \log \pi(y_i | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}_0) \xrightarrow{a.s.} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log g(\mathbf{Y}^{(i)}, \theta_0, \mathcal{M}_0), \text{ as } n \rightarrow \infty. \quad (8.4.4)$$

Combining (8.4.3) and (8.4.4) yields

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log IPBF^{(n)}(\mathcal{M}, \mathcal{M}_0) \stackrel{a.s.}{=} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log \left[\frac{g(\mathbf{Y}^{(i)}, \tilde{\theta}, \mathcal{M})}{g(\mathbf{Y}^{(i)}, \theta_0, \mathcal{M}_0)} \right] = -h^*(\tilde{\theta}).$$

Hence, the result is proved. ■

Remark 32 Observe that $h^*(\tilde{\theta})$ in Theorem 31 does not correspond to the KL-divergence rate given by (4.A1.2), even though in the forward context, Theorem 29 shows almost convergence of $\frac{1}{n} \log FPBF^{(n)}$ to $-h(\tilde{\theta})$, where $h(\tilde{\theta})$ is the bona fide KL-divergence rate.

In Theorem 31 we have assumed that for cross-validation even in the true model \mathcal{M}_0 , x_i is assumed unknown, and that a prior has been placed on the corresponding unknown random quantity \tilde{x}_i . If, on the other hand, x_i is considered known for cross-validation in \mathcal{M}_0 , then we have the following theorem, which is an appropriately modified version of Theorem 31.

Theorem 33 Assume conditions (S1)–(S7) of Shalizi for models \mathcal{M}_0 and \mathcal{M} , and let the infimum of $h(\theta)$ over Θ be attained at $\tilde{\theta} \in \Theta$, where $\tilde{\theta} \neq \theta_0$. Also assume that Θ and Θ_0 are complete separable metric spaces and that for $i \geq 1$, $g(\mathbf{Y}^{(i)}, \theta, \mathcal{M})$ and $f(y_i | \theta, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_0)$ are bounded and continuous in θ . Then the following result holds if x_i is assumed known for cross-validation with respect to \mathcal{M}_0 :

$$\frac{1}{n} \log IPBF^{(n)}(\mathcal{M}, \mathcal{M}_0) = \frac{1}{n} \log \left[\frac{\prod_{i=1}^n \pi(y_i | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M})}{\prod_{i=1}^n \pi(y_i | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}_0)} \right] \xrightarrow{a.s.} -h^*(\tilde{\theta}), \text{ as } n \rightarrow \infty, \quad (8.4.5)$$

where, for any θ ,

$$h^*(\theta) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log \left[\frac{f(y_i | \theta_0, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_0)}{g(\mathbf{Y}^{(i)}, \theta, \mathcal{M})} \right],$$

provided that the limit exists.

Proof. In this case, for the true model \mathcal{M}_0 , the cross-validation posterior $\pi(y_i | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}_0)$ is of the same form as (8.2.3) and hence, (8.3.7) holds. The rest of the proof remains the same as that of Theorem 31. ■

Remark 34 Observe that $h^*(\tilde{\theta})$ in Theorem 33 is a genuine KL-divergence rate. However, this is not the same as $h(\tilde{\theta})$ of Theorem 29, which is the KL-divergence rate between \mathcal{M} and \mathcal{M}_0 when all the x_i are known. Since cross-validation with all x_i known can occur only in the forward regression setup, convergence rates of pseudo-Bayes factors in inverse regression problems can never be associated with h , even though the conditions of Theorem 33 show that $\tilde{\theta}$ is the minimizer of h .

Theorem 35 For models \mathcal{M}_0 , \mathcal{M}_1 and \mathcal{M}_2 with complete separable parameter spaces Θ_0 , Θ_1 and Θ_2 , assume conditions (S1)–(S7) of Shalizi, and for $j = 1, 2$, let the infimum of $h_j(\theta)$ over Θ_j be attained at $\tilde{\theta}_j \in \Theta_j$, where $\tilde{\theta}_j \neq \theta_0$. Also assume that for $i \geq 1$, $g(\mathbf{Y}^{(i)}, \theta, \mathcal{M}_j)$; $j = 1, 2$, and $f(y_i | \theta, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_0)$ are bounded and continuous in θ . Then, if x_i is assumed known for cross-validation with respect to \mathcal{M}_0 , the following holds:

$$\frac{1}{n} \log IPBF^{(n)}(\mathcal{M}_1, \mathcal{M}_2) = \frac{1}{n} \log \left[\frac{\prod_{i=1}^n \pi(y_i | \mathbf{Y}_{n,-i}, \mathbf{X}_n, \mathcal{M}_1)}{\prod_{i=1}^n \pi(y_i | \mathbf{Y}_{n,-i}, \mathbf{X}_n, \mathcal{M}_2)} \right] \xrightarrow{a.s.} -[h_1^*(\tilde{\theta}_1) - h_2^*(\tilde{\theta}_2)], \text{ as } n \rightarrow \infty, \quad (8.4.6)$$

where, for $j = 1, 2$, and for any θ ,

$$h_j^*(\theta) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log \left[\frac{f(y_i | \theta_0, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_0)}{g(\mathbf{Y}^{(i)}, \theta, \mathcal{M}_j)} \right], \quad (8.4.7)$$

provided the limit exists.

Proof. The proof follows by noting that

$$\frac{1}{n} \log IPBF^{(n)}(\mathcal{M}_1, \mathcal{M}_2) = \frac{1}{n} \log IPBF^{(n)}(\mathcal{M}_1, \mathcal{M}_0) - \frac{1}{n} \log IPBF^{(n)}(\mathcal{M}_2, \mathcal{M}_0),$$

and then using (8.4.5) for $\frac{1}{n} \log IPBF^{(n)}(\mathcal{M}_1, \mathcal{M}_0)$ and $\frac{1}{n} \log IPBF^{(n)}(\mathcal{M}_2, \mathcal{M}_0)$. ■

Remark 36 Note that the result of Theorem 35 holds without the assumption that Θ_0 is complete separable and $f(y_i|\theta, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_0)$ is bounded and continuous in θ , irrespective of whether or not x_i is treated as known in the case of cross-validation with respect to the true model \mathcal{M}_0 . Indeed, assuming the rest of the conditions of Theorem 35, it holds that

$$\frac{1}{n} \log IPBF^{(n)}(\mathcal{M}_1, \mathcal{M}_2) = \frac{1}{n} \log \left[\frac{\prod_{i=1}^n \pi(y_i|\mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}_1)}{\prod_{i=1}^n \pi(y_i|\mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}_2)} \right] \xrightarrow{a.s.} -h^*(\tilde{\theta}_1, \tilde{\theta}_2), \text{ as } n \rightarrow \infty,$$

where, for any θ_1, θ_2 ,

$$h^*(\theta_1, \theta_2) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log \left[\frac{g(\mathbf{Y}^{(i)}, \theta_2, \mathcal{M}_2)}{g(\mathbf{Y}^{(i)}, \theta_1, \mathcal{M}_1)} \right],$$

provided that the limit exists. The proof follows in the same way as in Theorem 31 by replacing \mathcal{M} and \mathcal{M}_0 with \mathcal{M}_1 and \mathcal{M}_2 . Note that $h^*(\tilde{\theta}_1, \tilde{\theta}_2)$ above is the same as $h^*(\tilde{\theta}_1) - h^*(\tilde{\theta}_2)$ of Theorem 35, but the latter is interpretable as the difference between limiting KL-divergence rates for \mathcal{M}_1 and \mathcal{M}_2 , while the former does not admit such desirable interpretation since without the assumptions Θ_0 is complete separable and $f(y_i|\theta, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_0)$ is bounded and continuous in θ , the convergence

$$\frac{1}{n} \sum_{i=1}^n \log \pi(y_i|\mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}_0) \xrightarrow{a.s.} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log f(y_i|\theta_0, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_0), \text{ as } n \rightarrow \infty,$$

need not hold, even if x_i is considered known for cross-validation with respect to \mathcal{M}_0 .

8.5 Convergence results for PBF in inverse regression: second setup

In the misspecified situation, $\theta_0 \notin \Theta$, and $\tilde{\theta}$ is the minimizer of the limiting KL-divergence rate from θ_0 . If θ is thus misspecified, then as $m \rightarrow \infty$, $B_{im}(\tilde{\theta}) \xrightarrow{a.s.} \{x_i^*\}$ for some non-random x_i^* ($\neq x_i$) depending upon both $\tilde{\theta}$ and θ_0 . In other words, the prior distribution of \tilde{x}_i given $\tilde{\theta}$ and \mathbf{y}_i concentrates around x_i^* , as $m \rightarrow \infty$. We now state and prove our result on IPBF convergence with respect to the prior (6.4.1).

Theorem 37 *Assume conditions (S1)–(S7) of Shalizi. Let the infimum of $h(\theta)$ over Θ be attained at $\tilde{\theta} \in \Theta$, where $\tilde{\theta} \neq \theta_0$. Assume that $\tilde{\theta}$ and θ_0 are one-to-one functions. Also assume that Θ and Θ_0 are complete separable metric spaces and that for $i \geq 1$ and $k \geq 1$, $f(y_{ik}|\theta, \tilde{x}_i, \mathbf{Y}_k^{(i-1)}, \mathcal{M})$ and $f(y_{ik}|\theta, \tilde{x}_i, \mathbf{Y}_k^{(i-1)}, \mathcal{M}_0)$ are bounded and continuous in (θ, \tilde{x}_i) . Then, for prior (6.4.1) on \tilde{x}_i , the following holds for any $k \geq 1$:*

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{1}{n} \log IPBF^{(n,m,k)}(\mathcal{M}, \mathcal{M}_0) = \lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{1}{n} \log \left[\frac{\prod_{i=1}^n \pi(y_{ik}|\mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M})}{\prod_{i=1}^n \pi(y_{ik}|\mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M}_0)} \right] \xrightarrow{a.s.} -h^*(\tilde{\theta}), \quad (8.5.1)$$

where

$$h^*(\tilde{\theta}) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log \left[\frac{f(y_{ik}|\theta_0, x_i, \mathbf{Y}_k^{(i-1)}, \mathcal{M}_0)}{f(y_{ik}|\tilde{\theta}, x_i^*, \mathbf{Y}_k^{(i-1)}, \mathcal{M})} \right],$$

provided that the limit exists.

Proof. It follows from (8.2.6) that $\pi(\tilde{x}_i, \theta | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M}) = \pi(\tilde{x}_i | \theta, \mathcal{M}) \pi(\theta | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M})$.

Hence, letting $U_i \times V$ be any neighborhood of $(x_i^*, \tilde{\theta})$, we have

$$\pi(\tilde{x}_i \in U_i, \theta \in V | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M}) = \int_V \pi(\tilde{x}_i \in U_i | \theta, \mathcal{M}) d\pi(\theta | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}). \quad (8.5.2)$$

Since $\pi(\cdot | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}) \xrightarrow{w} \delta_{\tilde{\theta}}(\cdot)$, as $n \rightarrow \infty$, for any $m \geq 1$, and since $\pi(\tilde{x}_i \in U_i | \theta, \mathcal{M})$ is bounded (since it is a probability) and continuous in θ by Lemma 17, by the

Portmanteau theorem it follows from (8.5.2) that for $m \geq 1$,

$$\pi(\tilde{x}_i \in U_i, \theta \in V | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M}) \xrightarrow{a.s.} \pi(\tilde{x}_i \in U_i | \tilde{\theta}, \mathcal{M}), \text{ as } n \rightarrow \infty. \quad (8.5.3)$$

Now, since $B_{im}(\tilde{\theta}) \xrightarrow{a.s.} \{x_i^*\}$ as $m \rightarrow \infty$ since $\tilde{\theta}$ is one-to-one, it follows that there exists $m_0 \geq 1$ such that for $m \geq m_0$, $B_{im}(\tilde{\theta}) \subset U_i$. Hence,

$$\pi(\tilde{x}_i \in U_i | \tilde{\theta}, \mathcal{M}) \xrightarrow{a.s.} 1, \text{ as } m \rightarrow \infty. \quad (8.5.4)$$

Combining (8.5.3) and (8.5.4) yields

$$\pi(\tilde{x}_i \in U_i, \theta \in V | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M}) \xrightarrow{a.s.} 1, \text{ as } m \rightarrow \infty, n \rightarrow \infty. \quad (8.5.5)$$

From (8.5.5) it follows thanks to complete separability of \mathcal{X} and Θ , that

$$\pi(\cdot | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M}) \xrightarrow{w} \delta_{(x_i^*, \tilde{\theta})}(\cdot), \text{ as } m \rightarrow \infty, n \rightarrow \infty. \quad (8.5.6)$$

Since $\pi(y_{ik} | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M}) = \int_{\mathcal{X}} \int_{\Theta} f(y_{ik} | \theta, \tilde{x}_i, \mathbf{Y}^{(i-1)}, \mathcal{M}) d\pi(\tilde{x}_i, \theta | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M})$, and $f(y_{ik} | \theta, \tilde{x}_i, \mathbf{Y}^{(i-1)}, \mathcal{M})$ is bounded and continuous in (θ, \tilde{x}_i) , it follows using (8.5.6) and the Portmanteau theorem, that

$$\pi(y_{ik} | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M}) \xrightarrow{a.s.} f(y_{ik} | \tilde{\theta}, x_i^*, \mathbf{Y}_k^{(i-1)}, \mathcal{M}), \text{ as } m \rightarrow \infty, n \rightarrow \infty. \quad (8.5.7)$$

Hence,

$$\frac{1}{n} \sum_{i=1}^n \log \pi(y_{ik} | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M}) \xrightarrow{a.s.} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log f(y_{ik} | \tilde{\theta}, x_i^*, \mathbf{Y}_k^{(i-1)}, \mathcal{M}), \text{ as } m \rightarrow \infty, n \rightarrow \infty. \quad (8.5.8)$$

In the same way,

$$\frac{1}{n} \sum_{i=1}^n \log \pi(y_{ik} | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M}_0) \xrightarrow{a.s.} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log f(y_{ik} | \theta_0, x_i, \mathbf{Y}_k^{(i-1)}, \mathcal{M}_0), \text{ as } m \rightarrow \infty, n \rightarrow \infty. \quad (8.5.9)$$

Combining (8.5.8) and (8.5.9) yields

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{1}{n} \log IPBF^{(n,m,k)}(\mathcal{M}, \mathcal{M}_0) \stackrel{a.s.}{=} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log \left[\frac{f(y_{ik} | \tilde{\theta}, x_i^*, \mathbf{Y}_k^{(i-1)}, \mathcal{M})}{f(y_{ik} | \theta_0, x_i, \mathbf{Y}_k^{(i-1)}, \mathcal{M}_0)} \right] = -h^*(\tilde{\theta}),$$

thereby proving the result. ■

Remark 38 Theorem 37 assumes that for \mathcal{M}_0 , cross-validation is carried out assuming x_i is unknown. However, as is clear from the proof, the same result continues to hold even if x_i is treated as known.

Theorem 39 For models \mathcal{M}_0 , \mathcal{M}_1 and \mathcal{M}_2 with complete separable parameter spaces Θ_0 , Θ_1 and Θ_2 , assume conditions (S1)–(S7) of Shalizi and for $j = 1, 2$, let the infimum of $h_j(\theta)$ over Θ_j be attained at $\tilde{\theta}_j \in \Theta_j$, where $\tilde{\theta}_j \neq \theta_0$. Consider the prior (6.4.1) on \tilde{x}_i and let $B_{im}(\tilde{\theta}_j) \xrightarrow{a.s.} \{x_{ij}^*\}$, for $j = 1, 2$. Also assume that for $i \geq 1$ and $k \geq 1$, $f(y_{ik} | \theta, \tilde{x}_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_j)$; $j = 1, 2$, and $f(y_{ik} | \theta, \tilde{x}_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_0)$ are bounded and continuous in (θ, \tilde{x}_i) , in addition to the conditions that θ_0 and $\tilde{\theta}_j$; $j = 1, 2$, are one-to-one.

Then, the following holds for any $k \geq 1$:

$$\begin{aligned} & \lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{1}{n} \log IPBF^{(n,m,k)}(\mathcal{M}_1, \mathcal{M}_2) \\ &= \lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{1}{n} \log \left[\frac{\prod_{i=1}^n \pi(y_{ik} | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M}_1)}{\prod_{i=1}^n \pi(y_{ik} | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M}_2)} \right] \stackrel{a.s.}{=} -[h_1^*(\tilde{\theta}_1) - h_2^*(\tilde{\theta}_2)], \end{aligned} \quad (8.5.10)$$

where, for $j = 1, 2$, and for any θ ,

$$h_j^*(\theta) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log \left[\frac{f(y_{ik} | \theta_0, x_i, \mathbf{Y}_k^{(i-1)}, \mathcal{M}_0)}{f(y_{ik} | \theta, x_{ij}^*, \mathbf{Y}_k^{(i-1)}, \mathcal{M}_j)} \right], \quad (8.5.11)$$

provided the limit exists.

Proof. The proof follows by noting that

$$\frac{1}{n} \log IPBF^{(n,m,k)}(\mathcal{M}_1, \mathcal{M}_2) = \frac{1}{n} \log IPBF^{(n,m,k)}(\mathcal{M}_1, \mathcal{M}_0) - \frac{1}{n} \log IPBF^{(n,m,k)}(\mathcal{M}_2, \mathcal{M}_0),$$

and then using (8.5.1) for $\frac{1}{n} \log IPBF^{(n,m,k)}(\mathcal{M}_1, \mathcal{M}_0)$ and $\frac{1}{n} \log IPBF^{(n,m,k)}(\mathcal{M}_2, \mathcal{M}_0)$.

■

Remark 40 As in Remark 36 note that the result of Theorem 39 holds without the assumption that Θ_0 is complete separable and $f(y_{ik}|\theta, \tilde{x}_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_0)$ is bounded and continuous in (θ, \tilde{x}_i) for $k \geq 1$, irrespective of whether or not x_i is treated as known for cross-validation with respect to \mathcal{M}_0 . In this case, assuming the rest of the conditions of Theorem 39, it holds for any $k \geq 1$, that

$$\begin{aligned} & \lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{1}{n} \log IPBF^{(n,m,k)}(\mathcal{M}_1, \mathcal{M}_2) \\ &= \lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{1}{n} \log \left[\frac{\prod_{i=1}^n \pi(y_{ik}|\mathbf{Y}_{nm,-i}, \mathbf{X}_{nm,-i}, \mathcal{M}_1)}{\prod_{i=1}^n \pi(y_{ik}|\mathbf{Y}_{nm,-i}, \mathbf{X}_{nm,-i}, \mathcal{M}_2)} \right] \stackrel{a.s.}{=} -h^*(\tilde{\theta}_1, \tilde{\theta}_2), \end{aligned}$$

where, for any θ_1, θ_2 ,

$$h^*(\theta_1, \theta_2) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log \left[\frac{f(y_{ik}|\theta_2, x_{i2}^*, \mathbf{Y}_k^{(i-1)}, \mathcal{M}_2)}{f(y_{ik}|\theta_1, x_{i1}^*, \mathbf{Y}_k^{(i-1)}, \mathcal{M}_1)} \right],$$

provided that the limit exists. As in Remark 36, again $h^*(\tilde{\theta}_1, \tilde{\theta}_2)$ above is the same as $h^*(\tilde{\theta}_1) - h^*(\tilde{\theta}_2)$ of Theorem 39, although, unlike the latter, the former need not be interpretable as the difference between limiting KL-divergence rates for \mathcal{M}_1 and \mathcal{M}_2 .

8.6 Illustrations of PBF convergence in forward regression problems

8.6.1 Forward linear regression model

Let

$$\mathcal{M}_1 : y_i = \alpha + \beta x_i + \epsilon_i; \quad i = 1, \dots, n, \quad (8.6.1)$$

where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ independently, for $i = 1, \dots, n$. Here $\theta = (\alpha, \beta, \sigma_\epsilon^2)$ is the unknown set of parameters. Let the parameter space be $\Theta = \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$. Clearly, Θ is complete and separable.

Also let

$$\mathcal{M}_0 : y_i = \eta_0(x_i) + \epsilon_i; \quad i = 1, \dots, n, \quad (8.6.2)$$

where $\eta_0(x)$ is the true, non-linear function of x , which is also continuous, and $\epsilon_i \sim N(0, \sigma_0^2)$ independently, for $i = 1, \dots, n$. In this

Let us assume that \mathcal{X} , the covariate space, is compact, under both \mathcal{M}_1 and \mathcal{M}_0 .

Verification of the assumptions

From (8.6.1) it is clear that $f(y_i|\theta, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_1) = f(y_i|\theta, x_i, \mathcal{M}_1)$ is bounded and continuous in θ , and the true model $f(y_i|x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_0) = f(y_i|x_i, \mathcal{M}_0)$ is devoid of any parameters. Consequently, in this case, $\pi(y_i|\mathbf{Y}_{n,-i}, \mathbf{X}_n, \mathcal{M}_0) \equiv f(y_i|x_i, \mathcal{M}_0)$.

We are now left to verify the seven assumptions of Shalizi. First note from the forms of (8.6.1) and (8.6.2) that measurability of $R_n(\theta)$ clearly holds, so that the first assumption of Shalizi, namely, (S1) is satisfied.

Now,

$$\begin{aligned} \frac{1}{n} \log \prod_{i=1}^n f(y_i | \theta, x_i, \mathcal{M}_1) &= -\frac{1}{2} \log 2\pi\sigma_\epsilon^2 - \frac{1}{2\sigma_\epsilon^2 n} \sum_{i=1}^n (y_i - \eta_0(x_i))^2 - \frac{1}{2\sigma_\epsilon^2 n} \sum_{i=1}^n (\eta_0(x_i) - \alpha - \beta x_i)^2 \\ &\quad - \frac{1}{\sigma_\epsilon^2 n} \sum_{i=1}^n (y_i - \eta_0(x_i))(\eta_0(x_i) - \alpha - \beta x_i). \end{aligned} \quad (8.6.3)$$

In (8.6.3),

$$\frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i))^2 \xrightarrow{a.s.} \sigma_0^2, \text{ as } n \rightarrow \infty, \quad (8.6.4)$$

and letting $|\mathcal{X}|$ denote the Lebesgue measure of the compact space \mathcal{X} ,

$$\frac{1}{n} \sum_{i=1}^n (\eta_0(x_i) - \alpha - \beta x_i)^2 \rightarrow |\mathcal{X}|^{-1} \int_{\mathcal{X}} (\eta_0(x) - \alpha - \beta x)^2 dx, \text{ as } n \rightarrow \infty, \quad (8.6.5)$$

since the former is a Riemann sum. Also, letting E_0 and V_0 denote the mean and variance under model \mathcal{M}_0 , we see that for all $n \geq 1$,

$$E_0 \left[\frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i))(\eta_0(x_i) - \alpha - \beta x_i) \right] = 0, \quad (8.6.6)$$

and

$$\sum_{i=1}^{\infty} \frac{V_0 [(y_i - \eta_0(x_i))(\eta_0(x_i) - \alpha - \beta x_i)]}{i^2} \leq \sigma_0^2 \sup_{x \in \mathcal{X}} (\eta_0(x) - \alpha - \beta x)^2 \sum_{i=1}^{\infty} \frac{1}{i^2} < \infty. \quad (8.6.7)$$

From (8.6.6) and (8.6.7), it follows from Kolmogorov's strong law of large numbers for independent but non-identical random variables,

$$\frac{1}{n} \sum_{i=1}^n (y_i - \eta_0(x_i))(\eta_0(x_i) - \alpha - \beta x_i) \xrightarrow{a.s.} 0, \text{ as } n \rightarrow \infty. \quad (8.6.8)$$

Applying (8.6.4), (8.6.5) and (8.6.8) to (8.6.3) yields

$$\frac{1}{n} \log \prod_{i=1}^n f(y_i|\theta, x_i, \mathcal{M}_1) \xrightarrow{a.s.} -\frac{1}{2} \log 2\pi\sigma_\epsilon^2 - \frac{|\mathcal{X}|^{-1}}{2\sigma_\epsilon^2} \int_{\mathcal{X}} (\eta_0(x) - \alpha - \beta x)^2 dx, \text{ as } n \rightarrow \infty. \quad (8.6.9)$$

Now observe that for the true model \mathcal{M}_0 ,

$$\frac{1}{n} \log \prod_{i=1}^n f(y_i|x_i, \mathcal{M}_0) = -\frac{1}{2} \log 2\pi\sigma_0^2 - \frac{1}{2\sigma_0^2 n} \sum_{i=1}^n (y_i - \eta_0(x_i))^2 \xrightarrow{a.s.} -\frac{1}{2} \log 2\pi\sigma_0^2 - \frac{1}{2}, \text{ as } n \rightarrow \infty. \quad (8.6.10)$$

From (8.6.9) and (8.6.10) we have, for $\theta \in \Theta$,

$$\frac{1}{n} \log R_n(\theta) \xrightarrow{a.s.} -h_1(\theta),$$

where

$$h_1(\theta) = \frac{1}{2} \log \left(\frac{\sigma_\epsilon^2}{\sigma_0^2} \right) + \frac{\sigma_0^2}{2\sigma_\epsilon^2} + \frac{|\mathcal{X}|^{-1}}{2\sigma_\epsilon^2} \int_{\mathcal{X}} (\eta_0(x) - \alpha - \beta x)^2 dx - \frac{1}{2}. \quad (8.6.11)$$

Hence, (S3) of Shalizi holds.

It is easy to see by taking the limits of the expectations of $\frac{1}{n} \log \prod_{i=1}^n f(y_i|\theta, x_i, \mathcal{M}_1)$ and $\frac{1}{n} \log \prod_{i=1}^n f(y_i|x_i, \mathcal{M}_0)$, that the following also holds:

$$\lim_{n \rightarrow \infty} \frac{1}{n} E_0 [\log R_n(\theta)] = -h_1(\theta).$$

In other words, (S2) holds.

Note that $h_1(\theta) < \infty$ almost surely if under the priors for $\alpha, \beta, \sigma_\epsilon^2$, $|\alpha| < \infty$, $|\beta| < \infty$ and $0 < \sigma_\epsilon^2 < \infty$, almost surely. Hence, (S4) holds.

Let

$$\mathcal{G}_n = \{ \theta \in \Theta : |\alpha| \leq \exp(\gamma n), |\beta| \leq \exp(\gamma n), \sigma_\epsilon^{-2} \leq \exp(\gamma n) \}, \quad (8.6.12)$$

where $\gamma > 2h(\Theta)$. Then $\mathcal{G}_n \uparrow \Theta$, as $n \rightarrow \infty$.

Let us assume that the prior for $(\alpha, \beta, \sigma_\epsilon^{-2})$ is such that the prior expectations $E(|\alpha|)$, $E(|\beta|)$ and $E(\sigma_\epsilon^{-2})$ are finite. Then under such priors, using Markov's inequality, the probabilities $P(|\alpha| > \exp(\gamma n))$, $P(|\beta| > \exp(\gamma n))$ and $P(\sigma_\epsilon^{-2} > \exp(\gamma n))$ are bounded above as follows:

$$P(|\alpha| > \exp(\gamma n)) < E(|\alpha|) \exp(-\gamma n); \quad (8.6.13)$$

$$P(|\beta| > \exp(\gamma n)) < E(|\beta|) \exp(-\gamma n); \quad (8.6.14)$$

$$P(\sigma_\epsilon^{-2} > \exp(\gamma n)) < E(\sigma_\epsilon^{-2}) \exp(-\gamma n). \quad (8.6.15)$$

From (8.6.12) and the inequalities (8.6.13), (8.6.14) and (8.6.15) it follows that

$$\begin{aligned} \pi(\mathcal{G}_n) &\geq 1 - (P(|\alpha| > \exp(\gamma n)) + P(|\beta| > \exp(\gamma n)) + P(\sigma_\epsilon^{-2} > \exp(\gamma n))) \\ &\geq 1 - (E(|\alpha|) + E(|\beta|) + E(\sigma_\epsilon^{-2})) \exp(-\gamma n). \end{aligned} \quad (8.6.16)$$

Thus, (S5)(1) holds.

The differential of $\frac{1}{n} \log R_n(\theta)$ is continuous in θ , and since \mathcal{X} is compact, it is easy to see that the differential is almost surely bounded on any compact subset G of Θ , as $n \rightarrow \infty$. That is, $\frac{1}{n} \log R_n(\theta)$ is almost surely Lipschitz, hence, equicontinuous on G . Since $\frac{1}{n} \log R_n(\theta)$ almost surely converges to $-h_1(\theta)$ pointwise, as $n \rightarrow \infty$, it holds due to the stochastic Ascoli lemma that

$$\limsup_{n \rightarrow \infty} \left| \frac{1}{n} \log R_n(\theta) + h_1(\theta) \right| = 0, \text{ almost surely.} \quad (8.6.17)$$

Since for any $n \geq 1$, \mathcal{G}_n is compact, (S5)(2) holds.

Since $h_1(\theta)$ is continuous in θ , \mathcal{G}_n is compact and $h(\mathcal{G}_n)$ is non-increasing in n , (S5)(3) holds. Also, for any set A such that $\pi(A) > 0$, since $\mathcal{G}_n \cap A$ increases to A , it follows due to continuity of $h_1(\theta)$ that $h(\mathcal{G}_n \cap A)$ decreases to $h_1(A)$, so that (S7) holds.

Regarding verification of (S6), recall that the aim of assumption (S6) is to ensure that (see the proof of Lemma 7 of Shalizi (2009)) for every $\varepsilon > 0$ and for all n sufficiently large,

$$\frac{1}{n} \log \int_{\mathcal{G}_n} R_n(\theta) d\pi(\theta) \leq -h(\mathcal{G}_n) + \varepsilon, \text{ almost surely.}$$

Since $h(\mathcal{G}_n) \rightarrow h(\Theta)$ as $n \rightarrow \infty$, it is enough to verify that for every $\varepsilon > 0$ and for all n sufficiently large,

$$\frac{1}{n} \log \int_{\mathcal{G}_n} R_n(\theta) d\pi(\theta) \leq -h(\Theta) + \varepsilon, \text{ almost surely.}$$

In other words, it is sufficient to verify that

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log \int_{\mathcal{G}_n} R_n(\theta) \pi(\theta) d\theta \leq -h(\Theta), \text{ almost surely.} \quad (8.6.18)$$

Theorem 55 stated and proved in Appendix 8.A1 provides sufficient conditions for (8.6.18) to hold in general with proper priors on the parameters. We now make use of Theorem 55 of Appendix 8.A1 to validate (S6) of Shalizi. For any function $g(x)$ on \mathcal{X} , let us consider the notation

$$E_X[g(X)] = |\mathcal{X}|^{-1} \int_{\mathcal{X}} g(x) dx. \quad (8.6.19)$$

Note that (8.6.19) is indeed the expectation of $g(X)$ with respect to the uniform distribution on the compact set \mathcal{X} .

Now observe that $h_1(\theta)$ is uniquely minimized by

$$\tilde{\beta} = \frac{E_X[(X - E_X(X))(\eta_0(X) - E(\eta_0(X)))]}{E_X(X - E_X(X))^2}, \quad (8.6.20)$$

$$\tilde{\alpha} = E_X(\eta_0(X)) - \tilde{\beta}E_X(X); \quad (8.6.21)$$

$$\tilde{\sigma}_\epsilon^2 = \sigma_0^2 + E_X \left(\eta_0(X) - \tilde{\alpha} - \tilde{\beta}X \right)^2. \quad (8.6.22)$$

Now, letting $\bar{x}_n = \frac{\sum_{i=1}^n x_i}{n}$, $\bar{y}_n = \frac{\sum_{i=1}^n y_i}{n}$ and $\bar{\eta}_{0n} = \frac{\sum_{i=1}^n \eta_0(x_i)}{n}$, we see that $\frac{1}{n} \log R_n(\theta)$ is

maximized at

$$\tilde{\beta}_n^* = \frac{\sum_{i=1}^n (y_i - \bar{y}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}; \quad (8.6.23)$$

$$\tilde{\alpha}_n^* = \bar{y}_n - \tilde{\beta}_n^* \bar{x}_n; \quad (8.6.24)$$

$$\begin{aligned} \tilde{\sigma}_n^{*2} &= \frac{1}{n} \left[\sum_{i=1}^n (y_i - \eta_0(x_i))^2 + \sum_{i=1}^n (\eta_0(x_i) - \tilde{\alpha}_n^* - \tilde{\beta}_n^* x_i)^2 \right. \\ &\quad \left. + 2 \sum_{i=1}^n (y_i - \eta_0(x_i)) (\eta_0(x_i) - \tilde{\alpha}_n^* - \tilde{\beta}_n^* x_i) \right]. \end{aligned} \quad (8.6.25)$$

Using Kolmogorov's strong law of large numbers and Riemann sum convergence, we see that

$$\tilde{\beta}_n^* \xrightarrow{a.s.} \tilde{\beta}, \quad (8.6.26)$$

where $\tilde{\beta}$ is given by (8.6.20).

By (8.6.26), and since $\bar{y}_n \xrightarrow{a.s.} E_X(\eta_0(X))$, $\bar{x}_n \rightarrow E_X(X)$, it follows that

$$\tilde{\alpha}_n^* \xrightarrow{a.s.} \tilde{\alpha}, \quad (8.6.27)$$

where $\tilde{\alpha}$ is given by (8.6.21).

For the convergence of $\tilde{\sigma}_n^{*2}$ given by (8.6.25), first observe that the first term on the right hand side of (8.6.25) converges almost surely to σ_0^2 . The i -th term of the second term on the right hand side converges to $(\eta_0(x_i) - \tilde{\alpha} - \tilde{\beta}x_i)^2$ almost surely, so that the second term converges to $E_X(\eta_0(X) - \tilde{\alpha} - \tilde{\beta}X)^2$. The i -th term of the third term on the right hand side converges almost surely to $2(y_i - \eta_0(x_i))(\eta_0(x_i) - \tilde{\alpha} - \tilde{\beta}x_i)$, so that the third term converges to zero almost surely due to (8.6.8). It follows that

$$\tilde{\sigma}_n^{*2} \xrightarrow{a.s.} \tilde{\sigma}_\epsilon^2, \quad (8.6.28)$$

where $\tilde{\sigma}_\epsilon^2$ is given by (8.6.22). Combining (8.6.26), (8.6.27) and (8.6.28) yields

$$\tilde{\theta}_n^* = \left(\tilde{\alpha}_n^*, \tilde{\beta}_n^*, \tilde{\sigma}_n^*{}^2 \right) \xrightarrow{a.s.} \left(\tilde{\alpha}, \tilde{\beta}, \tilde{\sigma}_\epsilon^2 \right) = \tilde{\theta}, \text{ as } n \rightarrow \infty. \quad (8.6.29)$$

In other words, we have shown that conditions (i) and (ii) of Theorem 55 hold. Since we have already shown pointwise almost sure convergence of $\frac{1}{n} \log R_n(\theta)$ to $-h_1(\theta)$ in the context of verifying (S3) and stochastic equicontinuity of $\frac{1}{n} \log R_n(\theta)$ on compact subsets of Θ in the context of verifying (S5)(2), all the conditions of Theorem 55 go through with proper prior for θ . Hence (8.6.18), and consequently, (S6), holds.

With these, it is seen that the conditions of Theorem 29 are satisfied, which leads to the following specialized version of the theorem:

Theorem 41 *Consider the linear regression model \mathcal{M}_1 given by (8.6.1) and the true, non-linear model \mathcal{M}_0 given by (8.6.2). Assume the parameter space Θ associated with model \mathcal{M}_1 be $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$, and let the covariate space \mathcal{X} be compact. Then (8.3.1) holds for $\frac{1}{n} \log FPBF^{(n)}(\mathcal{M}_1, \mathcal{M}_0)$, where for $\theta \in \Theta$, $h(\theta) = h_1(\theta)$ is given by (8.6.11), and $\tilde{\theta} = (\tilde{\alpha}, \tilde{\beta}, \tilde{\sigma}_\epsilon^2)$, where $\tilde{\alpha}$, $\tilde{\beta}$ and $\tilde{\sigma}_\epsilon^2$ are given by (8.6.21), (8.6.20) and (8.6.22), respectively.*

8.6.2 Forward quadratic regression model

Now consider the following model on quadratic regression which may be regarded as a competitor to linear regression:

$$\mathcal{M}_2 : y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i; \quad i = 1, \dots, n, \quad (8.6.30)$$

where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ independently, for $i = 1, \dots, n$. Here $\theta = (\alpha, \beta_1, \beta_2, \sigma_\epsilon^2)$ is the unknown set of parameters, and the parameter space is $\Theta = \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$.

In this case,

$$\frac{1}{n} \log R_n(\theta) \xrightarrow{a.s.} -h_2(\theta),$$

where

$$h_2(\theta) = \frac{1}{2} \log \left(\frac{\sigma_\epsilon^2}{\sigma_0^2} \right) + \frac{\sigma_0^2}{2\sigma_\epsilon^2} + \frac{|\mathcal{X}|^{-1}}{2\sigma_\epsilon^2} \int_{\mathcal{X}} (\eta_0(x) - \alpha - \beta_1 x - \beta_2 x^2)^2 dx - \frac{1}{2}. \quad (8.6.31)$$

It is easy to see that $h_2(\theta)$ is uniquely minimized at $\tilde{\vartheta} = (\tilde{\alpha}, \tilde{\beta}_1, \tilde{\beta}_2)$, given by

$$\tilde{\vartheta} = A^{-1}b, \quad (8.6.32)$$

where

$$A = \begin{pmatrix} 1 & E_X(X) & E_X(X^2) \\ E_X(X) & E_X(X^2) & E_X(X^3) \\ E_X(X^2) & E_X(X^3) & E_X(X^4) \end{pmatrix} \text{ and } b = \begin{pmatrix} E_X(\eta_0(X)) \\ E_X(X\eta_0(X)) \\ E_X(X^2\eta_0(X)) \end{pmatrix}, \quad (8.6.33)$$

and

$$\tilde{\sigma}_\epsilon^2 = \sigma_0^2 + E_X \left(\eta_0(X) - \tilde{\alpha} - \tilde{\beta}_1 X - \tilde{\beta}_2 X^2 \right)^2. \quad (8.6.34)$$

That A in (8.6.33) is invertible, will be shown shortly.

The maximizer of $\frac{1}{n} \log R_n(\theta)$ here is given by the least squares estimators $\tilde{\vartheta}_n^* = (\tilde{\alpha}_n^*, \tilde{\beta}_{1n}^*, \tilde{\beta}_{2n}^*)$ given by

$$\tilde{\vartheta}_n^* = A_n^{-1}b_n, \quad (8.6.35)$$

where

$$A_n = n^{-1} \begin{pmatrix} n & \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 \\ \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^4 \end{pmatrix} \text{ and } b_n = n^{-1} \begin{pmatrix} \sum_{i=1}^n \eta_0(x_i) \\ \sum_{i=1}^n x_i \eta_0(x_i) \\ \sum_{i=1}^n x_i^2 \eta_0(x_i) \end{pmatrix}, \quad (8.6.36)$$

and

$$\begin{aligned}\tilde{\sigma}^*_n &= \frac{1}{n} \left[\sum_{i=1}^n (y_i - \eta_0(x_i))^2 + \sum_{i=1}^n \left(\eta_0(x_i) - \tilde{\alpha}_n^* - \tilde{\beta}_{1n}^* x_i - \tilde{\beta}_{2n}^* x_i^2 \right)^2 \right. \\ &\quad \left. + 2 \sum_{i=1}^n (y_i - \eta_0(x_i)) \left(\eta_0(x_i) - \tilde{\alpha}_n^* - \tilde{\beta}_{1n}^* x_i - \tilde{\beta}_{2n}^* x_i^2 \right) \right].\end{aligned}\quad (8.6.37)$$

Now note that A_n in (8.6.36) corresponds to the so-called Vandermonde design matrix (see, for example, Macon and Spitzbart (1958)) associated with the least squares quadratic regression. The design matrix is of full rank if all the x_i are distinct, which we assume. Hence, for all $n \geq 3$, A_n is invertible, which makes the least squares estimators $\tilde{\vartheta}_n^*$, given by (8.6.35), well-defined, for all $n \geq 3$. Now observe that by Riemann sum convergence,

$$A_n \xrightarrow{a.s.} A, \text{ as } n \rightarrow \infty, \text{ and} \quad (8.6.38)$$

$$b_n \xrightarrow{a.s.} b, \text{ as } n \rightarrow \infty. \quad (8.6.39)$$

Since A_n is invertible for every $n \geq 3$, A must also be invertible, since (8.6.38) holds. Hence, $\tilde{\vartheta}$ given by (8.6.32), is well-defined.

Now, thanks to (8.6.38) and (8.6.39), we have

$$\tilde{\vartheta}_n^* \xrightarrow{a.s.} \tilde{\vartheta}, \text{ as } n \rightarrow \infty,$$

and also in the same way as for model \mathcal{M}_1 , here also,

$$\tilde{\sigma}_n^2 \xrightarrow{a.s.} \tilde{\sigma}_\epsilon^2, \text{ as } n \rightarrow \infty.$$

In other words,

$$\tilde{\theta}_n^* \xrightarrow{a.s.} \tilde{\theta}, \text{ as } n \rightarrow \infty,$$

even for model \mathcal{M}_2 .

For this quadratic regression model, let

$$\mathcal{G}_n = \left\{ \theta \in \Theta : |\alpha| \leq \exp(\gamma n), |\beta_1| \leq \exp(\gamma n), |\beta_2| \leq \exp(\gamma n), \sigma_\epsilon^{-2} \leq \exp(\gamma n) \right\},$$

where $\gamma > 2h(\Theta)$. Then $\mathcal{G}_n \uparrow \Theta$, as $n \rightarrow \infty$, and the rest of the assumptions of Shalizi are easily seen to be satisfied. The condition of boundedness and continuity of $f(y_i|\theta, x_i, \mathcal{M}_2)$ are also clearly satisfied.

We summarize our results on FPBF consistency in favour of \mathcal{M}_0 when the data is modeled by \mathcal{M}_2 as follows.

Theorem 42 *Consider the quadratic regression model \mathcal{M}_2 given by (8.6.30) and the true, non-linear model \mathcal{M}_0 given by (8.6.2). Assume the parameter space Θ associated with model \mathcal{M}_2 be $\mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$, and let the covariate space \mathcal{X} be compact. Also assume that $x_i; i \geq 1$ are all distinct. Then (8.3.1) holds for $\frac{1}{n} \log FPBF^{(n)}(\mathcal{M}_2, \mathcal{M}_0)$, where for $\theta \in \Theta$, $h(\theta) = h_2(\theta)$ is given by (8.6.31), and $\tilde{\theta} = (\tilde{\alpha}, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\sigma}_\epsilon^2)$, where $\tilde{\alpha}$, $\tilde{\beta}_1$, $\tilde{\beta}_2$ and $\tilde{\sigma}_\epsilon^2$ are given by (8.6.32) and (8.6.34).*

8.6.3 Asymptotic comparison of forward linear and quadratic models with FPBF

Theorems 41 and 42 show almost sure exponential convergence of FPBF in favour of the true model \mathcal{M}_0 given by (8.6.2) when the postulated models are either the forward linear or quadratic regression model. Now, if the goal is to make asymptotic comparison between the linear and quadratic regression models, then the aforementioned theorems ensure the following result:

Theorem 43 *Let the true model be given by \mathcal{M}_0 formulated in (8.6.2). Assuming that the covariate observations $x_i; i \geq 1$ are all distinct and that the covariate space \mathcal{X} is compact, consider comparison of the linear and quadratic regression models \mathcal{M}_1 and \mathcal{M}_2 given by (8.6.1) and (8.6.30), respectively. Let $\tilde{\theta}_1$ and $\tilde{\theta}_2$ be the unique minimizers of h_1*

and h_2 . Then,

$$\frac{1}{n} \log FPBF^{(n)}(\mathcal{M}_1, \mathcal{M}_2) \xrightarrow{a.s.} -\left(h_1(\tilde{\theta}_1) - h_2(\tilde{\theta}_2)\right), \text{ as } n \rightarrow \infty.$$

8.6.4 FPBF asymptotics for variable selection in autoregressive time series regression

Let us consider the following first order autoregressive (AR(1)) time series linear regression as model \mathcal{M}_1 :

$$y_t = \rho_1 y_{t-1} + \beta_1 x_t + \epsilon_{1t}; \quad t = 1, \dots, n, \quad (8.6.40)$$

where $y_0 \equiv 0$, $x_t; t = 1, \dots, n$ are covariate observations associated with variable x and $\epsilon_{1t} \stackrel{iid}{\sim} N(0, \sigma_1^2)$. Here $\theta_1 = (\rho_1, \beta_1, \sigma_1^2)$ is the set of unknown parameters and $\Theta_1 = \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$ is the parameter space. We might wish to compare this model with another AR(1) regression model with covariate z different from x . This model, which we refer to as \mathcal{M}_2 , is given as follows:

$$y_t = \rho_2 y_{t-1} + \beta_2 z_t + \epsilon_{2t}; \quad t = 1, \dots, n, \quad (8.6.41)$$

where $y_0 \equiv 0$, $z_t; t = 1, \dots, n$ are observations associated with covariate z different from x and $\theta_2 = (\rho_2, \beta_2, \sigma_2^2)$ is the set of parameters and the parameter space $\Theta_2 = \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$ remains the same as Θ_1 . Here, for $t = 1, \dots, n$, $\epsilon_{2t} \stackrel{iid}{\sim} N(0, \sigma_2^2)$. Let the true model \mathcal{M}_0 be given by

$$y_t = \rho_0 y_{t-1} + \beta_0(x_t + z_t) + \epsilon_{0t}; \quad t = 1, \dots, n, \quad (8.6.42)$$

where $|\rho_0| < 1$ and $\epsilon_{0t} \stackrel{iid}{\sim} N(0, \sigma_0^2)$, for $t = 1, \dots, n$.

Our goal in this example is to compare models \mathcal{M}_1 and \mathcal{M}_2 using FPBF. Note that if we use the same priors for θ_1 and θ_2 , this boils down to selection of either covariate x or z in the AR(1) regression. Hence, variable selection constitutes an important ingredient in this FPBF convergence example. Note that both the models \mathcal{M}_1 and \mathcal{M}_2 are wrong

with respect to the true model \mathcal{M}_0 which consists of both x and z . The purpose of variable selection here is then to select the more important variable among x and z when none of the available models considers both x and z .

We make the following assumptions that are analogous to the AR(1) regression example considered in [Chandra and Bhattacharya \(2020a\)](#):

(A1)

$$\begin{aligned} & \frac{1}{n} \sum_{t=1}^n x_t \rightarrow 0, \quad \frac{1}{n} \sum_{t=1}^n z_t \rightarrow 0; \\ & \frac{1}{n} \sum_{t=1}^n x_t z_t \rightarrow 0; \quad \frac{1}{n} \sum_{t=1}^n x_{t+k} z_t \rightarrow 0; \quad \frac{1}{n} \sum_{t=1}^n x_t z_{t+k} \rightarrow 0 \text{ for any } k \geq 1; \\ & \frac{1}{n} \sum_{t=1}^n x_{t+k} x_t \rightarrow 0 \text{ and } \frac{1}{n} \sum_{t=1}^n z_{t+k} z_t \rightarrow 0 \text{ for any } k \geq 1; \\ & \frac{1}{n} \sum_{t=1}^n x_t^2 \rightarrow \sigma_x^2 \text{ and } \frac{1}{n} \sum_{t=1}^n z_t^2 \rightarrow \sigma_z^2, \end{aligned}$$

as $n \rightarrow \infty$. In the above, σ_x^2 and σ_z^2 are positive quantities.

(A2) $\sup_{t \geq 1} |x_t \beta_0| < C$ and $\sup_{t \geq 1} |z_t \beta_0| < C$, for some $C > 0$.

Let $\frac{1}{n} \log R_n^{(1)}(\theta)$ and $\frac{1}{n} \log R_n^{(2)}(\theta)$ stand for $\frac{1}{n} \log R_n(\theta)$ for models \mathcal{M}_1 and \mathcal{M}_2 , respectively. Also let $\sigma_{x+z}^2 = \sigma_x^2 + \sigma_z^2$. Then proceeding in the same way as in [Chandra and Bhattacharya \(2020a\)](#) it can be shown that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log R_n^{(1)}(\theta) \stackrel{a.s.}{=} -h_1(\theta), \quad \text{for all } \theta \in \Theta_1; \quad (8.6.43)$$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log R_n^{(2)}(\theta) \stackrel{a.s.}{=} -h_2(\theta), \quad \text{for all } \theta \in \Theta_2, \quad (8.6.44)$$

and the above convergences are uniform on compact subsets of Θ_1 and Θ_2 , respectively.

In the above,

$$\begin{aligned}
 h_1(\theta) = & \log\left(\frac{\sigma}{\sigma_0}\right) + \left(\frac{1}{2\sigma^2} - \frac{1}{2\sigma_0^2}\right) \left(\frac{\sigma_0^2}{1-\rho_0^2} + \frac{\beta_0^2\sigma_{x+z}^2}{1-\rho_0^2}\right) + \left(\frac{\rho^2}{2\sigma^2} - \frac{\rho_0^2}{2\sigma_0^2}\right) \left(\frac{\sigma_0^2}{1-\rho_0^2} + \frac{\beta_0^2\sigma_{x+z}^2}{1-\rho_0^2}\right) \\
 & + \frac{1}{2\sigma^2}\beta^2\sigma_{x+z}^2 - \frac{1}{2\sigma_0^2}\beta_0^2\sigma_{x+z}^2 - \left(\frac{\rho}{\sigma^2} - \frac{\rho_0}{\sigma_0^2}\right) \left(\frac{\rho_0\sigma_0^2}{1-\rho_0^2} + \frac{\rho_0\beta_0^2\sigma_{x+z}^2}{1-\rho_0^2}\right) - \left(\frac{\beta}{\sigma^2} - \frac{\beta_0}{\sigma_0^2}\right) \sigma_{x+z}^2\beta_0 + \frac{\sigma_z^2\beta(2\beta_0 - \beta)}{2\sigma^2}.
 \end{aligned} \tag{8.6.45}$$

and

$$\begin{aligned}
 h_2(\theta) = & \log\left(\frac{\sigma}{\sigma_0}\right) + \left(\frac{1}{2\sigma^2} - \frac{1}{2\sigma_0^2}\right) \left(\frac{\sigma_0^2}{1-\rho_0^2} + \frac{\beta_0^2\sigma_{x+z}^2}{1-\rho_0^2}\right) + \left(\frac{\rho^2}{2\sigma^2} - \frac{\rho_0^2}{2\sigma_0^2}\right) \left(\frac{\sigma_0^2}{1-\rho_0^2} + \frac{\beta_0^2\sigma_{x+z}^2}{1-\rho_0^2}\right) \\
 & + \frac{1}{2\sigma^2}\beta^2\sigma_{x+z}^2 - \frac{1}{2\sigma_0^2}\beta_0^2\sigma_{x+z}^2 - \left(\frac{\rho}{\sigma^2} - \frac{\rho_0}{\sigma_0^2}\right) \left(\frac{\rho_0\sigma_0^2}{1-\rho_0^2} + \frac{\rho_0\beta_0^2\sigma_{x+z}^2}{1-\rho_0^2}\right) - \left(\frac{\beta}{\sigma^2} - \frac{\beta_0}{\sigma_0^2}\right) \sigma_{x+z}^2\beta_0 + \frac{\sigma_x^2\beta(2\beta_0 - \beta)}{2\sigma^2}.
 \end{aligned} \tag{8.6.46}$$

For $i = 1, 2$, for model \mathcal{M}_i , let

$$\mathcal{G}_n^{(i)} = \{\theta \in \Theta_i : |\rho| \leq \exp(\gamma_i n), |\beta| \leq \exp(\gamma_i n), \sigma_\epsilon^{-2} \leq \exp(\gamma_i n)\}, \tag{8.6.47}$$

where $\gamma_i > 2h_i(\Theta_i)$. Then $\mathcal{G}_n^{(i)} \uparrow \Theta_i$, as $n \rightarrow \infty$. Let us assume that under both \mathcal{M}_1 and \mathcal{M}_2 , the prior for $(\rho, \beta, \sigma_\epsilon^{-2})$ is such that the prior expectations $E(|\rho|)$, $E(|\beta|)$ and $E(\sigma_\epsilon^{-2})$ are finite.

With these, conditions (S1)–(S5) and (S7) of Shalizi hold for \mathcal{M}_1 and \mathcal{M}_2 in the same way as the AR(1) regression example of Chandra and Bhattacharya (2020a). Thus verification of (S6) only remains, for which we begin with the following result.

Theorem 44 *The functions $\frac{1}{n} \log R_n^{(1)}(\theta)$ and $\frac{1}{n} \log R_n^{(2)}(\theta)$ are asymptotically concave in θ .*

Proof. The proof follows in the same line as that of Theorem 17 of Chandra and Bhattacharya (2020a). ■

It is also easy to see that both $h_1(\theta)$ and $h_2(\theta)$ given by (8.6.45) and (8.6.46) are convex in θ . Hence, there exist unique minimizers $\tilde{\theta}_1$ and $\tilde{\theta}_2$, respectively, of h_1 and h_2 . Theorem 45 shows consistency of the unique roots of $\frac{1}{n} \log R_n^{(1)}(\theta)$ and $\frac{1}{n} \log R_n^{(2)}(\theta)$ for $\tilde{\theta}_1$ and $\tilde{\theta}_2$, respectively.

Theorem 45 *Given any $\eta > 0$, $\frac{1}{n} \log R_n^{(1)}(\theta)$ and $\frac{1}{n} \log R_n^{(2)}(\theta)$ have their unique roots in the η -neighbourhood of $\tilde{\theta}_1$ and $\tilde{\theta}_2$, respectively, almost surely, for large n .*

Proof. See Appendix 8.A2. ■

For $i = 1, 2$, let $\tilde{\theta}_n^{(i)}$ stand for the unique maximizer of $\frac{1}{n} \log R_n^{(i)}(\theta)$. By Theorem 45

$$\tilde{\theta}_n^{(i)} \xrightarrow{a.s.} \tilde{\theta}^{(i)}, \text{ for } i = 1, 2,$$

which, in turn implies thanks to Theorem 55, that (8.6.18), and hence (S6) of Shalizi, holds for both \mathcal{M}_1 and \mathcal{M}_2 .

In other words, models \mathcal{M}_1 and \mathcal{M}_2 satisfy conditions (S1)–(S7) of Shalizi. We summarize below our results on variable selection in forward AR(1) regression framework.

Theorem 46 (FPBF consistency for \mathcal{M}_1 versus \mathcal{M}_0) *Consider the AR(1) regression models \mathcal{M}_1 and \mathcal{M}_0 given by (8.6.40) and (8.6.42). Then under assumptions (A1) and (A2),*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log FPBF^{(n)}(\mathcal{M}_1, \mathcal{M}_0) \stackrel{a.s.}{=} -h_1(\tilde{\theta}_1),$$

where h_1 is given by (8.6.45) and $\tilde{\theta}_1$ is its unique minimizer.

Theorem 47 (FPBF consistency for \mathcal{M}_2 versus \mathcal{M}_0) *Consider the AR(1) regression models \mathcal{M}_2 and \mathcal{M}_0 given by (8.6.41) and (8.6.42). Then under assumptions (A1) and (A2),*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log FPBF^{(n)}(\mathcal{M}_2, \mathcal{M}_0) \stackrel{a.s.}{=} -h_2(\tilde{\theta}_2),$$

where h_2 is given by (8.6.46) and $\tilde{\theta}_2$ is its unique minimizer.

Theorem 48 (FPBF convergence for \mathcal{M}_1 versus \mathcal{M}_2) Consider the AR(1) regression models \mathcal{M}_1 and \mathcal{M}_2 given by (8.6.40) and (8.6.41) and the true model \mathcal{M}_0 given by (8.6.42). Then under assumptions (A1) and (A2),

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log FPBF^{(n)}(\mathcal{M}_1, \mathcal{M}_2) \stackrel{a.s.}{=} - \left(h_1(\tilde{\theta}_1) - h_2(\tilde{\theta}_2) \right),$$

where h_1 and h_2 are given by (8.6.45) and (8.6.46) and $\tilde{\theta}_1$ and $\tilde{\theta}_2$ are their respective unique minimizers.

8.7 Illustrations of PBF convergence in inverse regression problems

First note that if $f(y_i|\theta, \tilde{x}_i, \mathbf{Y}^{(i-1)}, \mathcal{M})$ is bounded and continuous in (θ, \tilde{x}_i) , then in inverse regression setups, $g(y_i, \theta, \mathcal{M})$ is bounded and continuous in θ if $\pi(\tilde{x}_i|\theta, \mathcal{M})$ is bounded and continuous in (θ, \tilde{x}_i) . Here continuity of $g(\mathbf{Y}^{(i)}, \theta, \mathcal{M})$ follows by the dominated convergence theorem. Thus, whenever $f(y_i|\theta, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M}_0)$ are also bounded and continuous in θ and conditions (S1)–(S7) of Shalizi are verified, almost sure exponential convergence of IPBF also hold, provided that $h^*(\tilde{\theta})$ exists. But existence of $h^*(\tilde{\theta})$ requires existence of the limit of $n^{-1} \sum_{i=1}^n g(\mathbf{Y}^{(i)}, \tilde{\theta}, \mathcal{M})$. Although this is expected to exist, it is not straightforward to guarantee this rigorously for general regression problems.

However, in practice, simple approximations may be used. For example, if \mathcal{M} stands for simple linear regression, then let us consider a uniform prior for \tilde{x}_i on $\mathcal{X} = [-a, a]$, for some $a > 0$. Then

$$g(\mathbf{Y}^{(i)}, \theta, \mathcal{M}) = \int_{-a}^a \frac{1}{\sigma_\epsilon \sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma_\epsilon^2} (y_i - \alpha - \beta \tilde{x}_i)^2 \right\} d\tilde{x}_i \xrightarrow{a.s.} |\beta|^{-1}, \text{ as } a \rightarrow \infty.$$

Thus for sufficiently large a , $g(\mathbf{Y}^{(i)}, \tilde{\theta}, \mathcal{M})$ can be approximated by $|\tilde{\beta}|^{-1}$, which is independent of i . Thus, for large enough a , the limit of $n^{-1} \sum_{i=1}^n \log g(\mathbf{Y}^{(i)}, \tilde{\theta}, \mathcal{M})$ can be

approximated by $|\beta|^{-1}$. But in general non-linear regression, such simple approximations are not available.

The setup where $\mathbf{y}_i = \{y_{i1}, \dots, y_{im}\}$, is far more flexible in this regard. Let us illustrate this with respect to the models \mathcal{M}_0 , \mathcal{M}_1 and \mathcal{M}_2 considered in Section 8.6. Assuming invertibility of η_0 in addition to continuity, we assume the prior

$$\pi(\tilde{x}_i | \eta_0, \mathcal{M}_0) \equiv U\left(B_{im}^{(0)}(\eta_0)\right) \quad (8.7.1)$$

under model \mathcal{M}_0 , where

$$B_{im}^{(0)}(\eta_0) = \left\{ x : \eta_0(x) \in \left[\bar{y}_i - \frac{cs_i}{\sqrt{m}}, \bar{y}_i + \frac{cs_i}{\sqrt{m}} \right] \right\}. \quad (8.7.2)$$

In the case of the linear regression model \mathcal{M}_1 , we set

$$\pi(\tilde{x}_i | \theta, \mathcal{M}_1) \equiv U\left(B_{im}^{(1)}(\theta)\right) \quad (8.7.3)$$

where

$$B_{im}^{(1)}(\theta) = \left[\frac{\bar{y}_i - \alpha}{\beta} - \frac{cs_i}{|\beta|\sqrt{m}}, \frac{\bar{y}_i - \alpha}{\beta} + \frac{cs_i}{|\beta|\sqrt{m}} \right]. \quad (8.7.4)$$

For the quadratic model \mathcal{M}_2 , note that even if the true model is quadratic, then it is not one-to-one. Hence the general form of the prior considered in Section 6.4 is not applicable here. In this case, we propose the following prior for \tilde{x}_i under the quadratic model \mathcal{M}_2 :

$$\pi(\tilde{x}_i | \theta, \mathcal{M}_2) \equiv U\left(B_{im}^{(2)}(\theta)\right) \quad (8.7.5)$$

where

$$B_{im}^{(2)}(\theta) = \left[\frac{\bar{y}_i - \alpha - \beta_2 x_i^2}{\beta_1} - \frac{cs_i}{|\beta_1|\sqrt{m}}, \frac{\bar{y}_i - \alpha - \beta_2 x_i^2}{\beta_1} + \frac{cs_i}{|\beta_1|\sqrt{m}} \right]. \quad (8.7.6)$$

Note that the prior depends upon x_i itself, which is the truth in this case. It is unusual in

Bayesian inference to make the prior depend upon the truth. Indeed, the true parameter is always unknown; had it been known, then one would give full prior probability to the true parameter. In our case x_i is actually known but a prior is needed for \tilde{x}_i for the sake of cross-validation. Moreover, the prior does not consider \tilde{x}_i to be known as long as the sample sizes n and m remain finite and θ is unknown or takes false values. The prior has substantial variance in these cases. Hence, although unusual, such a prior on \tilde{x}_i is not untenable for inverse cross-validation.

Now observe that $\tilde{\theta}_1$ and $\tilde{\theta}_2$ associated with models \mathcal{M}_1 and \mathcal{M}_2 remain the same as those in Section 8.6. Also note that when the true model is \mathcal{M}_0 and when $\tilde{\theta}_1$ is associated with \mathcal{M}_1 , then

$$B_{im}^{(1)}(\tilde{\theta}_1) \xrightarrow{a.s.} \{x_{i1}^*\}, \text{ as } m \rightarrow \infty,$$

where

$$x_{i1}^* = \frac{\eta_0(x_i) - \tilde{\alpha}}{\tilde{\beta}}. \quad (8.7.7)$$

Similarly, when the true model is \mathcal{M}_0 and when $\tilde{\theta}_2$ is associated with \mathcal{M}_2 , then

$$B_{im}^{(2)}(\tilde{\theta}_2) \xrightarrow{a.s.} \{x_{i2}^*\}, \text{ as } m \rightarrow \infty,$$

where

$$x_{i2}^* = \frac{\eta_0(x_i) - \tilde{\alpha} - \tilde{\beta}_2 x_i^2}{\tilde{\beta}_1}. \quad (8.7.8)$$

Let \mathcal{X}_1 and \mathcal{X}_2 be the co-domains of $\frac{\eta_0(x) - \tilde{\alpha}}{\tilde{\beta}}$ and $\frac{\eta_0(x) - \tilde{\alpha} - \tilde{\beta}_2 x^2}{\tilde{\beta}_1}$ for $x \in \mathcal{X}$. Since these functions are both continuous in x , the asymptotic calculations of $\frac{1}{n} \log \prod_{i=1}^n f(y_{ik} | \tilde{\theta}_1, x_{i1}^*, \mathcal{M}_1)$ and $\frac{1}{n} \log \prod_{i=1}^n f(y_{ik} | \tilde{\theta}_1, x_{i2}^*, \mathcal{M}_2)$ remain the same as $\frac{1}{n} \log \prod_{i=1}^n f(y_i | \tilde{\theta}_1, x_i, \mathcal{M}_1)$ and $\frac{1}{n} \log \prod_{i=1}^n f(y_i | \tilde{\theta}_1, x_i, \mathcal{M}_2)$, respectively, detailed in Section 8.6, with \mathcal{X} replaced with \mathcal{X}_1 and \mathcal{X}_2 , respectively. Hence, the final asymptotic results for IPBF remain the same for FPBF with respect to the models \mathcal{M}_0 , \mathcal{M}_1 and \mathcal{M}_2 , only with \mathcal{X} replaced by \mathcal{X}_1 and

\mathcal{X}_2 , respectively. Also note that here the cross-validation posterior for \mathcal{M}_0 is given by

$$\pi(y_{ik}|\mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}) = \int_{\mathcal{X}} f(y_{ik}|\tilde{x}_i, \mathcal{M}_0) d\pi(\tilde{x}_i|\mathcal{M}_0) \xrightarrow{a.s.} f(y_{ik}|x_i), \text{ as } m \rightarrow \infty,$$

since $B_{im}^{(0)}(\eta_0) \xrightarrow{a.s.} \{x_i\}$, as $m \rightarrow \infty$. Hence, the final asymptotic results do not depend upon whether or not x_i is considered known or the prior $\pi(\tilde{x}_i|\eta_0, \mathcal{M}_0)$ is used treating it as unknown, when cross-validating for \mathcal{M}_0 . Appealing to Theorem 37, Remark 38 and Theorem 39 we thus summarize our results for IPBF concerning \mathcal{M}_0 , \mathcal{M}_1 and \mathcal{M}_2 as follows.

Theorem 49 (IPBF convergence for linear regression) *Assume the setup where data $\{y_{ij}; i = 1, \dots, n; j = 1, \dots, m\}$ are available. In this setup consider the linear regression model \mathcal{M}_1 given by (8.6.1) and the true, non-linear model \mathcal{M}_0 given by (8.6.2). Let the parameter space Θ associated with model \mathcal{M}_1 be $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$, and let the covariate space \mathcal{X} be compact. Assume the priors (8.7.1) and (8.7.3) on \tilde{x}_i under the models \mathcal{M}_0 and \mathcal{M}_1 , respectively. Then*

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{1}{n} \log IPBF^{(n,m,k)}(\mathcal{M}_1, \mathcal{M}_0) \xrightarrow{a.s.} -h_1(\tilde{\theta}_1),$$

where for $\theta \in \Theta$, $h_1(\theta)$ is given by (8.6.11) with \mathcal{X} replaced by \mathcal{X}_1 , and $\tilde{\theta}_1 = (\tilde{\alpha}, \tilde{\beta}, \tilde{\sigma}_\epsilon^2)$, where $\tilde{\alpha}$, $\tilde{\beta}$ and $\tilde{\sigma}_\epsilon^2$ are given by (8.6.24), (8.6.23) and (8.6.25), respectively. The result remains unchanged if x_i is treated as known for cross-validation with respect to \mathcal{M}_0 .

Theorem 50 (IPBF convergence for quadratic regression) *Assume the setup where data $\{y_{ij}; i = 1, \dots, n; j = 1, \dots, m\}$ are available. In this setup consider the quadratic regression model \mathcal{M}_2 given by (8.6.30) and the true, non-linear model \mathcal{M}_0 given by (8.6.2). Let the parameter space Θ associated with model \mathcal{M}_2 be $\mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$, and let the covariate space \mathcal{X} be compact. Also assume that $x_i; i \geq 1$ are all distinct. Assume*

the priors (8.7.1) and (8.7.5) on \tilde{x}_i under the models \mathcal{M}_0 and \mathcal{M}_2 , respectively. Then

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{1}{n} \log IPBF^{(n,m,k)}(\mathcal{M}_2, \mathcal{M}_0) \stackrel{a.s.}{=} -h_2(\tilde{\theta}_2),$$

where for $\theta \in \Theta$, $h_2(\theta)$ is given by (8.6.31) with \mathcal{X} replaced by \mathcal{X}_2 , and $\tilde{\theta}_2 = (\tilde{\alpha}, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\sigma}_\epsilon^2)$, where $\tilde{\alpha}$, $\tilde{\beta}_1$, $\tilde{\beta}_2$ and $\tilde{\sigma}_\epsilon^2$ are given by (8.6.32), (8.6.33) and (8.6.34). The result remains unchanged if x_i is treated as known for cross-validation with respect to \mathcal{M}_0 .

Theorem 51 (Comparison between linear and quadratic regressions) Assume the setup where data $\{y_{ij}; i = 1, \dots, n; j = 1, \dots, m\}$ are available. Let the true model be given by \mathcal{M}_0 formulated in (8.6.2). Assuming that the covariate observations $x_i; i \geq 1$ are all distinct and that the covariate space \mathcal{X} is compact, consider comparison of the linear and quadratic regression models \mathcal{M}_1 and \mathcal{M}_2 given by (8.6.1) and (8.6.30), respectively, using IPBF. Assume the priors (8.7.1), (8.7.3) and (8.7.5) on \tilde{x}_i under the models \mathcal{M}_0 , \mathcal{M}_1 and \mathcal{M}_2 , respectively. Then,

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{1}{n} \log IPBF^{(n,m,k)}(\mathcal{M}_1, \mathcal{M}_2) \stackrel{a.s.}{=} -\left(h_1(\tilde{\theta}_1) - h_2(\tilde{\theta}_2)\right)$$

where $h_1(\tilde{\theta}_1)$ and $h_2(\tilde{\theta}_2)$ are the same as in Theorems 49 and 50, with \mathcal{X} replaced by \mathcal{X}_1 and \mathcal{X}_2 , respectively. The result remains unchanged if \tilde{x}_i is treated as known for cross-validation with respect to \mathcal{M}_0 .

8.7.1 IPBF asymptotics for variable selection in AR(1)

Now let us reconsider the AR(1) regression setup described by the competing models \mathcal{M}_1 (8.6.40), \mathcal{M}_2 (8.6.41) and the true model \mathcal{M}_0 (8.6.42), along with assumptions (A1) and (A2). But now we reformulate the models as follows to suit the second setup of inverse regression.

$$y_{tj} = \rho_1 y_{t-1,j} + \beta_1 x_t + \epsilon_{1t,j}; \quad t = 1, \dots, n; \quad j = 1, \dots, m, \tag{8.7.9}$$

where $y_{0j} \equiv 0$ for $j = 1, \dots, m$ and $\epsilon_{1t,j} \stackrel{iid}{\sim} N(0, \sigma_1^2)$, for $t = 1, \dots, n$ and $j = 1, \dots, m$.

Similarly, \mathcal{M}_2 is now given by

$$y_{tj} = \rho_2 y_{t-1,j} + \beta_2 z_t + \epsilon_{2t,j}; \quad t = 1, \dots, n; \quad j = 1, \dots, m, \quad (8.7.10)$$

where $\epsilon_{2t,j} \stackrel{iid}{\sim} N(0, \sigma_2^2)$, for $t = 1, \dots, n$ and $j = 1, \dots, m$.

The true model \mathcal{M}_0 be given by

$$y_{tj} = \rho_0 y_{t-1,j} + \beta_0(x_t + z_t) + \epsilon_{0t,j}; \quad t = 1, \dots, n; \quad j = 1, \dots, m, \quad (8.7.11)$$

where $|\rho_0| < 1$ and $\epsilon_{0t,j} \stackrel{iid}{\sim} N(0, \sigma_0^2)$, for $t = 1, \dots, n$ and $j = 1, \dots, m$.

For $t = 1, \dots, n$, let $\bar{y}_t = \frac{\sum_{j=1}^m y_{tj}}{m}$ and $s_t^2(\rho) = \frac{1}{m} [(y_{tj} - \bar{y}_t) - \rho(y_{t-1,j} - \bar{y}_{t-1})]^2$. We consider the following priors for \tilde{x}_t and \tilde{z}_t associated with \mathcal{M}_1 and \mathcal{M}_2 :

$$\pi(\tilde{x}_t | \theta_1, \mathcal{M}_1) \equiv U\left(B_{tm}^{(1)}(\theta_1)\right); \quad (8.7.12)$$

$$\pi(\tilde{z}_t | \theta_2, \mathcal{M}_2) \equiv U\left(B_{tm}^{(2)}(\theta_2)\right), \quad (8.7.13)$$

where

$$B_{tm}^{(1)}(\theta_1) = \left[\frac{\bar{y}_t - \rho_1 \bar{y}_{t-1}}{\beta_1} - \frac{cst(\rho_1)}{|\beta_1| \sqrt{m}}, \frac{\bar{y}_t - \rho_1 \bar{y}_{t-1}}{\beta_1} + \frac{cst(\rho_1)}{|\beta_1| \sqrt{m}} \right]; \quad (8.7.14)$$

$$B_{tm}^{(2)}(\theta_2) = \left[\frac{\bar{y}_t - \rho_2 \bar{y}_{t-1}}{\beta_2} - \frac{cst(\rho_2)}{|\beta_2| \sqrt{m}}, \frac{\bar{y}_t - \rho_2 \bar{y}_{t-1}}{\beta_2} + \frac{cst(\rho_2)}{|\beta_2| \sqrt{m}} \right]. \quad (8.7.15)$$

Note that

$$B_{tm}^{(1)}(\tilde{\theta}_1) \xrightarrow{a.s.} \{x_t^*\} \text{ as } m \rightarrow \infty; \quad (8.7.16)$$

$$B_{tm}^{(2)}(\tilde{\theta}_2) \xrightarrow{a.s.} \{z_t^*\} \text{ as } m \rightarrow \infty, \quad (8.7.17)$$

where

$$x_t^* = \frac{\beta_0 \sum_{k=1}^t \rho^{t-k} x_k - \beta_0 \tilde{\rho}_1 \sum_{k=1}^{t-1} \rho_0^{t-k} x_k}{\tilde{\beta}_1}; \quad (8.7.18)$$

$$z_t^* = \frac{\beta_0 \sum_{k=1}^t \rho^{t-k} z_k - \beta_0 \tilde{\rho}_2 \sum_{k=1}^{t-1} \rho_0^{t-k} z_k}{\tilde{\beta}_2}. \quad (8.7.19)$$

Direct calculations reveal that

$$\frac{1}{n} \sum_{t=1}^n x_t^* \rightarrow 0; \quad \frac{1}{n} \sum_{t=1}^n x_t^{*2} \rightarrow \sigma_{x^*}^2 = \sigma_x^2 \frac{\beta_0^2 (1 - \tilde{\rho}_1)^2}{\tilde{\beta}_1^2 (1 - \rho_0^2)}, \text{ as } n \rightarrow \infty; \quad (8.7.20)$$

$$\frac{1}{n} \sum_{t=1}^n z_t^* \rightarrow 0; \quad \frac{1}{n} \sum_{t=1}^n z_t^{*2} \rightarrow \sigma_{z^*}^2 = \sigma_z^2 \frac{\beta_0^2 (1 - \tilde{\rho}_2)^2}{\tilde{\beta}_2^2 (1 - \rho_0^2)}, \text{ as } n \rightarrow \infty. \quad (8.7.21)$$

Hence, for the final IPBF calculations associated with h_1 and h_2 for this example, we need to replace x_t , z_t , σ_x^2 and σ_z^2 in (A1) with x_t^* , z_t^* , $\sigma_{x^*}^2$ and $\sigma_{z^*}^2$, respectively, for models \mathcal{M}_1 and \mathcal{M}_2 . In this regard, let

$$\begin{aligned} h_1^*(\theta) = & \log \left(\frac{\sigma}{\sigma_0} \right) + \left(\frac{1}{2\sigma^2} - \frac{1}{2\sigma_0^2} \right) \left(\frac{\sigma_0^2}{1 - \rho_0^2} + \frac{\beta_0^2 \sigma_{x+z}^2}{1 - \rho_0^2} \right) + \left(\frac{\rho^2}{2\sigma^2} - \frac{\rho_0^2}{2\sigma_0^2} \right) \left(\frac{\sigma_0^2}{1 - \rho_0^2} + \frac{\beta_0^2 \sigma_{x+z}^2}{1 - \rho_0^2} \right) \\ & + \frac{1}{2\sigma^2} \beta^2 \sigma_{x+z}^2 - \frac{1}{2\sigma_0^2} \beta_0^2 \sigma_{x+z}^2 - \left(\frac{\rho}{\sigma^2} - \frac{\rho_0}{\sigma_0^2} \right) \left(\frac{\rho_0 \sigma_0^2}{1 - \rho_0^2} + \frac{\rho_0 \beta_0^2 \sigma_{x+z}^2}{1 - \rho_0^2} \right) - \left(\frac{\beta}{\sigma^2} - \frac{\beta_0}{\sigma_0^2} \right) \sigma_{x+z}^2 \beta_0 \\ & + \frac{\sigma_z^2 \beta (\beta_0 - \beta)}{\sigma^2} + \frac{\beta^2}{2\sigma^2} \left(\sigma_{x+z}^2 + \sigma_{x^*}^2 - \frac{2\beta_0 \sigma_x^2}{\tilde{\beta}_1} \right). \end{aligned} \quad (8.7.22)$$

and

$$\begin{aligned} h_2^*(\theta) = & \log \left(\frac{\sigma}{\sigma_0} \right) + \left(\frac{1}{2\sigma^2} - \frac{1}{2\sigma_0^2} \right) \left(\frac{\sigma_0^2}{1 - \rho_0^2} + \frac{\beta_0^2 \sigma_{x+z}^2}{1 - \rho_0^2} \right) + \left(\frac{\rho^2}{2\sigma^2} - \frac{\rho_0^2}{2\sigma_0^2} \right) \left(\frac{\sigma_0^2}{1 - \rho_0^2} + \frac{\beta_0^2 \sigma_{x+z}^2}{1 - \rho_0^2} \right) \\ & + \frac{1}{2\sigma^2} \beta^2 \sigma_{x+z}^2 - \frac{1}{2\sigma_0^2} \beta_0^2 \sigma_{x+z}^2 - \left(\frac{\rho}{\sigma^2} - \frac{\rho_0}{\sigma_0^2} \right) \left(\frac{\rho_0 \sigma_0^2}{1 - \rho_0^2} + \frac{\rho_0 \beta_0^2 \sigma_{x+z}^2}{1 - \rho_0^2} \right) - \left(\frac{\beta}{\sigma^2} - \frac{\beta_0}{\sigma_0^2} \right) \sigma_{x+z}^2 \beta_0 \\ & + \frac{\sigma_x^2 \beta (\beta_0 - \beta)}{\sigma^2} + \frac{\beta^2}{2\sigma^2} \left(\sigma_{x+z}^2 + \sigma_{z^*}^2 - \frac{2\beta_0 \sigma_z^2}{\tilde{\beta}_2} \right). \end{aligned} \quad (8.7.23)$$

If cross-validation is considered with respect to the true model \mathcal{M}_0 with a prior on

the covariates, then since x_t and z_t are not separately identifiable in \mathcal{M}_0 , let $u_t = x_t + z_t$ and consider a prior on \tilde{u}_t as follows:

$$\pi(\tilde{u}_t | \theta_0, \mathcal{M}_0) \equiv U\left(B_{tm}^{(0)}(\theta_0)\right), \quad (8.7.24)$$

where

$$B_{tm}^{(0)}(\theta_0) = \left[\frac{\bar{y}_t - \rho_0 \bar{y}_{t-1}}{\beta_0} - \frac{c s_t(\rho_0)}{|\beta_0| \sqrt{m}}, \frac{\bar{y}_t - \rho_1 \bar{y}_{t-1}}{\beta_0} + \frac{c s_t(\rho_0)}{|\beta_0| \sqrt{m}} \right]. \quad (8.7.25)$$

Note that $B_{tm}^{(0)}(\theta_0) \xrightarrow{a.s.} \{u_t\}$, as $m \rightarrow \infty$. Let $\mathbf{U}_{n,-t} = \{u_1, \dots, u_n\} \setminus \{u_t\}$. As before, it follows that $\pi(y_{tk} | \mathbf{Y}_{nm,-t}, \mathbf{U}_{n,-t}) \xrightarrow{a.s.} f(y_{tk} | u_t, y_{t-1,k})$, as $m \rightarrow \infty$. Hence, the final asymptotic results do not depend upon whether or not u_t is considered known or the prior (8.7.24) is used for \tilde{u}_t treating u_t it as unknown, when cross-validating for the true model \mathcal{M}_0 .

We summarize our results on variable selection in the inverse AR(1) regression framework as follows.

Theorem 52 (IPBF consistency for \mathcal{M}_1 versus \mathcal{M}_0) *Consider comparing model \mathcal{M}_1 (8.7.9) against the true model \mathcal{M}_0 (8.7.11). Assume the priors (8.7.12) and (8.7.24) on \tilde{x}_t and \tilde{u}_t under the models \mathcal{M}_1 and \mathcal{M}_0 , respectively. Then*

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{1}{n} \log IPBF^{(n,m,k)}(\mathcal{M}_1, \mathcal{M}_0) \stackrel{a.s.}{=} -h_1^*(\tilde{\theta}_1),$$

where for $\theta \in \Theta_1$, $h_1^*(\theta)$ is given by (8.7.22), and $\tilde{\theta}_1$ is the unique minimizer of h_1 given by (8.6.45). The result remains unchanged if u_t is treated as known for cross-validation with respect to \mathcal{M}_0 .

Theorem 53 (IPBF consistency for \mathcal{M}_2 versus \mathcal{M}_0) *Consider comparing model \mathcal{M}_2 (8.7.10) against the true model \mathcal{M}_0 (8.7.11). Assume the priors (8.7.13) and*

(8.7.24) on \tilde{z}_t and \tilde{u}_t under the models \mathcal{M}_2 and \mathcal{M}_0 , respectively. Then

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{1}{n} \log IPBF^{(n,m,k)}(\mathcal{M}_1, \mathcal{M}_0) \stackrel{a.s.}{=} -h_2^*(\tilde{\theta}_2),$$

where for $\theta \in \Theta_2$, $h_2^*(\theta)$ is given by (8.7.23), and $\tilde{\theta}_2$ is the unique minimizer of h_2 given by (8.6.46). The result remains unchanged if u_t is treated as known for cross-validation with respect to \mathcal{M}_0 .

Theorem 54 (IPBF convergence for \mathcal{M}_1 versus \mathcal{M}_2) Consider comparing models \mathcal{M}_1 (8.7.9) against model \mathcal{M}_2 (8.7.10). Assume the priors (8.7.12) and (8.7.13) on \tilde{x}_t and \tilde{z}_t under the models \mathcal{M}_1 and \mathcal{M}_2 , respectively. Then

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{1}{n} \log IPBF^{(n,m,k)}(\mathcal{M}_1, \mathcal{M}_2) \stackrel{a.s.}{=} - \left(h_1^*(\tilde{\theta}_1) - h_2^*(\tilde{\theta}_2) \right),$$

where h_1^* and h_2^* are given by (8.7.22) and (8.7.23). In the above, $\tilde{\theta}_1$ and $\tilde{\theta}_2$ are the unique minimizers of h_1 of h_2 given by (8.6.45) and (8.6.46), respectively. The result remains unchanged if u_t is treated as known for cross-validation with respect to \mathcal{M}_0 .

8.7.2 Discussion of FPBF and IPBF convergence for nonparametric regression models

In Chapter 4 investigate posterior convergence for Gaussian and general stochastic process regression under suitable assumptions while posterior convergence for binary and Poisson nonparametric regression based on Gaussian process modeling of the regression function are addressed in Chapter 5. In all these nonparametric setups, we verified assumptions (S1)–(S7) of Shalizi. Here it is important to point out that Theorem 55 used to verify assumption (S6) of Shalizi in our parametric setups, is not valid in infinite-dimensional nonparametric models since without further assumptions on model sparsity, $\tilde{\theta}_n^*$ can not converge to $\tilde{\theta}$. That is, condition (ii) of Theorem 55 does not hold in general for nonparametric models. Moreover, enforcing sparsity conditions to general

stochastic processes, such as Gaussian processes, need not be desirable. Chapter 4 (see also Chapter 5) introduces a general sufficient condition for verification of (S6) of Shalizi, which is appropriate for nonparametric models, and we use that condition for our purposes.

The point of the above discussion is that assumptions (S1)–(S7) are already verified in Chapters 4 and 5 for nonparametric Bayesian regression models, and since boundedness and continuity of $f(y_i|\theta, \mathcal{M})$ also hold for such models \mathcal{M} , our asymptotic results on almost sure exponential convergence of FPBF and IPBF are directly applicable to such models. For IPBF convergence in nonparametric situations, the priors for \tilde{x}_i proposed in Section 6.4.1 for nonparametric cases (ii)–(iv) are appropriate.

Note that parametric and nonparametric models can also be compared asymptotically using our FPBF and IPBF theory.

8.8 Simulation experiments

So far we have investigated large sample properties of FPBF and IPBF. However, for all practical purposes it is important to provide insights into small sample behaviours of such versions of pseudo-Bayes factor. In this section we undertake such small sample study with the help of simulation experiments. Specifically, we set $n = m = 10$ and generate data from relevant Poisson distribution with the log-linear link function and consider modeling the data with Poisson and geometric distributions with log, logit and probit links for linear models as well as nonparametric regression modeled by Gaussian process having linear mean function and squared exponential covariance. We also consider variable selection in these setups with respect to two different covariates. We report both FPBF and IPBF results for the experiments. Details follow.

8.8.1 Poisson versus geometric linear and nonparametric regression models when the true model is Poisson linear regression

True distribution

Let us first consider the case where the true data-generating distribution is $y_{ij} \sim Poisson(\lambda(x_i))$, with $\lambda(x) = \exp(\alpha_0 + \beta_0 x)$. We generate the data by simulating $\alpha_0 \sim U(-1, 1)$, $\beta_0 \sim U(-1, 1)$ and $x_i \sim U(-1, 1)$; $i = 1, \dots, n$, and then finally simulating $y_{ij} \sim Poisson(\lambda(x_i))$; $j = 1, \dots, m$, $i = 1, \dots, n$.

To model the data generated from the true distribution, we consider both Poisson and geometric distributions and both linear and Gaussian process based nonparametric regression for such models. Let us begin with the Poisson setup.

8.8.2 Competing forward and inverse Poisson regression models

Forward Poisson linear regression model

In this setup we model the data as follows: $y_{ij} \sim Poisson(\lambda(x_i))$, with $\lambda(x) = \exp(\alpha + \beta x)$, and set the prior $\pi(\alpha, \beta) = 1$, for $-\infty < \alpha, \beta < \infty$. For the forward setup, this completes the model and prior specifications. Denoting this by model \mathcal{M} , we compute the forward cross-validation posterior of the form

$$\pi(y_{i1} | \mathbf{Y}_{n,-i}, \mathbf{X}_n, \mathcal{M}) = \int_{\Theta} f(y_{i1} | \theta, x_i, \mathbf{Y}_1^{(i-1)}, \mathcal{M}) d\pi(\theta | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M}), \quad (8.8.1)$$

by taking Monte Carlo averages of $f(y_{i1} | \theta, x_i, \mathbf{Y}^{(i-1)}, \mathcal{M})$ over realizations of θ from $\pi(\theta | \mathbf{Y}_{n,-i}, \mathbf{X}_{n,-i}, \mathcal{M})$. In our case this is the Monte Carlo average of the relevant Poisson probability of y_{i1} given x_i over realizations of $\theta = (\alpha, \beta)$. Samples of θ are obtained approximately from the posterior distribution of $\pi(\theta | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i})$ by first generating realizations from the “importance sampling density” $\pi(\theta | \mathbf{Y}_{nm}, \mathbf{X}_n)$ using transformation based Markov chain Monte Carlo (TMCMC) ([Dutta and Bhattacharya \(2014\)](#)) and then

re-using the realizations with importance weights to obtain the desired Monte Carlo averages. The rationale behind the choice of the full posterior $\pi(\theta|\mathbf{Y}_{nm}, \mathbf{X}_n)$ associated with the full data set as the importance sampling density is that it is not significantly different from the posterior $\pi(\theta|\mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i})$ associated with leaving out a single data point. This choice is also quite popular in the literature; see, for example, Gelfand (1996). In our examples, we generate 30,000 TMCMC samples from $\pi(\theta|\mathbf{Y}_{nm}, \mathbf{X}_n)$ of which we discard the first 10,000 as burn-in, and re-sample 1000 θ -realizations without replacement from the remaining 20,000 realizations. We re-use each re-sampled θ -value 100 times and compute the Monte Carlo average over such $1000 \times 100 = 100,000$ realizations. The re-use of each re-sampled θ -value corresponds to importance re-sampling MCMC (IRMCMC) of Bhattacharya and Haslett (2007). Although IRMCMC is meant for cross-validation in inverse problems, the idea carries over to forward problems as well. We finally compute $\frac{1}{n} \sum_{i=1}^n \log \pi(y_{i1}|\mathbf{Y}_{nm,-i}, \mathbf{X}_n, \mathcal{M})$ for model \mathcal{M} .

Inverse Poisson linear regression model

With the same Poisson linear regression model as in the forward case, we now put a prior on \tilde{x}_i corresponding to x_i . In our case, it follows from Section 6.4.1 that $\pi(\tilde{x}_i|\alpha, \beta) \equiv U(a, b)$, where

$$a = \min \left\{ \beta^{-1} \left(\log \left(\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}} \right) - \alpha \right), \beta^{-1} \left(\log \left(\bar{y}_i + \frac{c_2 s_i}{\sqrt{m}} \right) - \alpha \right) \right\} \quad (8.8.2)$$

and

$$b = \max \left\{ \beta^{-1} \left(\log \left(\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}} \right) - \alpha \right), \beta^{-1} \left(\log \left(\bar{y}_i + \frac{c_2 s_i}{\sqrt{m}} \right) - \alpha \right) \right\}. \quad (8.8.3)$$

We set $c_1 = 1$ and $c_2 = 100$, for ensuring positive value of $\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}}$ (so that logarithm of this quantity is well-defined) and a reasonably large support of the prior for \tilde{x}_i . We then

compute

$$\pi(y_{i1}|\mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M}) = \int_{\mathcal{X}} \int_{\Theta} f(y_{i1}|\theta, \tilde{x}_i, \mathbf{Y}_1^{(i-1)}, \mathcal{M}) d\pi(\tilde{x}_i, \theta | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M})$$

by Monte Carlo averaging of the relevant Poisson probability of y_{i1} over realizations of $(\tilde{x}_i, \theta) = (\tilde{x}_i, \alpha, \beta)$ generated from $\pi(\tilde{x}_i, \theta | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M})$. Since it follows from (8.2.6) that $\pi(\tilde{x}_i, \theta | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M}) = \pi(\tilde{x}_i | \theta, \mathcal{M})\pi(\theta | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M})$, and since realizations of θ from $\pi(\theta | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M})$ are already available in the forward context, we simply generate \tilde{x}_i given θ from the prior for \tilde{x}_i to obtain realizations from $\pi(\tilde{x}_i, \theta | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M})$. Note that for different i , only sub-samples of θ of size 1000 from the original sample of size 20,000 from the full posterior of θ are available, and each θ is repeated 100 times. However, realizations of \tilde{x}_i are all distinct in spite of repetitions of θ -values.

Once for each $i = 1, \dots, n$, the Monte Carlo estimates of $\pi(y_{i1} | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M})$ are available, we finally obtain the estimate of $\frac{1}{n} \sum_{i=1}^n \log \pi(y_{i1} | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M})$ using the individual Monte Carlo estimates.

Forward Poisson nonparametric regression model

We now consider the case where $y_{ij} \sim \text{Poisson}(\lambda(x_i))$, where $\lambda(x) = \exp(\eta(x))$, where $\eta(\cdot)$ is a Gaussian process with mean function $\mu(x) = \alpha + \beta x$ and covariance $\text{Cov}(\eta(x_1), \eta(x_2)) = \sigma^2 \exp\{-(x_1 - x_2)^2\}$, where σ is unknown. For our convenience, we reparameterize σ^2 as $\exp(\omega)$, where $-\infty < \omega < \infty$. For the prior on the parameters, we set $\pi(\alpha, \beta, \omega) = 1$, for $-\infty < \alpha, \beta, \omega < \infty$.

In the inverse case, for the reason of prior specification, we linearize $\eta(\tilde{x}_i)$ as $\alpha + \beta \tilde{x}_i$; see Section 8.8.2. Hence, for comparability with the inverse counterpart, we set $\eta(x_i) = \alpha + \beta x_i$. Thus, in the forward case, $\theta = (\alpha, \beta, \eta(x_1), \dots, \eta(x_{i-1}), \eta(x_{i+1}), \dots, \eta(x_n), \omega)$. We obtain $\frac{1}{n} \sum_{i=1}^n \log \pi(y_{i1} | \mathbf{Y}_{nm,-i}, \mathbf{X}_n, \mathcal{M})$ using the same method of Monte Carlo averaging described in Section 8.8.2, where θ is again first generated using TMCMC

from the full posterior of θ by discarding the first 10,000 iterations and retaining the next 20,000 for inference, which are re-used to approximate the desired posteriors $\pi(\theta|\mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M})$. As before, we obtain Monte Carlo averages over 100,000 realizations of θ .

Inverse Poisson nonparametric regression model

The model in this case remains the same as that in Section 8.8.2, but now a prior on \tilde{x}_i is needed. However, note that the prior for \tilde{x}_i , which is uniform on $B_{im}(\eta) = \left\{x : \eta(x) \in \log \left\{ \left[\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}}, \bar{y}_i + \frac{c_2 s_i}{\sqrt{m}} \right] \right\}\right\}$, does not have a closed form, since the form of $\eta(x)$ is unknown. However, if m is large, the interval $\log \left\{ \left[\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}}, \bar{y}_i + \frac{c_2 s_i}{\sqrt{m}} \right] \right\}$ is small, and $\eta(x)$ falling in this small interval can be reasonably well-approximated by a straight line. Hence, we set $\eta(x) = \mu(x) = \alpha + \beta x$, for $\eta(x)$ falling in this interval. Thus it follows that $\pi(\tilde{x}_i|\eta) \equiv U(a, b)$, where a and b are given by (8.8.2) and (8.8.3), respectively. Hence, we obtain the same prior for \tilde{x}_i as in the case of linear Poisson regression described in Section 8.8.2. As before we set $c_1 = 1$ and $c_2 = 100$.

The method for obtaining $\frac{1}{n} \sum_{i=1}^n \log \pi(y_{i1}|\mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M})$ remains the same as discussed in Section 8.8.2.

8.8.3 Competing forward and inverse geometric regression models

We also report results of our simulation experiments where data generated from Poisson linear regression is modeled by geometric regression models of the form

$$f(y_{ij}|\theta, x_i) = (1 - p(x_i))^{y_{ij}} p(x_i), \quad (8.8.4)$$

where $p(x_i)$ is modeled as logit or probit linear or nonparametric regression. In other words, we consider the following possibilities of modeling $p(x)$:

$$\log \left(\frac{p(x)}{1 - p(x)} \right) = \alpha + \beta x; \quad \log \left(\frac{p(x)}{1 - p(x)} \right) = \eta(x); \\ p(x) = \Phi(\alpha + \beta x); \quad p(x) = \Phi(\eta(x)),$$

where Φ is the cumulative distribution function of the standard normal distribution. In the above, η is again modeled by a Gaussian process with mean function $\mu(x) = \alpha + \beta x$ and covariance function given by $Cov(\eta(x_1), \eta(x_2)) = \sigma^2 \exp\{-(x_1 - x_2)^2\}$. We again set $\sigma^2 = \exp(\omega)$, where $-\infty < \omega < \infty$, and consider the prior $\pi(\alpha, \beta, \omega) = 1$ for $-\infty < \alpha, \beta, \omega < \infty$.

In the inverse setup we assign prior on \tilde{x}_i such that the mean of the geometric distribution, namely, $\frac{1-p(x)}{p(x)}$, lies in $\left[\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}}, \bar{y}_i + \frac{c_2 s_i}{\sqrt{m}} \right]$. Using the same principles as before it follows that for the logit link, either for linear or Gaussian process regression, the prior for \tilde{x}_i is $U(a_1, b_1)$, where

$$a_1 = \min \left\{ -\beta^{-1} \left(\log \left(\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}} \right) + \alpha \right), -\beta^{-1} \left(\log \left(\bar{y}_i + \frac{c_2 s_i}{\sqrt{m}} \right) + \alpha \right) \right\} \quad (8.8.5)$$

and

$$b_1 = \max \left\{ -\beta^{-1} \left(\log \left(\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}} \right) + \alpha \right), -\beta^{-1} \left(\log \left(\bar{y}_i + \frac{c_2 s_i}{\sqrt{m}} \right) + \alpha \right) \right\}. \quad (8.8.6)$$

We set $c_1 = 1$ and $c_2 = 100$, as before.

For geometric probit regression, first let $\ell_{im} = \bar{y}_i - \frac{c_1 s_i}{\sqrt{m}}$ and $u_{im} = \bar{y}_i + \frac{c_2 s_i}{\sqrt{m}}$. Let

$$a_2 = \min \left\{ \frac{\Phi^{-1} \left(\frac{1}{u_{im}+1} \right) - \alpha}{\beta}, \frac{\Phi^{-1} \left(\frac{1}{\ell_{im}+1} \right) - \alpha}{\beta} \right\}; \quad (8.8.7)$$

$$b_2 = \max \left\{ \frac{\Phi^{-1} \left(\frac{1}{u_{im}+1} \right) - \alpha}{\beta}, \frac{\Phi^{-1} \left(\frac{1}{\ell_{im}+1} \right) - \alpha}{\beta} \right\}. \quad (8.8.8)$$

Then the prior for \tilde{x}_i is $U(a_2, b_2)$, for both linear and Gaussian process based geometric probit regression.

The rest of the methodology for computing FPBF and IPBF for geometric regression remains the same as for Poisson regression described in Section 8.8.2.

Results of the simulation experiment for model selection

For $n = m = 10$, when the true model is Poisson with log-linear regression, the last two columns of Table 8.8.1 provide the forward and inverse estimates of $\frac{1}{n} \sum_{i=1}^n \log \pi(y_{i1} | \mathbf{Y}_{nm,-i}, \mathbf{X}_n, \mathcal{M})$ and $\frac{1}{n} \sum_{i=1}^n \log \pi(y_{i1} | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M})$, respectively, for Poisson and geometric linear and Gaussian process regression with different link functions, using which the models can be easily compared with respect to both forward and inverse perspectives using FPBF and IPBF. Note that forward and inverse perspectives can also be compared.

Observe that the forward Poisson log-linear regression turns out to be the best model as expected, since this corresponds to the true, data-generating distribution. The Gaussian process based Poisson inverse regression model is the next best, followed closely by the Poisson log-linear inverse regression model, and then comes the Gaussian process based Poisson forward regression model. This order of model selection can be explained as follows. First, the inverse cases involve more uncertainties than the corresponding forward models, since these cases treat x_i as unknown. Hence, expectedly the Poisson log-linear forward regression model outperforms the inverse counterpart. But the inverse Gaussian process regression performs marginally better than the inverse linear model

and more significantly better than the forward Gaussian process model. This merits an interesting explanation. Recall that in the inverse Gaussian process model $\eta(\tilde{x}_i)$ has been linearized for constructing the prior for \tilde{x}_i , so that this part is equivalent to the linear model, which explains why the difference between the inverse linear and Gaussian process models is not significant. However, the linear part of the Gaussian process model is of course influenced by the additional Gaussian process part associated with the other data points, unlike the linear regression models. The posterior dependence structure, in conjunction with the posterior distribution of \tilde{x}_i , can yield better regression estimates $\eta(\tilde{x}_i)$ for the i -th data point in a substantial number of Monte Carlo iterations. Since the Gaussian process model includes the linear model as a special case (that is, it is not a case of misspecification), this explains why the inverse Gaussian process regression performs marginally better than the inverse linear model. In the forward Gaussian process regression, even though we have linearized $\eta(x_i)$ for comparability with the inverse model, x_i is fixed. Thus, when the i -th regression part is not well-estimated in the Monte Carlo simulations, there is no further scope for improvement in this part. However, in the inverse Gaussian process regression, x_i is replaced with the random \tilde{x}_i , which, though its posterior simulations, can improve upon the i -th regression part with positive probability, even if the regression coefficients are not well-estimated. Thus, the inverse Gaussian process regression model can significantly outperform the forward counterpart, as we observe here.

The geometric logit and probit linear and Gaussian process regressions are examples of model misspecifications since the true, data-generating model is the Poisson log-linear regression model. Accordingly, both the forward and inverse setups perform worse than the Poisson regression setups. Among the forward and inverse cases for geometric regression, the probit linear model performs the best, followed closely by the logit linear model, then by the forward logit Gaussian process and then by the forward probit Gaussian process – all the inverse regression models perform worse than the worse of

the forward regression models. This is not surprising since all these models are cases of misspecifications and given the data generated from the true model, the inverse models here only increase the uncertainty regarding x_i compared to the forward models without any positive effect. However, note that the inverse logit Gaussian process model significantly outperforms the inverse logit linear model thanks to its better flexibility and similar prior structure for \tilde{x}_i as in the case of the true log-linear Poisson regression whose positive effects carry over to this case from the first two rows of the last column of Table 8.8.1. But the same phenomenon of superiority of the inverse probit Gaussian process over inverse probit linear model is not at all visible since the prior structure of \tilde{x}_i in this misspecified case is completely different from that of the true Poisson log-linear model, and indeed, inconsistent.

Table 8.8.1

Results of our simulation study for model selection using FPBF and IPBF. The last two columns show the estimates of $\frac{1}{n} \sum_{i=1}^n \log \pi(y_{i1} | \mathbf{Y}_{nm,-i}, \mathbf{X}_n, \mathcal{M})$ and $\frac{1}{n} \sum_{i=1}^n \log \pi(y_{i1} | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M})$, respectively, for forward and inverse setups.

Model	Link function	Regression form	Forward	Inverse
<i>Poisson</i> ($\lambda(x_i)$)	log	linear	-7.913	-8.440
<i>Poisson</i> ($\lambda(x_i)$)	log	Gaussian process	-8.503	-8.409
<i>Geometric</i> ($p(x_i)$)	logit	linear	-9.176	-18.247
<i>Geometric</i> ($p(x_i)$)	logit	Gaussian process	-9.529	-14.766
<i>Geometric</i> ($p(x_i)$)	probit	linear	-9.348	-14.434
<i>Geometric</i> ($p(x_i)$)	probit	Gaussian process	-10.915	-23.733

8.8.4 Variable selection in Poisson and geometric linear and nonparametric regression models when true model is Poisson linear regression

Rather than a single covariate x in the previous examples, let us now consider covariates x and z , where the true data-generating distribution is $y_{ij} \sim \text{Poisson}(\lambda(x_i, z_i))$, with $\lambda(x, z) = \exp(\alpha_0 + \beta_0 x + \gamma_0 z)$. We generate the data by simulating $\alpha_0, \beta_0, \gamma_0 \sim U(-1, 1)$, independently; independently simulating $x_i \sim U(-1, 1)$, $z_i \sim U(0, 2)$; $i = 1, \dots, n$,

and then by finally simulating $y_{ij} \sim Poisson(\lambda(x_i, z_i))$; $j = 1, \dots, m$, $i = 1, \dots, n$, independently.

We model the data y_{ij} ; $i = 1, \dots, n$; $j = 1, \dots, m$ with both Poisson and geometric models as before with the regression part consisting of either x or z , or both. We denote the linear regression coefficients of the intercept, x and z as α , β and γ , respectively, and give the improper prior density 1 to (α, β) , (α, γ) and (α, β, γ) when the models consist of these combinations of parameters. For Gaussian process regression with both x and z , we let $\eta(x, z)$ be the regression function modeled by a Gaussian process with mean $\mu(x, z) = \alpha + \beta x + \gamma z$ and covariance function $Cov(\eta(x_1, z_1), \eta(x_2, z_2)) = \exp(\omega) \exp[-\{(x_1 - x_2)^2 + (z_1 - z_2)^2\}]$, and we assign prior mass 1 to (α, β, ω) , (α, γ, ω) and $(\alpha, \beta, \gamma, \omega)$ when the models consist of the covariates x , z or both. Using FPBF and IPBF we then compare the different models, along with the covariates associated with them. In the inverse cases, where the model consists of the single covariate x or z , then the priors for \tilde{x}_i and \tilde{z}_i remain the same as in the previous cases.

But wherever the models consist of both the covariates x and z , we need to assign priors for both \tilde{x}_i and \tilde{z}_i , in addition to requiring that $E(y_{ij}|\theta, x_i, z_i)$ under the postulated model fall in $[\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}}, \bar{y}_i + \frac{c_2 s_i}{\sqrt{m}}]$. The same priors for \tilde{x}_i and \tilde{z}_i as the previous situations where the models consisted of single covariates, will not be consistent in these situations. For consistent priors we adopt the following strategy. Letting α be the intercept, β and γ the coefficients of x_i and z_i respectively in the regression forms, we envisage the following priors for \tilde{x}_i and \tilde{z}_i .

Prior for \tilde{x}_i and \tilde{z}_i for Poisson regression

For the Poisson linear or Gaussian process regression model with log link consisting of both the covariates x and z , we set $\tilde{x}_i \sim U(a_x^{(1)}, b_x^{(1)})$ and $\tilde{z}_i \sim U(a_z^{(1)}, b_z^{(1)})$, where

$$a_x^{(1)} = \min \left\{ \beta^{-1} \left(\log \left(\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}} \right) - \alpha - \gamma z_i \right), \beta^{-1} \left(\log \left(\bar{y}_i + \frac{c_2 s_i}{\sqrt{m}} \right) - \alpha - \gamma z_i \right) \right\},$$

$$b_x^{(1)} = \max \left\{ \beta^{-1} \left(\log \left(\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}} \right) - \alpha - \gamma z_i \right), \beta^{-1} \left(\log \left(\bar{y}_i + \frac{c_2 s_i}{\sqrt{m}} \right) - \alpha - \gamma z_i \right) \right\},$$

$$a_z^{(1)} = \min \left\{ \gamma^{-1} \left(\log \left(\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}} \right) - \alpha - \beta x_i \right), \gamma^{-1} \left(\log \left(\bar{y}_i + \frac{c_2 s_i}{\sqrt{m}} \right) - \alpha - \beta x_i \right) \right\}$$

and

$$b_z^{(1)} = \max \left\{ \gamma^{-1} \left(\log \left(\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}} \right) - \alpha - \beta x_i \right), \gamma^{-1} \left(\log \left(\bar{y}_i + \frac{c_2 s_i}{\sqrt{m}} \right) - \alpha - \beta x_i \right) \right\}.$$

Note that the priors for \tilde{x}_i and \tilde{z}_i depend upon z_i and x_i respectively. This is somewhat in keeping with (8.7.6) where the prior for \tilde{x}_i depends upon x_i itself. The discussion following (8.7.6) is enough to justify that the priors for \tilde{x}_i and \tilde{z}_i in the current situation do make sense, apart from ensuring consistency.

Prior for \tilde{x}_i and \tilde{z}_i for geometric regression with logit link

For the geometric linear or Gaussian process regression model with logit link consisting of both the covariates x and z , we set $\tilde{x}_i \sim U(a_x^{(2)}, b_x^{(2)})$ and $\tilde{z}_i \sim U(a_z^{(2)}, b_z^{(2)})$, where

$$a_x^{(2)} = \min \left\{ -\beta^{-1} \left(\log \left(\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}} \right) + \alpha + \gamma z_i \right), -\beta^{-1} \left(\log \left(\bar{y}_i + \frac{c_2 s_i}{\sqrt{m}} \right) + \alpha + \gamma z_i \right) \right\},$$

$$b_x^{(2)} = \max \left\{ -\beta^{-1} \left(\log \left(\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}} \right) + \alpha + \gamma z_i \right), -\beta^{-1} \left(\log \left(\bar{y}_i + \frac{c_2 s_i}{\sqrt{m}} \right) + \alpha + \gamma z_i \right) \right\},$$

$$a_z^{(2)} = \min \left\{ -\gamma^{-1} \left(\log \left(\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}} \right) + \alpha + \beta x_i \right), -\gamma^{-1} \left(\log \left(\bar{y}_i + \frac{c_2 s_i}{\sqrt{m}} \right) + \alpha + \beta x_i \right) \right\}$$

and

$$b_z^{(2)} = \max \left\{ -\gamma^{-1} \left(\log \left(\bar{y}_i - \frac{c_1 s_i}{\sqrt{m}} \right) + \alpha + \beta x_i \right), -\gamma^{-1} \left(\log \left(\bar{y}_i + \frac{c_2 s_i}{\sqrt{m}} \right) + \alpha + \beta x_i \right) \right\}.$$

Prior for \tilde{x}_i and \tilde{z}_i for geometric regression with probit link

For the geometric linear or Gaussian process regression model with probit link consisting of both the covariates x and z , we set $\tilde{x}_i \sim U(a_x^{(3)}, b_x^{(3)})$ and $\tilde{z}_i \sim U(a_z^{(3)}, b_z^{(3)})$, where

$$a_x^{(3)} = \min \left\{ \frac{\Phi^{-1} \left(\frac{1}{u_{im}+1} \right) - \alpha - \gamma z_i}{\beta}, \frac{\Phi^{-1} \left(\frac{1}{\ell_{im}+1} \right) - \alpha - \gamma z_i}{\beta} \right\},$$

$$b_x^{(3)} = \max \left\{ \frac{\Phi^{-1} \left(\frac{1}{u_{im}+1} \right) - \alpha - \gamma z_i}{\beta}, \frac{\Phi^{-1} \left(\frac{1}{\ell_{im}+1} \right) - \alpha - \gamma z_i}{\beta} \right\},$$

$$a_z^{(3)} = \min \left\{ \frac{\Phi^{-1} \left(\frac{1}{u_{im}+1} \right) - \alpha - \beta x_i}{\gamma}, \frac{\Phi^{-1} \left(\frac{1}{\ell_{im}+1} \right) - \alpha - \beta x_i}{\gamma} \right\}$$

and

$$b_z^{(3)} = \max \left\{ \frac{\Phi^{-1} \left(\frac{1}{u_{im}+1} \right) - \alpha - \beta x_i}{\gamma}, \frac{\Phi^{-1} \left(\frac{1}{\ell_{im}+1} \right) - \alpha - \beta x_i}{\gamma} \right\}.$$

Results of the simulation experiment for model and variable selection

For $n = m = 10$, when the true model is Poisson with log-linear regression on both the covariates x and z , the last two columns of Table 8.8.2 provide the estimates of $\frac{1}{n} \sum_{i=1}^n \log \pi(y_{i1} | \mathbf{Y}_{nm,-i}, \mathbf{X}_n, \mathcal{M})$ and $\frac{1}{n} \sum_{i=1}^n \log \pi(y_{i1} | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M})$ for Poisson and geometric linear and Gaussian process regression on either x_i or z_i or both, with different link functions. Thus, the models, along with the associated covariates can be compared with respect to both forward and inverse perspectives.

Table 8.8.2 shows that the correct Poisson log-linear model with both the covariates x and z has turned out to be the third best, after the inverse Poisson log-linear model with covariate x and the forward Poisson log-linear model with covariate z . However, the difference between the latter and the correct model is not substantial and may perhaps be attributed to Monte Carlo sampling fluctuations. So, considering only the forward setup, it is difficult to rule out the possibility of the correct Poisson log-linear model with both the covariates x and z from being the best.

That the inverse Poisson log-linear model with covariate x seems to perform so well can be attributed to significant variability of the prior for \tilde{x}_i which goes on to account for the missing z_i as well in the additive model. Since the additive model is not identifiable when both x_i and z_i are unknown, the significant prior variability of \tilde{x}_i compensates for non-inclusion of z_i in the model, given the data that has arisen from the true model consisting of both x and z . The same argument is valid for good performance of the inverse Poisson log-linear model with covariate z , where the prior variance for \tilde{z}_i compensates for non-inclusion of x_i . However, note that the performance of the inverse Poisson log-linear model deteriorates significantly when the regression consists of both x and z . This is of course the consequence of the priors for both \tilde{x}_i and \tilde{z}_i , whose variances get added up in the linear model. For small n and m as in our examples, the true values x_i and z_i fail to get enough posterior weight, an issue that gets reflected in the Monte Carlo simulations where the true regression is not represented in sufficiently large

proportion.

For Poisson Gaussian process regression, the inverse models outperform their forward counterparts by large margins. This admits similar explanation provided in Section 8.8.3 for the superiority of the inverse Poisson Gaussian process model compared to its forward counterpart as visible in Table 8.8.1.

For geometric linear regression, the forward models emerge the winners in all the cases, as opposed to the inverse counterparts and also outperform the Gaussian geometric process regression models. Among the geometric models, the probit linear model with both the covariates x and z , turns out to be the best. That the corresponding inverse counterparts perform worse can be explained as in Section 8.8.3 that these are instances of model misspecification, and here the inverse models only increase uncertainty by treating x_i and z_i as unknown, without any beneficial effect.

In geometric Gaussian process regression, the inverse models perform better than the corresponding forward ones in most cases. In these cases, given the data generated from the true model, the Gaussian process dependence combined with the prior variability render the inverse models somewhat less misspecified than the forward models with no prior associated with the covariates.

Also observe that given either forward or inverse setups, the linear models perform better than the corresponding Gaussian process models, for both Poisson and geometric cases. Since the true regression is linear, this seems to provide an internal consistency. However, this phenomenon is somewhat different from that observed in Table 8.8.1 where the Gaussian process model performed better than the linear regression model for Poisson and geometric logit models. The reason for this is inconsistency of the prior for \tilde{x}_i when covariate z is ignored and that of the prior for \tilde{z}_i when covariate x is ignored in the postulated model. Indeed, Table 8.8.2 shows that in these cases, the inverse linear models outperform the Gaussian process models by considerably large margins. In these cases the Gaussian process priors only increase uncertainties without adding

any value, since the priors for \tilde{x}_i and \tilde{z}_i are inconsistent. On the other hand, note that when both x and z are incorporated in the inverse models, the linear models perform only marginally better than the Gaussian process models in the cases of inverse Poisson and inverse geometric logit models. This is because the priors of \tilde{x}_i and \tilde{z}_i are consistent in such cases, and moreover, the prior structures of \tilde{x}_i and \tilde{z}_i are similar for Poisson and geometric logit regressions. For geometric probit regression, the prior structures are entirely different from those of the correct Poisson model and in fact inconsistent, and as in Table 8.8.1, here also inverse geometric probit Gaussian process regression performs much worse than inverse geometric probit linear regression.

Table 8.8.2

Results of our simulation study for model and variable selection using FPBF and IPBF. The last two columns show the estimates of $\frac{1}{n} \sum_{i=1}^n \log \pi(y_{i1} | \mathbf{Y}_{nm,-i}, \mathbf{X}_n, \mathcal{M})$ and $\frac{1}{n} \sum_{i=1}^n \log \pi(y_{i1} | \mathbf{Y}_{nm,-i}, \mathbf{X}_{n,-i}, \mathcal{M})$, respectively, for forward and inverse setups.

Covariates	Model	Link function	Regression form	Forward	Inverse
x_i	$Poisson(\lambda(x_i))$	log	linear	-8.618	-8.388
z_i	$Poisson(\lambda(z_i))$	log	linear	-8.834	-8.739
(x_i, z_i)	$Poisson(\lambda(x_i, z_i))$	log	linear	-8.686	-13.257
x_i	$Poisson(\lambda(x_i))$	log	Gaussian process	-31.831	-9.136
z_i	$Poisson(\lambda(z_i))$	log	Gaussian process	-31.213	-10.052
(x_i, z_i)	$Poisson(\lambda((x_i, z_i)))$	log	Gaussian process	-17.712	-13.363
x_i	$Geometric(p(x_i))$	logit	linear	-9.810	-10.526
z_i	$Geometric(p(z_i))$	logit	linear	-9.673	-12.629
(x_i, z_i)	$Geometric(p(x_i, z_i))$	logit	linear	-11.806	-15.478
x_i	$Geometric(p(x_i))$	logit	Gaussian process	-26.232	-21.161
z_i	$Geometric(p(z_i))$	logit	Gaussian process	-19.391	-29.388
(x_i, z_i)	$Geometric(p(x_i, z_i))$	logit	Gaussian process	-17.128	-15.686
x_i	$Geometric(p(x_i))$	probit	linear	-9.543	-11.671
z_i	$Geometric(p(z_i))$	probit	linear	-9.401	-16.183
(x_i, z_i)	$Geometric(p(x_i, z_i))$	probit	linear	-9.060	-13.839
x_i	$Geometric(p(x_i))$	probit	Gaussian process	-23.538	-16.460
z_i	$Geometric(p(z_i))$	probit	Gaussian process	-20.522	-17.099
(x_i, z_i)	$Geometric(p(x_i, z_i))$	probit	Gaussian process	-20.102	-20.501

8.9 Summary and future direction

The importance of PBF in Bayesian model and variable selection seems to have been overlooked in the statistical literature. In this chapter we have pointed out the theoretical and computational advantages of PBF over BF, and investigated the asymptotic convergence properties of PBF in general forward and inverse regression setups. Since the inverse regression problem requires a prior on the covariate value to be predicted, this makes the treatise of PBF distinct from the forward regression problems. Specifically, we considered two setups for inverse regression. One setup is the same as that of forward regression except a prior for the relevant covariate value \tilde{x}_i . Although the priors in this case can not guarantee consistency of the posterior for \tilde{x}_i , we show that the corresponding PBF still converges exponentially and almost surely in favour of the better model, in the same way as for forward regression. However, for the inverse case, the convergence depends upon an integrated version of the KL-divergence, rather than KL-divergence as in the forward case. In another inverse regression setup, we consider m responses corresponding to each covariate value, and assign the general prior for \tilde{x}_i constructed in Chapter 6. This prior guarantees consistency for the posterior of \tilde{x}_i when m tends to infinity, along with the sample size. For this inverse setup, PBF has convergence results similar to that of forward regression which is also applicable to this setup, except that no prior is associated with the covariates.

Our results on PBF for forward regression are in agreement with the general BF convergence theory established in Chapter 7, as both are the same almost sure exponential convergence depending upon the KL-divergence from the true model. Now there might arise the question if PBF and BF convergence agree even for inverse regression setups. To clarify, first recall that BF is the ratio of the marginal densities of the data. Now for forward regression, the marginal density of the data \mathbf{Y}_n depends upon the observed covariates \mathbf{X}_n . For model \mathcal{M}_j ; $j = 1, 2$, let us denote this marginal by $m(\mathbf{Y}_n | \mathbf{X}_n, \mathcal{M}_j)$. In the inverse setup, we need to treat \mathbf{X}_n as unknown, and replace this with $\tilde{\mathbf{X}}_n =$

$(\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n)$ having some relevant prior, which may even follow from some stochastic process specification for $\tilde{\mathbf{X}}_\infty = (\tilde{x}_1, \tilde{x}_2, \dots)$. If $L(\theta_j | \mathbf{Y}_n, \mathbf{X}_n, \mathcal{M}_j)$ denotes the likelihood of θ_j for fully observed data, then the marginal density of \mathbf{Y}_n in the inverse situation is given by

$$\begin{aligned}\tilde{m}(\mathbf{Y}_n | \mathcal{M}_j) &= \int_{\Theta_j} \int_{\mathcal{X}^n} L(\theta_j | \mathbf{Y}_n, \tilde{\mathbf{X}}_n, \mathcal{M}_j) d\pi(\tilde{\mathbf{X}}_n | \theta_j, \mathcal{M}_j) d\pi(\theta_j | \mathcal{M}_j) \\ &= \int_{\Theta_j} \tilde{L}(\theta_j | \mathbf{Y}_n, \mathcal{M}_j) d\pi(\theta_j | \mathcal{M}_j),\end{aligned}$$

where

$$\tilde{L}(\theta_j | \mathbf{Y}_n, \mathcal{M}_j) = \int_{\mathcal{X}^n} L(\theta_j | \mathbf{Y}_n, \tilde{\mathbf{X}}_n, \mathcal{M}_j) d\pi(\tilde{\mathbf{X}}_n | \theta_j, \mathcal{M}_j).$$

Letting

$$\tilde{\pi}(\theta_j | \mathbf{Y}_n, \mathcal{M}_j) = \frac{\tilde{L}(\theta_j | \mathbf{Y}_n, \mathcal{M}_j) \pi(\theta_j | \mathcal{M}_j)}{\tilde{m}(\mathbf{Y}_n | \mathcal{M}_j)}$$

we have for all $\theta_j \in \Theta_j$,

$$\log \tilde{m}(\mathbf{Y}_n | \mathcal{M}_j) = \log \tilde{L}(\theta_j | \mathbf{Y}_n, \mathcal{M}_j) + \log \pi(\theta_j | \mathcal{M}_j) - \log \tilde{\pi}(\theta_j | \mathbf{Y}_n, \mathcal{M}_j),$$

which reduces the inverse marginal to the same form as that used in Chapter 7 for establishing the almost sure exponential BF convergence result which depends explicitly on the KL-divergence rate between the postulated and the true models. Hence, even in both the inverse setups that we consider, our PBF and BF convergence results agree.

We have illustrated our general asymptotic results for PBF with several theoretical examples, including linear, quadratic, AR(1) regression and variable selection, providing the explicit theoretical calculations for both forward and inverse setups. Our AR(1) regression results validate our general PBF convergence theory in a dependent data setup.

We also conducted extensive simulation experiments with small simulated datasets comparing Poisson log regression and geometric logit and probit regressions, where the

regressions are modeled by straight lines as well as Gaussian process based nonparametric functions. Both forward and inverse setups are undertaken, which include, in addition, variable selection among two possible covariates. Among several insightful revelations, our results demonstrate that the inverse regression can outperform the forward counterpart when the regression considered is nonparametric.

Thus, overall the premise for PBF investigation seems promising enough to pursue further research. In particular, we shall address PBF based variable selection in both forward and inverse regression contexts in the so-called “large p , small n ” framework, where the number of variables considered increases with sample size with various rates, crucially, at rates faster than the sample size. Various complex and high-dimensional real data based applications shall also be considered for model and variable selection using forward and inverse PBF. More sophisticated computational methods combining advanced versions of TMCMC, bridge sampling and path sampling may need to be created for accurate estimations of PBF in such real situations. These ideas will be communicated elsewhere.

Appendix

8.A1 A result on sufficient condition for (S6) of Shalizi

Theorem 55 Consider the following assumptions:

- (i) Let $\tilde{\theta} = \arg \min_{\theta \in \Theta} h(\theta)$ be the unique minimizer of $h(\theta)$ on Θ .
- (ii) Let $\tilde{\theta}_n^* = \arg \max_{\theta \in \Theta} \frac{1}{n} \log R_n(\theta)$, and assume that $\tilde{\theta}_n^* \xrightarrow{a.s.} \tilde{\theta}$, as $n \rightarrow \infty$.
- (iii) $\frac{1}{n} \log R_n(\theta)$ is stochastically equicontinuous on compact subsets of Θ .
- (iv) For all θ in such compact subsets,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log R_n(\theta) = h(\theta), \text{ almost surely.} \quad (8.A1.1)$$

- (v) The prior π on Θ is proper.

Then (8.6.18) holds.

Proof. Note that

$$\begin{aligned} \frac{1}{n} \log \int_{\mathcal{G}_n} R_n(\theta) \pi(\theta) d\theta &\leq \frac{1}{n} \log \left(\sup_{\theta \in \mathcal{G}_n} R_n(\theta) \right) + \frac{1}{n} \log \pi(\mathcal{G}_n) \\ &= \sup_{\theta \in \mathcal{G}_n} \frac{1}{n} \log R_n(\theta) + \frac{1}{n} \log \pi(\mathcal{G}_n) \\ &= \frac{1}{n} \log R_n(\tilde{\theta}_n^*) + \frac{1}{n} \log \pi(\mathcal{G}_n). \end{aligned} \quad (8.A1.2)$$

Since by condition (ii), $\tilde{\theta}_n^* \xrightarrow{a.s.} \tilde{\theta}$ as $n \rightarrow \infty$, for any $\epsilon > 0$, there exists $n_0(\epsilon) \geq 1$ such that for $n \geq n_0(\epsilon)$,

$$\tilde{\theta}_n^* \in (\tilde{\theta} - \epsilon, \tilde{\theta} + \epsilon), \text{ almost surely.} \quad (8.A1.3)$$

Conditions (iii) and (iv) validate the stochastic Ascoli lemma, and hence, for any compact subset G of Θ that contains $(\tilde{\theta} - \epsilon, \tilde{\theta} + \epsilon)$,

$$\limsup_{n \rightarrow \infty, \theta \in G} \left| \frac{1}{n} \log R_n(\theta) + h(\theta) \right| = 0, \text{ almost surely.}$$

Hence, for any $\xi > 0$, for all $\theta \in G$, almost surely,

$$\frac{1}{n} \log R_n(\theta) \leq -h(\theta) + \eta \leq -h(\Theta) + \eta, \text{ for sufficiently large } n. \quad (8.A1.4)$$

Since G contains $(\tilde{\theta} - \epsilon, \tilde{\theta} + \epsilon)$, which, in turn contains $\tilde{\theta}_n^*$ for sufficiently large n , due to (8.A1.3), it follows from (8.A1.4), that for any $\xi > 0$,

$$\frac{1}{n} \log R_n(\tilde{\theta}_n^*) \leq -h(\Theta) + \eta, \text{ for sufficiently large } n. \quad (8.A1.5)$$

The proof follows by combining (8.A1.2) and (8.A1.5), and noting that $\frac{1}{n} \log \pi(\mathcal{G}_n) < 0$ for all $n \geq 1$, since $0 < \pi(\mathcal{G}_n) < 1$ for proper priors.

■

8.A2 Proof of Theorem 47

Our proof uses concepts that are broadly similar to that of Theorem 10 of Chandra and Bhattacharya (2020a). Here we shall provide the proof for $\frac{1}{n} \log R_n^{(1)}(\theta)$ since that for $\frac{1}{n} \log R_n^{(2)}(\theta)$ is exactly the same. For notational convenience, we denote $\frac{1}{n} \log R_n^{(1)}(\theta)$ by $\frac{1}{n} \log R_n(\theta)$, $h_1(\theta)$ by $h(\theta)$, $\tilde{\theta}_1$ by $\tilde{\theta}$ and Θ_1 by Θ .

Since $h(\theta)$ is convex, $\tilde{\theta}$ must be an interior point of Θ . Hence, there exists a compact

set $G \subset \Theta$ such that $\tilde{\theta}$ is interior to G . From convergence (8.6.43) which is also uniform on compact sets, it follows that

$$\lim_{n \rightarrow \infty} \sup_{\theta \in G} \left| \frac{1}{n} \log R_n(\theta) + h(\theta) \right| = 0. \quad (8.A2.1)$$

For any $\eta > 0$, we define

$$N_\eta(\tilde{\theta}) = \{\theta : \|\tilde{\theta} - \theta\| < \eta\}; \quad N'_\eta(\tilde{\theta}) = \{\theta : \|\tilde{\theta} - \theta\| = \eta\}; \quad \overline{N}_\eta(\tilde{\theta}) = \{\theta : \|\tilde{\theta} - \theta\| \leq \eta\}.$$

Note that for sufficiently small η , $\overline{N}_\eta(\tilde{\theta}) \subset G$. Let $H = \inf_{\theta \in N'_\eta(\tilde{\theta})} h(\theta)$. Since $h(\theta)$ is minimum at $\theta = \tilde{\theta}$, $H > 0$. Let us fix an ε such that $0 < \varepsilon < H$. Then by (8.A2.1), for large enough n all $\theta \in N'_\eta(\tilde{\theta})$,

$$\frac{1}{n} \log R_n(\theta) < -h(\theta) + \varepsilon < -h(\tilde{\theta}) + \varepsilon. \quad (8.A2.2)$$

Since by (8.6.43) $\frac{1}{n} \log R_n(\tilde{\theta}) > -h(\tilde{\theta}) - \varepsilon$ for sufficiently large n , it follows from this and (8.A2.2) that

$$\frac{1}{n} \log R_n(\theta) < \frac{1}{n} \log R_n(\tilde{\theta}) + 2\varepsilon, \quad (8.A2.3)$$

for sufficiently large n . Since $0 < \varepsilon < H$ is arbitrary, it follows that for all $\theta \in N'_\eta(\tilde{\theta})$, for large enough n ,

$$\frac{1}{n} \log R_n(\theta) < \frac{1}{n} \log R_n(\tilde{\theta}), \quad (8.A2.4)$$

which shows that for large enough n , the maximum of $\frac{1}{n} \log R_n(\theta)$ is not attained at the boundary $N'_\eta(\tilde{\theta})$. Hence, the maximum must occur in the interior of $\overline{N}_\eta(\tilde{\theta})$ when n is sufficiently large. That the maximizer is unique is guaranteed by Theorem 44. Hence, the result is proved.

9

A Bayesian Multiple Testing Paradigm for Model Selection in Inverse Regression Problems

9.1 Introduction

As already mentioned, model selection in inverse regression setups is non-existent in the statistical literature. In this regard, we considered Bayes and pseudo-Bayes factors for such purpose in Chapters 7 and 8, respectively. Notably, although the Bayes factor approach is arguably the most principled and coherent approach to model comparison, they are usually difficult to compute in practice and suffer from numerical instability. Moreover, they are well-known to suffer from the so-called Lindley's paradox. The cross-validation idea proposed by Geisser and Eddy (1979) is to replace the marginal density of

the entire dataset in Bayes factors with products of cross-validation densities of individual data points. This constitutes the pseudo-Bayes factors which are computationally far simpler and numerically much more stable than the corresponding Bayes factors. Furthermore, they are also immune to Lindley’s paradox. Recognizing the importance, in Chapter 8 we established the asymptotic theory for pseudo-Bayes factors for both forward and inverse parametric and nonparametric regression problems in a very general setup that allows for dependent data and misspecified models, and showed that the results are in agreement with our corresponding asymptotic theory of Bayes factors, established in Chapter 7. We illustrated the pseudo-Bayes factor results with various theoretical examples and simulation experiments for small samples that even include simultaneous selection of models and covariates. However, the results of the simulation experiments, although interesting and insightful, do leave the scope for further improvement. In this chapter, we introduce and develop a novel Bayesian hypothesis testing formulation, incorporating the principle of the IRD approach, for model and variable selection in inverse regression setups. We show that this multiple testing strategy indeed provides improvement upon the approach based on pseudo-Bayes factor.

The area of multiple hypotheses testing can be envisaged as a promising alternative to Bayes factors for model selection if properly formulated, is expected to be very useful in inverse model selection. Unfortunately, in spite of rising popularity of the multiple testing paradigm for general testing problems, its applicability and utility in general model selection problems remain yet to be thoroughly investigated. In the classical multiple comparison context, Shimodaira (1998) use the sampling error of the Akaike Information Criterion (AIC) to select a “confidence set of models” rather than a single model. The method requires computation of standardized difference of AIC for every pair of models. Since every pair of models is involved, clearly, for even a moderate number of competing models the computation becomes infeasible, and reliability of the proposed normal approximation need not be unquestionable in general situations. We

are not aware of any other significant research on model selection in the multiple testing framework. Furthermore, multiple testing based model selection in inverse setups has not been hitherto even perceived.

In this chapter, for the first time ever, we propose and develop a Bayesian multiple testing paradigm for inverse model selection problems. Our starting point is the inverse reference distribution approach to Bayesian assessment of adequacy of inverse models introduced by [Bhattacharya \(2013\)](#). In a nutshell, the inverse model adequacy assessment idea is as follows. Given response data $\mathbf{Y}_n = \{y_1, \dots, y_n\}$, covariate data $\mathbf{X}_n = \{x_1, \dots, x_n\}$, and the Bayesian model for the data, consider the inverse leave-one-out cross-validation setup where for each $i = 1, \dots, n$, x_i needs to be predicted from the rest of the data and the underlying Bayesian model. Letting \tilde{x}_i denote the random variable corresponding to x_i when the latter is treated as unknown, the interest is then in the cross-validation posteriors $\pi(\tilde{x}_i | \mathbf{X}_{n,-i}, \mathbf{Y}_n); i = 1, \dots, n$, where $\mathbf{X}_{n,-i} = \{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n\}$. Letting $\tilde{\mathbf{X}}_n = \{\tilde{x}_1, \dots, \tilde{x}_n\}$, [Bhattacharya \(2013\)](#) considers the ‘inverse reference distribution’ of some suitable discrepancy measure $T(\tilde{\mathbf{X}}_n)$ where $\tilde{x}_i \sim \pi(\cdot | \mathbf{X}_{n,-i}, \mathbf{Y}_n); i = 1, \dots, n$. If the observed discrepancy measure $T(\mathbf{X}_n)$ falls within the desired $100(1 - \alpha)\%$ credible interval of $T(\tilde{\mathbf{X}}_n)$ where $\alpha \in (0, 1)$, then the underlying Bayesian model fits the data and not otherwise. [Bhattacharya \(2013\)](#) provides a Bayesian decision theoretic formalization of the above idea and investigates its theoretical and methodological properties, pointing out its advantages over existing ideas on forward Bayesian model assessment. The encouraging results obtained in simulation experiments and real data analyses reported in [Bhattacharya \(2013\)](#), [Bhattacharya \(2006\)](#) and [Mukhopadhyay and Bhattacharya \(2013\)](#) demonstrate the worth of the inverse model assessment idea using inverse reference distributions of appropriate discrepancy measures. Typical examples of discrepancy measures are given, for any n -dimensional vector

$\mathbf{v}_n = (v_1, \dots, v_n)$, by

$$T_1(\mathbf{v}_n) = \sum_{i=1}^n \frac{|v_i - E(\tilde{x}_i | \mathbf{X}_{n,-i}, \mathbf{Y}_n)|}{\sqrt{Var(\tilde{x}_i | \mathbf{X}_{n,-i}, \mathbf{Y}_n)}} \quad (9.1.1)$$

and

$$T_2(\mathbf{v}_n) = \sum_{i=1}^n \frac{(v_i - E(\tilde{x}_i | \mathbf{X}_{n,-i}, \mathbf{Y}_n))^2}{Var(\tilde{x}_i | \mathbf{X}_{n,-i}, \mathbf{Y}_n)}. \quad (9.1.2)$$

Since the inverse reference distribution approach turned out to be useful for assessing adequacy of inverse models, it is natural to discern that such an approach would be valuable even for inverse model selection. This very perception provided the motivation for our Bayesian multiple testing approach to inverse model selection using inverse reference distributions. The key idea is to embed all the competing inverse regression models in a mixture setting to constitute a single model needed for multiple testing. In simple terms, each hypothesis of the multiple testing procedure then essentially tests if the inverse reference distribution of the corresponding inverse regression model gives high posterior probability to appropriate regions containing the observed discrepancy measure for the model, in addition to testing if the posterior model probability is sufficiently high. The best inverse model is expected to have the highest posterior probability with respect to the above and our multiple testing formalism is so designed that it renders this idea precise with relevant coherent supports.

Our theoretical and methodological development deals with parametric and nonparametric inverse competing models, allowing dependent data as well as misspecified models. In this highly general framework we show that our multiple testing procedure almost surely selects the best possible model, as the sample size tends to infinity. Here “best” is in terms of the minimizer of the minimum Kullback-Leibler (KL) divergence from the true model, concepts that will be subsequently clarified. Our investigation also brings out the desirable results that the error rates, namely, relevant versions of the false

discovery rate and the false non-discovery rate, asymptotically converge to zero almost surely. Insightful theoretical results on asymptotic α -control of versions of the false discovery rate and its impact on the convergence of versions of the false non-discovery rate, are also presented.

Monte Carlo based computations of the model-specific posterior probabilities associated with the inverse reference distributions proceed via fast and efficient Importance Resampling Markov Chain Monte Carlo (IRMCMC) ([Bhattacharya and Haslett \(2007\)](#)) aided by Transformation based Markov Chain Monte Carlo (TMCMC) ([Dutta and Bhattacharya \(2014\)](#)) for generation of MCMC samples from the cross-validation posterior distributions having excellent mixing properties. The posterior model probabilities are based on an efficient Gibbs sampling scheme that utilizes the forward pseudo-Bayes factors for sampling from the relevant full conditional distributions of the model indices. Thus, our entire computational methodology is fast and efficient, more so because each hypothesis is associated with a single inverse model, and pairwise comparison as in [Shimodaira \(1998\)](#) is ruled out.

Recalling that one of our objectives behind development of this multiple testing paradigm is to obtain superior inverse model selection results compared to those obtained in Chapter 8 using pseudo-Bayes factors, we apply our multiple testing formalism to the same simulation experiments with the same datasets as in Chapter 8. The simulation experiments consist of two sets. In one set small sample based selection among inverse Poisson log regression and inverse geometric logit and probit regression is considered, where the regressions are either linear or based on Gaussian processes. In the other set, variable selection among two covariates is considered in addition to the aforementioned inverse model selection problem. We conduct the experiments in both non-misspecified and misspecified situations. Not only does our multiple testing procedure succeeds in selecting the best inverse models and variables in all the cases, it significantly outperforms the results yielded by the pseudo-Bayes factors.

The rest of this chapter is structured as follows. In Section 9.2 we introduce and develop our Bayesian multiple testing paradigm for inverse model selection. We progress towards a general asymptotic theory by establishing in Section 9.3 the asymptotic properties of the posterior probabilities of the alternative hypotheses. Asymptotic optimality theory for our multiple testing procedure is then provided in Section 9.4, followed by convergence theory of the measures of error in Section 9.5. In Section 9.6 we recommend some judicious modifications of the hypotheses to suit practical implementation, and in Sections 9.7 and 9.8 we provide details on two sets of simulation experiments with small samples involving Poisson and geometric linear and Gaussian process regression for relevant link functions, the second set also including in addition the problem of variable selection involving two covariates. Non-misspecified and misspecified situations are addressed in both the simulation experiments. Finally, in Section 6.9, we summarize our contributions and discuss selection of inverse models in the context of two palaeoclimate reconstruction problems, recasting our previous results on inverse model assessment in the current multiple testing context.

For theoretical purpose, we shall throughout assume that the space of covariates \mathcal{X} is compact, although such assumption is not necessary in practice.

9.2 A multiple testing framework for model selection in inverse regression problems

Let us consider models \mathcal{M}_k ; $k = 1, \dots, K$, from among which the best model needs to be selected respecting the inverse perspective. In this work, we assume that $1 < K < \infty$. We allow the provision that the true, data-generating model is not contained in the set of models being considered. For $k = 1, \dots, K$, let θ_k and Θ_k denote the parameter set and the parameter space associated with model \mathcal{M}_k . Let $\pi(\theta_k | \mathcal{M}_k)$ denote the prior for θ_k under model \mathcal{M}_k .

For our multiple testing treatise, we shall consider the second inverse regression setup detailed in Section 8.2.4. As such, for $n > 1$ and $m > 1$, let \mathbf{Y}_{nm} be generated from the marginal distribution of \mathcal{M}_0 , the true model having parameters θ_0 with prior $\pi(\theta_0|\mathcal{M}_0)$ on parameter space Θ_0 . Note that $\pi(\theta_0|\mathcal{M}_0)$ may even be the point mass on some element of Θ_0 . The dimensions of the parameter spaces $\Theta_0, \Theta_1, \dots, \Theta_K$ may all be different. We shall consider the consistent prior for \tilde{x}_i detailed in Section 6.4.

Now, for $k = 1, \dots, K$, let $f(\mathbf{Y}_{nm}|\mathbf{X}_n, \theta_k, \mathcal{M}_k)$ denote the density of \mathbf{Y}_{nm} under model \mathcal{M}_k . We combine the competing models in the following mixture form:

$$f(\mathbf{Y}_{nm}|\mathbf{X}_n, \theta) = \sum_{k=1}^K p_k f(\mathbf{Y}_{nm}|\mathbf{X}_n, \theta_k, \mathcal{M}_k), \quad (9.2.1)$$

where $\theta = (\theta_1, \dots, \theta_K)$, $0 \leq p_k \leq 1$, for $k = 1, \dots, K$ and $\sum_{k=1}^K p_k = 1$. Letting ζ denote the allocation variable (model index), with $P(\zeta = k) = p_k$, note that $f(\mathbf{Y}_{nm}|\mathbf{X}_n, \theta, \zeta = k) = f(\mathbf{Y}_{nm}|\mathbf{X}_n, \theta_k, \mathcal{M}_k)$. Now let $\tilde{\Theta}_k$ be a proper subset of Θ_k assumed to contain the minimizer of the KL-divergence from the true model \mathcal{M}_0 .

Let $\pi(\tilde{x}_i|\theta_k, \mathcal{M}_k)$ be the prior for \tilde{x}_i given θ_k , under \mathcal{M}_k . This yields the familiar (see, for example, Bhattacharya and Haslett (2007); see also Chapter 6) inverse cross-validation posterior for \tilde{x}_i given $\mathbf{X}_{n,-i}$ and \mathbf{Y}_{nm} given by

$$\pi(\tilde{x}_i|\mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k) = \int_{\Theta_k} \pi(\tilde{x}_i|\theta_k, \mathbf{y}_i, \mathcal{M}_k) d\pi(\theta_k|\mathbf{X}_{n,-i}, \mathbf{Y}_{nm}).$$

However, if θ_k is restricted to $\tilde{\Theta}_k$, then we obtain the following $\tilde{\Theta}_k$ -restricted inverse cross-validation posterior for \tilde{x}_i given $\mathbf{X}_{n,-i}$ and \mathbf{Y}_n :

$$\pi(\tilde{x}_i|\mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k, \tilde{\Theta}_k) = \frac{\int_{\tilde{\Theta}_k} \pi(\tilde{x}_i|\theta_k, \mathbf{y}_i, \mathcal{M}_k) d\pi(\theta_k|\mathbf{X}_{n,-i}, \mathbf{Y}_{nm})}{\pi(\tilde{\Theta}_k|\mathbf{X}_{n,-i}, \mathbf{Y}_{nm})}. \quad (9.2.2)$$

In the misspecified situation, $\theta_0 \notin \Theta_k$, and $\tilde{\theta}_k$ is the minimizer of the limiting KL-divergence rate from \mathcal{M}_0 . Thus, in the case of misspecification of θ_k , $B_{im}(\tilde{\theta}_k) \xrightarrow{a.s.} \{x_{ik}^*\}$

as $m \rightarrow \infty$, for some non-random x_{ik}^* ($\neq x_i$), depending upon model \mathcal{M}_k . In other words, the prior distribution of \tilde{x}_i given $\tilde{\theta}_k$ and \mathbf{y}_i concentrates around x_{ik}^* , as $m \rightarrow \infty$. In Theorem 58 we show that the cross-validation posterior of \tilde{x}_i also concentrates around x_{ik}^* . Note that x_{ik}^* depends upon both $\tilde{\theta}_k$ and θ_0 , apart from x_i (and perhaps x_j for some $j \neq i$).

For any n -dimensional vector $\mathbf{v}_n = (v_1, \dots, v_n)$, and for some $c > 0$, define

$$T_1^{(k)}(\mathbf{v}_n) = \frac{1}{n} \sum_{i=1}^n \frac{|v_i - E(\tilde{x}_i | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k, \tilde{\Theta}_k)|}{\sqrt{Var(\tilde{x}_i | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k, \tilde{\Theta}_k) + c}}. \quad (9.2.3)$$

Similarly, let

$$T_2^{(k)}(\mathbf{v}_n) = \frac{1}{n} \sum_{i=1}^n \frac{(v_i - E(\tilde{x}_i | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k, \tilde{\Theta}_k))^2}{Var(\tilde{x}_i | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k, \tilde{\Theta}_k) + c}. \quad (9.2.4)$$

In (9.2.3) and (9.2.4), \tilde{x}_i has the cross-validation posterior distribution (9.2.2), for $i = 1, \dots, n$. The positive constant c is not only needed for asymptotics, it plays the role of maintaining stability of the discrepancy measures when $Var(\tilde{x}_i | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k, \tilde{\Theta}_k)$ is close to zero for some $i \geq 1$. Various other measures of discrepancy can be defined (see [Bhattacharya \(2013\)](#) for a discussion on such discrepancy measures; see also [Mukhopadhyay and Bhattacharya \(2013\)](#)), but for brevity we focus on these two measures in our work.

For a given discrepancy measure $T^{(k)}$, let $[\tilde{\ell}_{knm}, \tilde{u}_{knm}]$ denote the $100(1-\alpha)\%$ credible interval for the posterior distribution of $T^{(k)}(\tilde{\mathbf{X}}_n)$ for any desired $\alpha \in (0, 1)$. In Theorem 61 we show that for any $\varepsilon > 0$, the posterior probability of the event

$$\left\{ T^{(k)}(\tilde{\mathbf{X}}_n) - T^{(k)}(\mathbf{X}_n) \in [\tilde{\ell}_{knm} - a_k - \varepsilon, \tilde{u}_{knm} - a_k + \varepsilon] \right\}$$

tends to one almost surely as $m \rightarrow \infty$ and $n \rightarrow \infty$. Here a_k are positive constants reflecting misspecification. If there is no misspecification, then $a_k = 0$.

With the above notions and ideas it seems reasonable to formulate the following multiple testing problem for inverse model selection. For given $\varepsilon > 0$ and $\eta > 0$, and given discrepancy measure $T^{(k)}$ associated with model \mathcal{M}_k , for $k = 1, \dots, K$, consider testing

$$H_{0k} : p_k > 1 - \eta, \theta_k \in \tilde{\Theta}_k, T^{(k)}(\tilde{\mathbf{X}}_n) - T^{(k)}(\mathbf{X}_n) \in [\tilde{\ell}_{knm} - a_k - \varepsilon, \tilde{u}_{knm} - a_k + \varepsilon]$$

versus

$$\begin{aligned} H_{1k} : & \{p_k \leq 1 - \eta\} \bigcup \left\{ p_k > 1 - \eta, \theta_k \in \tilde{\Theta}_k^c \right\} \\ & \bigcup \left\{ p_k > 1 - \eta, \theta_k \in \tilde{\Theta}_k, T^{(k)}(\tilde{\mathbf{X}}_n) - T^{(k)}(\mathbf{X}_n) \in [\tilde{\ell}_{knm} - a_k - \varepsilon, \tilde{u}_{knm} - a_k + \varepsilon]^c \right\}. \end{aligned}$$

The positive constants a_k in the hypotheses should be perceived as analogous to a_{1k} and a_{2k} in (9.3.15) and (9.3.16).

However, the above multiple testing formulation depends upon the choice of η . More importantly, even though the posterior probability of $\zeta = \tilde{k}$ goes to 1 asymptotically for the best model $\mathcal{M}_{\tilde{k}}$, that of $\{p_k > 1 - \eta\}$, for any $\eta > 0$, does not tend to one for any prior on (p_1, \dots, p_K) . For example, for a Dirichlet prior with parameters $(\alpha_1, \dots, \alpha_K)$, where $\alpha_k > 0$ for $k = 1, \dots, K$, the posterior distribution of (p_1, \dots, p_K) given ζ , the other parameters and the data, is Dirichlet with parameters $(\alpha_1 + I(\zeta = 1), \dots, \alpha_K + I(\zeta = K))$, where for any k , $I(\zeta = k) = 1$ if $\zeta = k$ and zero otherwise. Thus, even if $\zeta = \tilde{k}$ with posterior probability tending to one, asymptotically the posterior distribution of $p_{\tilde{k}}$ does not converge to one. It is thus necessary to modify the above multiple testing formulation, replacing the statements involving p_k with those involving ζ . Specifically, we re-write the hypotheses as follows:

$$H_{0k} : \zeta = k, \theta_k \in \tilde{\Theta}_k, T^{(k)}(\tilde{\mathbf{X}}_n) - T^{(k)}(\mathbf{X}_n) \in [\tilde{\ell}_{knm} - a_k - \varepsilon, \tilde{u}_{knm} - a_k + \varepsilon] \quad (9.2.5)$$

versus

$$\begin{aligned} H_{1k} : & \{\zeta \neq k\} \bigcup \left\{ \zeta = k, \theta_k \in \tilde{\Theta}_k^c \right\} \\ & \bigcup \left\{ \zeta = k, \theta_k \in \tilde{\Theta}_k, T^{(k)}(\tilde{\mathbf{X}}_n) - T^{(k)}(\mathbf{X}_n) \in [\tilde{\ell}_{knm} - a_k - \varepsilon, \tilde{u}_{knm} - a_k + \varepsilon]^c \right\}. \end{aligned} \quad (9.2.6)$$

Henceforth, unless stated otherwise, we shall refer to (9.2.5) and (9.2.6) for our multiple testing purpose.

9.2.1 Further discussion of the multiple testing formulation

To select the best model from an inverse perspective we first need to choose a model $f(\mathbf{Y}_{nm}|\mathbf{X}_n, \theta_{\tilde{k}}, \mathcal{M}_{\tilde{k}})$ indexed by $\zeta = \tilde{k}$ which has high marginal posterior probability. But this is not enough as the inverse context is not reflected in this selection. Indeed, such a selection is the same as in the forward context.

Thus, in addition to selecting such a \tilde{k} , we demand that for such model

$$T^{(k)}(\tilde{\mathbf{X}}_n) - T^{(k)}(\mathbf{X}_n) \in [\tilde{\ell}_{knm} - a_k - \varepsilon, \tilde{u}_{knm} - a_k + \varepsilon]. \quad (9.2.7)$$

This reflects the inverse perspective. We further demand that this holds for $\tilde{\mathbf{X}}_n$ associated with some region $\tilde{\Theta}_{\tilde{k}}$ of the parameter space that contains the minimizer of the KL-divergence of $f(\mathbf{Y}_{nm}|\mathbf{X}_n, \theta_{\tilde{k}}, \mathcal{M}_{\tilde{k}})$ from the true model. The reason for this is that $\tilde{\Theta}_{\tilde{k}}$ is the region that has the highest posterior probability, at least asymptotically, which we shall subsequently establish. Moreover, it follows from Chapter 6 that $\pi(\theta_k|\mathbf{X}_n, \mathbf{Y}_{nm}, \mathcal{M}_k)$ and $\pi(\theta_k|\mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k)$ are asymptotically the same for any $i \geq 1$, for any $m \geq 1$. Hence the event (9.2.7) associated with $\tilde{\Theta}_{\tilde{k}}$ for $k = \tilde{k}$, is expected to be reliable.

We shall also show that asymptotically the posterior probability of the best model, $\zeta = \tilde{k}$, tends to 1 almost surely. As already mentioned, here the notion the best model is

with respect to minimization of the minimum KL-divergence rate from the true model. We shall show that for this \tilde{k} , the posterior probability of $H_{0\tilde{k}}$ goes to 1 asymptotically, for any $\varepsilon > 0$ in (9.2.7). That is, asymptotically, only one inverse model, namely, the best inverse model satisfying the conditions of $H_{0\tilde{k}}$, will be selected.

It is useful to remark here that the KL-divergence rate referred to above is completely in the forward sense, where all the x_i ; $i \geq 1$, are assumed to be known. Hence, the above arguments and our subsequent theoretical underpinnings show that the asymptotic theory is dominated by the forward perspective. In fact, any consistent prior for \tilde{x}_i would asymptotically lead to the best forward model. However, the above can not be guaranteed in any non-asymptotic sense. The model $\mathcal{M}_{\tilde{k}}$ with high posterior probability of $\{\zeta = \tilde{k}\}$ may have low posterior probability of $T^{(k)}(\tilde{\mathbf{X}}_n) - T^{(k)}(\mathbf{X}_n) \in [\tilde{\ell}_{knm} - a_k - \varepsilon, \tilde{u}_{knm} - a_k + \varepsilon]$, which may result in overall lower posterior probability of $H_{0\tilde{k}}$ compared to H_{0k} for several $k \neq \tilde{k}$. In such situations, $\mathcal{M}_{\tilde{k}}$ will not be the best choice non-asymptotically. Thus, the inverse perspective is particularly important in realistic, non-asymptotic situations. An appropriate Bayesian multiple testing procedure is expected to yield the best possible inference regarding inverse model selection in both asymptotic and non-asymptotic situations, which we now devise.

9.2.2 The Bayesian multiple testing procedure

[Chandra and Bhattacharya \(2019\)](#) proposed a novel Bayesian non-marginal testing procedure for testing general dependent hypotheses. We first briefly discuss their method and then consider a special case of their idea to be applied to inverse model selection context.

Let

$$d_k = \begin{cases} 1 & \text{if the } k\text{-th hypothesis is rejected;} \\ 0 & \text{otherwise;} \end{cases}$$

$$r_k = \begin{cases} 1 & \text{if } H_{1k} \text{ is true;} \\ 0 & \text{if } H_{0k} \text{ is true.} \end{cases}$$

Let G_k be the set of hypotheses (including hypothesis k) where the parameters are dependent on the k -th hypothesis. In the new procedure, the decision of each hypothesis is penalized by incorrect decisions regarding other dependent parameters. Thus a compound criterion where all the decisions in G_k deterministically depends upon each other. Define the following quantity

$$z_k = \begin{cases} 1 & \text{if } H_{d_j,j} \text{ is true for all } j \in G_k \setminus \{k\}; \\ 0 & \text{otherwise.} \end{cases} \quad (9.2.8)$$

If, for any $k \in \{1, \dots, K\}$, $G_k = \{k\}$, a singleton, then we define $z_k = 1$. The notion of true positives (TP) are modified as the following

$$TP = \sum_{k=1}^K d_k r_k z_k, \quad (9.2.9)$$

The posterior expectation of TP is maximized subject to controlling the posterior expectation of the error term

$$E = \sum_{k=1}^K d_k (1 - r_k z_k). \quad (9.2.10)$$

It follows that the decision configuration can be obtained by minimizing the function

$$\begin{aligned}\xi(\mathbf{d}) &= -\sum_{k=1}^K d_k E(r_k z_k | \mathbf{X}_n, \mathbf{Y}_{nm}) + \lambda_{nm} \sum_{k=1}^K d_k E[(1 - r_k z_k) | \mathbf{X}_n, \mathbf{Y}_{nm}] \\ &= -(1 + \lambda_{nm}) \sum_{k=1}^K d_k \left(w_{knm}(\mathbf{d}) - \frac{\lambda_{nm}}{1 + \lambda_{nm}} \right),\end{aligned}$$

with respect to all possible decision configurations of the form $\mathbf{d} = \{d_1, \dots, d_K\}$, where $\lambda_{nm} > 0$, and

$$w_{knm}(\mathbf{d}) = E(r_k z_k | \mathbf{X}_n, \mathbf{Y}_{nm}) = \pi(H_{1k} \cap \{\cap_{j \neq k, j \in G_k} H_{d_j, j}\} | \mathbf{X}_n, \mathbf{Y}_{nm})$$

is the posterior probability of the decision configuration $\{d_1, \dots, d_{k-1}, 1, d_{k+1}, \dots, d_K\}$ being correct. Letting $\beta_{nm} = \lambda_{nm}/(1 + \lambda_{nm})$, one can equivalently maximize

$$f_{\beta_{nm}}(\mathbf{d}) = \sum_{k=1}^K d_k (w_{knm}(\mathbf{d}) - \beta_{nm}) \tag{9.2.11}$$

with respect to \mathbf{d} and obtain the optimal decision configuration.

Definition 56 *Let \mathbb{D} be the set of all m -dimensional binary vectors denoting all possible decision configurations. Define*

$$\hat{\mathbf{d}} = \arg \max_{\mathbf{d} \in \mathbb{D}} f_{\beta}(\mathbf{d})$$

where $0 < \beta < 1$. Then $\hat{\mathbf{d}}$ is the optimal decision configuration obtained as the solution of the non-marginal multiple testing method.

Note that in the definitions of both TP and E , d_i is penalized by incorrect decisions in the same group. This forces the decisions to be jointly taken also adjudging other dependent parameters.

9.2.3 Specialization of the general multiple testing procedure to inverse model selection problems

In our inverse model selection problem note that since the models \mathcal{M}_k ; $k = 1, \dots, K$, are independent, so are $\tilde{\mathbf{X}}_n$ associated with the different models. Thus, the hypotheses are dependent only through the relation $\sum_{k=1}^K I(\zeta = k) = 1$. As we shall show, the posterior probability of the event $\{\zeta = \tilde{k}\}$ converges to one *a posteriori* as the sample size tends to infinity, irrespective of any other dependence among $(I(\zeta = 1), \dots, I(\zeta = K))$ induced through (p_1, \dots, p_K) . Hence, there is not enough reason to consider the hypotheses as dependent. Thus, for our purpose, we simply set $G_k = \{k\}$. Consequently, (9.2.11) in our case reduces to

$$f_{\beta_{nm}}(\mathbf{d}) = \sum_{k=1}^K d_k (v_{knm} - \beta_{nm}), \quad (9.2.12)$$

where

$$v_{knm} = E(r_k | \mathbf{X}_n, \mathbf{Y}_{nm}) = \pi(H_{1k} | \mathbf{X}_n, \mathbf{Y}_{nm}).$$

In this case, the optimal decision configuration $\hat{\mathbf{d}}$ is given by the following: for $k = 1, \dots, K$,

$$\hat{d}_k = \begin{cases} 1 & \text{if } v_{knm} > \beta_{nm}; \\ 0 & \text{otherwise.} \end{cases} \quad (9.2.13)$$

Hence, although our formulation of the multiple hypothesis test for inverse model selection is novel, the Bayesian procedure for testing parallels that of Müller *et al.* (2004) (see also Guindani *et al.* (2009)), which is a special case of the general procedure proposed in Chandra and Bhattacharya (2019).

9.2.4 Error measures in multiple testing

Storey (2003) advocated *positive False Discovery Rate (pFDR)* as a measure of Type-I error in multiple testing. Let $\delta(\mathbf{d} | \mathbf{X}_n, \mathbf{Y}_{nm})$ be the probability of choosing \mathbf{d} as the

optimal decision configuration given data $(\mathbf{X}_n, \mathbf{Y}_{nm})$ when a given multiple testing method is employed. Then $pFDR$ is defined as:

$$pFDR_{nm} = E_{\mathbf{Y}_{nm}|\mathbf{X}_n} \left[\sum_{\mathbf{d} \in \mathbb{D}} \frac{\sum_{k=1}^K d_k(1 - r_k)}{\sum_{k=1}^K d_k} \delta(\mathbf{d}|\mathbf{X}_n, \mathbf{Y}_{nm}) \middle| \delta(\mathbf{d} = \mathbf{0}|\mathbf{X}_n, \mathbf{Y}_{nm}) = 0 \right]. \quad (9.2.14)$$

Analogous to Type-II error, the *positive False Non-discovery Rate* ($pFNR$) is defined as

$$pFNR_{nm} = E_{\mathbf{Y}_{nm}|\mathbf{X}_n} \left[\sum_{\mathbf{d} \in \mathbb{D}} \frac{\sum_{k=1}^K (1 - d_k)r_k}{\sum_{k=1}^K (1 - d_k)} \delta(\mathbf{d}|\mathbf{X}_n, \mathbf{Y}_{nm}) \middle| \delta(\mathbf{d} = \mathbf{1}|\mathbf{X}_n, \mathbf{Y}_{nm}) = 0 \right]. \quad (9.2.15)$$

Under prior $\pi(\cdot)$, Sarkar *et al.* (2008) defined posterior FDR and FNR . The measures are given as following:

$$\text{posterior } FDR_{nm} = E \left[\sum_{\mathbf{d} \in \mathbb{D}} \frac{\sum_{k=1}^K d_k(1 - r_k)}{\sum_{k=1}^K d_k \vee 1} \delta(\mathbf{d}|\mathbf{X}_n, \mathbf{Y}_{nm}) \middle| \mathbf{X}_n, \mathbf{Y}_{nm} \right] \quad (9.2.16)$$

$$= \sum_{\mathbf{d} \in \mathbb{D}} \frac{\sum_{k=1}^K d_k(1 - v_{knm})}{\sum_{k=1}^K d_k \vee 1} \delta(\mathbf{d}|\mathbf{X}_n, \mathbf{Y}_{nm}); \quad (9.2.17)$$

$$\text{posterior } FNR_{nm} = E \left[\sum_{\mathbf{d} \in \mathbb{D}} \frac{\sum_{k=1}^K (1 - d_k)r_k}{\sum_{k=1}^K (1 - d_k) \vee 1} \delta(\mathbf{d}|\mathbf{X}_n, \mathbf{Y}_{nm}) \middle| \mathbf{X}_n, \mathbf{Y}_{nm} \right] \quad (9.2.18)$$

$$= \sum_{\mathbf{d} \in \mathbb{D}} \frac{\sum_{k=1}^K (1 - d_k)v_{knm}}{\sum_{k=1}^K (1 - d_k) \vee 1} \delta(\mathbf{d}|\mathbf{X}_n, \mathbf{Y}_{nm}). \quad (9.2.19)$$

Also under any non-randomized decision rule, $\delta(\mathbf{d}|\mathbf{X}_n, \mathbf{Y}_{nm})$ is either 1 or 0 depending on data $(\mathbf{X}_n, \mathbf{Y}_{nm})$. Given $(\mathbf{X}_n, \mathbf{Y}_{nm})$, we denote these error measures conditional on the data by conditional FDR ($cFDR_{nm}$) and conditional FNR ($cFNR_{nm}$) respectively.

The positive Bayesian FDR ($pBFDR_{nm}$) and FNR ($pBFNR_{nm}$) are the expectations of $cFDR_{nm}$ and $cFNR_{nm}$ respectively, with respect to the distribution of \mathbf{Y}_{nm} given \mathbf{X}_n .

For our Bayesian purpose, we shall consider the Bayesian measures $cFDR_{nm}$, $pBFDR_{nm}$, $cFNR_{nm}$ and $pBFNR_{nm}$, and investigate their asymptotic properties. Chandra and Bhattacharya (2019) and Chandra and Bhattacharya (2020a) particularly recommend $cFDR_{nm}$ and $cFNR_{nm}$, since they are conditioned on the observed data $(\mathbf{X}_n, \mathbf{Y}_{nm})$ and hence qualify as *bona fide* Bayesian measures.

Let us now proceed towards development of the asymptotic theory for our proposed multiple testing strategy. The issue of misspecification will play a crucial role in this context. Suppose that the true data-generating parameter θ_0 is not contained in Θ , the parameter space considered. This is a case of misspecification that we must incorporate in our asymptotic theory. Indeed, we shall build a general asymptotic framework that allows for possibly infinite-dimensional parameters, dependent data as well as misspecification. In this regard, the approach presented in Shalizi (2009) seems to be very appropriate. Before proceeding further, we first provide a brief overview of this approach, which we conveniently exploit for our purpose.

9.3 Asymptotic properties of the posterior probabilities of the alternative hypotheses

9.3.1 Posterior convergence to the best model

Theorem 57 *Assume that for $k = 1, \dots, K$, \mathcal{M}_k satisfies conditions (S1)–(S6) of Shalizi, and that the competing models as well as the true model have densities with respect to some common σ -finite measure. Also assume that the posterior associated with \mathcal{M}_k is dominated by the prior, which is again absolutely continuous with respect to some appropriate σ -finite measure, and that the priors satisfy $\pi(\theta_k | \mathcal{M}_k) > 0$ for all $\theta_k \in \Theta_k$.*

Let $h_{\tilde{k}}(\Theta_{\tilde{k}}) = \min\{h_k(\Theta_k) : k = 1, \dots, K\}$. Then for any $m \geq 1$,

$$\lim_{n \rightarrow \infty} \pi(\zeta = k | \mathbf{X}_n, \mathbf{Y}_{nm}) \stackrel{a.s.}{=} \begin{cases} 1 & \text{if } k = \tilde{k} \\ 0 & \text{if } k \neq \tilde{k}. \end{cases} \quad (9.3.1)$$

Proof. For any $k_1, k_2 \in \{1, \dots, K\}$, let $BF^{(nm)}(\mathcal{M}_{k_1}, \mathcal{M}_{k_2})$ denote the Bayes factor of model \mathcal{M}_{k_1} against model \mathcal{M}_{k_2} . Then as a direct consequence of Theorem 25 of Chapter 7, the following holds for any $m \geq 1$:

$$\frac{1}{n} \log BF^{(nm)}(\mathcal{M}_k, \mathcal{M}_0) \rightarrow -h_k(\Theta_k), \text{ as } n \rightarrow \infty, \quad (9.3.2)$$

almost surely with respect to the true model \mathcal{M}_0 . In the above, $h_k(\Theta_k)$ corresponds to (4.1.1), (4.A1.2) and (4.A1.3) for model \mathcal{M}_k with parameter space Θ_k .

Now, since $h_{\tilde{k}}(\Theta_{\tilde{k}}) = \min\{h_k(\Theta_k) : k = 1, \dots, K\}$, it follows from (9.3.2) that as $n \rightarrow \infty$, for any $m \geq 1$,

$$\frac{1}{n} \log BF^{(nm)}(\mathcal{M}_k, \mathcal{M}_{\tilde{k}}) \rightarrow -[h_k(\Theta_k) - h_{\tilde{k}}(\Theta_{\tilde{k}})],$$

so that as $n \rightarrow \infty$, for any $m \geq 1$,

$$BF^{(nm)}(\mathcal{M}_k, \mathcal{M}_{\tilde{k}}) = \begin{cases} 1 & \text{if } k = \tilde{k} \\ \xrightarrow{a.s.} 0, & \text{if } k \neq \tilde{k}. \end{cases} \quad (9.3.3)$$

Now note that (see, for example, Liang *et al.* (2008))

$$\pi(\zeta = k | \mathbf{X}_n, \mathbf{Y}_{nm}, p_1, \dots, p_K) = \frac{p_k BF^{(nm)}(\mathcal{M}_k, \mathcal{M}_{\tilde{k}})}{\sum_{\ell=1}^K p_\ell BF^{(nm)}(\mathcal{M}_\ell, \mathcal{M}_{\tilde{k}})}. \quad (9.3.4)$$

Hence it follows by applying (9.3.3) to (9.3.4) that the following holds:

$$\lim_{n \rightarrow \infty} \pi(\zeta = k | \mathbf{X}_n, \mathbf{Y}_{nm}, p_1, \dots, p_K) \stackrel{a.s.}{=} \begin{cases} 1 & \text{if } k = \tilde{k} \\ 0 & \text{if } k \neq \tilde{k}. \end{cases} \quad (9.3.5)$$

Now note that $\pi(\zeta = k | \mathbf{X}_n, \mathbf{Y}_{nm}) = E[\pi(\zeta = k | \mathbf{X}_n, \mathbf{Y}_{nm}, p_1, \dots, p_K)]$, the expectation being over the posterior distribution of (p_1, \dots, p_K) given \mathbf{X}_n and \mathbf{Y}_{nm} . Since $\pi(\zeta = k | \mathbf{X}_n, \mathbf{Y}_{nm}, p_1, \dots, p_K) \leq 1$ almost surely, it follows by uniform integrability and (9.3.5), that

$$\lim_{n \rightarrow \infty} \pi(\zeta = k | \mathbf{X}_n, \mathbf{Y}_{nm}) = E[\pi(\zeta = k | \mathbf{X}_n, \mathbf{Y}_{nm}, p_1, \dots, p_K)] \stackrel{a.s.}{=} \begin{cases} 1 & \text{if } k = \tilde{k} \\ 0 & \text{if } k \neq \tilde{k}. \end{cases}$$

■

9.3.2 Convergence of the cross-validation posteriors of \tilde{x}_i

Theorem 58 For model \mathcal{M}_k assume conditions (S1)–(S7) of Shalizi, and let the infimum of $h_k(\theta_k)$ over $\Theta_{\tilde{k}}$ be attained at $\tilde{\theta}_k \in \tilde{\Theta}_k$, where $\tilde{\theta}_k \neq \theta_0$. Also assume that Θ_k and Θ_0 are complete separable metric spaces. Then, with the prior (6.4.1), under further assumptions that $\pi(\tilde{x}_i | \theta_k, \mathbf{y}_i, \mathcal{M}_k)$ is continuous in θ_k , $f(\mathbf{y}_i | \tilde{\theta}_k, \tilde{x}_i, \mathcal{M}_k)$ is continuous in \tilde{x}_i , for $i \geq 1$ and $\tilde{\eta}_k$ is a one-to-one function, the following holds:

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \pi(\tilde{x}_i \in V_{ik}^c | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k, \tilde{\Theta}_k) = 0, \text{ almost surely,} \quad (9.3.6)$$

for any neighborhood V_{ik} of x_{ik}^* .

Proof. By the hypotheses, (4.1.2) holds, from which it follows that for any $\epsilon > 0$, and for any $m \geq 1$,

$$\lim_{n \rightarrow \infty} \pi(\mathbb{N}_{k,\epsilon}^c | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k) = 0, \quad (9.3.7)$$

where $\mathbb{N}_{k,\epsilon} = \{\theta_k : h_k(\theta_k) \leq h_k(\Theta_k) + \epsilon\}$.

Now, by hypothesis, the infimum of $h_k(\theta_k)$ over Θ_k is attained at $\tilde{\theta}_k \in \Theta_k$, where $\tilde{\theta}_k \neq \theta_0$. Then by (9.3.7), the posterior of θ_k given $\mathbf{X}_{n,-i}$ and \mathbf{Y}_{nm} , concentrates around $\tilde{\theta}_k$, the minimizer of the limiting KL-divergence rate from the true distribution. Formally, given any neighborhood U_k of $\tilde{\theta}_k$, the set $\mathbb{N}_{k,\epsilon}$ is contained in U_k for sufficiently small ϵ . It follows that for any neighborhood U_k of $\tilde{\theta}_k$, $\pi(U_k | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k) \rightarrow 1$, almost surely, as $n \rightarrow \infty$. Since Θ_k is a complete, separable metric space, it follows that (see, for example, Ghosh and Ramamoorthi (2003), Ghosal and van der Vaart (2017))

$$\pi(\cdot | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k) \xrightarrow{w} \delta_{\tilde{\theta}_k}(\cdot), \text{ almost surely, as } n \rightarrow \infty, \text{ for any } m \geq 1. \quad (9.3.8)$$

In the above, $\delta_{\tilde{\theta}_k}(\cdot)$ denotes point mass at $\tilde{\theta}_k$.

Now since $\tilde{\Theta}_k^c \subset \Theta_k$, $h_k(\tilde{\Theta}_k^c) > h_k(\Theta_k)$. Hence, from (4.1.2) it follows that for any $m \geq 1$,

$$\pi\left(\theta_k \in \tilde{\Theta}_k^c | \mathbf{X}_n, \mathbf{Y}_{nm}, \mathcal{M}_k\right) \xrightarrow{a.s.} 0, \text{ as } n \rightarrow \infty. \quad (9.3.9)$$

Also note that since $\pi(\tilde{x}_i | \theta_k, \mathbf{y}_i, \mathcal{M}_k)$ is continuous in θ_k by assumption, it follows by Scheffe's theorem that any probability associated with $\pi(\tilde{x}_i | \theta_k, \mathbf{y}_i, \mathcal{M}_k)$ is continuous in θ_k (see Lemma 18 of Chapter 6). Hence, for any neighborhood V_{ik} of x_{ik}^* , the probability $\pi(\tilde{x}_i \in V_{ik}^c | \theta_k, \mathbf{y}_i, \mathcal{M}_k)$ is continuous in θ_k . Moreover, since it is a probability, it is bounded. Hence, by the Portmanteau theorem, weak convergence of $\pi(\theta_k | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k)$, and (9.3.9) it holds almost surely that

$$\begin{aligned} \pi(\tilde{x}_i \in V_{ik}^c | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k) &= \frac{\int_{\tilde{\Theta}_k} \pi(\tilde{x}_i \in V_{ik}^c | \theta_k, \mathbf{y}_i, \mathcal{M}_k) d\pi(\theta_k | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k)}{\pi(\tilde{\Theta}_k | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k)} \\ &\xrightarrow{a.s.} \pi(\tilde{x}_i \in V_{ik}^c | \tilde{\theta}_k, \mathbf{y}_i, \mathcal{M}_k), \text{ as } n \rightarrow \infty, \text{ for any } m \geq 1. \end{aligned}$$

That $\pi(\tilde{x}_i \in V_{ik}^c | \tilde{\theta}_k, \mathbf{y}_i, \mathcal{M}_k) \xrightarrow{a.s.} 0$, as $m \rightarrow \infty$, follows in the same way as the proof of Theorem 21 of Chapter 6 by replacing θ_0 with $\tilde{\theta}_k$. ■

9.3.3 Posterior convergence of the discrepancy measures

Theorem 59 *Under the conditions of Theorem 58, the following holds for any $\varepsilon > 0$:*

$$\pi \left(T^{(k)}(\tilde{\mathbf{X}}_n) > \varepsilon | \mathbf{X}_n, \mathbf{Y}_{mn}, \mathcal{M}_k, \tilde{\Theta}_k \right) \xrightarrow{a.s.} 0, \text{ as } m \rightarrow \infty, n \rightarrow \infty, \quad (9.3.10)$$

where $T^{(k)} = T_k^{(1)}$ or $T_k^{(2)}$.

Proof. For $i \geq 1$, Theorem 58 implies almost sure weak convergence of the i -th cross-validation posterior of \tilde{x}_i for model \mathcal{M}_k to $\delta_{x_{ik}^*}$, as $m \rightarrow \infty$ and $n \rightarrow \infty$. This is equivalent to convergence in (cross-validation posterior) distribution of \tilde{x}_i to the degenerate quantity x_{ik}^* , almost surely. Degeneracy guarantees that this is equivalent to convergence in probability, almost surely. In other words, with respect to the cross-validation posterior distribution of \tilde{x}_i for model \mathcal{M}_k , almost surely, as $m \rightarrow \infty, n \rightarrow \infty$,

$$\tilde{x}_i \xrightarrow{P} x_{ik}^*. \quad (9.3.11)$$

Now note that $T^{(k)}(\tilde{\mathbf{X}}_n)$ is an average of n terms, the i -th term being $\frac{|\tilde{x}_i - E(\tilde{x}_i | \mathbf{X}_n, \mathbf{Y}_{mn}, \mathcal{M}_k, \tilde{\Theta}_k)|}{\sqrt{Var(\tilde{x}_i | \mathbf{X}_n, \mathbf{Y}_{mn}, \mathcal{M}_k, \tilde{\Theta}_k) + c}}$ or its square. Since $\tilde{x}_i \in \mathcal{X}$ for $i \geq 1$ and \mathcal{X} is compact, (9.3.11) and uniform integrability entails that

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} E \left(\tilde{x}_i | \mathbf{X}_n, \mathbf{Y}_{mn}, \mathcal{M}_k, \tilde{\Theta}_k \right) \xrightarrow{a.s.} x_{ik}^*; \quad (9.3.12)$$

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} Var \left(\tilde{x}_i | \mathbf{X}_n, \mathbf{Y}_{mn}, \mathcal{M}_k, \tilde{\Theta}_k \right) \xrightarrow{a.s.} 0. \quad (9.3.13)$$

It follows from (9.3.12) and (9.3.13) that with respect to the cross-validation posterior distribution of \tilde{x}_i for model \mathcal{M}_k , almost surely, as $m \rightarrow \infty, n \rightarrow \infty$,

$$\frac{|\tilde{x}_i - E \left(\tilde{x}_i | \mathbf{X}_n, \mathbf{Y}_{mn}, \mathcal{M}_k, \tilde{\Theta}_k \right)|}{\sqrt{Var \left(\tilde{x}_i | \mathbf{X}_n, \mathbf{Y}_{mn}, \mathcal{M}_k, \tilde{\Theta}_k \right) + c}} \xrightarrow{P} 0, \text{ for all } i \geq 1. \quad (9.3.14)$$

Hence, by Theorem 7.15 of Schervish (1995) (page 398), it follows that with respect to the cross-validation posterior distributions of $\{\tilde{x}_i; i \geq 1\}$, for model \mathcal{M}_k , almost surely, as $m \rightarrow \infty, n \rightarrow \infty$,

$$T^{(k)}(\tilde{\mathbf{X}}_n) \xrightarrow{P} 0,$$

which is equivalent to (9.3.10). ■

Theorem 60 *Assume the conditions of Theorem 59. Also assume that for $i \geq 1$, x_{ik}^* is a continuous function of $\{x_1, x_2, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_{i+\ell}\}$, for some non-negative integer ℓ . Then there exist positive constants a_{1k} and a_{2k} such that*

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} T_1^{(k)}(\mathbf{X}_n) = a_{1k}; \quad (9.3.15)$$

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} T_2^{(k)}(\mathbf{X}_n) = a_{2k}. \quad (9.3.16)$$

Proof. It follows from (9.3.12) and (9.3.13) that

$$T_1^{(k)}(\mathbf{X}_n) \xrightarrow{a.s} \lim_{n \rightarrow \infty} \frac{1}{n\sqrt{c}} \sum_{i=1}^n |x_i - x_{ik}^*|; \quad (9.3.17)$$

$$T_2^{(k)}(\mathbf{X}_n) \xrightarrow{a.s} \lim_{n \rightarrow \infty} \frac{1}{nc} \sum_{i=1}^n (x_i - x_{ik}^*)^2. \quad (9.3.18)$$

Now, by our assumption, x_{ik}^* is a continuous function of $\{x_1, x_2, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_{i+\ell}\}$, for some non-negative integer ℓ . Hence, letting $u_{ik} = x_i - x_{ik}^*$, it follows by Riemann sum convergence that

$$\lim_{n \rightarrow \infty} \frac{1}{n\sqrt{c}} \sum_{i=1}^n |x_i - x_{ik}^*| = c^{-\frac{1}{2}} |\tilde{\mathcal{X}}_k|^{-1} \int_{\tilde{\mathcal{X}}_k} |u| du; \quad (9.3.19)$$

$$\lim_{n \rightarrow \infty} \frac{1}{nc} \sum_{i=1}^n (x_i - x_{ik}^*)^2 = c^{-1} |\tilde{\mathcal{X}}_k|^{-1} \int_{\tilde{\mathcal{X}}_k} u^2 du, \quad (9.3.20)$$

where $\tilde{\mathcal{X}}_k$ is the appropriate compact co-domain of u_{ik} induced by the transformation

$u_{ik} = x_i - x_{ik}^*$ and the original compact covariate space \mathcal{X} , and $|\tilde{\mathcal{X}}_k|$ stands for the Lebesgue measure of $\tilde{\mathcal{X}}_k$.

Since the right hand sides of (9.3.19) and (9.3.20) are well-defined positive quantities, the proof follows by combining (9.3.17) – (9.3.20). ■

Theorem 61 *Assume the conditions of Theorem 60. Then the following holds for any $\varepsilon > 0$, where $T^{(k)} = T_1^{(k)}$ or $T_2^{(k)}$ and respectively, $a_k = a_{1k}$ or a_{2k} :*

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \pi \left(T^{(k)}(\tilde{\mathbf{X}}_n) - T^{(k)}(\mathbf{X}_n) \in [\tilde{\ell}_{knm} - a_k - \varepsilon, \tilde{u}_{knm} - a_k + \varepsilon]^c \middle| \mathbf{X}_n, \mathbf{Y}_{nm}, \mathcal{M}_k, \tilde{\Theta}_k \right) \stackrel{a.s.}{=} 0. \quad (9.3.21)$$

Proof. First, observe that since for $i = 1, \dots, n$, $\tilde{x}_i \in \mathcal{X}$ almost surely, where \mathcal{X} is compact, $\frac{|\tilde{x}_i - E(\tilde{x}_i | \mathbf{X}_n, \mathbf{Y}_{mn}, \mathcal{M}_k, \tilde{\Theta}_k)|}{\sqrt{Var(\tilde{x}_i | \mathbf{X}_n, \mathbf{Y}_{mn}, \mathcal{M}_k, \tilde{\Theta}_k) + c}}$ are almost surely uniformly bounded. Hence, $T_1^{(k)}(\tilde{\mathbf{X}}_n)$ and $T_2^{(k)}(\tilde{\mathbf{X}}_n)$ are almost surely bounded. Consequently, using (9.3.10) of Theorem 59 and uniform integrability it follows that

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} E \left(T_1^{(k)}(\tilde{\mathbf{X}}_n) | \mathbf{X}_n, \mathbf{Y}_{nm}, \mathcal{M}_k, \tilde{\Theta}_k \right) \stackrel{a.s.}{=} 0; \quad (9.3.22)$$

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} E \left(T_2^{(k)}(\tilde{\mathbf{X}}_n) | \mathbf{X}_n, \mathbf{Y}_{nm}, \mathcal{M}_k, \tilde{\Theta}_k \right) \stackrel{a.s.}{=} 0; \quad (9.3.23)$$

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} Var \left(T_1^{(k)}(\tilde{\mathbf{X}}_n) | \mathbf{X}_n, \mathbf{Y}_{nm}, \mathcal{M}_k, \tilde{\Theta}_k \right) \stackrel{a.s.}{=} 0; \quad (9.3.24)$$

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} Var \left(T_2^{(k)}(\tilde{\mathbf{X}}_n) | \mathbf{X}_n, \mathbf{Y}_{nm}, \mathcal{M}_k, \tilde{\Theta}_k \right) \stackrel{a.s.}{=} 0. \quad (9.3.25)$$

The limits (9.3.22) – (9.3.25) imply that

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \tilde{\ell}_{knm} \stackrel{a.s.}{=} 0; \quad (9.3.26)$$

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \tilde{u}_{knm} \stackrel{a.s.}{=} 0. \quad (9.3.27)$$

Due to (9.3.26) and Theorem 60, given any $\varepsilon > 0$, for sufficiently large m and n ,

$\tilde{\ell}_{knm} - a_k + T^{(k)}(\mathbf{X}_n) - \varepsilon < 0$. Since $T^{(k)}(\tilde{\mathbf{X}}_n) > 0$ with probability one, we thus have

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \pi \left(T^{(k)}(\tilde{\mathbf{X}}_n) > \tilde{\ell}_{knm} - a_k + T^{(k)}(\mathbf{X}_n) - \varepsilon \middle| \mathbf{X}_n, \mathbf{Y}_{nm}, \mathcal{M}_k, \tilde{\Theta}_k \right) \stackrel{a.s.}{=} 1. \quad (9.3.28)$$

Also, due to (9.3.27) and Theorem 60, given any $\varepsilon > 0$, for sufficiently large m and n , $\tilde{u}_{knm} - a_k + T^{(k)}(\mathbf{X}_n) + \varepsilon > 0$. Hence, given any $\varepsilon > 0$, for sufficiently large m and n , we have by Markov's inequality,

$$\begin{aligned} & \pi \left(T^{(k)}(\tilde{\mathbf{X}}_n) > \tilde{u}_{knm} - a_k + T^{(k)}(\mathbf{X}_n) + \varepsilon \middle| \mathbf{X}_n, \mathbf{Y}_{nm}, \mathcal{M}_k, \tilde{\Theta}_k \right) \\ & < \left(\tilde{u}_{knm} - a_k + T^{(k)}(\mathbf{X}_n) + \varepsilon \right)^{-2} \\ & \quad \times \left[\text{Var} \left(T^{(k)}(\tilde{\mathbf{X}}_n) \middle| \mathbf{X}_n, \mathbf{Y}_{nm}, \mathcal{M}_k, \tilde{\Theta}_k \right) + \left\{ E \left(T^{(k)}(\tilde{\mathbf{X}}_n) \middle| \mathbf{X}_n, \mathbf{Y}_{nm}, \mathcal{M}_k, \tilde{\Theta}_k \right) \right\}^2 \right]. \end{aligned} \quad (9.3.29)$$

Taking limits of both sides of (9.3.29) and using (9.3.22) – (9.3.25) we obtain

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \pi \left(T^{(k)}(\tilde{\mathbf{X}}_n) > \tilde{u}_{knm} - a_k + T^{(k)}(\mathbf{X}_n) + \varepsilon \middle| \mathbf{X}_n, \mathbf{Y}_{nm}, \mathcal{M}_k, \tilde{\Theta}_k \right) \stackrel{a.s.}{=} 0. \quad (9.3.30)$$

Combining (9.3.28) and (9.3.30) yields

$$\begin{aligned} & \lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \pi \left(T^{(k)}(\tilde{\mathbf{X}}_n) - T^{(k)}(\mathbf{X}_n) \in [\tilde{\ell}_{knm} - a_k - \varepsilon, \tilde{u}_{knm} - a_k + \varepsilon] \middle| \mathbf{X}_n, \mathbf{Y}_{nm}, \mathcal{M}_k, \tilde{\Theta}_k \right) \\ & = \lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \pi \left(T^{(k)}(\tilde{\mathbf{X}}_n) > \tilde{\ell}_{knm} - a_k + T^{(k)}(\mathbf{X}_n) - \varepsilon \middle| \mathbf{X}_n, \mathbf{Y}_{nm}, \mathcal{M}_k, \tilde{\Theta}_k \right) \\ & \quad - \lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \pi \left(T^{(k)}(\tilde{\mathbf{X}}_n) > \tilde{u}_{knm} - a_k + T^{(k)}(\mathbf{X}_n) + \varepsilon \middle| \mathbf{X}_n, \mathbf{Y}_{nm}, \mathcal{M}_k, \tilde{\Theta}_k \right) \\ & \stackrel{a.s.}{=} 1, \end{aligned}$$

thus proving (9.3.21). ■

Remark 62 In all the examples provided in Chapter 8, it has been shown that the conditions of Theorem 60 are satisfied. Hence, Theorem 61 holds for all the examples presented in Chapter 8.

9.3.4 Convergence of the posterior probabilities of H_{1k}

Theorem 63 Assume that for $k = 1, \dots, K$, \mathcal{M}_k satisfies conditions (S1)–(S7) of Shalizi, and that the competing models as well as the true model have densities with respect to some common σ -finite measure. Also assume that the posterior associated with \mathcal{M}_k is dominated by the prior, which is again absolutely continuous with respect to some appropriate σ -finite measure, and that the priors satisfy $\pi(\theta_k | \mathcal{M}_k) > 0$ for all $\theta_k \in \Theta_k$. Let $h_{\tilde{k}}(\Theta_{\tilde{k}}) = \min\{h_k(\Theta_k) : k = 1, \dots, K\}$. Then

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} v_{knm} \stackrel{a.s.}{=} \begin{cases} 1 & \text{if } k \neq \tilde{k} \\ 0 & \text{if } k = \tilde{k}. \end{cases} \quad (9.3.31)$$

Proof. First, let $k \neq \tilde{k}$. Then

$$\begin{aligned} v_{knm} &= \pi(\zeta \neq k | \mathbf{X}_n, \mathbf{Y}_{nm}) + \pi\left(\zeta = k, \theta_k \in \tilde{\Theta}_k^c | \mathbf{X}_n, \mathbf{Y}_{nm}\right) \\ &\quad + \pi\left(\zeta = k, \theta_k \in \tilde{\Theta}_k, T^{(k)}(\tilde{\mathbf{X}}_n) - T^{(k)}(\mathbf{X}_n) \in [\tilde{\ell}_{knm} - a_k - \varepsilon, \tilde{u}_{knm} - a_k + \varepsilon]^c \middle| \mathbf{X}_n, \mathbf{Y}_{nm}\right). \end{aligned} \quad (9.3.32)$$

Since $k \neq \tilde{k}$, it follows due to (9.3.1) that for any $m \geq 1$, as $n \rightarrow \infty$,

$$\pi(\zeta \neq k | \mathbf{X}_n, \mathbf{Y}_{nm}) = \pi\left(\zeta = \tilde{k} | \mathbf{X}_n, \mathbf{Y}_{nm}\right) + \sum_{j \neq k, \tilde{k}} \pi(\zeta \neq k | \mathbf{X}_n, \mathbf{Y}_{nm}) \xrightarrow{a.s.} 1. \quad (9.3.33)$$

Using (9.3.1) again it follows that for any $m \geq 1$,

$$\pi(\zeta = k, \theta_k \in \tilde{\Theta}_k^c | \mathbf{X}_n, \mathbf{Y}_{nm}) \leq \pi(\zeta = k | \mathbf{X}_n, \mathbf{Y}_{nm}) \xrightarrow{a.s.} 0, \text{ as } n \rightarrow \infty \quad (9.3.34)$$

and

$$\begin{aligned} & \pi\left(\zeta = k, \theta_k \in \tilde{\Theta}_k, T^{(k)}(\tilde{\mathbf{X}}_n) - T^{(k)}(\mathbf{X}_n) \in [\tilde{\ell}_{knm} - a_k - \varepsilon, \tilde{u}_{knm} - a_k + \varepsilon]^c \middle| \mathbf{X}_n, \mathbf{Y}_{nm}\right) \\ & \leq \pi(\zeta = k | \mathbf{X}_n, \mathbf{Y}_{nm}) \xrightarrow{a.s.} 0, \text{ as } n \rightarrow \infty. \end{aligned} \quad (9.3.35)$$

Results (9.3.33), (9.3.34) and (9.3.35) imply that if $k \neq \tilde{k}$, then for any $m \geq 1$,

$$v_{knm} \xrightarrow{a.s.} 1, \text{ as } n \rightarrow \infty. \quad (9.3.36)$$

Now let us obtain the limit of v_{knm} when $k = \tilde{k}$. By (9.3.1),

$$\pi(\zeta \neq \tilde{k} | \mathbf{X}_n, \mathbf{Y}_{nm}) \xrightarrow{a.s.} 0, \text{ as } n \rightarrow \infty. \quad (9.3.37)$$

For any $m \geq 1$, using (9.3.9) we obtain

$$\pi(\zeta = \tilde{k}, \theta_{\tilde{k}} \in \tilde{\Theta}_{\tilde{k}}^c | \mathbf{X}_n, \mathbf{Y}_{nm}) \leq \pi(\theta_{\tilde{k}} \in \tilde{\Theta}_{\tilde{k}}^c | \mathbf{X}_n, \mathbf{Y}_{nm}) \xrightarrow{a.s.} 0, \text{ as } n \rightarrow \infty. \quad (9.3.38)$$

Now note that

$$\begin{aligned}
 & \pi \left(\zeta = \tilde{k}, \theta_{\tilde{k}} \in \tilde{\Theta}_{\tilde{k}}, T^{(\tilde{k})}(\tilde{\mathbf{X}}_n) - T^{(\tilde{k})}(\mathbf{X}_n) \in [\tilde{\ell}_{\tilde{k}nm} - a_{\tilde{k}} - \varepsilon, \tilde{u}_{\tilde{k}nm} - a_{\tilde{k}} + \varepsilon]^c \middle| \mathbf{X}_n, \mathbf{Y}_{nm} \right) \\
 &= \pi \left(\zeta = \tilde{k} | \mathbf{X}_n, \mathbf{Y}_{nm} \right) \\
 &\quad \times \pi \left(\theta_{\tilde{k}} \in \tilde{\Theta}_{\tilde{k}}, T^{(\tilde{k})}(\tilde{\mathbf{X}}_n) - T^{(\tilde{k})}(\mathbf{X}_n) \in [\tilde{\ell}_{\tilde{k}nm} - a_{\tilde{k}} - \varepsilon, \tilde{u}_{\tilde{k}nm} - a_{\tilde{k}} + \varepsilon]^c \middle| \mathbf{X}_n, \mathbf{Y}_{nm}, \zeta = \tilde{k} \right) \\
 &\leq \pi \left(T^{(\tilde{k})}(\tilde{\mathbf{X}}_n) - T^{(\tilde{k})}(\mathbf{X}_n) \in [\tilde{\ell}_{\tilde{k}nm} - a_{\tilde{k}} - \varepsilon, \tilde{u}_{\tilde{k}nm} - a_{\tilde{k}} + \varepsilon]^c \middle| \mathbf{X}_n, \mathbf{Y}_{nm}, \zeta = \tilde{k} \right) \\
 &\xrightarrow{a.s.} 0, \text{ as } m \rightarrow \infty, n \rightarrow \infty, \text{ due to (9.3.21)}. \tag{9.3.39}
 \end{aligned}$$

From (9.3.37), (9.3.38) and (9.3.39) it follows that

$$v_{\tilde{k}nm} \xrightarrow{a.s.} 0, \text{ as } m \rightarrow \infty, n \rightarrow \infty. \tag{9.3.40}$$

The limits (9.3.36) and (9.3.40) show that (9.3.31) holds. ■

9.4 Asymptotic optimality theory for our multiple testing procedure

Let $h_{\tilde{k}}(\Theta_{\tilde{k}}) = \min\{h_k(\Theta_k) : k = 1, \dots, K\}$. Also let us define $\tilde{\mathbf{d}} = (\tilde{d}_1, \dots, \tilde{d}_K)$, where

$$\tilde{d}_k = \begin{cases} 1 & \text{if } k \neq \tilde{k} \\ 0 & \text{if } k = \tilde{k}. \end{cases} \tag{9.4.1}$$

Definition 64 A multiple testing method for the inverse model selection is said to be asymptotically optimal for which

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \delta(\tilde{\mathbf{d}} | \mathbf{X}_n, \mathbf{Y}_{nm}) \xrightarrow{a.s.} 1.$$

Recall the constant β_{nm} in (9.2.11), which is the penalizing constant between the error E and true positives TP . For consistency of the non-marginal procedure, we need certain conditions on β_n , which we state below. These conditions will also play important roles in the asymptotic studies of the different versions of FDR and FNR that we consider.

(A1) We assume that the sequence β_{nm} is neither too small nor too large, that is,

$$\underline{\beta} = \liminf_{m \geq 1, n \geq 1} \beta_{nm} > 0; \quad (9.4.2)$$

$$\bar{\beta} = \limsup_{m \geq 1, n \geq 1} \beta_{nm} < 1. \quad (9.4.3)$$

With this conditions we propose and prove the following results.

Theorem 65 *Let $\delta(\cdot | \mathbf{X}_n, \mathbf{Y}_{nm})$ denote the decision rule given data \mathbf{X}_n and \mathbf{Y}_{nm} . Assume the conditions of Theorem 63 and condition (A1) on β_{nm} . Then the decision procedure is asymptotically optimal.*

Proof. Due to (A1), given $\epsilon_1 > 0$, there exist $m_0 \geq 1$ and $n_0 \geq 1$ such that for $m \geq m_0$ and $n \geq n_0$,

$$0 < \underline{\beta} - \epsilon_1 < \beta_{nm} < \bar{\beta} + \epsilon_1 < 1. \quad (9.4.4)$$

By (9.3.31), for any $0 < \epsilon_2 < 1 - \bar{\beta} - \epsilon_1$, for $k \neq \tilde{k}$, there exist $m_k \geq 1$ and $n_k \geq 1$ such that for $m \geq m_k$ and $n \geq n_k$,

$$v_{knm} > 1 - \epsilon_2 > \bar{\beta} + \epsilon_1. \quad (9.4.5)$$

Also, for $0 < \epsilon_3 < \underline{\beta} - \epsilon_1$, there exist $m_{\tilde{k}} \geq 1$ and $n_{\tilde{k}} \geq 1$ such that for $m \geq m_{\tilde{k}}$ and $n \geq n_{\tilde{k}}$,

$$v_{\tilde{k}nm} < \epsilon_3 < \underline{\beta} - \epsilon_1. \quad (9.4.6)$$

Let $\tilde{m} = \max\{m_0, m_1, \dots, m_K\}$ and $\tilde{n} = \max\{n_0, n_1, \dots, n_K\}$. Then it can be seen from (9.4.4), (9.4.5) and (9.4.6) that for $m \geq \tilde{m}$ and $n \geq \tilde{n}$ the following hold almost surely:

$$v_{knm} > \beta_{nm}, \text{ if } k \neq \tilde{k}; \quad (9.4.7)$$

$$v_{knm} < \beta_{nm}, \text{ if } k = \tilde{k}. \quad (9.4.8)$$

Using (9.4.7) and (9.4.8) in (9.2.13) shows that for $m \geq \tilde{m}$ and $n \geq \tilde{n}$,

$$\hat{d}_k = \begin{cases} 1 & \text{if } k \neq \tilde{k}; \\ 0 & \text{if } k = \tilde{k}. \end{cases} \quad (9.4.9)$$

In other words, almost surely, $\hat{\mathbf{d}} = \tilde{\mathbf{d}}$ for $m \geq \tilde{m}$ and $n \geq \tilde{n}$. This completes the proof. ■

Remark 66 Since $\delta(\cdot | \mathbf{X}_n, \mathbf{Y}_{nm})$ is an indicator function, the following also holds:

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} E_{\mathbf{Y}_{nm} | \mathbf{X}_n} [\delta(\tilde{\mathbf{d}} | \mathbf{X}_n, \mathbf{Y}_{nm})] = 1.$$

9.5 Asymptotic theory of the error measures

9.5.1 Convergence of versions of FDR and FNR

Theorem 67 Assume the conditions of Theorem 63 and condition (A1) on β_{nm} . Then

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} cFDR_{nm} \stackrel{a.s.}{=} 0; \quad (9.5.1)$$

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} pBFDR_{nm} = 0. \quad (9.5.2)$$

Proof. From (9.2.17) observe that

$$cFDR_{nm} = \frac{\sum_{k=1}^K \tilde{d}_k (1 - v_{knm})}{\sum_{k=1}^K \tilde{d}_k \vee 1} \delta(\tilde{\mathbf{d}} | \mathbf{X}_n, \mathbf{Y}_{nm}) + \sum_{\mathbf{d} \neq \tilde{\mathbf{d}} \in \mathbb{D}} \frac{\sum_{k=1}^K d_k (1 - v_{knm})}{\sum_{k=1}^K d_k \vee 1} \delta(\mathbf{d} | \mathbf{X}_n, \mathbf{Y}_{nm}) \quad (9.5.3)$$

The proof of Theorem 65 shows that there exist $\tilde{m} \geq 1$ and $\tilde{n} \geq 1$ such that $\delta(\tilde{\mathbf{d}}|\mathbf{X}_n, \mathbf{Y}_{nm}) = 1$ almost surely for $m \geq \tilde{m}$ and $n \geq \tilde{n}$. This, combined with (9.5.3) shows that for $m \geq \tilde{m}$ and $n \geq \tilde{n}$, almost surely,

$$cFDR_{nm} = \frac{\sum_{k=1}^K \tilde{d}_k (1 - v_{knm})}{\sum_{k=1}^K \tilde{d}_k \vee 1} = \frac{\sum_{k \neq \tilde{k}} (1 - v_{knm})}{K - 1}. \quad (9.5.4)$$

Applying (9.3.31) to the right most side of (9.5.4) shows that

$$cFDR_{nm} \xrightarrow{a.s.} 0, \text{ as } m \rightarrow \infty, n \rightarrow \infty,$$

establishing (9.5.1).

Since $cFDR_{nm} < 1$ almost surely, (9.5.2) follows from (9.5.1) by uniform integrability.

■

Theorem 68 *Assume the conditions of Theorem 63 and condition (A1) on β_{nm} . Then*

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} cFNR_{nm} \xrightarrow{a.s.} 0; \quad (9.5.5)$$

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} pBFNR_{nm} = 0. \quad (9.5.6)$$

Proof. It follows from (9.2.19) and the proof of Theorem 65 that there exist $\tilde{m} \geq 1$ and $\tilde{n} \geq 1$ such that for $m \geq \tilde{m}$ and $n \geq \tilde{n}$, almost surely,

$$\begin{aligned} cFNR_{nm} &= \frac{\sum_{k=1}^K (1 - \tilde{d}_k) v_{knm}}{\sum_{k=1}^K (1 - \tilde{d}_k) \vee 1} \delta(\tilde{\mathbf{d}}|\mathbf{X}_n, \mathbf{Y}_{nm}) + \sum_{d \neq \tilde{d} \in \mathbb{D}} \frac{\sum_{k=1}^K (1 - d_k) v_{knm}}{\sum_{k=1}^K (1 - d_k) \vee 1} \delta(d|\mathbf{X}_n, \mathbf{Y}_{nm}) \\ &= \frac{\sum_{k=1}^K (1 - \tilde{d}_k) v_{knm}}{\sum_{k=1}^K (1 - \tilde{d}_k) \vee 1} = v_{\tilde{k}nm}. \end{aligned} \quad (9.5.7)$$

Application of (9.3.31) to the right most side of (9.5.7) yields

$$cFNR_{nm} \xrightarrow{a.s.} 0, \text{ as } m \rightarrow \infty, n \rightarrow \infty,$$

establishing (9.5.5).

Again, (9.5.6) follows from (9.5.5) by uniform integrability, since $cFNR_{nm}$ is almost surely bounded above by one.

■

9.5.2 Convergence of versions of FNR when versions of FDR are α -controlled

Theorem 69 *Assume the conditions of Theorem 63. Then $\alpha = K^{-1}$ is the only asymptotic FDR control possible in the sense that there exist sequences $\beta_{nm} \rightarrow 0$ as $m \rightarrow \infty$ and $n \rightarrow \infty$ such that the following hold:*

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} cFDR_{nm} \stackrel{a.s.}{=} K^{-1}; \quad (9.5.8)$$

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} pBFDR_{nm} = K^{-1}. \quad (9.5.9)$$

Proof.

It follows from Chandra and Bhattacharya (2019) (see also Chandra and Bhattacharya (2020a)) that $pBFDR_{nm}$ is continuous and decreasing in β_{nm} , for any given $m \geq 1$ and $n \geq 1$. Hence, the maximum error given any $m \geq 1$ and $n \geq 1$ occurs when $\beta_{nm} = 0$. Hence, in this case, for any given $m \geq 1$ and $n \geq 1$, for our multiple testing procedure we must maximize $\sum_{k=1}^K d_k v_{knm}$ with respect to \mathbf{d} . This of course yields $\hat{d}_k = 1$, for $k = 1, \dots, K$. For this decision $\hat{\mathbf{d}}$, we obtain using (9.3.31):

$$cFDR_{nm} = \frac{\sum_{k=1}^K \hat{d}_k (1 - v_{knm})}{\sum_{k=1}^K \hat{d}_k \vee 1} = \frac{\sum_{k=1}^K (1 - v_{knm})}{K} \xrightarrow{a.s.} K^{-1}, \text{ as } m \rightarrow \infty, n \rightarrow \infty. \quad (9.5.10)$$

Uniform integrability and (9.5.10) shows that when $\beta_{nm} = 0$ for any $m \geq 1$ and $n \geq 1$,

$$pBFDR_{nm} \rightarrow K^{-1}, \text{ as } m \rightarrow \infty, n \rightarrow \infty. \quad (9.5.11)$$

Now consider any sequence β_{nm} that yields any decision $\hat{\mathbf{d}}$ such that $\hat{d}_{\tilde{k}} = 1$ almost surely, for sufficiently large m and n . Note that $\hat{d}_{\tilde{k}} = 1$ can occur only if $v_{\tilde{k}nm} > \beta_{nm}$. Since $v_{\tilde{k}nm} \xrightarrow{a.s.} 0$ by (9.3.31), we must have $\beta_{nm} \rightarrow 0$ as $m \rightarrow \infty$ and $n \rightarrow \infty$ in such cases. Also since $v_{knm} \xrightarrow{a.s.} 1$ for $k \neq \tilde{k}$ due to (9.3.31), it follows that $\hat{d}_k = 1$ almost surely for large enough m and n , for $k \neq \tilde{k}$. Hence, the limits (9.5.10) and (9.5.11) continue to hold in all cases such that $\hat{d}_{\tilde{k}} = 1$, for sufficiently large m and n .

On the other hand, for any sequence β_{nm} that yields any decision $\hat{\mathbf{d}}$ such that $\hat{d}_{\tilde{k}} = 0$ almost surely for sufficiently large m and n , it is easily seen that $cFDR_{nm} \xrightarrow{a.s.} 0$ and $pBFDR_{nm} \rightarrow 0$, as $m \rightarrow \infty$ and $n \rightarrow \infty$.

In other words, asymptotic control of $cFDR_{nm}$ and $pBFDR_{nm}$ is possible only at $\alpha = K^{-1}$. ■

Theorem 70 *Assume that either of $cFDR_{nm}$ or $pBFDR_{nm}$ is asymptotically controlled at $\alpha = K^{-1}$. Then for sufficiently large m and n ,*

$$cBFNR_{nm} \xrightarrow{a.s.} 0; \quad (9.5.12)$$

$$pBFNR_{nm} = 0. \quad (9.5.13)$$

Proof. From the proof of Theorem 69, recall that for asymptotic control of $cFDR_{nm}$ or $pBFDR_{nm}$ at $\alpha = K^{-1}$, we must obtain decision $\hat{\mathbf{d}}$ where $\hat{d}_k = 1$, for $k = 1, \dots, K$, for large enough m and n . Hence, (9.5.12) and (9.5.13) follow simply from the definitions of $cBFNR_{nm}$ and $pBFNR_{nm}$ with $\mathbf{d} = \hat{\mathbf{d}}$ for sufficiently large m and n . ■

Remark 71 *Theorem 70 shows that $cBFNR_{nm}$ and $pBFNR_{nm}$ are exactly zero for large enough m and n . Needless to mention, these are far stronger results than convergence to zero in the limit. In other words, essentially in keeping with the classical hypothesis testing paradigm, α -control of the Type-I error actually minimizes the Type-II error for sufficiently large m and n .*

9.6 Modification of the multiple testing procedure for practical implementation

Note that the constants a_k in (9.2.5) and (9.2.6), which depend upon the true parameter(s) θ_0 , are unknown, since θ_0 is unknown. The constants a_k also depend upon $\tilde{\theta}_k$, the minimizer of the KL-divergence of model \mathcal{M}_k from the true model. Since the true model itself is generally unknown, $\tilde{\theta}_k$ is usually unknown. Estimation of these parameters need not be reliable unless assumptions regarding the true model is accurate enough.

In practice, the considered models $\mathcal{M}_k; k = 1, \dots, K$, are expected to be carefully chosen for final model selection so that misspecifications, if any, are not expected to be severe. Hence, for finite samples, where the variability of $T^{(k)}(\tilde{\mathbf{X}}_n)$, and hence the desired credible intervals, are reasonably large, a_k is not expected to play significant role. In such cases, it makes sense to set $a_k = 0$. Similarly, setting $\varepsilon = 0$ also makes sense.

Also in practice, one might set $\tilde{\Theta}_k = \Theta_k$ since accurate specification of a small set containing $\tilde{\theta}_k$ is not possible without knowledge of $\tilde{\theta}_k$. With these, for practical purposes we re-formulate (9.2.5) and (9.2.6) as follows:

$$H_{0k} : \zeta = k, T^{(k)}(\tilde{\mathbf{X}}_n) - T^{(k)}(\mathbf{X}_n) \in [\tilde{\ell}_{knm}, \tilde{u}_{knm}] \quad (9.6.1)$$

versus

$$H_{1k} : \{\zeta \neq k\} \bigcup \left\{ \zeta = k, T^{(k)}(\tilde{\mathbf{X}}_n) - T^{(k)}(\mathbf{X}_n) \in [\tilde{\ell}_{knm}, \tilde{u}_{knm}]^c \right\}. \quad (9.6.2)$$

We shall consider the above hypotheses for our applications.

9.7 First simulation study: selection among Poisson and geometric parametric and nonparametric inverse regression models

For our simulation experiments we consider the same data and models considered in Chapter 8 for the forward and inverse pseudo-Bayes factor illustration. Specifically, we set $n = m = 10$ and generate data from relevant Poisson distribution with the log-linear link function and consider modeling the data with Poisson and geometric distributions with log, logit and probit links for linear regression as well as nonparametric regression modeled by Gaussian process having linear mean function and squared exponential covariance. We also consider variable selection in these setups with respect to two different covariates.

Here we demonstrate that the forward and inverse pseudo-Bayes factor results obtained in Chapter 8 for both the experiments involving model selection and variable selection can be significantly improved with our inverse multiple testing framework.

Let us begin with the model selection framework. In this context, the details of the true, data-generating distribution and the competing inverse regression models are the same as in Chapter 8.8.1.

9.7.1 Implementation of our multiple testing procedure for inverse model selection

We now briefly discuss our strategy for implementing our multiple testing procedure for hypotheses (9.6.1) and (9.6.2). We set $\tilde{\Theta}_k$ to Θ_k , so we shall denote $\pi(\tilde{x}_i | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k, \tilde{\Theta}_k)$ by $\pi(\tilde{x}_i | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k)$.

**Obtaining the posterior distributions of the discrepancy measures using
IRMCMC and TMCMC**

For each competing model \mathcal{M}_k ; $k = 1, \dots, K$, we obtain samples from the cross-validation posterior distribution $\pi(\tilde{x}_i | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k)$, for $i = 1, \dots, n$, using fast and efficient IRMCMC. The key idea is to first generate realizations of size N from some appropriate ‘importance sampling density’ of the form $\pi(\tilde{x}_{i^*}, \theta_k | \mathbf{X}_{n,-i^*}, \mathbf{Y}_{nm}, \mathcal{M}_k)$, for some $i^* \in \{1, \dots, n\}$ using TMCMC. Note that a major advantage of TMCMC over regular MCMC is that it effectively reduces the dimensionality of the parameters to a single dimension, thus drastically improving the acceptance rate and computational speed, while ensuring good mixing properties at the same time. Appropriate choice of i^* , which is equivalent to appropriate choice of the importance sampling density, has been proposed in Bhattacharya and Haslett (2007). For $i \in \{1, \dots, n\}$, a sub-sample of the realizations of θ_k (but not of \tilde{x}_{i^*}) of size M ($< N$) is selected without replacement with importance weights proportional to the ratio of $\pi(\tilde{x}_i, \theta_k | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k)$ and $\pi(\tilde{x}_{i^*}, \theta_k | \mathbf{X}_{n,-i^*}, \mathbf{Y}_{nm}, \mathcal{M}_k)$. For each member θ_k of the sub-sampled realizations, R realizations of \tilde{x}_i are generated using TMCMC from $\pi(\tilde{x}_i | \theta_k, \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k)$, to yield a total of $R \times M$ realizations from $\pi(\tilde{x}_i | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k)$.

In our examples, we generate 30,000 TMCMC samples from $\pi(\tilde{x}_{i^*}, \theta_k | \mathbf{X}_{n,-i^*}, \mathbf{Y}_{nm}, \mathcal{M}_k)$ of which we discard the first 10,000 as burn-in, and re-sample 1000 θ_k -realizations without replacement from the remaining 20,000 realizations with importance weights proportional to the ratio of $\pi(\tilde{x}_i, \theta_k | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k)$ and $\pi(\tilde{x}_{i^*}, \theta_k | \mathbf{X}_{n,-i^*}, \mathbf{Y}_{nm}, \mathcal{M}_k)$. For each re-sampled θ_k -value, we generate 100 TMCMC realizations of \tilde{x}_i . We discard the first 10,000 realizations of \tilde{x}_i as burn-in for the first re-sampled θ_k -realization, and for the subsequent θ_k -realizations, we set the final value of \tilde{x}_i of the previous value of θ_k as the initial value for \tilde{x}_i given the current θ_k -value, and continue TMCMC without any further burn-in. We thus obtain $1000 \times 100 = 100,000$ realizations of \tilde{x}_i for each $i = 1, \dots, n$. In all our examples, the above IRMCMC strategy, in conjunction with

efficient implementation of additive TMCMD, has led to excellent mixing properties.

Using the 100,000 IRMCMC samples, we obtain the posterior distribution of any given discrepancy measure $T^{(k)}(\tilde{\mathbf{X}}_n)$.

Obtaining the posterior model probabilities using Gibbs sampling

To obtain the posterior distribution of ζ , we first need to specify a prior for (p_1, \dots, p_K) . We consider the Dirichlet prior with parameters $(\alpha_1, \dots, \alpha_K)$, where $\alpha_k > 0$, for $k = 1, \dots, K$. Given ζ , the posterior distribution of (p_1, \dots, p_K) is again a Dirichlet distribution with parameters $(\alpha_1 + I(\zeta = 1), \dots, \alpha_K + I(\zeta = K))$. In other words,

$$\pi(p_1, \dots, p_K | \mathbf{X}_n, \mathbf{Y}_{nm}, \zeta) \equiv \text{Dirichlet}(\alpha_1 + I(\zeta = 1), \dots, \alpha_K + I(\zeta = K)). \quad (9.7.1)$$

Given (p_1, \dots, p_K) , the posterior distribution of ζ is given by (9.3.4), which is a function of the Bayes factors $BF^{(nm)}(\mathcal{M}_k, \mathcal{M}_{\tilde{k}})$; $k = 1, \dots, K$. In Chapter 8 we have shown that the corresponding pseudo-Bayes factors $PBF^{(nm)}(\mathcal{M}_k, \mathcal{M}_{\tilde{k}})$; $k = 1, \dots, K$, have the same asymptotic properties as the Bayes factors and are computationally far more efficient. Moreover, unlike Bayes factors, pseudo-Bayes factors do not suffer from Lindley's paradox. Thus, it seems reasonable to replace $BF^{(nm)}(\mathcal{M}_k, \mathcal{M}_{\tilde{k}})$ in (9.3.4) with the corresponding $PBF^{(nm)}(\mathcal{M}_k, \mathcal{M}_{\tilde{k}})$. In other words, we approximate the posterior probability $\pi(\zeta = k | \mathbf{X}_n, \mathbf{Y}_{nm}, p_1, \dots, p_K)$ as

$$\pi(\zeta = k | \mathbf{X}_n, \mathbf{Y}_{nm}, p_1, \dots, p_K) \approx \frac{p_k PBF^{(nm)}(\mathcal{M}_k, \mathcal{M}_{\tilde{k}})}{\sum_{\ell=1}^K p_\ell PBF^{(nm)}(\mathcal{M}_\ell, \mathcal{M}_{\tilde{k}})}; \quad k = 1, \dots, K. \quad (9.7.2)$$

Since the model probabilities are associated with the forward part, that is, where all the covariate values are treated as fixed, we consider the forward, or the traditional pseudo-Bayes factor in (9.7.2). In our examples, the values of $PBF^{(nm)}(\mathcal{M}_k, \mathcal{M}_{\tilde{k}})$; $k = 1, \dots, K$, are already available from Chapter 8 who provide estimates of $\frac{1}{n} \sum_{i=1}^n \log \pi(y_{i1} | \mathbf{Y}_{nm,-i}, \mathbf{X}_n, \mathcal{M}_k)$

in the second last column of Table 8.8.1. Note that

$$\frac{1}{n} \log PBF^{(nm)}(\mathcal{M}_k, \mathcal{M}_{\tilde{k}}) = \frac{1}{n} \sum_{i=1}^n \log \pi(y_{i1} | \mathbf{Y}_{nm,-i}, \mathbf{X}_n, \mathcal{M}_k) - \frac{1}{n} \sum_{i=1}^n \log \pi(y_{i1} | \mathbf{Y}_{nm,-i}, \mathbf{X}_n, \mathcal{M}_{\tilde{k}}).$$

Here $\tilde{k} = \arg \max_{k=1, \dots, K} \frac{1}{n} \sum_{i=1}^n \log \pi(y_{i1} | \mathbf{Y}_{nm,-i}, \mathbf{X}_n, \mathcal{M}_k)$.

Using the full conditional distributions (9.7.1) and (9.7.2), we obtain 100,000 realizations from the posterior distribution of (ζ, p_1, \dots, p_K) using Gibbs sampling, after discarding the first 10,000 iterations as burn-in.

Obtaining the posterior probabilities of the alternative hypotheses H_{1k}

Note that for $k = 1, \dots, K$, the posterior probability of H_{1k} is given by

$$\begin{aligned} v_{knm} &= 1 - \pi \left(\zeta = k, T^{(k)}(\tilde{\mathbf{X}}_n) - T^{(k)}(\mathbf{X}_n) \in [\tilde{\ell}_{knm}, \tilde{u}_{knm}] \mid \mathbf{X}_n, \mathbf{Y}_{nm} \right) \\ &= 1 - \pi \left(\zeta = k \mid \mathbf{X}_n, \mathbf{Y}_{nm} \right) \pi \left(T^{(k)}(\tilde{\mathbf{X}}_n) - T^{(k)}(\mathbf{X}_n) \in [\tilde{\ell}_{knm}, \tilde{u}_{knm}] \mid \zeta = k, \mathbf{X}_n, \mathbf{Y}_{nm} \right). \end{aligned} \quad (9.7.3)$$

Once we obtain realizations from the posteriors of $T^{(k)}(\tilde{\mathbf{X}}_n)$ for $k = 1, \dots, K$, and (ζ, p_1, \dots, p_K) , evaluation of the posterior probabilities of H_{1k} , denoted by v_{knm} ; $k = 1, \dots, K$, follows simply by Monte Carlo averaging associated with the two factors of (9.7.3).

9.7.2 Results of the simulation experiment for model selection

Non-misspecified situation

It is clear that for this experiment, $K = 6$, when no misspecification is considered. We set $\alpha_k = 1$; $k = 1, \dots, K$, for the parameters of the Dirichlet prior for (p_1, \dots, p_K) . That is, we assume a uniform prior distribution for (p_1, \dots, p_K) on the simplex. We report our results with respect to this prior, but our experiments with other values of $(\alpha_1, \dots, \alpha_K)$

did not yield different results.

For $n = m = 10$, the $cFDR_{nm}$ and $cFNR_{nm}$, for $\beta_{nm} \in [0.01, 0.99]$ are provided in Figure 9.7.1. The red and green colours correspond to $T_1^{(k)}(\tilde{\mathbf{X}}_n) - T_1^{(k)}(\mathbf{X}_n)$ and $T_2^{(k)}(\tilde{\mathbf{X}}_n) - T_2^{(k)}(\mathbf{X}_n)$, respectively. In the plots we denote these red and green coloured cFDRs as cFDR1 and cFDR2, respectively. Similarly, cFNR1 and cFNR2 denote the red and green coloured cFNRs. When $T_1^{(k)}(\tilde{\mathbf{X}}_n) - T_1^{(k)}(\mathbf{X}_n)$ is considered, $cFDR_{nm} = 0.024$ for $\beta_{nm} < 0.86$ and equals 9.023×10^{-6} for $\beta_{nm} \geq 0.86$. On the other hand, for $T_2^{(k)}(\tilde{\mathbf{X}}_n) - T_2^{(k)}(\mathbf{X}_n)$, $cFDR_{nm} = 0.087$ for $0.01 \leq \beta_{nm} < 0.48$ and falls to 5.444×10^{-5} for $0.48 \leq \beta_{nm} \leq 0.99$. In the first case, the multiple testing procedure selects H_{1k} for $k = 1, \dots, K$ when $0.01 \leq \beta_{nm} < 0.86$. When $0.86 < \beta_{nm} \leq 0.99$, the method selects $H_{0\tilde{k}}$ and H_{1k} for $k \neq \tilde{k}$. Here \tilde{k} corresponds to the true data-generating model, namely, the Poisson log-linear regression model. In the second case, all the alternative hypotheses are selected when $0.01 \leq \beta_{nm} < 0.48$; the true null and remaining alternative hypotheses are chosen for $0.48 \leq \beta_{nm} \leq 0.99$. Thus, for both the discrepancy measures, the correct model is selected for appropriate values of β_{nm} . However, cFDR2 falls close to zero much faster than cFDR1, and from the point onwards where the true decision occurs, cFNR2 is much lesser than cFNR1. These demonstrate that $T_2^{(k)}(\tilde{\mathbf{X}}_n) - T_2^{(k)}(\mathbf{X}_n)$ is a more efficient choice compared to $T_1^{(k)}(\tilde{\mathbf{X}}_n) - T_1^{(k)}(\mathbf{X}_n)$.

Here is an important point regarding comparison with our multiple testing result with that of inverse pseudo-Bayes factor reported in the last column of Table 8.8.1 of Chapter 8. The column shows that the inverse pseudo-Bayes factor identifies the true Poisson log-linear regression model as only the second best. However our multiple testing procedure correctly identifies the true model as the best one, for appropriate values of β_{nm} .

It is also important to remark in this context that the posterior probabilities of $T^{(k)}(\tilde{\mathbf{X}}_n) - T^{(k)}(\mathbf{X}_n) \in [\tilde{\ell}_{knm}, \tilde{u}_{knm}]$ when k is the true model, is significantly smaller than several other models. That the true model still turns out to be the best is due

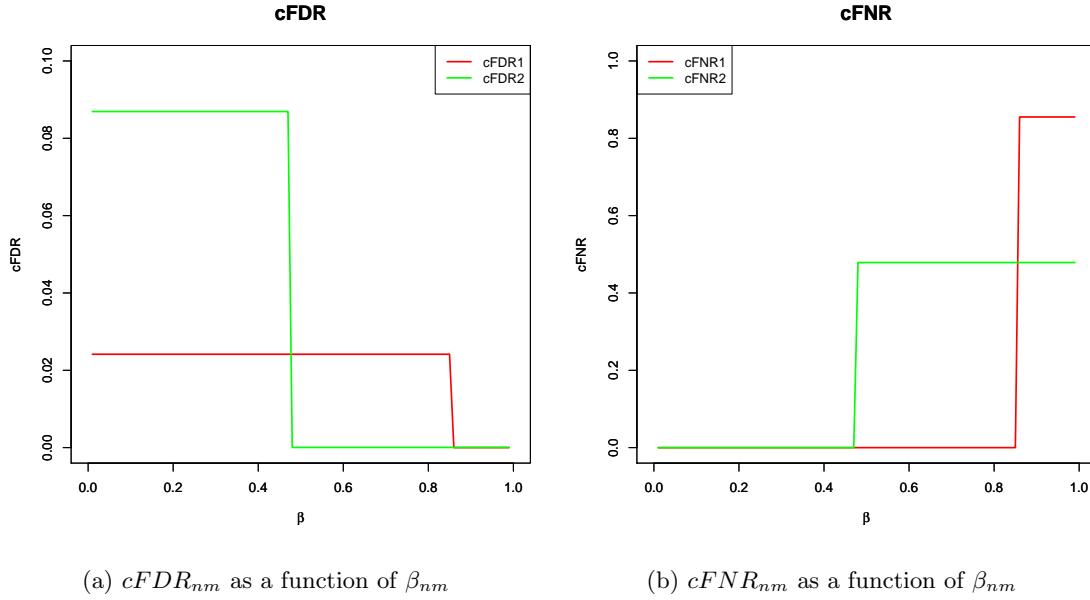


Figure 9.7.1: $cFDR_{nm}$ and $cFNR_{nm}$ as functions of β_{nm} in the non-misspecified case.

to its much larger posterior model probability compared to the others. The point is that even the true data-generating model need not have large posterior probabilities associated with the inverse discrepancy measure, and if the corresponding posterior model probability is not significantly large, then any other model can turn out to be the best on the basis of its stronger inverse perspective.

Misspecified situation

Let us now consider the case of misspecification, that is, when the true Poisson log-linear model is left out from consideration among the competing models. Thus, $K = 5$ in this case. The remaining setup is the same as in the non-misspecified scenario. Figure 9.7.2 display the cFDRs and cFNRs for this situation, each associated with both $T_1^{(k)}(\tilde{\mathbf{X}}_n) - T_1^{(k)}(\mathbf{X}_n)$ and $T_2^{(k)}(\tilde{\mathbf{X}}_n) - T_2^{(k)}(\mathbf{X}_n)$. In this case, for both the discrepancy measures, the correct decision, namely, the null hypothesis for the Poisson log-Gaussian

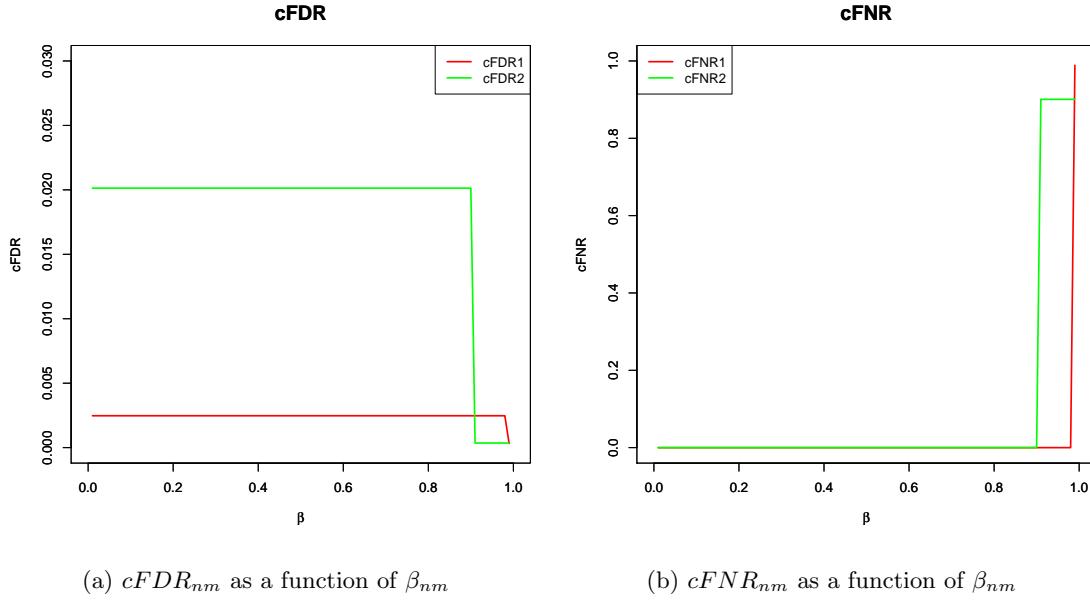


Figure 9.7.2: $cFDR_{nm}$ and $cFNR_{nm}$ as functions of β_{nm} in the misspecified case.

process and the alternative hypotheses for the remaining models, is reached for relatively large values of β_{nm} . Indeed, $cFDR1 = 0.002$ for $0.01 \leq \beta_{nm} < 0.99$ and 0.0003 for $\beta_{nm} = 0.99$ and $cFDR2 = 0.020$ for $0.01 \leq \beta_{nm} < 0.91$ and 0.0003 for $0.91 \leq \beta_{nm} \leq 0.99$. Again, $T_2^{(k)}(\tilde{\mathbf{X}}_n) - T_2^{(k)}(\mathbf{X}_n)$ performs better than $T_1^{(k)}(\tilde{\mathbf{X}}_n) - T_1^{(k)}(\mathbf{X}_n)$ in terms of faster decrease of $cFDR_{mn}$ towards zero and lesser value of $cFNR_{nm}$ once the right decision has been obtained.

Here the multiple testing procedure turns out to be consistent with both forward and inverse pseudo-Bayes factor, since the last two columns of Table 8.8.1 of Chapter 8 show that if the Poisson log-linear model is not considered among the competing models, then the Poisson log-Gaussian process model is the best. Here the corresponding posterior probability of $T^{(k)}(\tilde{\mathbf{X}}_n) - T^{(k)}(\mathbf{X}_n) \in [\tilde{\ell}_{knm}, \tilde{u}_{knm}]$ is higher than those of the other models, in addition to higher posterior model probability.

9.8 Second simulation study: variable selection in Poisson and geometric linear and nonparametric regression models when true model is Poisson linear regression

Here, the true and competing inverse regression models in the variable selection context, are as described in Chapter 8.8.4.

9.8.1 Discrepancy measure and Dirichlet prior parameters for more than one covariate

In models where both the covariates are considered, for any two n -dimensional vectors $\mathbf{v}_{1n} = (v_{11}, \dots, v_{1n})$ and $\mathbf{v}_n = (v_{21}, \dots, v_{2n})$, letting $\mathbf{v}_i = (v_{1i}, v_{2i})^T$, $\mathbf{V}_n = (\mathbf{v}_1, \dots, \mathbf{v}_n)$ and denoting the posterior mean vector and covariance matrix of $\tilde{\mathbf{u}}_i = (\tilde{x}_i, \tilde{z}_i)^T$ by $E_k(\tilde{\mathbf{u}}_i)$ and $Var_k(\tilde{\mathbf{u}}_i)$ respectively, for $i = 1, \dots, n$, we set

$$T_3^{(k)}(\mathbf{V}_n) = \frac{1}{n} \sum_{i=1}^n (\tilde{\mathbf{v}}_i - E_k(\tilde{\mathbf{u}}_i))^T (Var_k(\tilde{\mathbf{u}}_i) + c\mathbb{I})^{-1} (\tilde{\mathbf{v}}_i - E_k(\tilde{\mathbf{u}}_i)), \quad (9.8.1)$$

where $c > 0$ and \mathbb{I} is the identity matrix. Here $E_k(\tilde{\mathbf{u}}_i)$ and $Var_k(\tilde{\mathbf{u}}_i)$ correspond to the cross-validation posterior $\pi(\tilde{\mathbf{u}}_i | \mathbf{X}_{n,-i}, \mathbf{Y}_{nm}, \mathcal{M}_k)$.

In our experiment, as before we shall compare the results corresponding to $T_1^{(k)}(\tilde{\mathbf{W}}_n) - T_1^{(k)}(\mathbf{W}_n)$ and $T_2^{(k)}(\tilde{\mathbf{W}}_n) - T_2^{(k)}(\mathbf{W}_n)$, where $\tilde{\mathbf{W}}_n$ is either $\tilde{\mathbf{X}}_n$ or $\tilde{\mathbf{Z}}_n$ and \mathbf{W}_n is either \mathbf{X}_n or \mathbf{Z}_n . But for any inverse model that consists of both the covariates x and z , we replace both $T_1^{(k)}(\tilde{\mathbf{W}}_n) - T_1^{(k)}(\mathbf{W}_n)$ and $T_2^{(k)}(\tilde{\mathbf{W}}_n) - T_2^{(k)}(\mathbf{W}_n)$ with $T_3^{(k)}(\tilde{\mathbf{V}}_n) - T_3^{(k)}(\mathbf{V}_n)$, where $\tilde{\mathbf{v}}_i = (\tilde{x}_i, \tilde{z}_i)^T$, $\tilde{\mathbf{V}}_n = (\tilde{\mathbf{v}}_1, \dots, \tilde{\mathbf{v}}_n)$, $\mathbf{v}_i = (x_i, z_i)^T$ and $\mathbf{V}_n = (\mathbf{v}_1, \dots, \mathbf{v}_n)$.

For models having both x and z as covariates, the corresponding discrepancy measures $T_3^{(k)}(\tilde{\mathbf{V}}_n) - T_3^{(k)}(\mathbf{V}_n)$ are associated with joint cross-validation posterior distributions of $(\tilde{x}_i, \tilde{z}_i)$, and hence the corresponding posterior probabilities of the hypotheses are expected to be much smaller than posterior probabilities of the hypotheses of the models

with single covariates. We make amends for this by setting the parameters α_k of the Dirichlet prior for (p_1, \dots, p_K) for any model \mathcal{M}_k with both covariates to be 5 times that of the remaining parameters. So, in our case, we set $\alpha_k = 5$ for those k associated with both the covariates, and set the remaining parameters to 1.

Note that in this experiment, $K = 18$, including the true inverse Poisson log-linear regression model with both the covariates x and z . The implementation details remain the same as described in Section 9.7.1.

9.8.2 Results of our multiple testing experiment for model and variable selection

Non-misspecified situation

For $n = m = 10$, when the true model is Poisson with log-linear regression on both the covariates x and z , Figure 9.8.1 shows $cFDR_{nm}$ and $cFNR_{nm}$ as functions of β_{nm} . In this case $cFDR1$ decreases towards zero slightly faster than $cFDR2$. The numerical values of step functions $cFDR1$ and $cFDR2$ are provided as follows:

$$cFDR1 = \begin{cases} 0.025 & \text{if } 0.01 \leq \beta_{nm} < 0.67; \\ 0.007 & \text{if } 0.67 \leq \beta_{nm} < 0.91; \\ 0.001 & \text{if } 0.91 \leq \beta_{nm} < 0.99; \\ 6.214 \times 10^{-7} & \text{if } \beta_{nm} = 0.99 \end{cases} \quad (9.8.2)$$

and

$$cFDR2 = \begin{cases} 0.032 & \text{if } 0.01 \leq \beta_{nm} < 0.67; \\ 0.014 & \text{if } 0.67 \leq \beta_{nm} < 0.80; \\ 0.002 & \text{if } 0.80 \leq \beta_{nm} < 0.98; \\ 5.767 \times 10^{-6} & \text{if } 0.98 \leq \beta_{nm} \leq 0.99. \end{cases} \quad (9.8.3)$$

Note that the first change point for both cFDR1 and cFDR2 occurs at $\beta_{mn} = 0.67$, and at this point, we obtain the decision configuration that selects the null hypothesis of the true, Poisson log-linear model with both covariates x and z , and alternative hypotheses of all other models. For $\beta_{mn} < 0.67$, for all the models, the alternative hypotheses are selected. Thus, the first change point associated with both cFDR1 and cFDR2 yields the correct decision configuration. The next change points $\beta_{nm} = 0.91$ and $\beta_{nm} = 0.80$ for cFDR1 and cFDR2 are associated with selecting the null hypothesis for the model with the Poisson log-linear model with covariate x , in addition to the null hypothesis of the true, Poisson log-linear model with both covariates x and z . The final change points $\beta_{nm} = 0.99$ and $\beta_{nm} = 0.98$ yield the decision configurations that select the null hypothesis for the model with the Poisson log-linear model with covariate z , in addition to the previous null hypotheses. Thus, cFDR1 and cFDR2 behave quite consistently in this example and there seems to be no obvious reason for preferring one discrepancy measure to the other. Observe in Figure 9.8.1 that cFNR1 and cFNR2 are also quite consistently behaved.

Again the important observation is that our multiple testing procedure seems to easily identify the true inverse model, while neither forward nor inverse pseudo-Bayes factor successfully identified the true inverse model, as shown in the last two columns of Table 8.8.2 of Chapter 8. The second and third best models, namely, the Poisson log-linear model with covariate x and the Poisson log-linear model with covariate z , respectively, are however, consistent with forward and inverse pseudo-Bayes factor results reported in Chapter 8.

Again we find that the posterior probabilities of $T^{(k)}(\tilde{\mathbf{X}}_n) - T^{(k)}(\mathbf{X}_n) \in [\tilde{\ell}_{knm}, \tilde{u}_{knm}]$ when k is the true model, is significantly smaller than most of the other models, but its much higher posterior model probability compared to the others succeeds in making it the winner. The above inverse posterior probabilities for the second and third best models are also not higher than the remaining ones.

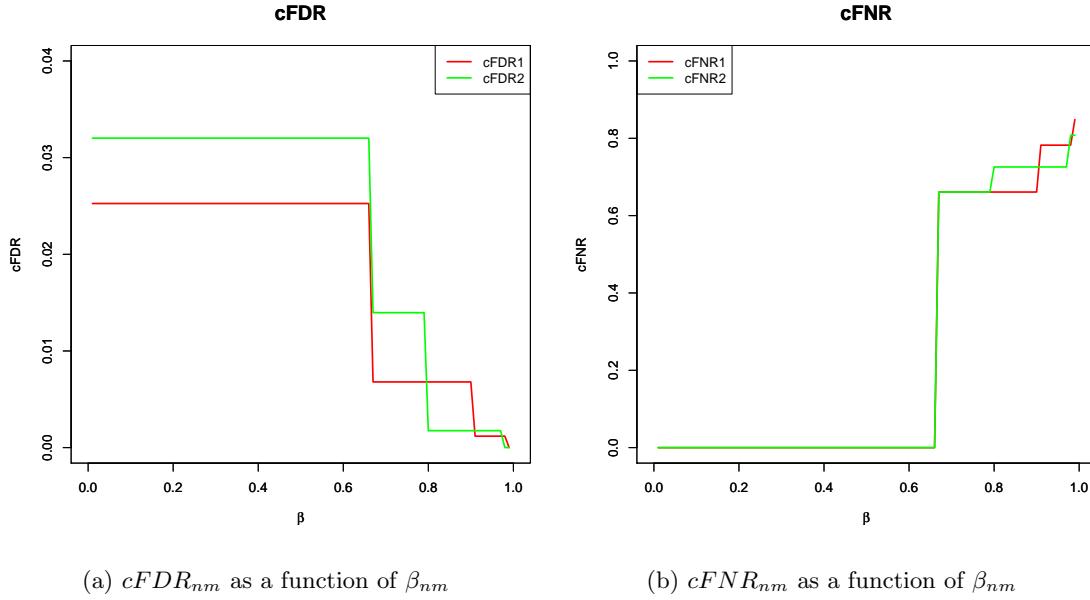


Figure 9.8.1: $cFDR_{nm}$ and $cFN R_{nm}$ as functions of β_{nm} in the non-misspecified situation of the model and variable selection problem.

Misspecified situation

In the misspecified situation we leave out the true Poisson log-linear model with both covariates x and z from among the competing models and implement our multiple testing procedure to obtain the best possible inverse models among the remaining ones. Figure 9.8.2 summarizes the results of our implementation in this direction. Both cFDR1 and cFDR2 yield the Poisson log-linear model with covariate x and the Poisson log-linear model with covariate z as the best and the next best inverse models, corresponding to the two change points observed in the graphs of cFDR1 and cFDR2. Recall that these were the second and the third best models in the non-misspecified situation, showing that our results for this misspecified case is very much coherent.

Observe that the best model in this case is detected by cFDR2 much earlier than cFDR1, and its value falls close to zero much earlier than that of cFDR1 in the process.

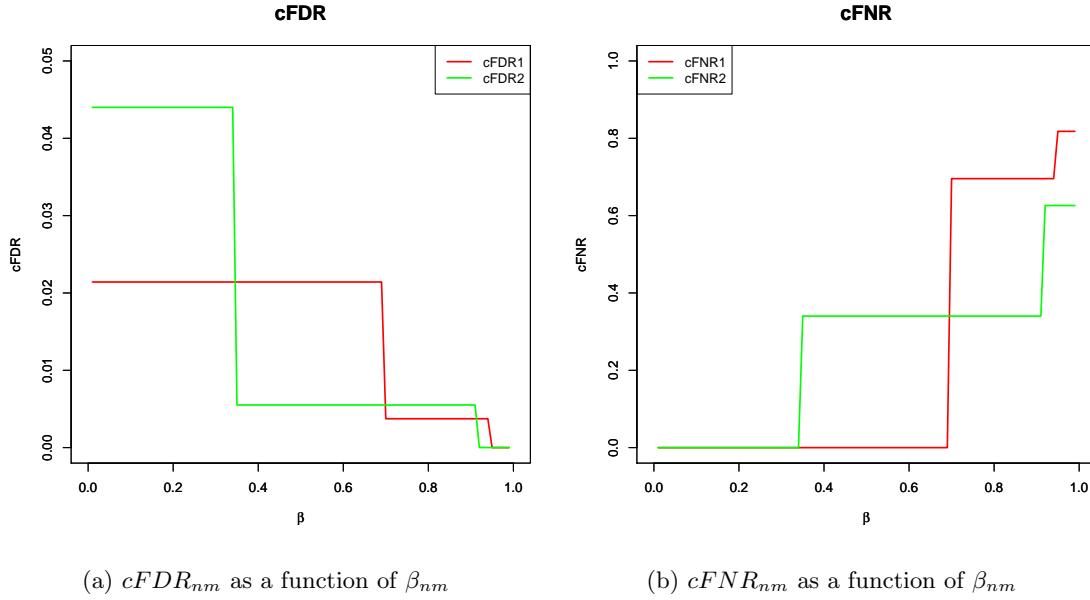


Figure 9.8.2: $cFDR_{nm}$ and $cFNR_{nm}$ as functions of β_{nm} in the misspecified situation of the model and variable selection problem.

The graphs for cFNR1 and cFNR2 show that at points where the best and the next best models are selected, cFNR2 is significantly smaller than cFNR1. Hence, in this misspecified situation, $T_2^{(k)}$ is again a better performer than $T_1^{(k)}$.

9.9 Summary and discussion

In this chapter we propose and develop a novel Bayesian multiple testing formulation for model and variable selection in inverse regression problems. Despite the relevance and elegance of the asymptotic theory, the real importance of our contribution lies in realistic, small sample situations where the inverse perspective of the competing models are expected to be most pronounced. The fast and efficient computational strategy that we employ for implementing our multiple testing procedure renders inverse model selection straightforward in the realistic finite sample context. Interestingly, the

forward pseudo-Bayes factor also features in our computational methodology, lending efficiency once it is available for the competing models. Most importantly, our simulation experiments demonstrate that our Bayesian multiple testing procedure can improve upon the results of both forward and inverse pseudo-Bayes factors.

Although we have exclusively considered the consistent prior for \tilde{x}_i developed in Chapter 6, at least for applications there is no bar to specifying any other sensible prior for \tilde{x}_i . Even though such priors need not lead to consistency of the inverse cross-validation posteriors, acceptable finite-sample based Bayesian inference can be obtained as in any other situations, for any $n > 1$ and $m \geq 1$.

Although we shall consider applications of our multiple testing procedure to various real data problems, let us present here some of our previous results on assessment of some palaeoclimate reconstruction models using the inverse reference distribution approach of [Bhattacharya \(2013\)](#) in the light of our new multiple testing strategy.

[Vasko *et al.* \(2000\)](#) reported a regular MCMC based inverse cross-validation exercise for a data set comprising multivariate counts y_i on $m = 52$ species of chironomid at $n = 62$ lakes (sites) in Finland. The unidimensional x_i denote mean July air temperature. As species respond differently to summer temperature, the variation in the composition provides the analyst with information on summer temperatures. This information is exploited to reconstruct past climates from count data derived from fossils in the lake sediment; see [Korhola *et al.* \(2002\)](#). The Bayesian model is a Multinomial-Dirichlet model for the species counts with a Gaussian response function of the species parameters. However, [Bhattacharya \(2013\)](#) showed that the posterior probabilities associated with the discrepancy measures T_1 and T_2 given by (9.1.1) and (9.1.2) were almost zero. [Bhattacharya \(2006\)](#) proposed an improved Bayesian model for the same dataset, by replacing the unimodal Gaussian response function with a Dirichlet process ([Ferguson \(1973\)](#)) based mixture of Gaussian functions, which very flexibly allows unknown number of climate preferences and tolerance levels for each species. Although this model brought

about marked improvement over that of Vasko *et al.* (2000) in terms of including significantly more x_i in the associated 95% highest posterior density credible intervals of the cross-validation posteriors, the posterior probabilities associated with T_1 and T_2 were still almost zero. A much improved palaeoclimate model was finally postulated by Mukhopadhyay and Bhattacharya (2013) by replacing the multinomial model with zero-inflated multinomial to account for excess zero species counts typically present in the data. The other features of the model are similar to that of Bhattacharya (2006). Not only does this model far surpasses the previous models in terms of including the percentage of x_i in the corresponding 95% highest posterior density credible intervals of the cross-validation posteriors (indeed, about 97% x_i are included in the respective intervals), inverse reference distributions for various discrepancy measures, including T_1 and T_2 , comfortably contain the observed discrepancy measures in their respective 95% highest posterior density credible intervals such that the relevant posterior probabilities associated with the discrepancy measures are significantly large. Recast in our multiple testing framework, the results show that irrespective of the posterior probabilities of the aforementioned three Bayesian models, the multiple testing method would select the model of Mukhopadhyay and Bhattacharya (2013) because of the overwhelming impact of its inverse regression part compared to the other two competing models.

In Haslett *et al.* (2006) pollen data was used, rather than chironomid data. The training data consisted of 7815 observations of two climate variables and 14 species of pollen. The model proposed by Haslett *et al.* (2006) is again a Multinomial-Dirichlet distribution, but the two-dimensional response surface is based on lattice Gaussian Markov Random Field (GMRF) (see, for example, Rue and Held (2005)) which is responsible for creation of a very large number of parameters. Indeed, their model consists of about 10,000 parameters. The other limitations of this model are summarized in Mukhopadhyay and Bhattacharya (2013). Applying the inverse reference distribution approach to this model and data Bhattacharya (2004) (Chapter 7) obtained almost

zero posterior probability of the inverse part. In fact, he demonstrated that this model overfits the pollen data; see also Mukhopadhyay and Bhattacharya (2013) who point out that such overfit is the consequence of the very large number of parameters and the GMRF assumption. The general zero-inflated Multinomial-Dirichlet model along with the Dirichlet process based bivariate Gaussian mixture model for the response functions proposed by Mukhopadhyay and Bhattacharya (2013) again turned out to be very successful in handling this pollen based palaeoclimate data. While including more than 94% of the two observed climate variables in their respective 95% highest posterior density credible intervals, the inverse reference distributions well-captured the observed discrepancy measures, so that again the posterior probability of the inverse part turned out to be emphatically pronounced. Thus, recast in our multiple testing paradigm, one can easily see that the zero-inflated Multinomial-Dirichlet model with the Dirichlet process based response function would emerge the clear winner.

10

How Ominous is the Future Global Warming Premonition?

10.1 Introduction

The gradual warming of the earth's average surface temperature, known as global warming, is perhaps the gravest concern for environmental scientists. Overwhelming evidence from multiple and independent sources of data have led the U.S. Global Change Research Program, the National Academy of Sciences, and the Intergovernmental Panel on Climate Change (IPCC) to independently conclude that global warming, particularly, in the recent decades, is undeniable. As per the records (see [IPCC \(2018\)](#), for example), compared to the pre-industrial baseline 1850 – 1900, the 2009 – 2015 time period was warmer by about 0.87°C , and that each decade is getting warmer by about 0.2°C . Such an alarming rate of increase is unprecedented, and even the prehistorical rates of global

warming, such as the Paleocene-Eocene Thermal Maximum, fail to match the current rate of global warming (see, for example, Masson-Delmotte *et al.* (2013)). However, see Idso *et al.* (2013b) and the references therein who argue, providing details on past temperature records, that this global warming is not unprecedented.

Such global warming is considered responsible for increasing droughts, heat waves, increase in extreme wet or dry events within the monsoon period in India and East Asia, increase in frequencies of hurricanes and typhoons, increase in global sea level as a result of melting glaciers, expansion of deserts and much more. According to the IPCC, “human influence on climate has been the dominant cause of observed warming since the mid-20th century”, and this conclusion has been upheld by all scientific bodies. In fact, human activities are estimated to have caused approximately 1.0°C of global warming above pre-industrial levels. Scientific investigations reveal that (see Olivier and Peters (2019)) the emission of greenhouse gases, with over 90% of the impact from carbon dioxide and methane, has been a major contributing factor to global warming by human activities such as fossil fuel burning, agricultural emissions and deforestation. But also see Idso *et al.* (2013b) who write “The empirical observations cited above reveal a relationship opposite of what is expected if carbon dioxide and methane were the powerful greenhouse gases the IPCC claims them to be. Clearly, if there is anything at all that is unusual, unnatural, or unprecedented about Earth’s current surface air temperature, it is that it is so *cold*.” and de Lange and Carter (2013) who mention in their key findings section “There appears to be nothing unusual about the extremes of wetness and dryness experienced during the twentieth century, or about recent changes in ocean circulation, sea level, or heat content, that would require atmospheric carbon dioxide forcing to be invoked as a causative factor. Natural variability in the frequency or intensity of precipitation extremes and sea-level change occurs largely on decadal and multidecadal time scales, and this variability cannot be discounted as a major cause of recent changes where they have occurred.”

The IPCC has warned that if the warming increases by 1.5°C compared to the pre-industrial era 1850 – 1900, then human and natural systems would be at grave risk. The concerning news is that under the current conditions global warming is projected to surpass 2.8°C by the year 2100 (see Climate Action Tracker (2019)).

The climate projections are performed by the general circulation models (GCMs) that attempt to model the major climate system components, namely, atmosphere, land surface, ocean and sea ice, and the interactions among them. Expressing great confidence in such models, the IPCC has claimed that (see Lupo *et al.* (2013)) “development of climate models has resulted in more realism in the representation of many quantities and aspects of the climate system,” adding, “it is extremely likely that human activities have caused more than half of the observed increase in global average surface temperature since the 1950s”. However, Lupo *et al.* (2013) write “Confidence in a model is further based on the careful evaluation of its performance, in which model output is compared against actual observations. A large portion of this chapter, therefore, is devoted to the evaluation of climate models against real-world climate and other biospheric data. That evaluation, summarized in the findings of numerous peer-reviewed scientific papers described in the different subsections of this chapter, reveals the IPCC is overestimating the ability of current state-of-the-art GCMs to accurately simulate both past and future climate. The IPCC’s stated confidence in the models, as presented at the beginning of this chapter, is likely exaggerated. The many and varied model deficiencies discussed in this chapter indicate much work remains to be done before model simulations can be treated with the level of confidence ascribed to them by the IPCC.” This was written seven years ago, and by now we expect the GCMs to have reduced their deficiencies and to yield more reliable climate projections.

The current GCM predictions by different GCMs available from the IPCC website http://www.ipcc-data.org/sim/gcm_global/index.html, under the assumptions of several future climate scenarios associated with greenhouse gas emissions, pertaining

to the Special Report on Emissions Scenarios (SRES), a report by the IPCC published in 2000. According to the IPCC Fourth Assessment Report (AR4), published in 2007, there are three SRES, namely, A1B, A2 and B1. Brief descriptions of the assumptions, obtained from the IPCC website, are reproduced below for the reader's convenience.

The key assumption for A1B is a future world of very rapid economic growth, low population growth and rapid introduction of new and more efficient technology. Major underlying themes are economic and cultural convergence and capacity building, with a substantial reduction in regional differences in per capita income. In this world, people pursue personal wealth rather than environmental quality.

SRES A2 corresponds to a very heterogeneous world. The underlying theme is that of strengthening regional cultural identities, with an emphasis on family values and local traditions, high population growth, and less concern for rapid economic development.

In SRES B1, a convergent world with the same global population as in the A1B is assumed but with rapid changes in economic structures toward a service and information economy, with reductions in materials intensity, and the introduction of clean and resource-efficient technologies.

Commitment is a non-SRES idealised scenario in which the atmospheric burdens of long-lived greenhouse gases are held fixed at AD2000 levels.

The scenarios A1B, A2, B1 and Commitment consist of 21, 17, 21 and 16 GCMs, respectively, each yielding a simulated global mean temperature time series in the duration 1900 – 2099. The HadCRUT4 observed near surface average global temperature dataset during the years 1850 – 2020 is also 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, we find reasons to doubt the reliability of the last few data points, and as such, we shall consider the dataset ranging from 1850 – 2016. This dataset pertains to temperature anomalies in degree celsius relative to the years 1961 – 1990. Now, the most widely quoted value for the global average temperature for the 1961 – 1990

period is 14°C, which has been developed by [Jones *et al.* \(1999\)](#). Hence, we convert the HadCRUT4 temperature anomalies data to (approximate) actual temperatures by adding 14°C to the anomalies. We also convert the GCM-simulated actual temperatures, originally available in Kelvin, to degree celsius.

Figure 10.1.1 presents the diagrams of the HadCRUT4 dataset (thick, black line) and the GCM predictions. Observe that the GCM based global temperatures seem to significantly underestimate the observed global temperatures during the years 1900 – 2016. Moreover, their rates of increase seem to be much faster than that of the observed dataset. Hence, the sharp increase of most of the GCM based future temperatures till the end of this century, is perhaps not unquestionable. Observe that the future predictions of the Commitment models are more stable compared to the others.

Perhaps the most important ingredient in any statistical learning is quantification of uncertainty. The GCM results displayed in Figure 10.1.1 are devoid of any uncertainty quantification; at least we are unable to find any in the IPCC website. In the observed HadCRUT4 data context, an ensemble of 100 time series are available, which has been recommended by climatologists to quantify uncertainty in the observations to some extent. It seems that ensembles can be obtained even for GCM models, provided they are run with different initial conditions. But the models are deterministically dynamic, and non-probabilistic, so that rigorous statistical ways of uncertainty quantification need not apply. It is thus not clear how believable the future global warming forecasts presented in Figure 10.1.1 are. In fact, as detailed in [Lupo *et al.* \(2013\)](#), the leading scientific experts have placed no faith in the GCMs. For instance, Freeman Dyson has written (see [Dyson \(2007\)](#)), “I have studied the climate models and I know what they can do. The models solve the equations of fluid dynamics, and they do a very good job of describing the fluid motions of the atmosphere and the oceans. They do a very poor job of describing the clouds, the dust, the chemistry, and the biology of fields and farms and forests. They do not begin to describe the real world that we live in”. [Green](#)

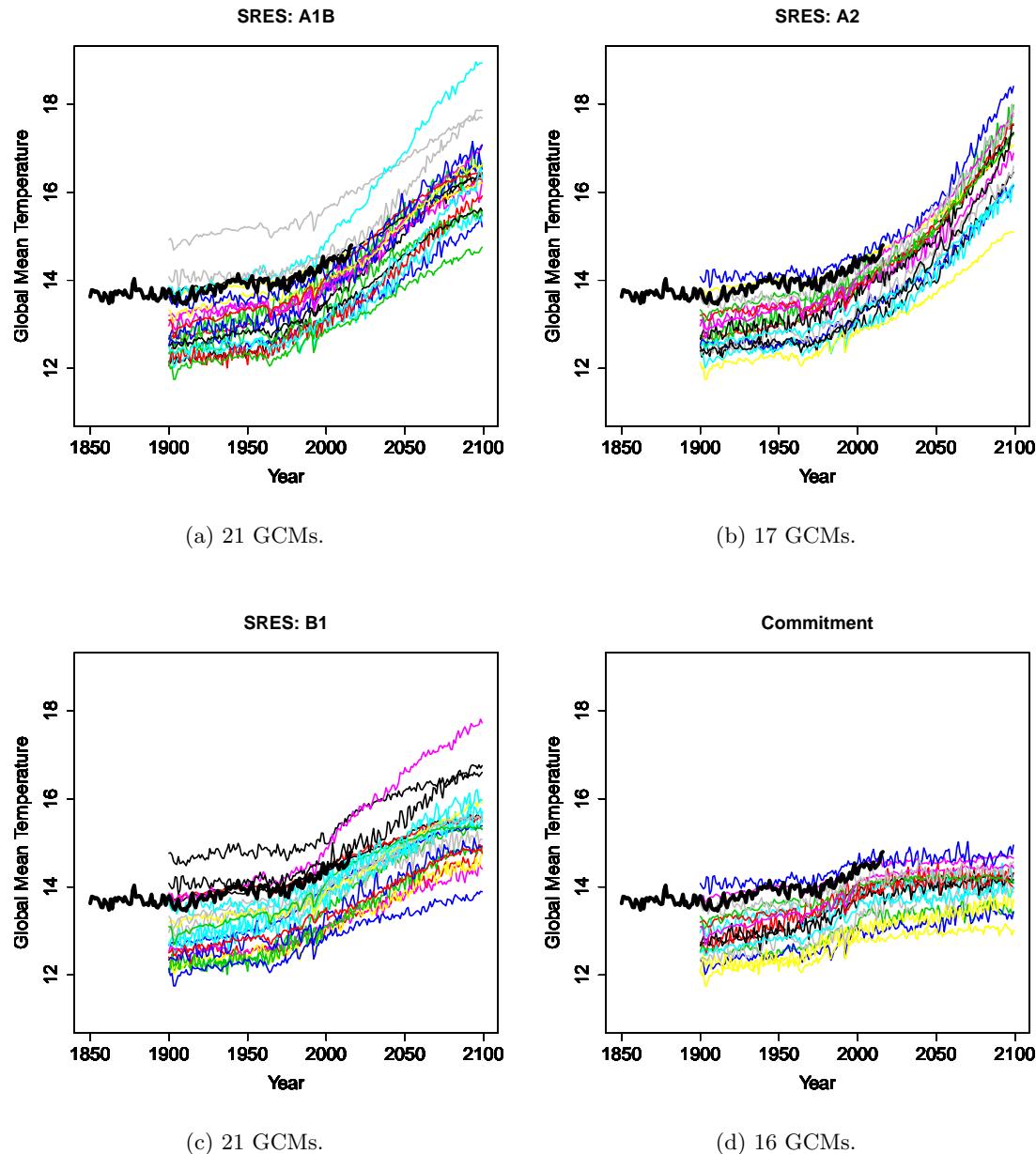


Figure 10.1.1: Visualization of the HadCRUT4 data (thick, black line) and the GCM based time series. The temperature is in $^{\circ}\text{C}$.

et al. (2009) tested whether the warming-trend forecasts used by the IPCC are more accurate than the standard benchmark forecast that there will be no change, using the historical HadCRUT3 observed dataset, which exhibited clear global warming till the present years. To their surprise, they found that the errors from the IPCC warming trend forecasts were nearly eight times greater than the errors from the no-change forecasts. Consequently, *Green et al.* (2009) recommend that the best policy is to do nothing about global warming.

The evaluation method of *Green et al.* (2009) was not based upon model based statistical or probabilistic methods and thus calls for more sophisticated analyses. In this work, we evaluate the global warming forecasts shown in Figure 10.1.1 in a rigorous footing using our recently-developed Bayesian methods. An important question in this regard is if the observed HadCRUT4 time series is plausible, given the GCM forecasts. This gives rise to an inverse regression problem in the following sense. The future temperature depends upon the present, our goal is to learn about the present, pretending it to be unknown, while the future is assumed to be known. Given each climate scenario, we then select the best GCM using our Bayesian multiple testing paradigm for model selection in inverse regression problems. The multiple testing procedure, it must be mentioned, not only considers the inverse aspect; it combines the inverse aspect with the forward in a coherent Bayesian compound decision theoretic sense, to compare the models under consideration. Once the best models are selected, we then show that even for such best GCMs, the Bayesian posterior time series for the current years (1850 – 2016) do not convincingly support the observed HadCRUT4 data, given the future forecasts for the years 2017 – 2099.

It is important to discern that the actual model for climate dynamics must be infeasibly complex and in fact unknown. Even the GCMs, which are complex computer models, are nothing but black boxes to us. The purpose of this discussion is to make it clear that standard time series models are inappropriate for climate dynamics. As such, we

consider modeling the logarithm of the global temperature at any year as a function of that at the previous year, plus some random error, where the function is assumed to be unknown and modeled appropriately by Gaussian process. The key idea has parallels with Bhattacharya (2007) and Ghosh *et al.* (2014).

Apart from the Bayesian model selection framework, we also treat the different GCM time series in any given climate scenario as an ensemble, and extend our univariate climate dynamics modeling to the multivariate situation, with multidimensional Gaussian processes replacing the previous one-dimensional Gaussian processes. The posterior distribution of the mean of the logarithm of the time series during 1850 – 2016, averaged over the dimensions (ensembles) in the corresponding climate scenario, is of interest in these cases. Our results in the multidimensional context very emphatically bear out that the HadCRUT4 data with its global warming trend must be highly implausible if the GCM forecasts are believed to be true.

Furthermore, given the observed HadCRUT4 data and our Gaussian process emulation model, we also provide Bayesian forecasts for the years 2017 – 2099, which show no evidence of drastic future global warming. Interestingly, as can be anticipated from panel (d) of Figure 10.1.1, only the forecasted time series by the best GCM model in the Commitment scenario fall in the high density regions of our Bayesian forecasted time series.

The general reader is likely to anticipate from the above discussions that computations associated with a study of such a proportion must be infeasibly complex. We assure this is not so. We wrote all our codes in the C language as efficiently as possible, parallelizing them using the Message Passing Interface (MPI) protocol whenever relevant, for example, in the case of the Bayesian multiple testing procedure. In such a case, we implemented the Gaussian process models associated with the large number of GCM forecasts in the parallel computing architecture (VMWare) available at our institution. Very efficient and time-saving computations are the results of our parallel processing. Details will be

presented in due course.

The rest of this chapter is structured as follows. In Section 10.2 we introduce our Gaussian process emulation model for climate dynamics, and discuss relevant prior choices in Section 10.3. The methods for Bayesian posterior inference regarding the current temperature time series given the future GCM simulations, and regarding future forecasts given the current temperature time series, are detailed in Section 10.4. In Section 10.5 we introduce our Bayesian multiple testing procedure in the context of best GCM selection in different climate scenarios, and provide details on our method of implementation in Section 10.6. The results of our best GCM selections and their detailed analyses are provided in Section 10.7. In Section 10.8, we model the ensemble of GCM-based future temperature time series in each climatic scenario as nonparametric multidimensional time series, driven by multidimensional Gaussian processes, and present the relevant theory and methods. The results and detailed analyses of our Bayesian multivariate Gaussian process emulation of climate dynamics are presented in Section 10.9. In Section 10.10 we forecast the future global temperature with our Bayesian Gaussian process approach, conditional on the HadCRUT4 data, and compare our results with the GCM forecasts as well as with the analysis of [Green *et al.* \(2009\)](#). Some discussion on existing works on climate model evaluation is provided in Section 10.11. Finally, in Section 10.12, we summarize our contributions, along with relevant discussions.

10.2 Gaussian process based emulation process for non-parametric climate dynamics

Let $\{x_t : t = 0, 1, 2, \dots\}$ denote a time series, here the logarithm of the global temperature time series. For time $t \geq 1$, we model x_t as

$$x_t = f_t(x_{t-1}) + \epsilon_t, \quad (10.2.1)$$

where $\epsilon_t \sim N(0, \sigma_\epsilon^2)$ independently, for $t \geq 1$. In this work, we assume that x_0 is known. Crucially, we assume that f_t is an unknown function dependent on time t . For any real z , we write $f_t(z) = f(t, z)$, where $f(\cdot)$ is considered an unknown function on $\mathbb{R}^+ \times \mathbb{R}$, which we shall model as a Gaussian process.

Using the notation $x_{t,u}^* = (t, x_u)$ following [Ghosh et al. \(2014\)](#), we re-write (10.2.1) as

$$x_t = f(x_{t,t-1}^*) + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_\epsilon^2) \text{ independently.} \quad (10.2.2)$$

Next, we provide details of the Gaussian process based model for f .

10.2.1 Modeling the unknown time-varying functions using Gaussian process

We model $f(\cdot)$ as a Gaussian processes with mean functions $\mu_f(\cdot) = \mathbf{h}(\cdot)' \boldsymbol{\beta}_f$ and with $\mathbf{h}(x^*) = (1, x^*)'$ for any x^* , and covariance function of the form $\sigma_f^2 c_f(\cdot, \cdot)$. Here σ_f^2 is the process variance and c_f is the correlations function. Typically, for any z_1, z_2 , $c_f(z_1, z_2) = \exp\{-(z_1 - z_2)' \mathbf{R}_f(z_1 - z_2)\}$, where \mathbf{R}_f is a 2×2 -dimensional diagonal matrix consisting of respective smoothness parameters $\{r_{1,f}, r_{2,f}\}$. These choices of the correlation functions imply that the functions, modeled by the process realizations, are infinitely smooth.

Our model is thus associated with the parameter set $\boldsymbol{\theta} = (\boldsymbol{\theta}_f, \sigma_\epsilon^2)$, where $\boldsymbol{\theta}_f =$

$(\boldsymbol{\beta}_f, \sigma_f^2, r_{1f}, r_{2f})$. The choice of the priors on the parameters will be discussed subsequently, but we shall assume that all the components of $\boldsymbol{\theta}$ are *a priori* independent. Henceforth, abusing notation, we shall denote densities and distributions using the notation $[\cdot]$ and conditional densities and distributions by $[\cdot|\cdot]$.

10.2.2 Hierarchical structure induced by our Gaussian process approach

Thus, for any $T > 1$, our modeling strategy can be described in the following hierarchical form, with GP standing for Gaussian process:

$$[x_t | f, \boldsymbol{\theta}_f, x_{t-1}] \sim N(f(x_{t,t-1}^*), \sigma_\epsilon^2); \quad t = 1, \dots, T; \quad (10.2.3)$$

$$[f(\cdot) | \boldsymbol{\theta}_f] \sim GP(\mathbf{h}(\cdot)' \boldsymbol{\beta}_f, \sigma_f^2 c_f(\cdot, \cdot)); \quad (10.2.4)$$

$$[\boldsymbol{\beta}_f, \sigma_f^2, \mathbf{R}_f, \sigma_\epsilon^2] = [\boldsymbol{\beta}_f][\sigma_f^2][r_{1f}][r_{2f}][\sigma_\epsilon^2], \quad (10.2.5)$$

where the components of $\boldsymbol{\beta}_f$ will also be considered independent *a priori*. Forms of the prior distributions in (10.2.5) are provided in Section 10.3.

10.2.3 Joint distribution of $\{x_t : t = 1, \dots, T\}$

Note that $[x_1 | x_0] \sim N(\mathbf{h}(x_0)' \boldsymbol{\beta}_f, \sigma_f^2 + \sigma_\epsilon^2)$, but $[x_2 | x_1, x_0] = [f(2, x_1) + \epsilon_2 | x_1, x_0] = [f(2, f(1, x_0) + \epsilon_1) + \epsilon_2 | f(1, x_0) + \epsilon_1, x_0]$. Hence, the conditional distribution of $[x_t | x_{t-1}, x_0]$, for $t \geq 2$, need not be straightforward to get hold of. In this regard, we adopt the procedure introduced by [Bhattacharya \(2007\)](#) which has also been successfully exploited in the nonparametric state-space modeling approach of [Ghosh *et al.* \(2014\)](#), to deal with this problem. The key idea is to conceptually simulate the entire function f modeled by the Gaussian process, and use the simulated process as a look-up table to obtain the conditional distributions of $\{x_t : t \geq 2\}$.

The key concept

For simplicity of illustration, let $x_t = f(x_{t,t-1}^*)$. Now consider a table with the first column $z \in \mathbb{R}^+ \times \mathbb{R}$ and the second column $f(z)$. Existence of this table hinges on the implicit assumption that the entire process $f(\cdot)$ is available. Given this table, conditional on $x_{t,t-1}^*$ (equivalently, conditional on x_{t-1}), $x_t = f(x_{t,t-1}^*)$ can be obtained by looking-up the input $x_{t,t-1}^*$ from the first column of the table and getting hold of the corresponding output value $f(x_{t,t-1}^*)$, located in the second column of the table. Thus, we refer to such a hypothetical table as a “look-up table”. In practice, we can construct a look-up table by simulating a realization of the Gaussian process f on a fine enough grid of inputs. Given this look-up table realization, simulation from the conditional distribution of $f(x_{t,t-1}^*)$, fixing $x_{t,t-1}^*$ as known, will approximate x_t as accurately as we desire by making the grid as fine as required, thanks to the well-known interpolation property of Gaussian processes. Formalization of this key concept leads to the following detailed steps.

Auxiliary variables for emulating the look-up table

Note that given x_0 we can simulate $x_1 = f(x_{1,0}^*) \sim N(\mathbf{h}(x_{1,0}^*)'\boldsymbol{\beta}_f, \sigma_f^2)$, the marginal distribution of the Gaussian process prior. To simulate the rest of the dynamic sequence, we first need to generate the rest of the process $\{f(x^*) : x^* \neq x_{1,0}^*\}$ for the look-up table approach.

In practice, it is not possible to have a simulation of this entire set $\{f(x^*) : x^* \neq x_{1,0}^*\}$. We only have available a set of grid points $\mathbf{G}_n = \{z_1, \dots, z_n\}$ obtained, perhaps, by Latin hypercube sampling (see, for example, [Santner et al. \(2003\)](#)) and a corresponding simulation of f , given by $\mathbf{D}_n^* = \{f(z_1), \dots, f(z_n)\}$, the latter having a joint multivariate normal distribution with mean

$$E[\mathbf{D}_n^* | \boldsymbol{\theta}_f] = \mathbf{H}_{D_n^*} \boldsymbol{\beta}_f \quad (10.2.6)$$

and covariance matrix

$$V[\mathbf{D}_n^* | \boldsymbol{\theta}_f] = \sigma_f^2 \mathbf{A}_{f,D_n^*}, \quad (10.2.7)$$

where $\mathbf{H}'_{D_n^*} = [\mathbf{h}(z_1), \dots, \mathbf{h}(z_n)]$ and \mathbf{A}_{f,D_n^*} is a correlation matrix with the (i,j) -th element $c_f(z_i, z_j)$.

Given $(x_0, f(x_{1,0}^*))$, we simulate \mathbf{D}_n^* from $[\mathbf{D}_n^* | \boldsymbol{\theta}_f, f(x_{1,0}^*), x_0]$. Note that the conditional $[\mathbf{D}_n^* | f(x_{1,0}^*), x_{1,0}^*]$ has an n -variate normal distribution with mean vector

$$E[\mathbf{D}_n^* | \boldsymbol{\theta}_f, f(x_{1,0}^*), x_0] = \boldsymbol{\mu}_{g,D_n^*} = \mathbf{H}_{D_n^*} \boldsymbol{\beta}_f + \mathbf{s}_{f,D_n^*}(x_{1,0}^*)(f(x_{1,0}^*) - \mathbf{h}(x_{1,0}^*)' \boldsymbol{\beta}_f) \quad (10.2.8)$$

and covariance matrix

$$V[\mathbf{D}_n^* | \boldsymbol{\theta}_f, f(x_{1,0}^*), x_0] = \sigma_f^2 \boldsymbol{\Sigma}_{f,D_n^*}, \quad (10.2.9)$$

where $\mathbf{s}_{f,D_n^*}(\cdot) = (c_f(\cdot, z_1), \dots, c_f(\cdot, z_n))'$ and

$$\boldsymbol{\Sigma}_{f,D_n^*} = \mathbf{A}_{f,D_n^*} - \mathbf{s}_{f,D_n^*}(x_{1,0}^*) \mathbf{s}_{f,D_n^*}(x_{1,0}^*)'. \quad (10.2.10)$$

Distribution of x_t given \mathbf{D}_n^*

Let us now deal with the conditional distribution $[x_t = f(x_{t,t-1}^*) | \mathbf{D}_n^*, x_{t-1}, x_{t-2}, \dots, x_1]$. Since the look-up table idea supports conditional independence, that is, given a simulation of the entire random function f , x_t depends only upon x_{t-1} via $x_t = f(x_{t,t-1}^*)$, it is sufficient to obtain the conditional distribution of $[f(x_{t,t-1}^*) | \mathbf{D}_n^*, x_{t-1}]$; see [Bhattacharya \(2007\)](#) and [Ghosh et al. \(2014\)](#) for detailed arguments. This distribution is of course normal with mean

$$\mu_t = \mathbf{h}(x_{t,t-1}^*)' \boldsymbol{\beta}_f + \mathbf{s}_{g,D_n^*}(x_{t,t-1}^*)' \mathbf{A}_{f,D_n^*}^{-1} (\mathbf{D}_n^* - \mathbf{H}_{D_n^*} \boldsymbol{\beta}_f) \quad (10.2.11)$$

and variance

$$\sigma_t^2 = \sigma_f^2 \left\{ 1 - \mathbf{s}_{f,D_n^*}(x_{t,t-1}^*)' \mathbf{A}_{f,D_n^*}^{-1} \mathbf{s}_{f,D_n^*}(x_{t,t-1}^*) \right\}. \quad (10.2.12)$$

For mathematical theory on the accuracy of the Markov approximation of the distributions of x_t given \mathbf{D}_n^* , see Ghosh *et al.* (2014).

Summary of the look-up table procedure

The look-up table idea involves the following steps, given that x_0 is known:

- (1) Draw $x_1 = f(x_{1,0}^*) \sim N(\mathbf{h}(x_{1,0}^*)' \boldsymbol{\beta}_f, \sigma_f^2)$.
- (2) Given x_0 , and $x_1 = f(x_{1,0}^*)$, draw $\mathbf{D}_n^* \sim [\mathbf{D}_n^* \mid \boldsymbol{\theta}_f, f(x_{1,0}^*), x_0]$.
- (3) For $t = 2, 3, \dots$, draw $x_t \sim [x_t = f(x_{t,t-1}^*) \mid \boldsymbol{\theta}_f, \mathbf{D}_n^*, x_{t-1}]$.

Joint distribution of $\{x_1, \dots, x_T, \mathbf{D}_n^*\}$

So far we have discussed the situations where $\epsilon_t = 0$, but our actual model (10.2.2) consists of non-zero ϵ_t which are normally distributed with mean zero and variance σ_ϵ^2 . In such case, once \mathbf{G}_n and \mathbf{D}_n^* are available, we write down the joint distribution of $\{x_1, \dots, x_T, \mathbf{D}_n^*\}$ conditional on the other parameters as

$$\begin{aligned} [x_1, \dots, x_T, \mathbf{D}_n^* \mid \boldsymbol{\theta}_f, \sigma_\epsilon^2] &= [x_1 = f(x_{1,0}^*) + \epsilon_1 \mid x_0, \sigma_\epsilon^2][\mathbf{D}_n^* \mid \boldsymbol{\theta}_f] \\ &\times \prod_{t=1}^{T-1} [x_{t+1} = f(x_{t+1,t}^*) + \epsilon_{t+1} \mid \mathbf{D}_n^*, x_t, \boldsymbol{\theta}_f, \sigma_\epsilon^2]. \end{aligned} \quad (10.2.13)$$

In (10.2.13), $[x_1 = f(x_{1,0}^*) + \epsilon_1 \mid x_0, \sigma_\epsilon^2] \sim N(\mathbf{h}(x_{1,0}^*)' \boldsymbol{\beta}_f, \sigma_f^2 + \sigma_\epsilon^2)$ and the distribution of \mathbf{D}_n^* is multivariate normal with mean and variance given by (10.2.6) and (10.2.7). The conditional distribution $[x_{t+1} = f(x_{t+1,t}^*) + \epsilon_{t+1} \mid \mathbf{D}_n^*, x_t, \boldsymbol{\theta}_f, \sigma_\epsilon^2]$ is normal with mean

$$\mu_{x_t} = \mathbf{h}(x_{t+1,t}^*)' \boldsymbol{\beta}_f + \mathbf{s}_{f,D_n^*}(x_{t+1,t}^*)' \mathbf{A}_{f,D_n^*}^{-1} (\mathbf{D}_n^* - \mathbf{H}_{D_n^*} \boldsymbol{\beta}_f) \quad (10.2.14)$$

and variance

$$\sigma_{x_t}^2 = \sigma_\epsilon^2 + \sigma_f^2 \left\{ 1 - \mathbf{s}_{f,D_n^*}(x_{t+1,t}^*)' \mathbf{A}_{f,D_n^*}^{-1} \mathbf{s}_{f,D_n^*}(x_{t+1,t}^*) \right\}. \quad (10.2.15)$$

Observe that in this case even if $x_{t+1,t}^* \in \mathbf{G}_n$, due to the presence of the additive error term ϵ_{t+1} , the conditional variance of x_{t+1} is non-zero, equalling $\sigma_{x_t}^2 = \sigma_\epsilon^2$, the error variance.

Non-Markovian dependence structure of $\{x_1, \dots, x_T\}$

Note that although conditionally on \mathbf{D}_n^* the variables x_t have a Markovian structure, if \mathbf{D}_n^* is integrated out from (10.2.13), then the marginalized distribution of $\{x_1, \dots, x_T\}$ is non-Markovian. In fact, the marginalized conditional distribution of x_{t+1} depends upon $\{x_k : k < t+1\}$; (see also Bhattacharya (2007) and Ghosh *et al.* (2014)). An important issue discussed in this context by Bhattacharya (2007) and Ghosh *et al.* (2014) is that this strong marginalized dependence structure is the root of all numerical instabilities associated with the model implementation. Essentially, by sample path continuity of the underlying Gaussian process, x_0, x_1, \dots, x_t will be often close to each other with high probability, particularly if σ_f^2 and σ_ϵ^2 are small. This would render the relevant correlation matrix almost singular, which would be difficult to invert. Since such inversions are required for every $t \in \{2, \dots, T\}$ and at every iteration of any Monte Carlo simulation method, progress would be almost impossible when T is relatively large, with increasing computational cost for each t further aggravating the situation.

In contrast, if \mathbf{D}_n^* is retained, it is required to deal with $[x_{t+1} | \mathbf{D}_n^*, x_t, \boldsymbol{\theta}_f, \sigma_f^2]$, which requires computation of $\mathbf{A}_{f,D_n^*}^{-1}$ only once, for all $t \geq 2$. This can be done even before beginning the simulation procedure. Moreover, invertibility of \mathbf{A}_{f,D_n^*} can be ensured by the user, since the (i,j) -th element of $\mathbf{A}_{f,D_n^*}^{-1}$ is of the form $c_f(z_i, z_j)$, where z_1, \dots, z_n are fixed constants, which can be judiciously chosen by the user to guarantee invertibility. Thus, retaining \mathbf{D}_n^* solves both the issues of numerical instability and computational

burden inherent in the marginalized distribution of $\{x_1, \dots, x_T\}$. It is hence no wonder that retaining \mathbf{D}_n^* in the model is the only sensible decision.

10.3 Prior distributions for θ_f and σ_ϵ^2

We assume the following forms of the prior distributions:

$$[\boldsymbol{\beta}_f] \sim N_3(\boldsymbol{\beta}_{f,0}, \boldsymbol{\Sigma}_{\beta_{f,0}}); \quad (10.3.1)$$

$$[\sigma_f^2] \propto (\sigma_f^2)^{-\left(\frac{\alpha_f+2}{2}\right)} \exp\left\{-\frac{\gamma_f}{2\sigma_f^2}\right\}; \quad \alpha_f, \gamma_f > 0; \quad (10.3.2)$$

$$[\sigma_\epsilon^2] \propto (\sigma_\epsilon^2)^{-\left(\frac{\alpha_\epsilon+2}{2}\right)} \exp\left\{-\frac{\gamma_\epsilon}{2\sigma_\epsilon^2}\right\}; \quad \alpha_\epsilon, \gamma_\epsilon > 0; \quad (10.3.3)$$

$$[\log(r_{i,f})] \sim N\left(\mu_{r_{i,f}}, \sigma_{r_{i,f}}^2\right); \quad \text{for } i = 1, 2. \quad (10.3.4)$$

All the prior parameters are assumed to be known. Now we discuss our approach to selecting the prior parameters for our application our Bayesian model in simulation studies and real data application in the univariate situations.

As per (10.3.1), we set the prior of $\boldsymbol{\beta}_f$ to be trivariate normal with the identity matrix as the variance, that is, we set $\boldsymbol{\Sigma}_{\beta_{f,0}} = \mathbf{I}_3$, where \mathbf{I}_3 is the 3-dimensional identity matrix. This choice turned out to be appropriate as larger variances in the diagonal caused the posterior time series to explode with increasing time. For the mean $\boldsymbol{\beta}_{f,0}$, except the first component associated with the intercept, we set the rest of the components to zero. We set the first component of $\boldsymbol{\beta}_{f,0}$ to be the mean of the underlying logarithm of the time series data to be modeled, after thinning by 5 observations. This ensures that the intercept corresponds to the overall mean of the log time series.

For the choice of the parameters of the priors of σ_f^2 and σ_ϵ^2 we first note that the mean is of the form $\gamma/(\alpha - 2)$ and the variance is of the form $2\gamma^2/\{(\alpha - 2)^2(\alpha - 4)\}$. Thus, if we set $\gamma/(\alpha - 2) = a$, then the variance becomes $2a^2/(\alpha - 4)$. Here we set $a = \hat{\sigma}^2/2$ for both σ_f^2 and σ_ϵ^2 , where $\hat{\sigma}^2$ is the variance of the underlying log time series obtained

after thinning by 5 observations. Again, this strategy is to ensure that the expected variability matches the data variability. For each of these priors we set $\alpha = 4.01$, so that the variance is of the form $200a^2$.

In order to choose the parameters of the log-normal priors of the smoothness parameters r_{1f} and r_{2f} , we set the mean of the log-normal prior with parameters μ and σ^2 , given by $\exp(\mu + \sigma^2/2)$, to 1. This yields $\mu = -\sigma^2/2$. Since the variance of this log-normal prior is given by $(\exp(\sigma^2) - 1) \exp(2\mu + \sigma^2)$, the relation $\mu = -\sigma^2/2$ implies that the variance is $\exp(\sigma^2) - 1 = \exp(-2\mu) - 1$. We set $\sigma^2 = 1$, so that $\mu = -0.5$. This implies that the mean is 1 and the variance is approximately 2, for the priors of each smoothness parameter $r_{i,f}$; $i = 1, 2$. This prevents the smoothness parameters from being too large or too small. Indeed, if the smoothness parameters are too large then $c_f(z_1, z_2) \approx 0$ for $z_1 \neq z_2$, so that the correlation matrix is rendered almost the identity matrix. On the other hand, if the smoothness parameters are close to zero, then $c_f(z_1, z_2) \approx 1$ for z_1, z_2 , making the correlation matrix almost singular. Both these undesirable situations are ruled out by our prior choice.

10.4 Posterior distributions of current and future time series in our dynamic Gaussian process approach

10.4.1 Posterior of current given the future

Let us assume that for any given GCM, the logarithms of the future mean global temperatures $\{x_t : t = T_0 + 1, \dots, T\}$ are given, where $1 \leq T_0 \leq T - 1$. In our case, the times $\{0, \dots, T_0\}$ correspond to the current years $\{1850, \dots, 2016\}$ and the times $\{T_0 + 1, \dots, T\}$ correspond to the future years $\{2017, \dots, 2099\}$. Then assuming that x_0 is known, we can obtain the posterior distribution of the logarithms of the current mean

global temperatures $\{x_t : t = 1, \dots, T_0\}$ as follows:

$$\begin{aligned} & [x_1, \dots, x_{T_0} | x_{T_0+1}, \dots, x_T] \\ &= \int [x_1, \dots, x_{T_0} | \mathbf{D}_n^*, x_{T_0+1}, \dots, x_T, \boldsymbol{\theta}_f, \sigma_\epsilon^2] d[\mathbf{D}_n^*, \boldsymbol{\theta}_f, \sigma_\epsilon^2 | x_{T_0+1}, \dots, x_T] \\ &\approx \int [x_1, \dots, x_{T_0} | \mathbf{D}_n^*, \boldsymbol{\theta}_f, \sigma_\epsilon^2] d[\mathbf{D}_n^*, \boldsymbol{\theta}_f, \sigma_\epsilon^2 | x_{T_0+1}, \dots, x_T]. \end{aligned} \quad (10.4.1)$$

The second approximate equality follows from the first equality since given \mathbf{D}_n^* , $\{x_1, \dots, x_{T_0}\}$ are conditionally approximately independent of $\{x_{T_0+1}, \dots, x_T\}$, “approximate” because x_{T_0} and x_{T_0+1} are not independent, even when \mathbf{D}_n^* is conditioned upon. This approximate conditional independence ensures $[x_1, \dots, x_{T_0} | \mathbf{D}_n^*, x_{T_0+1}, \dots, x_T, \boldsymbol{\theta}_f, \sigma_\epsilon^2] \approx [x_1, \dots, x_{T_0} | \mathbf{D}_n^*, \boldsymbol{\theta}_f, \sigma_\epsilon^2]$. In our practical applications, however, we shall replace this approximate equality with equality. For well-chosen fine enough grid \mathbf{G}_n this is not at all a serious issue.

Hence, if we can have simulations from the posterior $[\mathbf{D}_n^*, \boldsymbol{\theta}_f, \sigma_\epsilon^2 | x_{T_0+1}, \dots, x_T]$, then we can easily simulate from (10.4.1) using

$$[x_1, \dots, x_{T_0} | \mathbf{D}_n^*, \boldsymbol{\theta}_f, \sigma_\epsilon^2] = \prod_{t=0}^{T_0-1} [x_{t+1} = f(x_{t+1,t}^*) + \epsilon_{t+1} | \mathbf{D}_n^*, x_t, \boldsymbol{\theta}_f, \sigma_\epsilon^2],$$

where $[x_{t+1} = f(x_{t+1,t}^*) + \epsilon_{t+1} | \mathbf{D}_n^*, x_t, \boldsymbol{\theta}_f, \sigma_\epsilon^2]$ is normally distributed with mean and variance given by (10.2.14) and (10.2.15), respectively, for $t = 0, 1, \dots, T_0 - 1$.

To obtain samples from the posterior

$$\begin{aligned} & [\mathbf{D}_n^*, \boldsymbol{\theta}_f, \sigma_\epsilon^2 | x_{T_0+1}, \dots, x_T] \\ & \propto [\mathbf{D}_n^* | \boldsymbol{\theta}_f][\boldsymbol{\theta}_f][\sigma_\epsilon^2][x_{T_0+1}, \dots, x_T | \mathbf{D}_n^*, \boldsymbol{\theta}_f, \sigma_\epsilon^2] \\ &= [\mathbf{D}_n^* | \boldsymbol{\theta}_f][\boldsymbol{\theta}_f][\sigma_\epsilon^2] \prod_{t=T_0}^{T-1} [x_{t+1} = f(x_{t+1,t}^*) + \epsilon_{t+1} | \mathbf{D}_n^*, x_t, \boldsymbol{\theta}_f, \sigma_\epsilon^2], \end{aligned}$$

we resort to Markov Chain Monte Carlo (MCMC) where we sample $\boldsymbol{\beta}_f$ and \mathbf{D}_n^* from their

respective full conditional distributions and the remaining parameters $\{r_{1f}, r_{2f}, \sigma_f^2, \sigma_\epsilon^2\}$ using Transformation based Markov Chain Monte Carlo (TMCMC) introduced by Dutta and Bhattacharya (2014). In particular, we use the additive transformation, with judicious choice of the tuning constants.

10.4.2 Posterior of future given the current

Now, given $\{x_1, \dots, x_{T_0}\}$, which may be interpreted as the current observed log global mean temperatures, we can obtain the posterior distribution of the future log global mean temperatures $\{x_{T_0+1}, \dots, x_T\}$ in a similar manner. That is,

$$\begin{aligned} & [x_{T_0+1}, \dots, x_T | x_1, \dots, x_{T_0}] \\ &= \int [x_{T_0+1}, \dots, x_T | \mathbf{D}_n^*, x_1, \dots, x_{T_0}, \boldsymbol{\theta}_f, \sigma_\epsilon^2] d[\mathbf{D}_n^*, \boldsymbol{\theta}_f, \sigma_\epsilon^2 | x_1, \dots, x_{T_0}] \\ &= \int [x_{T_0+1}, \dots, x_T | \mathbf{D}_n^*, x_{T_0}, \boldsymbol{\theta}_f, \sigma_\epsilon^2] d[\mathbf{D}_n^*, \boldsymbol{\theta}_f, \sigma_\epsilon^2 | x_1, \dots, x_{T_0}]. \end{aligned} \quad (10.4.2)$$

Thus, after obtaining MCMC samples from

$$\begin{aligned} & [\mathbf{D}_n^*, \boldsymbol{\theta}_f, \sigma_\epsilon^2 | x_1, \dots, x_{T_0}] \\ & \propto [\mathbf{D}_n^* | \boldsymbol{\theta}_f][\boldsymbol{\theta}_f][\sigma_\epsilon^2][x_1, \dots, x_{T_0} | \mathbf{D}_n^*, \boldsymbol{\theta}_f, \sigma_\epsilon^2] \\ & = [\mathbf{D}_n^* | \boldsymbol{\theta}_f][\boldsymbol{\theta}_f][\sigma_\epsilon^2] \prod_{t=0}^{T_0-1} [x_{t+1} = f(x_{t+1,t}^*) + \epsilon_{t+1} | \mathbf{D}_n^*, x_t, \boldsymbol{\theta}_f, \sigma_\epsilon^2] \end{aligned}$$

using the same techniques as for $[\mathbf{D}_n^*, \boldsymbol{\theta}_f, \sigma_\epsilon^2 | x_{T_0+1}, \dots, x_T]$, we simulate from

$$[x_{T_0+1}, \dots, x_T | \mathbf{D}_n^*, x_{T_0}, \boldsymbol{\theta}_f, \sigma_\epsilon^2] = \prod_{t=T_0}^{T-1} [x_{t+1} = f(x_{t+1,t}^*) + \epsilon_{t+1} | \mathbf{D}_n^*, x_t, \boldsymbol{\theta}_f, \sigma_\epsilon^2],$$

where $[x_{t+1} = f(x_{t+1,t}^*) + \epsilon_{t+1} | \mathbf{D}_n^*, x_t, \boldsymbol{\theta}_f, \sigma_\epsilon^2]$ is normally distributed with mean and variance given by (10.2.14) and (10.2.15), respectively, for $t = T_0, 1, \dots, T - 1$. This

yields simulations from (10.4.2).

10.5 A Bayesian multiple testing framework for GCM selection in any given climate scenario

Given any climate scenario, let us consider GCMs \mathcal{M}_k ; $k = 1, \dots, K$, from among which the best model needs to be selected. For our purpose, we adopt and extend the novel Bayesian multiple testing procedure for model selection introduced in Chapter 9 that respects the inverse regression perspective of the models, in coherence with the forward aspect.

It is important to mention that in statistics, model selection pertains to choosing the best model from among a set of models that attempt to fit a single dataset. However, in our present GCM case, there are K datasets generated by K GCMs in a given climate scenario. Our strategy will be to combine the K datasets into a single dataset by taking averages over the K GCMs for each time point, and then to invoke our Gaussian process based dynamics for the averaged time series, where the hyperparameters of the model are fixed using the mean and variance of the original GCM-specific simulated time series. This yields K different Gaussian process based models for the averaged time series, inheriting the main characteristics of the GCM-specific time series. The design of our Bayesian multiple testing procedure ensures that the Gaussian process models will be compared with respect to their abilities to fit the averaged simulated future global temperature data in the forward sense, as well as their abilities to capture the HadCRUT4 data given the averaged GCM-simulated future global temperature data, in the inverse sense. Details follow.

Let us denote the logarithms of the observed current global mean temperatures (the HadCRUT4 data) by $\{x_t^{(0)} : t = 1, \dots, T_0\}$. For GCM \mathcal{M}_k , let $\{x_t^{(k)} : t = 0, 1, \dots\}$ denote the logarithms of its simulated global mean temperature time series, for $k =$

$1, \dots, K$. For $t = 0, 1, \dots$, let $\bar{x}_t = K^{-1} \sum_{k=1}^K x_t^{(k)}$, and let this averaged time series $\{\bar{x}_t : t = 0, 1, \dots\}$ be also modeled by the Gaussian process emulation procedure given by (10.2.3), (10.2.4) and (10.2.5), with parameters denoted by $\boldsymbol{\theta}^{(k)} = (\boldsymbol{\theta}_f^{(k)}, \sigma_\epsilon^{(k)2})$. The rationale behind this modeling strategy is simple: if the functional forms $f(\cdot)$ associated with the individual time series $\{x_t^{(k)} : t = 0, 1, \dots\}$ are unknown, then the functional form driving the dynamics of their average must also be unknown, which is again best modeled by a Gaussian process. In this regard, let $[\bar{x}_{T_0+1}, \dots, \bar{x}_T | \boldsymbol{\theta}^{(k)}, \mathcal{M}_k]$ denote the density of the logarithms of the future global mean temperatures, averaged over all the models in the climate scenario under Gaussian process emulation model \mathcal{M}_k , with its associated parameters $\boldsymbol{\theta}^{(k)}$.

We combine the competing models in the following mixture form:

$$[\bar{x}_{T_0+1}, \dots, \bar{x}_T | \boldsymbol{\theta}] = \sum_{k=1}^K p_k [\bar{x}_{T_0+1}, \dots, \bar{x}_T | \boldsymbol{\theta}^{(k)}, \mathcal{M}_k], \quad (10.5.1)$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(K)})$, $0 \leq p_k \leq 1$, for $k = 1, \dots, K$ and $\sum_{k=1}^K p_k = 1$. Letting ζ denote the allocation variable (model index), with $P(\zeta = k) = p_k$, note that $[\bar{x}_{T_0+1}, \dots, \bar{x}_T | x_1^{(0)}, \dots, x_{T_0}^{(0)}, \boldsymbol{\theta}, \zeta = k] = [\bar{x}_{T_0+1}, \dots, \bar{x}_T | x_1^{(0)}, \dots, x_{T_0}^{(0)}, \boldsymbol{\theta}^{(k)}, \mathcal{M}_k]$. We consider the Dirichlet prior for (p_1, \dots, p_K) with parameters $(\alpha_1, \dots, \alpha_K)$, where $\alpha_k > 0$, for $k = 1, \dots, K$. In our problem, we shall set $\alpha_k = 1$, for all $k = 1, \dots, K$, for all the climate scenarios. Thus, the prior is uniform over the simplex, indicating no preference for any specific GCM *a priori*. The priors for the parameters $\boldsymbol{\theta}^{(k)}$ remain the same as described in Section 10.3. Since for different k the prior depends upon the mean and variance of the underlying entire k -th GCM-simulated time series, the priors are all very distinct from one another. In fact, the distinctions among the priors induces distinctions among the competing Bayesian models, since otherwise all of them have the same dynamic structure driven by Gaussian processes, started at the same known initial value x_0 .

We let $\{\bar{x}_t : t = 1, \dots, T_0\}$ stand for the random quantities corresponding to $\{x_t^{(0)} : t = 1, \dots, T_0\}$, whose posterior distribution will be of interest to us. In particular, it is of interest in evaluating how well this posterior captures the observed current log global mean temperatures, which we shall formalize in our multiple testing procedure. Towards this goal, for any T_0 -dimensional vector $\mathbf{v}_{T_0} = (v_1, \dots, v_{T_0})$, and for some $c > 0$, let us define the following discrepancy measures in the spirit of Chapter 9:

$$S_1^{(k)}(\mathbf{v}_{T_0}) = \frac{1}{T_0} \sum_{t=1}^{T_0} \frac{|v_t - M(\bar{x}_t | \bar{x}_{T_0+1}, \dots, \bar{x}_T, \mathcal{M}_k)|}{\sqrt{Var(\bar{x}_t | \bar{x}_{T_0+1}, \dots, \bar{x}_T, \mathcal{M}_k)} + c}, \quad (10.5.2)$$

where $M(\bar{x}_t | \bar{x}_{T_0+1}, \dots, \bar{x}_T, \mathcal{M}_k)$ stands for the posterior mode of $[\bar{x}_t | \bar{x}_{T_0+1}, \dots, \bar{x}_T, \mathcal{M}_k]$.

Similarly, let

$$S_1^{(k)}(\mathbf{v}_{T_0}) = \frac{1}{T_0} \sum_{t=1}^{T_0} \frac{(v_t - M(\bar{x}_t | \bar{x}_{T_0+1}, \dots, \bar{x}_T, \mathcal{M}_k))^2}{Var(\bar{x}_t | \bar{x}_{T_0+1}, \dots, \bar{x}_T, \mathcal{M}_k) + c}. \quad (10.5.3)$$

In our examples, we set $c = 0.01$. Although various other measures of discrepancy can be defined, for brevity we focus on these two measures in this work. Importantly, using the discrepancy measures and the IRD approach to Bayesian model assessment in inverse regression problems, we shall assess goodness-of-fit of the best GCMs with respect to fitting the HadCRUT4 data, conditioned on the future GCM projections and our Bayesian dynamic Gaussian process emulation strategy.

With $\bar{\mathbf{x}}_{T_0} = (\bar{x}_1, \dots, \bar{x}_{T_0})$ and $\mathbf{x}_{T_0}^{(0)} = (x_1^{(0)}, \dots, x_{T_0}^{(0)})$, for a given discrepancy measure $S^{(k)}$, let $[\bar{\ell}_k, \bar{u}_k]$ denote the $100(1 - \alpha)\%$ credible interval for the posterior distribution of $S^{(k)}(\bar{\mathbf{x}}_{T_0})$ for any desired $\alpha \in (0, 1)$; in our application, we set $\alpha = 0.05$. Following the recommendation in Section 9.6 for practical purposes we now define the appropriate multiple hypotheses that we shall test for our Bayesian model selection purpose. For $k = 1, \dots, K$,

$$H_{0k} : \zeta = k, S^{(k)}(\bar{\mathbf{x}}_{T_0}) - S^{(k)}(\mathbf{x}_{T_0}^{(0)}) \in [\bar{\ell}_k, \bar{u}_k] \quad (10.5.4)$$

versus

$$H_{1k} : \{\zeta \neq k\} \bigcup \left\{ \zeta = k, S^{(k)}(\bar{x}_{T_0}) - S^{(k)}(\mathbf{x}_{T_0}^{(0)}) \in [\bar{\ell}_k, \bar{u}_k]^c \right\}, \quad (10.5.5)$$

where, for any set A , A^c stands for its complement.

The hypotheses are so designed that the best model is chosen on the basis of both forward and inverse perspectives. To elucidate, note that to select the best model we first need to choose a model $[\bar{x}_{T_0+1}, \dots, \bar{x}_T | \boldsymbol{\theta}^{(k)}, \mathcal{M}_k]$ indexed by $\zeta = k$ which has high marginal posterior probability. This reflects the forward perspective of the model selection problem. Indeed, the posterior probability of $\{\zeta = k\}$ is proportional to its corresponding marginal density $[\bar{x}_{T_0+1}, \dots, \bar{x}_T | \mathcal{M}_k] = \int [\bar{x}_{T_0+1}, \dots, \bar{x}_T | \boldsymbol{\theta}^{(k)}, \mathcal{M}_k] d[\boldsymbol{\theta}^{(k)}]$ (see (10.6.3) for details). This marginal density has interpretation in the forward sense only since it is not associated with the posterior distribution $[\bar{x}_1, \dots, \bar{x}_{T_0} | \bar{x}_{T_0+1}, \dots, \bar{x}_T, \mathcal{M}_k]$, the latter to be interpreted as the inverse aspect of the problem.

The inverse sense in our multiple testing formalization is made explicit in the following way. In addition to selecting $\zeta = k$ with high marginal posterior probability, we demand that for such model

$$S^{(k)}(\bar{x}_{T_0}) - S^{(k)}(\mathbf{x}_{T_0}^{(0)}) \in [\bar{\ell}_k, \bar{u}_k] \quad (10.5.6)$$

is also satisfied. Roughly, this condition demands that for \mathcal{M}_k to qualify as a good inverse regression model, the observed discrepancy measure $S^{(k)}(\mathbf{x}_{T_0}^{(0)})$ must be included in the desired credible intervals of the reference discrepancy measure $S^{(k)}(\bar{x}_{T_0})$. This reflects the inverse perspective since the reference discrepancy measure explicitly deals with the posterior $[\bar{x}_1, \dots, \bar{x}_{T_0} | \bar{x}_{T_0+1}, \dots, \bar{x}_T, \mathcal{M}_k]$ associated with the inverse regression problem. The key idea of the Bayesian goodness-of-fit test devised by [Bhattacharya \(2013\)](#) is based on the same principle.

Note that our Bayesian multiple hypotheses formulation (10.5.4) and (10.5.5) does not involve cross-validation, even though in Chapter 9 we formulated the general Bayesian

multiple testing framework for model and variable selection in problems involving covariates using inverse leave-one-out cross-validation with respect to posteriors associated with the covariates (see also [Bhattacharya \(2013\)](#)). Indeed, as must be evident from the very beginning, our current global climate change problem is not the traditional model selection problem. However, our Bayesian multiple testing procedure is based on similar principles introduced in Chapter 9.

10.5.1 The Bayesian multiple testing procedure

Let

$$d_k = \begin{cases} 1 & \text{if the } k\text{-th null hypothesis is rejected;} \\ 0 & \text{otherwise;} \end{cases}$$

$$r_k = \begin{cases} 1 & \text{if } H_{1k} \text{ is true;} \\ 0 & \text{if } H_{0k} \text{ is true.} \end{cases}$$

Following Chapter 9, let us define the true positives as

$$TP = \sum_{k=1}^K d_k r_k, \quad (10.5.7)$$

the posterior expectation of which is to be maximized subject to controlling the posterior expectation of the error term

$$E = \sum_{k=1}^K d_k (1 - r_k). \quad (10.5.8)$$

From the above notions it is clear that the optimal decision configuration can be obtained by minimizing the function

$$\begin{aligned}\xi(\mathbf{d}) &= -\sum_{k=1}^K d_k E(r_k | \bar{x}_{T_0+1}, \dots, \bar{x}_T) + \lambda \sum_{k=1}^K d_k E[(1 - r_k) | \bar{x}_{T_0+1}, \dots, \bar{x}_T] \\ &= -(1 + \lambda) \sum_{k=1}^K d_k \left(v_k - \frac{\lambda}{1 + \lambda} \right),\end{aligned}$$

with respect to all possible decision configurations of the form $\mathbf{d} = \{d_1, \dots, d_K\}$, where $\lambda > 0$, and

$$v_k = E(r_k | \bar{x}_{T_0+1}, \dots, \bar{x}_T) = [H_{1k} | \bar{x}_{T_0+1}, \dots, \bar{x}_T]$$

is the posterior probability of the k -th alternative hypothesis. Letting $\beta = \lambda/(1 + \lambda)$ denote the penalizing constant, one can equivalently maximize

$$f_\beta(\mathbf{d}) = \sum_{k=1}^K d_k (v_k - \beta) \tag{10.5.9}$$

with respect to \mathbf{d} and obtain the optimal decision configuration. In this case, the optimal decision configuration $\hat{\mathbf{d}} = \{\hat{d}_1, \dots, \hat{d}_K\}$ is given by the following: for $k = 1, \dots, K$,

$$\hat{d}_k = \begin{cases} 1 & \text{if } v_k > \beta; \\ 0 & \text{otherwise.} \end{cases} \tag{10.5.10}$$

In our model selection setup, the least value of the penalty $\beta \in (0, 1)$ for which the decision configuration $\hat{d}_{\tilde{k}} = 0$ and $\hat{d}_k = 1$ for all $k \in \{1, \dots, K\} \setminus \{\tilde{k}\}$ is obtained, for some $\tilde{k} \in \{1, \dots, K\}$, yields the best model $\mathcal{M}_{\tilde{k}}$. This is because in such a case, $v_{\tilde{k}} \leq \beta$, even though β is reasonably small, suggesting that $H_{0\tilde{k}}$ has significant posterior probability. Since $v_k > \beta$ for all $k \in \{1, \dots, K\} \setminus \{\tilde{k}\}$, the posterior probabilities of H_{0k} for $k \in \{1, \dots, K\} \setminus \{\tilde{k}\}$ are less substantial compared to that of $H_{0\tilde{k}}$. This indicates

that $\mathcal{M}_{\tilde{k}}$ is the best model among \tilde{M}_k ; $k = 1, \dots, K$. This key intuition is rigorously formalized in our Bayesian multiple testing procedure.

10.5.2 Error measures for our Bayesian multiple testing procedure

To discuss appropriate measures of error for our Bayesian multiple testing procedure, first let us define $\delta(\mathbf{d}|\bar{x}_{T_0+1}, \dots, \bar{x}_T)$ to be the probability of choosing \mathbf{d} as the optimal decision configuration given data $\bar{x}_{T_0+1}, \dots, \bar{x}_T$ when a given multiple testing method is employed. Also, let \mathbb{D} be the set of all K -dimensional binary vectors, standing for all possible decision configurations.

As suitable posterior measures of Type-I and Type-II errors, [Sarkar et al. \(2008\)](#) defined posterior false discovery rate and false non-discovery rate, respectively, which we denote as conditional false discovery rate (cFDR) and conditional false non-discovery rate (cFNR). The measures, in our current setup, are given as the following:

$$\begin{aligned} cFDR &= E \left[\sum_{\mathbf{d} \in \mathbb{D}} \frac{\sum_{k=1}^K d_k(1 - r_k)}{\sum_{k=1}^K d_k \vee 1} \delta(\mathbf{d}|\bar{x}_{T_0+1}, \dots, \bar{x}_T) \middle| \bar{x}_{T_0+1}, \dots, \bar{x}_T \right] \\ &= \sum_{\mathbf{d} \in \mathbb{D}} \frac{\sum_{k=1}^K d_k(1 - v_k)}{\sum_{k=1}^K d_k \vee 1} \delta(\mathbf{d}|\bar{x}_{T_0+1}, \dots, \bar{x}_T); \\ cFNR &= E \left[\sum_{\mathbf{d} \in \mathbb{D}} \frac{\sum_{k=1}^K (1 - d_k)r_k}{\sum_{k=1}^K (1 - d_k) \vee 1} \delta(\mathbf{d}|\bar{x}_{T_0+1}, \dots, \bar{x}_T) \middle| \bar{x}_{T_0+1}, \dots, \bar{x}_T \right] \\ &= \sum_{\mathbf{d} \in \mathbb{D}} \frac{\sum_{k=1}^K (1 - d_k)v_k}{\sum_{k=1}^K (1 - d_k) \vee 1} \delta(\mathbf{d}|\bar{x}_{T_0+1}, \dots, \bar{x}_T). \end{aligned}$$

Note that since in our multiple testing method the decision rule is non-randomized, $\delta(\mathbf{d}|\bar{x}_{T_0+1}, \dots, \bar{x}_T)$ is either 1 or 0 depending on data $\{\bar{x}_{T_0+1}, \dots, \bar{x}_T\}$.

For our Bayesian purpose, following Chapter 9, we shall consider the Bayesian measures *cFDR* and *cFNR* as Bayesian multiple testing error rates. These measures are also

recommended by Chandra and Bhattacharya (2019) and Chandra and Bhattacharya (2020a) since they are conditioned on the observed data and hence qualify as *bona fide* Bayesian measures.

The above error measures also point towards the best model yielded by our multiple testing procedure. Recall from the discussion toward the end of Section 10.5.1 that the least value of $\beta \in (0, 1)$ such that the decision configuration $\hat{d}_{\tilde{k}} = 0$ and $\hat{d}_k = 1$ for all $k \in \{1, \dots, K\} \setminus \{\tilde{k}\}$ is obtained, for some $\tilde{k} \in \{1, \dots, K\}$, yields the best model $\mathcal{M}_{\tilde{k}}$. Now, since cFDR and cFNR are step functions of β , it is clear that the first jump of the graph of either of the functions cFDR or cFNR corresponds to the same best model.

10.6 Implementation of the Bayesian multiple testing procedure

10.6.1 Parallel computation of $[\bar{x}_1, \dots, \bar{x}_{T_0} | \bar{x}_{T_0+1}, \dots, \bar{x}_T, \mathcal{M}_k]$ for different GCMs and climate scenarios

Note that for conducting the Bayesian multiple hypotheses tests, we need to obtain samples from the posteriors $[\bar{x}_1, \dots, \bar{x}_{T_0} | \bar{x}_{T_0+1}, \dots, \bar{x}_T, \mathcal{M}_k]$, for all $k = 1, \dots, K$, for any given climate scenario. These are required to evaluate the posterior probabilities of (10.5.6), associated with the inverse perspective.

The method of obtaining posterior samples from the above distributions is the same as described in Section 10.4.1, with the priors discussed in Section 10.3, but we need to select the grid \mathbf{G}_n appropriately for creating the Gaussian process based look-up table. Note that the input grid \mathbf{G}_n is a two-dimensional grid, the first component being the time component and the second being the real line. In our case, we re-label the times 1850 – 2099 as 0 – 249 and further divide the re-labeled times by 250 to have them lie in $[0, 1]$. We then divide up the interval $[0, 1]$ into $n = 50$ equal sub-intervals and randomly simulate a value from each sub-interval. For the second component of \mathbf{G}_n , gridding the

interval $[0, 5]$ instead of a large interval turned out to be more than adequate for our problem, particularly because we consider the logarithms of the time series rather than the actual time series. We divide up the interval $[0, 5]$ into $n = 50$ equal sub-intervals and randomly simulate a value from each sub-interval. Thus, we construct \mathbf{G}_n using component-wise Latin hypercube sampling with $n = 50$.

For each model $[\bar{x}_1, \dots, \bar{x}_{T_0} | \bar{x}_{T_0+1}, \dots, \bar{x}_T, \mathcal{M}_k]$, $k = 1, \dots, K$, we obtain 60,000 samples of $\{\bar{x}_1, \dots, \bar{x}_{T_0}\}$ following the method described in Section 10.4.1, discarding the first 10,000 as burn-in. Now recall that the climate scenarios A1B, A2, B1 and Commitment consist of 21, 17, 21 and 16 GCMs, respectively. That is, in all, there are 75 posteriors of the form $[\bar{x}_1, \dots, \bar{x}_{T_0} | \bar{x}_{T_0+1}, \dots, \bar{x}_T, \mathcal{M}_k]$, and from each of them 60,000 realizations are to be simulated. This is an infeasible task if the models are implemented separately. However, we implement our code, written in C in accordance with the MPI protocol, in a parallel architecture associated with a VMWare consisting of 100 cores, running at 2.80 GHz speed, and having 1 TB memory. Specifically, we parallelize our computation by splitting 75 model implementations into 75 separate cores of our VMWare. The entire exercise takes less than an hour in our parallel implementation.

10.6.2 Obtaining the posterior model probabilities using Gibbs sampling

Recall that our multiple testing approach also requires computation of the posterior model probabilities $[\zeta = k | \bar{x}_{T_0+1}, \dots, \bar{x}_T]$. As in Chapter 9, we propose Gibbs sampling for simulation-based computations of these probabilities, by sampling from the full conditionals $[\zeta | \bar{x}_{T_0+1}, \dots, \bar{x}_T, p_1, \dots, p_K]$ and $[p_1, \dots, p_K | \bar{x}_{T_0+1}, \dots, \bar{x}_T, \zeta]$ successively.

Note that given ζ , the posterior distribution of (p_1, \dots, p_K) is again a Dirichlet distribution with parameters $(\alpha_1 + I(\zeta = 1), \dots, \alpha_K + I(\zeta = K))$. In other words, since

$\alpha_k = 1$ for $k = 1, \dots, K$, we have

$$[p_1, \dots, p_K | \bar{x}_{T_0+1}, \dots, \bar{x}_T, \zeta] \equiv Dirichlet(1 + I(\zeta = 1), \dots, 1 + I(\zeta = K)). \quad (10.6.1)$$

Given (p_1, \dots, p_K) , the posterior distribution of ζ is given by

$$[\zeta = k | \bar{x}_{T_0+1}, \dots, \bar{x}_T, p_1, \dots, p_K] = \frac{p_k[\bar{x}_{T_0+1}, \dots, \bar{x}_T | \mathcal{M}_k]}{\sum_{\ell=1}^K p_\ell[\bar{x}_{T_0+1}, \dots, \bar{x}_T | \mathcal{M}_\ell]}, \quad k = 1, \dots, K, \quad (10.6.2)$$

where for any k , letting $\mathbf{D}_n^{*(k)}$ denote the look-up table associated with model \mathcal{M}_k ,

$$\begin{aligned} & [\bar{x}_{T_0+1}, \dots, \bar{x}_T | \mathcal{M}_k] \\ &= \int [\bar{x}_{T_0+1}, \dots, \bar{x}_T | \mathbf{D}_n^{*(k)}, \boldsymbol{\theta}^{(k)}, \mathcal{M}_k] d[\boldsymbol{\theta}^{(k)}] d[\mathbf{D}_n^{*(k)}] \\ &= \int \prod_{t=T_0}^{T-1} [\bar{x}_{t+1} = f(t+1, \bar{x}_t) + \epsilon_{t+1} | \mathbf{D}_n^{*(k)}, \bar{x}_t, \boldsymbol{\theta}^{(k)}, \mathcal{M}_k] d[\boldsymbol{\theta}^{(k)}] d[\mathbf{D}_n^{*(k)}] \\ &\approx \frac{1}{N} \sum_{i=1}^N \prod_{t=T_0}^{T-1} [\bar{x}_{t+1} = f(t+1, \bar{x}_t) + \epsilon_{t+1} | \mathbf{D}_{ni}^{*(k)}, \bar{x}_t, \boldsymbol{\theta}_i^{(k)}, \mathcal{M}_k], \end{aligned} \quad (10.6.3)$$

where $\left\{ (\mathbf{D}_{ni}^{*(k)}, \boldsymbol{\theta}_i^{(k)}) : i = 1, \dots, N \right\}$, for sufficiently large N , is a set of simulations from the prior distributions of $\boldsymbol{\theta}^{(k)}$ and the distribution of the look-up table $\mathbf{D}_n^{*(k)}$.

In practice, rather than simulating from the priors, we simulate $\left\{ (\mathbf{D}_{ni}^{*(k)}, \boldsymbol{\theta}_i^{(k)}) : i = 1, \dots, N \right\}$ from the posterior distributions of $\boldsymbol{\theta}^{(k)}$ and $\mathbf{D}_n^{*(k)}$. The reason for this is the following. Simulating from the priors would lead to many realizations that are not well-supported by the data $\{\bar{x}_{T_0+1}, \dots, \bar{x}_T\}$, and these realizations would render the density $[\bar{x}_{T_0+1}, \dots, \bar{x}_T | \mathbf{D}_n^{*(k)}, \boldsymbol{\theta}^{(k)}, \mathcal{M}_k]$ extremely small, thus significantly reducing the effective simulation size. This issue is clearly much alleviated if the simulations correspond to the posterior distributions $[\mathbf{D}_n^{*(k)}, \boldsymbol{\theta}^{(k)} | \bar{x}_{T_0+1}, \dots, \bar{x}_T, \mathcal{M}_k]$, since such realizations are well-supported by the data that has been conditioned upon. This strategy also led to numerically stable estimates of the marginal densities in all our cases.

Using the full conditional distributions (10.6.1) and (10.6.2), along with the aforementioned posterior-based computation of (10.6.3), we obtain 100,000 realizations from the posterior distribution of (ζ, p_1, \dots, p_K) using Gibbs sampling, after discarding the first 10,000 iterations as burn-in.

10.6.3 Obtaining the posterior probabilities of the alternative hypotheses H_{1k}

Note that for $k = 1, \dots, K$, the posterior probability of H_{1k} is given by

$$\begin{aligned} v_k &= 1 - \left[\zeta = k, S^{(k)}(\bar{\boldsymbol{x}}_{T_0}) - S^{(k)}(\boldsymbol{x}_{T_0}^{(0)}) \in [\bar{\ell}_k, \bar{u}_k] \mid \bar{x}_{T_0+1}, \dots, \bar{x}_T \right] \\ &= 1 - \left[\zeta = k \mid \bar{x}_{T_0+1}, \dots, \bar{x}_T \right] \left[S^{(k)}(\bar{\boldsymbol{x}}_{T_0}) - S^{(k)}(\boldsymbol{x}_{T_0}^{(0)}) \in [\bar{\ell}_k, \bar{u}_k] \mid \zeta = k, \bar{x}_{T_0+1}, \dots, \bar{x}_T \right]. \end{aligned} \quad (10.6.4)$$

Hence, once we obtain realizations from the posteriors of $S^{(k)}(\bar{\boldsymbol{x}}_{T_0})$ for $k = 1, \dots, K$, and (ζ, p_1, \dots, p_K) , evaluation of v_k ; $k = 1, \dots, K$, follows simply by Monte Carlo averaging associated with the two factors of (10.6.4).

10.7 GCM selection results

We implemented our Bayesian multiple testing procedure with both the discrepancy measures $S_1^{(k)}$ and $S_2^{(k)}$ given by (10.5.2) and (10.5.3), respectively. We denote the corresponding cFDRs by cFDR1 and cFDR2 and the corresponding cFNRs by cFNR1 and cFNR2, respectively. Figures 10.7.1 and 10.7.2 depict these Bayesian error measures as functions of the penalty β , for all the four climate scenarios A1B, A2, B1 and Commitment, with respect to both the discrepancy measures $S_1^{(k)}$ (red line) and $S_2^{(k)}$ (green line).

The discussions toward the ends of Section 10.5.1 and 10.5.2 point out that the first jump occurring in either of the graphs of cFDR or cFNR as functions of β , corresponds

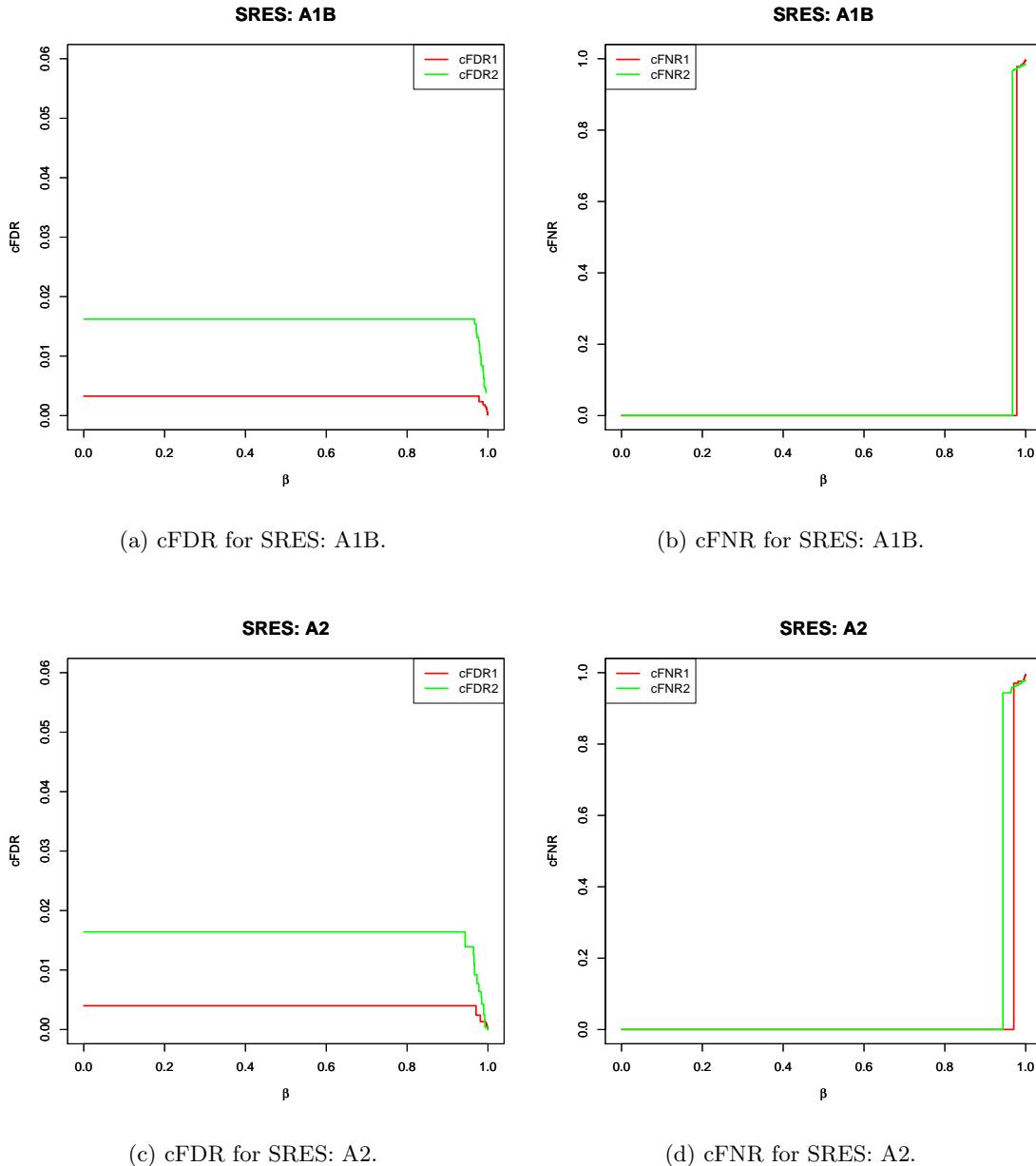


Figure 10.7.1: cFDR and cFNR for GCM selection in the climate scenarios A1B and A2 using Bayesian multiple testing.

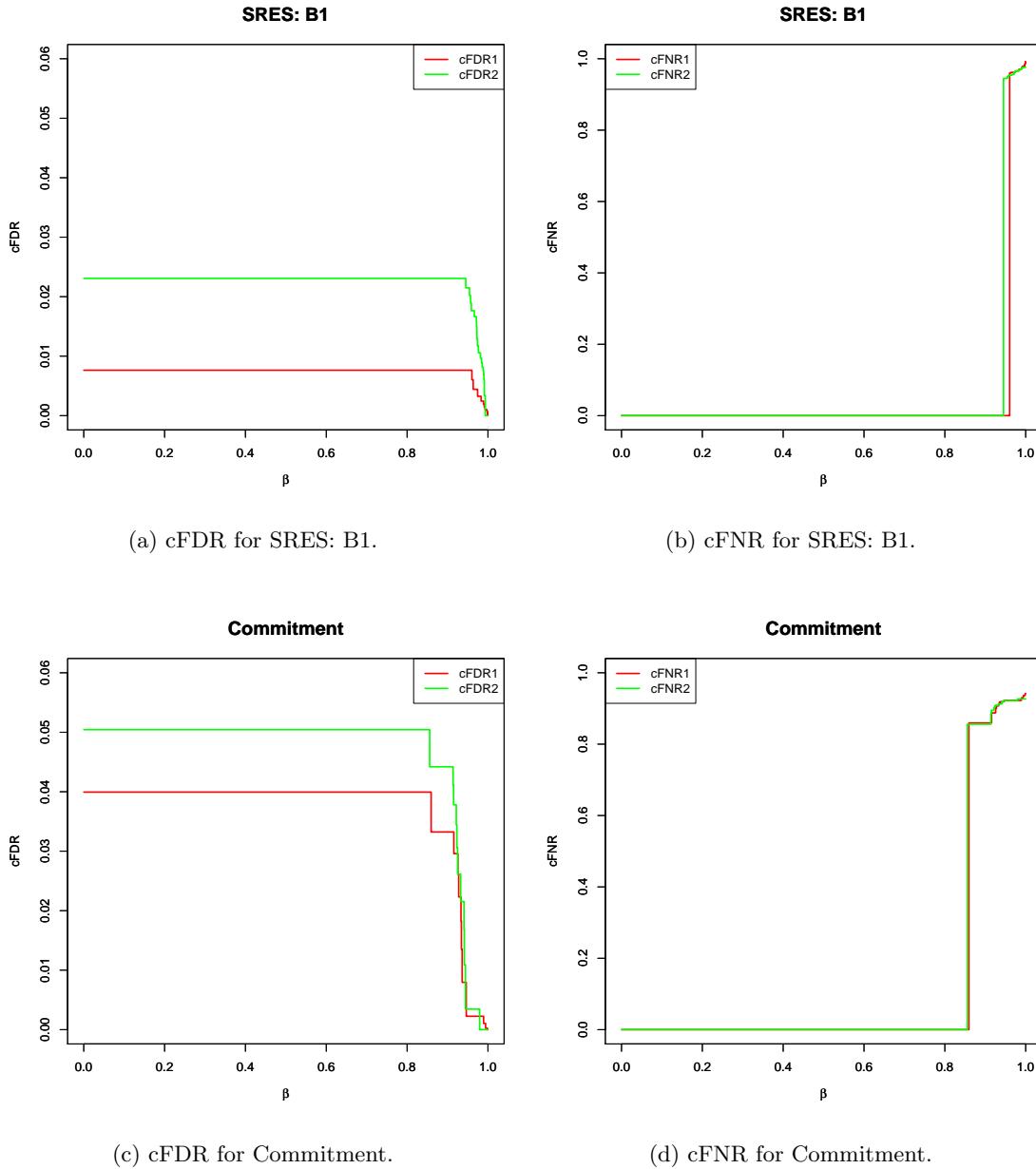


Figure 10.7.2: cFDR and cFNR for GCM selection in the climate scenarios B1 and Commitment using Bayesian multiple testing.

to the best model. In this regard, Figures 10.7.1 and 10.7.2 show that values of the penalty β close to one are required to obtain the first jumps of cFDR and cFNR for both the discrepancy measures $S_1^{(k)}$ and $S_2^{(k)}$, for all the four climate scenarios. Thus, none of the selected models seem to be satisfactory. Also, all the jumps occur close to each other in all the cases, indicating that the best models are not significantly good compared to the other competing models.

In all the cases, $S_2^{(k)}$ performs relatively better than $S_1^{(k)}$ in the sense that the value of β required for $S_2^{(k)}$ is somewhat less than the $S_1^{(k)}$ counterpart for selecting the best model. Among all the four climate scenarios, the Commitment scenario turns out to be the best since here the best model is selected for a value of β that is lesser than those of the other scenarios.

In the case of A1B, $S_1^{(k)}$ and $S_2^{(k)}$ yielded two different best models, csiro_mk3_0 and immcm3_0, respectively. In the remaining climate scenarios, both the discrepancy measures $S_1^{(k)}$ and $S_2^{(k)}$ yielded the same best models. The best GCMs selected for the scenarios A2, B1 and Commitment, are ukmo_hadgem1, gfdl_cm2_0 and cnrm_cm3, respectively.

Figure 10.7.3 displays the posterior distribution of the time series $[\bar{x}_0, \bar{x}_1, \dots, \bar{x}_{T_0} | \bar{x}_{T_0+1}, \dots, \bar{x}_T]$ (note that $\bar{x}_0 = x_0$, since x_0 is assumed to be known) corresponding to the aforementioned best GCMs selected by our Bayesian multiple testing procedure, as colour plots. The progressively higher densities are represented by progressively intense colours. The thick black line is the HadCRUT4 data, which is the current global temperature (CGT) and the dashed line is the model based global temperature (MBGT), the simulated global temperatures by the underlying GCM. The other starred line stands for the average model based global temperature (AMBGT), which is the average over all the GCM based simulated time series in the respective climate scenario. All the time series are in degree celsius and in the log scale. Recall that the HadCRUT4 data is associated with the years 1850 – 2016 and the GCMs are associated with 1900 – 2099, which is why the

time scales for the HadCRUT4 data and the GCM based simulated data are different.

Observe that except for B1 and Commitment most part of the observed HadCRUT4 data is not included in the high density regions of the corresponding posterior time series associated with the best GCMs. In fact, except the case of Commitment, all other posteriors strongly support lower temperatures than HadCRUT4. This is not surprising since Figure 10.1.1 show that the GCM-simulated time series significantly underestimate the HadCRUT4 data during the relevant time period, and so must the averaged GCM time series, and this is broadly consistent with the observations on Figure 10.7.3. Also observe that MBGT and AMBGT lie closer to the high density regions compared to CGT, which is again not unexpected as Figure 10.1.1 indicates.

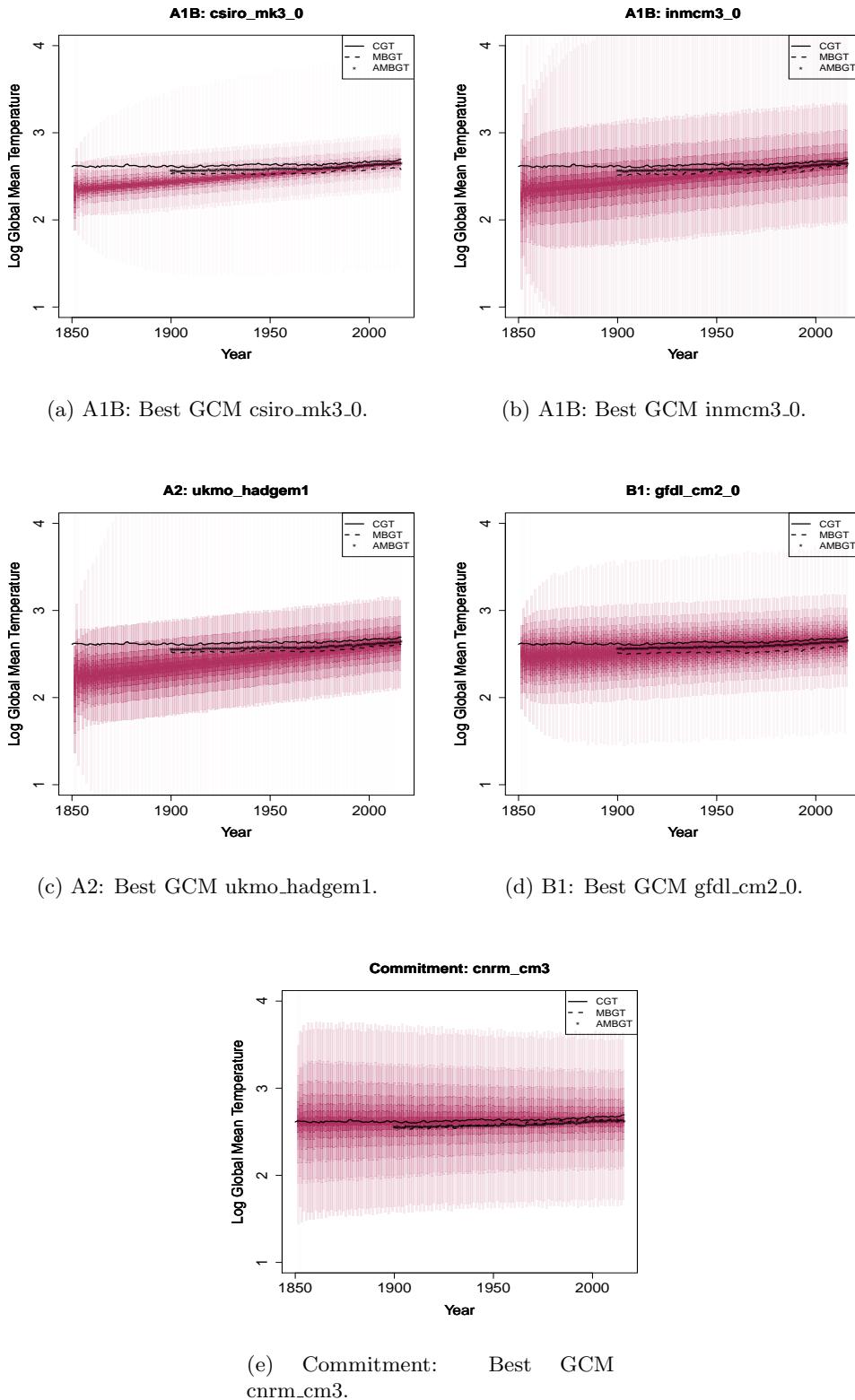


Figure 10.7.3: The posteriors corresponding to the HadCRUT4 data or the current global temperature (CGT) conditional on GCM-based average time series are shown as colour plots with progressively higher densities depicted by progressively intense colours. Also shown are the HadCRUT4 data (CGT), GCM based time series (MBGT) and the average of GCM based time series (AMBGT). The temperature is in °C and in the log-scale.

Table 10.7.1

Goodness-of-fit check for the best GCMs with respect to averaged time series. Here 95% BCI stands for 95% Bayesian credible intervals.

Model	$S_1^{(k)}(\bar{\mathbf{x}}_{T_0}^{(0)})$	95% BCI of $S_1^{(k)}(\bar{\mathbf{x}}_{T_0})$	$S_2^{(k)}(\bar{\mathbf{x}}_{T_0}^{(0)})$	95% BCI of $S_2^{(k)}(\bar{\mathbf{x}}_{T_0})$
A1B (csiro_mk3_0)	0.126	[0.104, 0.281]	0.024	[0.015, 0.424]
A1B (inmcm3_0)	0.001	[5×10^{-4} , 0.023]	126×10^{-6}	[7.324×10^{-6} , 0.048]
A2 (ukmo_hadgem1)	0.006	[0.003, 0.048]	0.001	[9.047×10^{-5} , 0.082]
B1 (gfdl_cm2_0)	0.142	[0.107, 1.048]	0.028	[0.017, 2.420]
Commit (cnrm_cm3)	0.039	[0.119, 1.599]	0.002	[0.021, 3.848]

Table 10.7.1, summarizing the goodness-of-fit of the posteriors to the HadCRUT4 data with respect to the discrepancy measures $S_1^{(k)}$ and $S_2^{(k)}$, tell a somewhat different story. The best GCM in the Commitment scenario seems to overfit the HadCRUT4 data in the sense that the observed discrepancies are too small to be included the 95% credible intervals of the reference discrepancy measures. Given the large variability of the time series as shown in panel (e) of Figure 10.7.3, which can also be gauged by the less colour intensities compared to the other panels, this result is not unexpected in retrospect. On the other hand, in the other cases, the observed discrepancies are included in the respective 95% credible intervals. Although again this seems surprising at the first glance, this is due the fact that the posterior time series relatively closer to the year 2017, where the GCM time series begins in our posterior formulation, well-captures the HadCRUT4 data, with relatively small posterior variability. Hence, even though the posteriors fail to perform well for the years closer to 1850, the overall goodness-of-fit still can not be declared as poor.

Figure 10.7.4 shows the posterior distributions of $[x_0, x_1, \dots, x_{T_0} | x_{T_0+1}, \dots, x_T]$ associated with the individual time series for the best GCM models, rather than the averaged time series as shown in Figure 10.7.3. The overall story, however, did not seem to be very different compared to that told by Figure 10.7.3. Table 10.7.2, evaluating goodness-of-fit for these posteriors using the discrepancy measures, also provide similar inference as Table 10.7.1, where $\mathbf{x}_{T_0} = (x_1, \dots, x_{T_0})$.

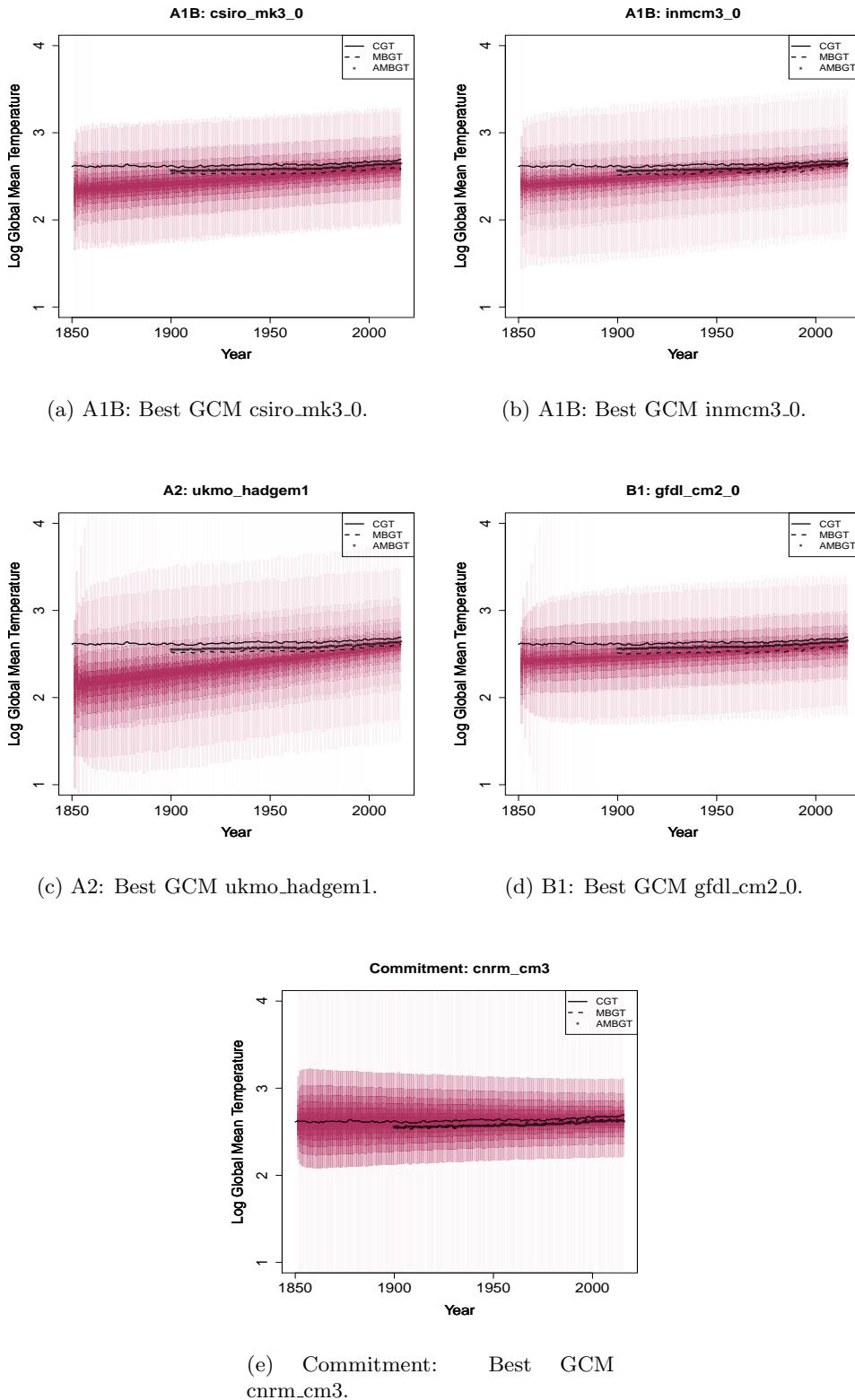


Figure 10.7.4: The posteriors corresponding to the HadCRUT4 data or the current global temperature (CGT) conditional on individual best GCM time series are shown as colour plots with progressively higher densities depicted by progressively intense colours. Also shown are the HadCRUT4 data (CGT), GCM based time series (MBGT) and the average of GCM based time series (AMBGT). The temperature is in °C and in the log-scale.

Table 10.7.2

Goodness-of-fit check for the best GCMs with respect to individual time series. Here 95% BCI stands for 95% Bayesian credible intervals.

Model	$S_1^{(k)}(\mathbf{x}_{T_0}^{(0)})$	95% BCI of $S_1^{(k)}(\mathbf{x}_{T_0})$	$S_2^{(k)}(\mathbf{x}_{T_0}^{(0)})$	95% BCI of $S_2^{(k)}(\mathbf{x}_{T_0})$
A1B (csiro_mk3_0)	0.127	[0.099,0.255]	0.028	[0.014,0.419]
A1B (inmcm3_0)	0.105	[0.098,0.163]	0.016	[0.013,0.179]
A2 (ukmo_hadgem1)	0.088	[0.077,0.152]	0.015	[0.009,0.186]
B1 (gfdl_cm2_0)	0.072	[0.062,0.202]	0.010	[0.005,0.348]
Commit (cnrm_cm3)	0.043	[0.179,1.496]	0.003	[0.049,3.504]

10.8 GCM simulations as ensembles: extension of our Gaussian process emulation approach to the multivariate situation

So far, the inference with our one-dimensional Gaussian process approach demonstrated that although even the best GCM models are not as adequate as desired, it is not very easy to discard them since Tables 10.7.1 and 10.7.2 demonstrate quantitatively that in general their overall performances in fitting the observed current global temperatures are not particularly poor. However, Figures 10.7.3 and 10.7.4 show that a large part of the current global temperature data, beginning from 1851, fails to lie in the high density regions of the relevant posterior, which is clearly very disconcerting. Even though the Commitment scenario includes almost the entire current temperature time series in its high posterior density region, the posterior variability turns out to be too high to render the fit satisfactory.

For further investigation we consider all the K GCM-based time series in any climate scenario as an ensemble of time series, and consider modeling them as multivariate (K -dimensional) time series, extending our one-dimensional Gaussian process emulation theory to multidimensional Gaussian process emulation. In this regard, for $t = 0, 1, 2, \dots$, let $\mathbf{x}_t = (x_1^{(1)}, \dots, x_t^{(K)})'$ be K -component vectors, corresponding to K different GCM

based log time series $x_t^{(k)}$; $k = 1, \dots, K$. With $\bar{x}_t = K^{-1} \sum_{k=1}^K x_t^{(k)}$, we shall be interested in the posterior $[\bar{x}_1, \dots, \bar{x}_{T_0} | \mathbf{x}_{T_0+1}, \dots, \mathbf{x}_T]$, for predicting the logarithm of the observed current temperature data. Note that $[\bar{x}_1, \dots, \bar{x}_{T_0} | \mathbf{x}_{T_0+1}, \dots, \mathbf{x}_T]$ is induced by $[\mathbf{x}_1, \dots, \mathbf{x}_{T_0} | \mathbf{x}_{T_0+1}, \dots, \mathbf{x}_T]$ as the former is obtained from the latter by simply taking the averages of the components of \mathbf{x}_t , for each $t = 1, \dots, T_0$. It is thus sufficient to build the multivariate Gaussian process emulation theory with respect to the K -dimensional vectors \mathbf{x}_t .

Our multivariate dynamic model is of the form

$$\mathbf{x}_t = \mathbf{f}(\mathbf{x}_{t-1}^*) + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim N_K(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon), \quad (10.8.1)$$

where $\mathbf{x}_0 = x_0 \mathbf{1}_K$ is assumed known. Here $\mathbf{1}_K$ is a K -dimensional vector with all components 1.

In the above, $\mathbf{f}(\cdot) = (f_1(\cdot), \dots, f_K(\cdot))'$ is a function with K components. We assume that $\mathbf{f}(\cdot)$ is a K -variate Gaussian process with mean $E[\mathbf{f}(\cdot)] = \mathbf{B}'_f \mathbf{h}(\cdot)$ and covariance function $cov(\mathbf{f}(\mathbf{z}_1), \mathbf{f}(\mathbf{z}_2)) = c_f(\mathbf{z}_1, \mathbf{z}_2) \boldsymbol{\Sigma}_f$, for any $(K+1)$ -dimensional inputs $\mathbf{z}_1, \mathbf{z}_2$. Here $\mathbf{h}(\cdot) = (h_1(\cdot), \dots, h_m(\cdot))'$ and $\mathbf{B}_f = (\boldsymbol{\beta}_{1,f}, \dots, \boldsymbol{\beta}_{K,f})$, where, for $j = 1, \dots, K$, $\boldsymbol{\beta}_{j,f}$ are m -dimensional column vectors. Note that $h_1(\cdot) \equiv 1$ corresponds to the intercept and $h_2(\cdot), \dots, h_m(\cdot)$ correspond to the components of $(K+1)$ -dimensional inputs \mathbf{z} . Hence, it is clear that $m = K+2$. Also, $c_f(\mathbf{z}_1, \mathbf{z}_2) = \exp\{-(\mathbf{z}_1 - \mathbf{z}_2)' \mathbf{R}_f (\mathbf{z}_1 - \mathbf{z}_2)\}$, where \mathbf{R}_f is a diagonal matrix consisting of $(K+1)$ smoothness parameters, denoted by $\{r_{1,f}, \dots, r_{(K+1),f}\}$.

10.8.1 Distributions of $\mathbf{f}(\mathbf{x}_{1,0}^*)$ and \mathbf{D}_n^*

Conditional on \mathbf{x}_0 , $\mathbf{f}(\mathbf{x}_{1,0}^*)$ is K -variate normal with mean $\mathbf{B}'_f \mathbf{h}(\mathbf{x}_{1,0}^*)$ and covariance matrix $\boldsymbol{\Sigma}_f$. Now, $\mathbf{D}_{z,nK} = (\mathbf{f}'(\mathbf{z}_1), \mathbf{f}'(\mathbf{z}_2), \dots, \mathbf{f}'(\mathbf{z}_n))'$ has an nK -variate normal

distribution with mean

$$E[\mathbf{D}_{z,nK} \mid \mathbf{B}_f, \boldsymbol{\Sigma}_f, \mathbf{R}_f] = \begin{pmatrix} \mathbf{B}'_f \mathbf{h}(\mathbf{z}_1) \\ \mathbf{B}'_f \mathbf{h}(\mathbf{z}_2) \\ \vdots \\ \mathbf{B}'_f \mathbf{h}(\mathbf{z}_n) \end{pmatrix} = \boldsymbol{\mu}_{D_{z,nK}} \quad (\text{say}) \quad (10.8.2)$$

and covariance matrix

$$V[\mathbf{D}_{z,nK} \mid \mathbf{B}_f, \boldsymbol{\Sigma}_f, \mathbf{R}_f] = \mathbf{A}_{f,D_n^*} \otimes \boldsymbol{\Sigma}_f = \boldsymbol{\Sigma}_{D_{z,nK}} \quad (\text{say}), \quad (10.8.3)$$

where “ \otimes ” denotes Kronecker product. Hence, the distribution of the $n \times K$ -dimensional matrix $\mathbf{D}_n^* = (\mathbf{f}(\mathbf{z}_1), \mathbf{f}(\mathbf{z}_2), \dots, \mathbf{f}(\mathbf{z}_n))'$ is matrix normal:

$$[\mathbf{D}_n^* \mid \mathbf{B}_f, \boldsymbol{\Sigma}_f, \mathbf{R}_f] \sim \mathcal{N}_{n,K}(\mathbf{H}_{D_n^*} \mathbf{B}_f, \mathbf{A}_{f,D_n^*}, \boldsymbol{\Sigma}_f). \quad (10.8.4)$$

Conditionally on $(\mathbf{x}_0, \mathbf{f}(\mathbf{x}_{1,0}^*))$, it follows that \mathbf{D}_n^* is $n \times K$ -dimensional matrix-normal:

$$[\mathbf{D}_n^* \mid \mathbf{f}(\mathbf{x}_{1,0}^*), \mathbf{x}_0, \mathbf{B}_f, \boldsymbol{\Sigma}_f, \mathbf{R}_f, \boldsymbol{\Sigma}_\epsilon] \sim \mathcal{N}_{n,K}(\boldsymbol{\mu}_{f,D_n^*}, \boldsymbol{\Sigma}_{f,D_n^*}, \boldsymbol{\Sigma}_f) \quad (10.8.5)$$

In (10.8.5) $\boldsymbol{\mu}_{f,D_n^*}$ is the mean matrix, given by

$$\boldsymbol{\mu}_{f,D_n^*} = \mathbf{H}_{D_n^*} \mathbf{B}_f + \mathbf{s}_{f,D_n^*}(\mathbf{x}_{1,0}^*)(\mathbf{f}(\mathbf{x}_{1,0}^*)' - \mathbf{h}(\mathbf{x}_{1,0}^*)' \mathbf{B}_f), \quad (10.8.6)$$

and

$$\boldsymbol{\Sigma}_{f,D_n^*} = \mathbf{A}_{f,D_n^*} - \mathbf{s}_{f,D_n^*}(\mathbf{x}_{1,0}^*) \mathbf{s}_{f,D_n^*}(\mathbf{x}_{1,0}^*)'. \quad (10.8.7)$$

Here we slightly abuse notation to denote both univariate and multivariate versions of the mean matrix and the right covariance matrix by $\boldsymbol{\mu}_{f,D_n^*}$ and $\boldsymbol{\Sigma}_{f,D_n^*}$, respectively.

10.8.2 Joint distribution of $\{\mathbf{x}_1, \dots, \mathbf{x}_T, \mathbf{D}_n^*\}$

Note that

$$[\mathbf{x}_1 \mid \mathbf{f}(\mathbf{x}_0), \mathbf{x}_0, \mathbf{B}_f, \Sigma_f] \sim N_K (\mathbf{f}(\mathbf{x}_{1,0}^*), \Sigma_\epsilon), \quad (10.8.8)$$

and for $t = 1, \dots, T$, the conditional distribution $[\mathbf{x}_{t+1} = \mathbf{f}(\mathbf{x}_{t+1,t}^*) + \boldsymbol{\epsilon}_{t+1} \mid \mathbf{D}_n^*, \mathbf{x}_t, \mathbf{B}_f, \Sigma_f, \mathbf{R}_f, \Sigma_\epsilon]$ is K -variate normal with mean

$$\mu_{x_t} = \mathbf{B}'_f \mathbf{h}(\mathbf{x}_{t+1,t}^*) + (\mathbf{D}_n^* - \mathbf{H}_{D_n^*} \mathbf{B}_f)' \mathbf{A}_{f,D_n^*}^{-1} \mathbf{s}_{f,D_n^*}(\mathbf{x}_{t+1,t}^*) \quad (10.8.9)$$

and variance

$$\Sigma_{x_t} = \left\{ 1 - \mathbf{s}_{f,D_n^*}(\mathbf{x}_{t+1,t}^*)' \mathbf{A}_{f,D_n^*}^{-1} \mathbf{s}_{f,D_n^*}(\mathbf{x}_{t+1,t}^*) \right\} \Sigma_f + \Sigma_\epsilon. \quad (10.8.10)$$

Since \mathbf{x}_0 is assumed to be known and the distribution of \mathbf{D}_n^* is given by (10.8.4), the joint distribution is obtained by taking products of the individual distributions.

10.8.3 Prior distributions

We assume the following forms of the prior distributions:

$$\begin{aligned} [\mathbf{B}_f \mid \Sigma_f] &\sim \mathcal{N}_{m,K} (\mathbf{B}_{f,0}, \Sigma_{B_f,0}, \psi \Sigma_f); \\ [\Sigma_f] &\propto |\Sigma_f|^{-\frac{\nu_f+K+1}{2}} \exp \left[-\frac{1}{2} \text{tr} (\Sigma_f^{-1} \Sigma_{f,0}) \right], \text{ with } \nu_f > K - 1; \\ [\Sigma_\epsilon] &\propto |\Sigma_\epsilon|^{-\frac{\nu_\epsilon+K+1}{2}} \exp \left[-\frac{1}{2} \text{tr} (\Sigma_\epsilon^{-1} \Sigma_{\epsilon,0}) \right], \text{ with } \nu_\epsilon > K - 1; \text{ and} \end{aligned}$$

for $i = 1, \dots, (K + 1)$,

$$[\log(r_{i,f})] \stackrel{iid}{\sim} N(\mu_{R_f}, \sigma_{R_f}^2).$$

For the prior of \mathbf{B}_f we set $\psi = 1$, and except the first column of $\mathbf{B}_{f,0}$, we set all other

columns of $\mathbf{B}_{f,0}$ to be null vectors. We set the first column of $\mathbf{B}_{f,0}$ to be the vector of means of the K GCM based time series thinned by 5 observations. Recall that these means are also used for the corresponding prior in the one-dimensional situation for model selection.

In the priors for Σ_f and Σ_ϵ , we set $\nu_f = K$ and $\nu_\epsilon = K$. For $\Sigma_{B_f,0}$ and $\Sigma_{\epsilon,0}$, we first let $\hat{\Sigma}$ to be the empirical covariance matrix for the K GCM-based time series, thinned by 5 observations. Then we set $\Sigma_{B_f,0} = \Sigma_{\epsilon,0} = \hat{\Sigma}/2$. Again, this choice is analogous to the previous one-dimensional setup.

For the log-normal priors of the smoothness parameters we set $\mu_{R_f} = -0.5$ and $\sigma_{R_f}^2 = 1$. The choices imply as in the one-dimensional situation that the prior mean and the prior variance of each of the smoothness parameters are, respectively, 1 and 2 (approximately).

Thus, these prior choices are in keeping with the one-dimensional situation and have similar rationale as before.

10.8.4 Choice of the input grid \mathbf{G}_n

To set up the $(K + 1)$ -dimensional grid \mathbf{G}_n for the model-fitting purpose, we considered $[-5, 5]^K$ to be a grid space for the K -dimensional variable \mathbf{z} . We divide $[-5, 5]$ into 50 equal sub-intervals and choose a point randomly from each of the 50 sub-intervals, in each dimension, yielding $n = 50$ K -dimensional points corresponding to \mathbf{z} . For the first component of the grid, corresponding to the time component, we follow the strategy in the one-dimensional Gaussian process situation. That is, we first re-label the times 1850 – 2099 as 0 – 249 and further divide the re-labeled times by 250 to have them lie in $[0, 1]$. Then, after dividing the interval $[0, 1]$ into 50 equal sub-intervals we randomly simulate a value from each sub-interval, to complete construction of the input grid \mathbf{G}_n . This grid choice turned out to be adequate for our purpose.

The rest of the multivariate Gaussian process emulation theory remains analogous

to the corresponding univariate case, but the full conditionals of \mathbf{B}_f and \mathbf{D}_n^* are no longer available in standard form for simulating in the MCMC context, which is not analogous to the univariate context discussed in Section 10.4; see the supplement of Ghosh *et al.* (2014) for details. We use additive TMCMC to update the unknowns in the multidimensional situation. For updating the positive definite matrices $\boldsymbol{\Sigma}_f$ and $\boldsymbol{\Sigma}_\epsilon$, we represent the matrices in the Cholesky decomposition forms \mathbf{CC}' , where \mathbf{C} is a lower triangular matrix, and use additive TMCMC to update the non-zero elements in a single block. We implement our codes, written in C, in our VMWare. The implementations associated with A1B, A2, B1 and Commitment took about 30 hours 37 minutes, 18 hours 52 minutes, 29 hours 12 minutes and 16 hours 59 minutes, respectively.

10.9 Results for the multivariate climate dynamics

For the four climate scenarios, the posterior distributions of $[\bar{x}_0, \bar{x}_1, \dots, \bar{x}_{T_0} | \mathbf{x}_{T_0+1}, \dots, \mathbf{x}_T]$ (where $\bar{x}_0 = x_0$, since x_0 is assumed to be known) are shown in Figure 10.9.1. Now, compared to the one-dimensional situations, severe under-estimation of the HadCRUT4 data by all the four climate scenarios is corroborated by this multivariate framework. And, Table 10.9.1 revealing severe underfits for all the four climate scenarios, confirms that even the discrepancy measures could not act as saviours this time.

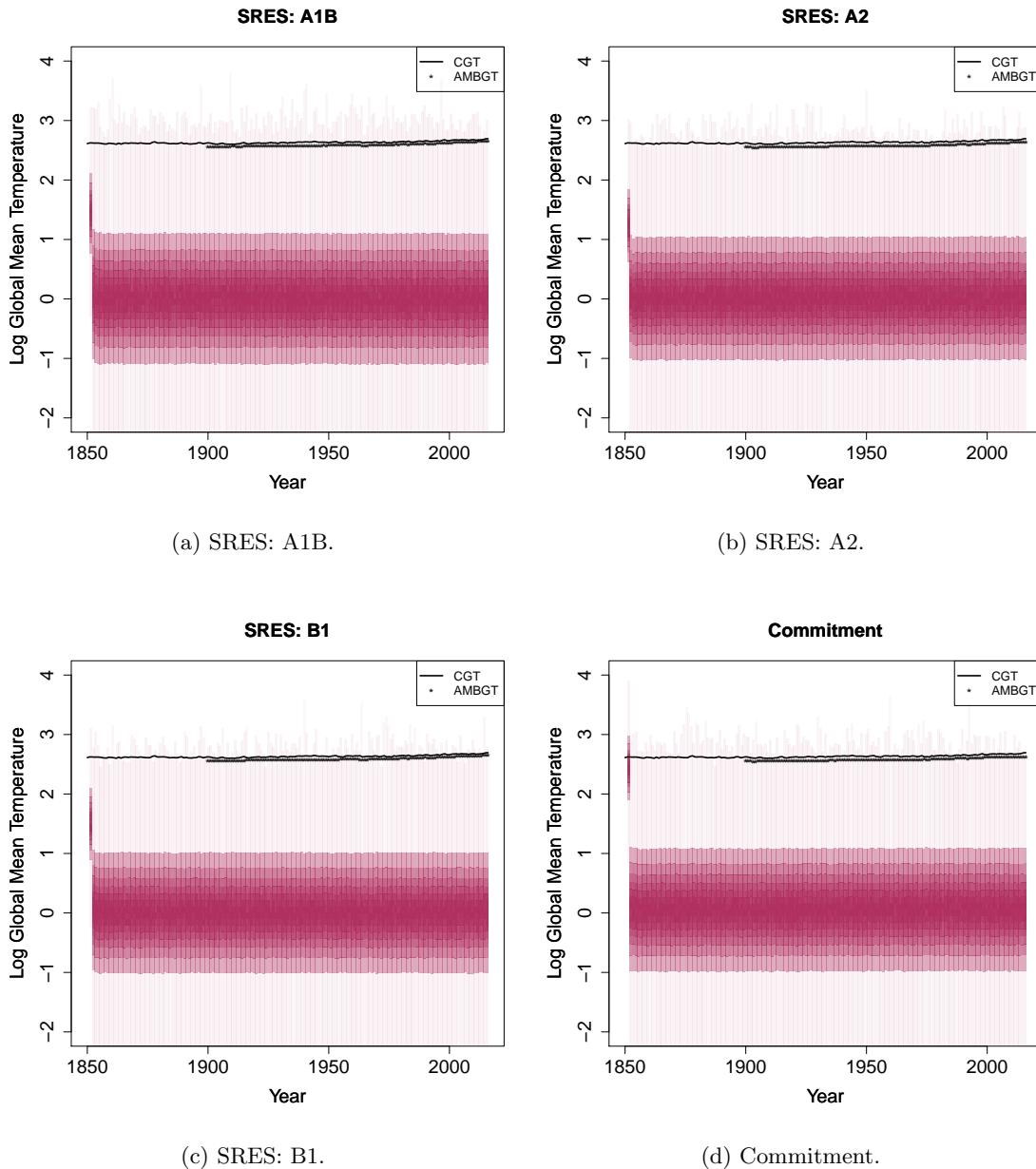


Figure 10.9.1: The posteriors $[\bar{x}_0, \bar{x}_1, \dots, \bar{x}_{T_0} | \mathbf{x}_{T_0+1}, \dots, \mathbf{x}_T]$ are shown as colour plots with progressively higher densities depicted by progressively intense colours, along with the HadCRUT4 data (CGT) and the average of GCM based time series (AMBGT). The temperature is in $^{\circ}\text{C}$ and in the log-scale.

Table 10.9.1

Goodness-of-fit check for ensembles of GCM time series with respect to $[\bar{x}_0, \bar{x}_1, \dots, \bar{x}_{T_0} | \mathbf{x}_{T_0+1}, \dots, \mathbf{x}_T]$. Here 95% BCI stands for 95% Bayesian credible intervals.

Model	$S_1^{(k)}(\mathbf{x}_{T_0}^{(0)})$	95% BCI of $S_1^{(k)}(\mathbf{x}_{T_0})$	$S_2^{(k)}(\mathbf{x}_{T_0}^{(0)})$	95% BCI of $S_2^{(k)}(\mathbf{x}_{T_0})$
A1B	3.580	[0.690, 0.872]	13.158	[0.763, 1.182]
A2	3.807	[0.689, 0.871]	14.909	[0.759, 1.179]
B1	3.872	[0.688, 0.870]	15.434	[0.758, 1.177]
Commit	3.711	[0.690, 0.870]	14.229	[0.760, 1.176]

Since $[\bar{x}_0, \bar{x}_1, \dots, \bar{x}_{T_0} | \mathbf{x}_{T_0+1}, \dots, \mathbf{x}_T]$ severely under-estimates the HadCRUT4 data, we now investigate how well the posterior $[x_0^{(max)}, x_1^{(max)}, \dots, x_{T_0}^{(max)} | \mathbf{x}_{T_0+1}, \dots, \mathbf{x}_T]$ can capture the observed current temperature data, where for $t = 0, 1, 2, \dots$, $x_t^{(max)}$ is the maximum of the components of \mathbf{x}_t . Figure 10.9.2 displays the relevant posterior time series as colour plots, along with the HadCRUT4 data (CGT) and the maximum of model based global temperature (MMGT) associated with the GCM simulations, in the log scales. Observe that CGT and MMGT are included in the supports, but it is doubtful how good the fits are, since the posterior variances are high and moreover for A2 and Commitment CGT and MMGT fall in low density regions. Table 10.9.2 shows that the fits are indeed not encouraging. Observe that A1B overfits with respect to both $S_1^{(k)}$ and $S_2^{(k)}$. With respect to $S_1^{(k)}$, A2 slightly underfits, while the fit is adequate with respect to $S_2^{(k)}$. Since $S_2^{(k)}$ is generally a better performer than $S_1^{(k)}$, one can consider the fit of A2 to be adequate. B1 seriously overfits with respect to both the discrepancy measures, while Commitment seriously underfits with respect to both $S_1^{(k)}$ and $S_2^{(k)}$.

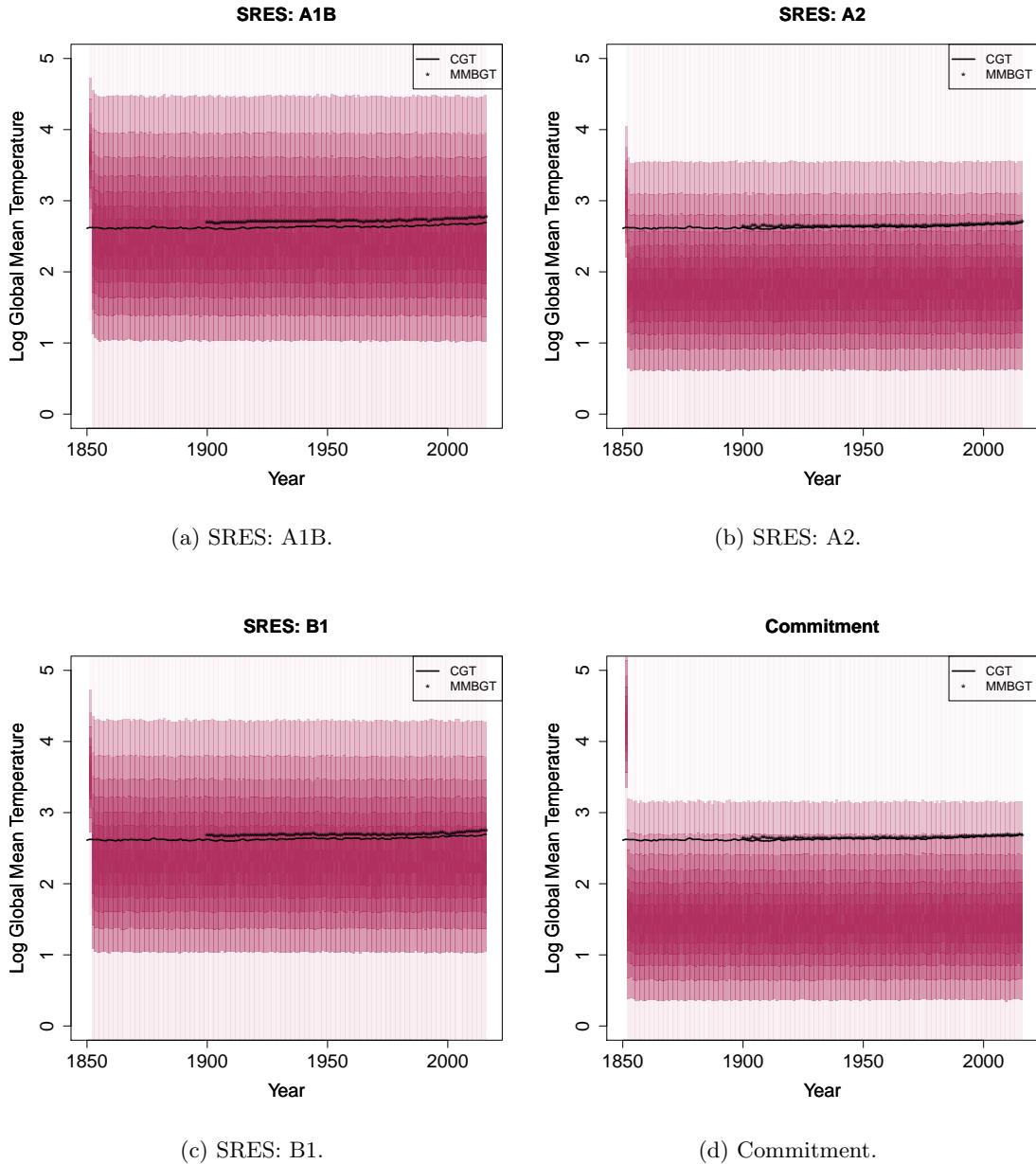


Figure 10.9.2: The posteriors $[x_0^{(max)}, x_1^{(max)}, \dots, x_{T_0}^{(max)} | \mathbf{x}_{T_0+1}, \dots, \mathbf{x}_T]$ are shown as colour plots with progressively higher densities depicted by progressively intense colours, along with the HadCRUT4 data (CGT) and the maximum of model based global temperature (MMBGT). The temperature is in $^{\circ}\text{C}$ and in the log-scale.

Table 10.9.2

Goodness-of-fit check for ensembles of GCM time series with respect to $[x_0^{(max)}, x_1^{(max)}, \dots, x_{T_0}^{(max)} | \mathbf{x}_{T_0+1}, \dots, \mathbf{x}_T]$. Here 95% BCI stands for 95% Bayesian credible intervals.

Model	$S_1^{(k)}(\mathbf{x}_{T_0}^{(0)})$	95% BCI of $S_1^{(k)}(\mathbf{x}_{T_0})$	$S_2^{(k)}(\mathbf{x}_{T_0}^{(0)})$	95% BCI of $S_2^{(k)}(\mathbf{x}_{T_0})$
A1B	0.216	[0.693,0.891]	0.061	[0.787,1.313]
A2	0.893	[0.692,0.888]	0.816	[0.786,1.318]
B1	0.303	[0.690,0.891]	0.104	[0.785,1.332]
Commit	1.256	[0.671,0.879]	1.617	[0.755,1.376]

10.10 Future climate forecast with our Bayesian Gaussian process dynamics model

Our detailed analyses of the GCM forecasts so far failed to justify their credibilities. This failure, however, seems to hold a great deal of positivity since the rapid future global warming foreboding that might eventually threaten life on earth, need not become the reality. However, it is not clear yet then what kind of climate change we can expect in the future. We attempt to answer this question, again with our Bayesian Gaussian process emulation theory, now forecasting the log global average temperature in the years 2017 – 2099 given the log HadCRUT4 dataset for the years 1850 – 2016, using the theory and strategies proposed in Section 10.4.2. Here we let the prior distributions remain the same as detailed in Section 10.3, except that the first component of $\beta_{f,0}$ and $\hat{\sigma}^2$ are now based upon thinning the log HadCRUT4 data by 5 observations. The input grid \mathbf{G}_n remains the same as in the one-dimensional setup detailed in Section 10.6.1.

Our future climate prediction results are presented in Figure 10.10.1, along with the posterior modes associated with the Gaussian process forecasted global temperature (GPFGT), the best GCM-specific model based forecasted global temperature (MBFGT) and average model based forecasted global temperature (AMBFGT). In stark contrast with MBFGT and AMBFGT which show steep increase in the temperature in panels (a)-(d), the high posterior density regions of our Bayesian forecasts do not support

increasing future global temperature. Only in the case of Commitment (panel (e)) MBFGT and AMBFGT tend to fall within the high posterior density regions of our Bayesian forecasts.

According to [Green et al. \(2009\)](#): “The benchmark forecast is that the global mean temperature for each year for the rest of this century will be within 0.5°C of the 2008 figure.” Thus, according to their prediction, the future global temperature should lie in the interval [13.895, 14.895]°C. This interval is included even within all the 50% credible intervals of our year-wise Bayesian posterior forecast distributions for 2017 – 2099. Thus, our results are broadly in agreement with the forecast of [Green et al. \(2009\)](#), and clearly do not support drastic global warming as projected by the GCMs.

10.11 A brief discussion of the existing works on climate model evaluation

It is important to mention that the existing literature has recorded several attempts to evaluate the GCM based climate projections. Some examples in this regard (see [Hausfather et al. \(2020\)](#)) are [Hansen et al. \(1988\)](#) National Aeronautics and Space Administration Goddard Institute for Space Studies model ([Hargreaves \(2010\)](#); [Rahmstorf et al. \(2007\)](#)), the [Stouffer et al. \(1989\)](#) Geophysical Fluid Dynamics Laboratory model ([Stouffer and Manabe \(2017\)](#)), the IPCC First Assessment Report ([IPCC \(FAR\) \(1990\)](#); [Frame and Stone \(2012\)](#)), and the IPCC Third and Fourth Assessment reports ([IPCC \(TAR\) \(2001\)](#); [IPCC \(AR4\) \(2007\)](#); [Rahmstorf et al. \(2012\)](#)).

[Hausfather et al. \(2020\)](#) attempt to evaluate the performance of several past climate models, such as The specific models projections evaluated were [Manabe \(1970\)](#), [Mitchell \(1970\)](#), [Benson \(1970\)](#), [Rasool and Schneider \(1971\)](#), [Sawyer \(1972\)](#), [Broecker \(1975\)](#), [Nordhaus \(1977\)](#), [Schneider and Thompson \(1981\)](#), [Hansen et al. \(1981\)](#), [Hansen et al. \(1988\)](#) and [Manabe and Stouffer \(1993\)](#). In their analyses, [Hausfather et al. \(2020\)](#) take

10.11. A BRIEF DISCUSSION OF THE EXISTING WORKS ON CLIMATE MODEL EVALUATION
363

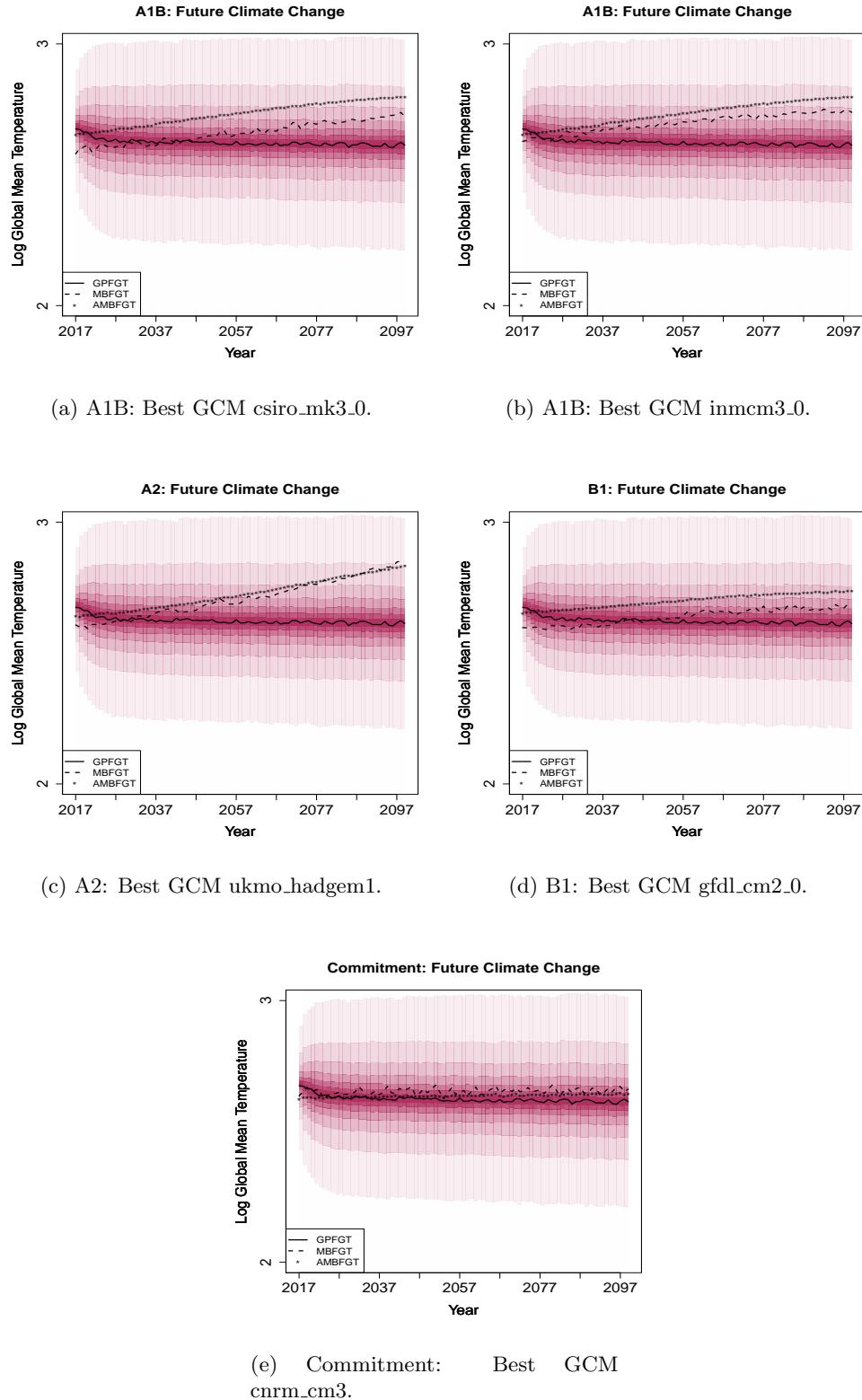


Figure 10.10.1: The posteriors $[x_{T_0+1}, \dots, x_T | x_1, \dots, x_{T_0}]$ for future climate prediction are shown as colour plots, along with the posterior modes of the Gaussian process forecasted global temperature (GPFGT), best GCM-specific model based forecasted global temperature (MBFGT) and average model based forecasted global temperature (AMBFGT). The temperature is in $^{\circ}\text{C}$ and in the log-scale.

into account the accuracy of projected changes in external forcing due to greenhouse gases and aerosols, as well as natural forcing such as solar or volcanic forcing. Indeed, they apply both a conventional assessment of the change in temperature over time and an assessment of the response of temperature to the change in forcing to assess the performance of future projections by past climate models compared to observations. Specifically, they compared observations to climate model projections over the model projection period using two approaches: change in temperature versus time and change in temperature versus change in radiative forcing.

Moreover, following the approach of Hargreaves (2010), Hausfather *et al.* (2020) calculate a skill score for each model for both temperature versus time and radiative forcing. The skill score is based on the root-mean-square errors of the model projection trend versus observations compared to a zero-change null-hypothesis projection.

The results of Hausfather *et al.* (2020) demonstrate that even when the temperature versus time yields inaccurate results, the corresponding temperature versus radiative forcing results may still be somewhat acceptable. For example, while Nordhaus (1977) and Schneider and Thompson (1981) projected more warming than observed, their radiative forcings are not unacceptable. Similarly, while a specific climate scenario of Hansen *et al.* (1981) (Scenario 2a) projects less warming than observed, its radiative forcings are not inconsistent with observations. The models of Manabe (1970), Mitchell (1970), Benson (1970), Broecker (1975) display radiative forcings on the high end of the observational ensemble-based range, but still provide somewhat acceptable temperature projections.

The temperature versus time skill scores are reasonably high for many of the 1970-era models turned out to be reasonably high, but not so for the more modern models. However, the skill score of Hansen *et al.* (1981) (Scenario 1) has been the highest (0.93). Regarding the skills with respect to temperature versus change in radiative forcing, the 1970-era models do not perform as well, and the more modern models seem to have

better skills. Here Hansen *et al.* (1981) (Scenario 2a) and Manabe and Stouffer (1993) models seem to possess the best skills, both having the score 0.97.

Although the methods used by Hausfather *et al.* (2020) for their evaluations may be deemed useful by climatologists, any well-trained statistician might feel less comfortable with such methods and may demand more sophisticated and probabilistically sound methods that properly quantify uncertainties. For instance, a high skill score indicates close agreement between observed and projected temperatures for the small time frame considered, but for longer time periods the differences between observed and projected temperatures may turn out to be increasing. Signs of this phenomenon may already be present even within the small time frames, which could not be captured by a single number, namely, the skill score. Thus, proper probabilistic quantifications, particularly respecting the time-series aspects, should be allowed to play the key role in responsible evaluations of climate models. The Bayesian paradigm immediately suggests itself in this context. Indeed, our methods are constructed from such a perspective, and it would be interesting to evaluate the performances of the models considered by Hausfather *et al.* (2020) using our methods. We reserve this endeavor for some interesting future works.

10.12 Summary and discussion

As stated in Lupo *et al.* (2013) (see also the references therein), “When physicists, biologists, and other scientists who are unaware of the rules of forecasting attempt to make climate predictions, their forecasts are at risk of being no more reliable than those made by non-experts, even when they are communicated through complex computer models”. The GCMs are indeed complex computer models built by physicists, biologists, and other scientists. The future global warming forecasts yielded by such models have great bearing on the current world and particularly on the IPCC policymakers. But as discussed by Lupo *et al.* (2013) in great detail, major scientists of the world do not find much reason to pin faith on the global warming foreboding, and most of them, based

on their experiments and experiences, are strongly critical of the abilities of GCM to adequately model so complex a system as world climate.

Despite the existing works on climate model evaluations discussed in Section 10.11, as elucidated in the same section, rigorous statistical research that evaluates the GCM-based global warming projections, is of utmost importance. Such a task, which is of global importance, must be seriously undertaken and no wonder statistics is the only discipline that can promise to make justice to such an issue where quantification of uncertainties (in the predictions by the GCMs) plays the most important role. It is also very well-established that the Bayesian statistical paradigm is the most well-equipped to coherently deal with uncertainty quantifications.

As such, we undertake the task, in the Bayesian framework, of evaluating the global warming forecasts, with observed current temperature data and GCM-simulated data obtained from the IPCC website. We first consider dynamically but nonparametrically modeling temperature using Gaussian process emulation procedure, borrowing ideas from [Bhattacharya \(2007\)](#) and [Ghosh *et al.* \(2014\)](#). Such a modeling strategy seems to quite appropriate for nonparametrically addressing the uncertainties in dynamic climate change, in the absence of any known model framework for either the GCMs or the observed current temperature.

With such a Gaussian process based model we then attempt to address the question of how to select the best GCM within any climate scenario using the principle of our Bayesian multiple testing procedure that provides rigorous assessment of the projected future dynamics (forward sense) of the GCMs as well as their abilities to predict the observed HadCRUT4 data (inverse regression sense), and yields the best model by comparing these combined abilities in a theoretically sound manner. The procedure, along with discrepancy measure based theory of Bayesian model adequacy test proposed in [Bhattacharya \(2013\)](#) provides the additional evaluation if the best models adequately satisfy the goodness-of-fit test with respect to fitting the HadCRUT4 data, given the

future projections.

Such evaluations are previously contemplated upon in the climate context by other researchers: for example, Lupo *et al.* (2013), quoting Reifen and Toumi (2009), write “Expounding on this principle, Reifen and Toumi (2009) note, “with the ever increasing number of models, the question arises of how to make a best estimate prediction of future temperature change.” That is to say, which model should one use? With respect to this question, they note, “one key assumption, on which the principle of performance-based selection rests, is that a model which performs better in one time period will continue to perform better in the future.” In other words, if a model predicts past climate fairly well, it should predict future climate fairly well. The principle sounds reasonable enough, but does it hold true?”

To our knowledge, there does not exist any other sound statistical analysis in this respect. We believe that ours is a significant contribution from this perspective, where the nonparametric Gaussian process dynamics substantially adds to the overall novelty. Employment of efficient C-coding and parallel computing architectures ensure quite cheap computation, in spite of evaluations of a large number of GCMs using intricate Bayesian nonparametric models and methods.

Our results on model selection and evaluation fail to provide any evidence in favour of global warming, challenging all our evaluated GCMs along the way. To further strengthen our results regarding these, we model the GCM forecasts in any given climate scenario as ensembles, which we model by extending our one-dimensional climate dynamics to multidimensional climate dynamics, driven by multidimensional Gaussian processes. Attempts to predict the observed HadCRUT4 data given the future ensembles of GCM forecasts, resulted in conspicuously poor fits. This strongly reinforces that future global warming, as projected by the GCMs, need not turn out to be the reality.

Finally, based on our one-dimensional Gaussian process based climate dynamics, we provide Bayesian climate forecasts into the future, conditioned on the HadCRUT4 data.

Our results quite persuasively demonstrate that the future does not hold the drastic global warming doom for the world, in stark contrast with the GCM warnings. An important finding in this respect is that the best model of the Commitment scenario at least falls closer to our high posterior density regions of our future predictions as well as those of current HadCRUT4 prediction given the future simulations, compared to the other GCMs. Since the greenhouse gases are held fixed in the Commitment case, it is probably not unreasonable to question the role of greenhouse gases as important drivers of climate change. Our result in this regard seems to be broadly supported by some scientific experiments conducted by Hansen *et al.* (1998). Quoting these authors and criticizing IPCC's faith in the state-of-the-art climate projections, Idso *et al.* (2013a) write “Hansen *et al.* (1998) examined the forcings of well-mixed greenhouse gases (CO₂, CH₄, N₂O, and CFCs), tropospheric ozone, stratospheric ozone, tropospheric aerosols, forced cloud changes, vegetation and other planetary surface alterations, solar variability, and volcanic aerosols. That examination revealed so many uncertainties in the forcings that the researchers concluded, “the forcings that drive long-term climate change are not known with an accuracy sufficient to define future climate change.” Nevertheless, the IPCC has expressed confidence in projections of future climate, saying the temperature sensitivity of Earth’s climate system in response to a doubling of atmospheric CO₂ concentrations “is *likely* to be in the range 2°C to 4.5°C with a best estimate of about 3°C, and is very *unlikely* to be less than 1.5°C [italics in the original]” (IPCC (2007)).”

In summary, our novel Bayesian models and methodologies provide a peek into the future world, which does not seem to be as gloomy as portrayed by the non-Bayesian scientists and the IPCC policymakers.

11

Summary and Future Directions

11.1 Summary

As an important part of this thesis, we have attempted to clarify the differences between the traditional inverse problems and the inverse regression problems. Although the so-called “ill-posed” inverse problems, essentially on function estimation, occupy significantly larger space in the literature compared to inverse regression problems, we have argued that strictly speaking, only the latter class of problems can be regarded as authentic inverse problems, and includes the traditional inverse problems as special cases when learning unknown covariates as well as unknown functions are of interest.

Our investigation of Bayesian inverse regression has led to the conclusion that posterior covariate consistency is not achievable for general priors. However, we have proved that for judiciously chosen data-driven priors, covariate consistency holds. And this holds quite generally, even for nonparametric Bayesian models involving unknown functions modeled

by appropriate stochastic processes, the Gaussian process being the most popular. The results and the detailed proofs in this regard that include rate of convergence and misspecification of the underlying functions, are of independent interest as well, apart from aiding the proofs on Bayesian covariate consistency in the LOO-CV setup. Bayesian covariate consistency finds further utilization in our proof of consistency of the IRD approach introduced in [Bhattacharya \(2013\)](#).

Asymptotic validity of the IRD approach to goodness-of-fit tests for Bayesian inverse models is gratifying but as it is, the IRD method is incapable of handling inverse model selection. The existing methods of model selection are not equipped with the ability to make appropriate selection among inverse models. To deal with inverse model and covariate selection, we began with the traditional, but arguably the most principled approach to model selection, namely, the Bayes factor, and established its asymptotic convergence properties in as much general terms as possible, without particular reference to inverse regression. The setup and the asymptotic theory are valid even for inverse model and covariate selection, as we clarified subsequently, in the context of pseudo-Bayes factors. Indeed, we developed the general asymptotic theory of pseudo-Bayes factors for both forward and inverse regression setups, and have shown that the final convergence results are in agreement with our Bayes factor asymptotic results, for both forward and inverse regression. This inheritance of the very desirable asymptotic properties of Bayes factor is an welcome addition to its general usefulness, since in practice, pseudo-Bayes factors already have some distinct advantages over Bayes factors in terms of alleviating its theoretical inadequacies as well as significantly improving its computational inefficiency.

Considering the Bayes factor and pseudo-Bayes factor approaches to inverse model selection, the essence of the IRD approach seems to be relegated to the background. However, it returns with an important role in our new Bayesian multiple testing procedure for inverse model selection. The discrepancy measures of the IRD approach, which now feature in the hypotheses, play the pivotal role in our multiple testing strategy.

Development of the asymptotic theory for this procedure required us to borrow strength from our previous asymptotic theories of Bayes and pseudo-Bayes factors and covariate consistency. Very importantly, our simulation experiments demonstrate that our Bayesian multiple testing procedure can improve upon the results of both forward and inverse pseudo-Bayes factors. Since there does not exist any other multiple testing method for model or covariate selection in the inverse setup, these interesting properties seem to make our Bayesian multiple testing strategy for inverse model and covariate selection all the more important.

11.2 Future directions

11.2.1 Past climate reconstruction

Since palaeoclimate reconstruction has played a motivating role behind this thesis, it is worth making a few remarks regarding applications of the methods developed in this thesis to such problems. First, recall from Chapter 9.9 that using the IRD approach, Bhattacharya (2013) (see also Bhattacharya (2004)) showed that the Bayesian palaeoclimate model of Vasko *et al.* (2000) underfits the modern chironomid data while the Bayesian model proposed in Haslett *et al.* (2006) overfits the modern pollen data, as established in Bhattacharya (2004). Hence, based on the respective models, the Holocene temperature reconstructions of Korhola *et al.* (2002) in northern Fennoscandia and Glendalough palaeoclimate reconstructions of Haslett *et al.* (2006), are not unquestionable. On the other hand, the general and flexible Bayesian semiparametric palaeoclimate model proposed in Mukhopadhyay and Bhattacharya (2013) convincingly fits both the modern chironomid and pollen data, as confirmed by the IRD approach. Hence, it makes sense to reconstruct these past climates using the model of Mukhopadhyay and Bhattacharya (2013). We anticipate that most past climate reconstructions reported in the literature can be significantly improved upon using the aforementioned modeling

approaches. Model and/or covariate selection, if necessary, can be reliably addressed by our inverse pseudo-Bayes factor and inverse multiple testing approaches, as demonstrated in Chapters 8 and 9. As noted in Chapter 9, the multiple testing approach is expected to be particularly useful in this regard.

11.2.2 Function optimization

The class of inverse regression problems finds an unlikely candidate in function optimization, thanks to the recent work of Roy and Bhattacharya (2020). We briefly clarify this below.

Roy and Bhattacharya (2020) propose and develop a novel Bayesian algorithm for optimization of functions whose first and second partial derivatives are available. Their approach is to embed the underlying function, 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 their methodology. In a nutshell, with data consisting of suitable choices of input points in the function domain and their function values, they first obtain the posterior derivative process corresponding to the original Gaussian process. Then they construct the posterior distribution of the solutions corresponding to setting random partial derivative functions to the null vector. This posterior emulates the stationary points of the objective function. They 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 emulate the true optima even if the dataset is not large enough.

The inverse regression context is evident in the step where the posterior distribution of the solutions corresponding to setting random partial derivative functions to the null vector is considered. That is, denoting the vector of random partial derivatives by \mathbf{g}' ,

here the interest lies in learning about the posterior distribution of $\{\mathbf{x} : \mathbf{g}'(\mathbf{x}) = \mathbf{0}\}$.

The theory and methods in this novel function optimization premise has yielded quite encouraging results, as reported in Roy and Bhattacharya (2020). It thus makes sense to extend this idea to multi-objective optimization problems involving multiple objective functions to be optimized simultaneously. Multi-objective optimization is important in engineering, economics and logistics where optimal decisions need to be taken in the presence of trade-offs between two or more conflicting objectives. Examples of multi-objective optimization are minimizing cost while maximizing comfort while buying a car, and maximizing performance while minimizing fuel consumption and emission of pollutants of a vehicle. For details on multi-objective optimization, see Deb (2001).

References

- Adler, R. J. (1981). *The Geometry of Random Fields*. John Wiley & Sons Ltd., New York.
- Adler, R. J. and Taylor, J. E. (2007). *Random Fields and Geometry*. Springer, New York.
- Arbogast, T. and Bona, J. L. (2008). Methods of Applied Mathematics. University of Texas at Austin.
- Aronszajn, N. (1950). Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, **68**, 337–404.
- Ash, R. B. and Gardner, M. F. (1975). *Topics in Stochastic Processes*. Academic Press, New York.
- Aster, R. C., Borchers, B., and Thurber, C. H. (2013). *Parameter Estimation and Inverse Problems*. Academic Press, Oxford, UK.
- Avenhaus, R., Höpfinger, E., and Jewell, W. S. (1980). Approaches to Inverse Linear Regression. Technical Report. Available at <https://publikationen.bibliothek.kit.edu/270015256/3812158>.
- Baker, C. T. H. (1977). *The Numerical Treatment of Integral Equations*. Clarendon Press, Oxford.
- Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, USA.

- Bartlett, M. (1957). A Comment on D. V. Lindley's Statistical Paradox. *Bometrika*, **44**, 533–534.
- Bayarri, M. J. and Berger, J. O. (2000). P-Values for Composite Null Models (with discussion). *Journal of the American Statistical Association*, **95**, 1127–1142.
- Bennett, G. (1962). Probability Inequalities for the Sums of Independent Random Variables. *Journal of the American Statistical Association*, **57**, 33–45.
- Benson, G. S. (1970). Carbon Dioxide and its Role in Climate Change. *Proceedings of the National Academy of Sciences*, **67**, 898–899.
- Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.
- Bhattacharya, S. (2004). *Importance Resampling MCMC: A Methodology for Cross-Validation in Inverse Problems and its Applications in Model Assessment*. Doctoral thesis, Department of Statistics, Trinity College Dublin.
- Bhattacharya, S. (2006). A Bayesian Semiparametric Model for Organism Based Environmental Reconstruction. *Environmetrics*, **17**, 763–776.
- Bhattacharya, S. (2007). A Simulation Approach to Bayesian Emulation of Complex Dynamic Computer Models. *Bayesian Analysis*, **2**, 783–815.
- Bhattacharya, S. (2008). Gibbs Sampling Based Bayesian Analysis of Mixtures with Unknown Number of Components. *Sankhya. Series B*, **70**, 133–155.
- Bhattacharya, S. (2013). A Fully Bayesian Approach to Assessment of Model Adequacy in Inverse Problems. *Statistical Methodology*, **12**, 71–83.
- Bhattacharya, S. and Haslett, J. (2007). Importance Resampling MCMC for Cross-Validation in Inverse Problems. *Bayesian Analysis*, **2**, 385–408.

- Billingsley, P. (1995). *Probability and Measure*. John Wiley and Sons, New York.
- Billingsley, P. (2013). *Convergence of Probability Measures*. John Wiley & Sons, New Jersey, USA.
- Blackwell, D. and Dubins, L. (1962). Merging of Opinions with Increasing Opinions. *The Annals of Mathematical Statistics*, **33**, 882–886.
- Broecker, W. S. (1975). Climatic Change: Are We on the Brink of a Pronounced Global Warming? *Science*, **189**, 460–463.
- Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer, New York.
- Bui-Thanh, T. (2012). A Gentle Tutorial on Statistical Inversion Using the Bayesian Paradigm. ICES Report 12-18. Available at <http://users.ices.utexas.edu/~tanbui/PublishedPapers/BayesianTutorial.pdf>.
- Calvetti, D. and Somersalo, E. (2007). *Introduction to Bayesian Scientific Computing: Ten Lectures on Subjective Computing*. Springer, New York.
- Chakrabarty, D., Biswas, M., and Bhattacharya, S. (2015). Bayesian Nonparametric Estimation of Milky Way Parameters Using Matrix-Variate Data, in a New Gaussian Process Based Method. *Electronic Journal of Statistics*, **9**, 1378–1403.
- Chandra, N. K. and Bhattacharya, S. (2019). Non-marginal Decisions: A Novel Bayesian Multiple Testing Procedure. *Electronic Journal of Statistics*, **13**(1), 489–535.
- Chandra, N. K. and Bhattacharya, S. (2020a). Asymptotic Theory of Dependent Bayesian Multiple Testing Procedures Under Possible Model Misspecification. *Annals of the Institute of Statistical Mathematics*. To appear.

- Chandra, N. K. and Bhattacharya, S. (2020b). High-dimensional Asymptotic Theory of Bayesian Multiple Testing Procedures Under General Dependent Setup and Possible Misspecification. ArXiv Preprint.
- Chib, S. (1995). Marginal Output from the Gibbs Output. *Jounal of the American Statistical Association*, **90**, 1313–1321.
- Chib, S. and Kuffner, T. A. (2016). Bayes Factor Consistency. Available at arXiv:1607.00292.
- Choi, T. (2009). Asymptotic Properties of Posterior Distributions in Nonparametric Regression with Non-Gaussian Errors. *Annals of the Institute of Statistical Mathematics*, **61**, 835–859.
- Choi, T. and Rousseau, J. (2015). A Note on Bayes Factor Consistency in Partial Linear Models. *Jounal of Statistical Planning and Inference*, **166**, 158–170.
- Choi, T. and Schervish, M. J. (2007). On Posterior Consistency in Nonparametric Regression Problems. *Journal of Multivariate Analysis*, **98**, 1969–1987.
- Choudhuri, N., Ghosal, S., and Roy, A. (2007). Nonparametric Binary Rgression using a Gaussian Process Prior. *Statistical Methodology*, **4**, 227–243.
- Climate Action Tracker (2019). Warming Projections Global Update. December 2019 (Report). Available at https://climateactiontracker.org/documents/698/CAT_2019-12-10_BriefingCOP25_WarmingProjectionsGlobalUpdate_Dec2019.pdf.
- Cramer, H. and Leadbetter, M. R. (1967). *Stationary and Related Stochastic Processes*. Wiley, New York.
- Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Wiley, New York.
- Dashti, M. and Stuart, A. M. (2015). The Bayesian Approach to Inverse Problems. eprint: arXiv:1302.6989.

- Dawid, A. P. (1992). Prequential Analysis, Stochastic Complexity and Bayesian Inference (with discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 109–125, Oxford. Oxford University Press.
- de Lange, W. and Carter, R. M. (2013). Observations: The Hydrosphere and Ocean. In C. D. Isdo, R. M. Carter, and S. F. Singer, editors, *Climate Change Reconsidered II: Physical Science*, pages 149–246, Chicago, IL: The Heartland Institute.
- Deb, K. (2001). *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, Inc, U. K.
- Dey, D. K., Müller, P., and Sinha, D. (2012). *Practical Nonparametric and Semiparametric Bayesian Statistics*. Springer, New York, USA.
- Annals of Statistics, **14**, 1–67.

Diaconis, P. and Freedman, D. A. (1993). Nonparametric Binary Regression: A Bayesian Approach. *The Annals of Statistics*, **21**, 2108–2137.

Duchon, J. (1977). Splines Minimizing Rotation-Invariant Semi-norms in Sobolev Spaces. In W. Schempp and K. Zellner, editors, *Constructive Theory of Functions of Several Variables*, pages 85–100, New York. Springer-Verlag.

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>.

Dyson, F. (2007). Heretical Thoughts About Science and Society. Edge: The Third Culture. August.

- Efromovich, S. (2008). *Nonparametric Curve Estimation: Methods, Theory, and Applications*. Springer, New York, USA.
- Engl, H. W., Hanke, M., and Neubauer, A. (1996). *Regularization of Inverse Problems*. Kluwer Academic Publishers Group, Dordrecht. Volume 375 of Mathematics and its Applications.
- Escobar, M. D. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, **90**(430), 577–588.
- Eubank, R. (1999). *Nonparametric Regression and Spline Smoothing*. CRC Press, New York, USA.
- Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, **1**, 209–230.
- Frame, D. J. and Stone, D. A. (2012). Assessment of the First Consensus Prediction on Climate Change. *Nature Climate Change*, **3**, 357–359.
- Geisser, S. and Eddy, W. F. (1979). A Predictive Approach to Model Selection. *Journal of the American Statistical Association*, **74**(365), 153–160.
- Gelfand, A. E. (1996). Model Determination Using Sampling-Based Methods. In W. Gilks, S. Richardson, and D. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, Interdisciplinary Statistics, pages 145–162, London. Chapman and Hall.
- Gelfand, A. E. and Dey, D. K. (1994). Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society B*, **56**(3), 501–514.
- Gelfand, A. E. and Kuo, L. (1991). Nonparametric Bayesian Bioassay Including Ordered Polytomous Response. *Biometrika*, **78**, 657–666.

- Gelman, A. and Meng, X.-L. (1998). Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling. *Statistical Science*, **13**(2), 163–185.
- Gelman, A., Meng, X. L., and Stern, H. S. (1996). Posterior Predictive Assessment of Model Fitness via Realized Discrepancies (with discussion). *Statistica Sinica*, **6**, 733–807.
- Geman, S. and Hwang, C. R. (1982). Nonparametric Maximum Likelihood Estimation by the Method of Sieves. *The Annals of Statistics*, **10**, 401–414.
- Ghosal, A. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, Cambridge, UK.
- Ghosal, S. and Roy, A. (2006). Posterior Consistency of Gaussian Process Prior for Nonparametric Binary Regression. *The Annals of Statistics*, **34**, 2413–2429.
- Ghosal, S., Lember, J., and van der Vaart, A. W. (2008). Nonparametric Bayesian Model Selection and Averaging. *Electronic Journal of Statistics*, **2**, 63–89.
- Ghosh, A., Mukhopadhyay, S., Roy, S., and Bhattacharya, S. (2014). Bayesian Inference in Nonparametric Dynamic State-Space Models. *Statistical Methodology*, **21**, 35–48.
- Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer-Verlag, New York.
- Giraud, C. (2015). *Introduction to High-Dimensional Statistics*. Chapman and Hall, New York.
- Green, K. C., Armstrong, J. S., and Soon, W. (2009). Validity of Climate Change Forecasting for Public Policy Decision Making. *International Journal of Forecasting*, **25**, 826–832.

- Green, P. J. and Silverman, B. W. (1993). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. CRC Press, New York, USA.
- Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, A., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., and Steingroever, H. (2017). A Tutorial on Bridge Sampling. *Journal of Mathematical Psychology*, **81**, 80–97.
- Guindani, M., Müller, P., and Zhang, S. (2009). A bayesian discovery procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**(5), 905–925.
- Hansen, J., Johnson, D., Lacis, A., Lebedeff, S., Lee, P., rind, D., and Russell, G. (1981). Climate Impact of Increasing Atmospheric Carbon Dioxide. *Science*, **213**, 957–966.
- Hansen, J., Fung, I., Lacis, A., Rind, D., Lebedeff, S., and Ruedy, R. (1988). Global Climate Changes as Forecast by Goddard Institute for Space Studies Three-Dimensional Model. *Journal of Geophysical Research*, **93**, 9341–9364.
- Hansen, J. E., Sato, M., Lacis, A., Ruedy, R., Tegan, I., and Matthews, E. (1998). Climate Forcings in the Industrial Era. *Proceedings of the National Academy of Sciences, U.S.A.*, **95**, 12753–12758.
- Härdle, W. K. (1990). *Applied Nonparametric Regression*. Cambridge University Press, Cambridge, UK.
- Härdle, W. K., Müller, M., Sperlich, S., and Werwatz, A. (2012). *Nonparametric and Semiparametric Models*. Springer, New York, USA.
- Hargreaves, J. C. (2010). Skill and Uncertainty in Climate Models. *Wiley Interdisciplinary Reviews: Climate Change*, **1**, 556–564.
- Haslett, J., Whiley, M., Bhattacharya, S., Salter-Townshend, M., Wilson, S. P., Allen, J. R. M., Huntley, B., and Mitchell, F. J. G. (2006). Bayesian Palaeoclimate Reconstruction

- (with discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **169**, 395–438.
- Hausfather, Z., Drake, H. F., Abbott, T., and Schmidt, G. A. (2020). Evaluating the Performance of Past Climate Model Projections. *Geophysical Research Letters*, **47**.
- Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge University Press, Cambridge, UK.
- Hoadley, B. (1970). A Bayesian Look at Inverse Linear Regression. *Journal of the American Statistical Association*, **65**, 356–369.
- Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, **58**, 13–30.
- Idso, C., Ball, T., and Segalstad, T. (2013a). forcings and Feedbacks. In C. D. Idso, R. M. Carter, and S. F. Singer, editors, *Climate Change Reconsidered II: Physical Science*, pages 149–246, Chicago, IL: The Heartland Institute.
- Idso, S., Idso, C., Singer, S. F., McKittrick, R., and Spencer, R. (2013b). Observations: Temperature Records. In C. D. Idso, R. M. Carter, and S. F. Singer, editors, *Climate Change Reconsidered II: Physical Science*, pages 149–246, Chicago, IL: The Heartland Institute.
- IPCC (2007). Climate Change 2007: The Physical Science Basis. In S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, editors, *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge, UK: Cambridge University Press.
- IPCC (2018). Summary for Policymakers. In V. P. Masson-Delmotte, P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P. R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan,

- R. Pidcock, S. Connors, J. B. R. Matthews, Y. Chen, X. Zhou, M. I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfield, editors, *Global Warming of 1.5° C. An IPCC Special Report on the Impacts of Global Warming of 1.5° C Above Pre-industrial Levels and Related Global Greenhouse Gas Emission Pathways, in the Context of Strengthening the Global Response to the Threat of Climate Change, Sustainable Development, and Efforts to Eradicate Poverty*. In Press. Available at https://www.ipcc.ch/site/assets/uploads/sites/2/2019/06/SR15_Full_Report_High_Res.pdf.
- IPCC (AR4) (2007). The Physical Scientific Basis, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. In S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, editors, *Climate Change 2007*, Cambridge, UK and New York, NY. Cambridge University Press.
- IPCC (FAR) (1990). Climate Change: The IPCC Scientific Assessment. Report Prepared by Working Group I. In J. T. Houghton, G. J. Jenkins, and J. J. Ephraums, editors, *Intergovernmental Panel on Climate Change*, page 365, Cambridge, UK and New York, NY. Cambridge University Press.
- IPCC (TAR) (2001). The Scientific Basis, Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change. In J. T. Houghton, Y. Ding, G. J. Griggs, M. Noguer, P. J. van der Linden, X. Dai, K. Maskell, and C. A. Johnson, editors, *Climate Change 2001*, Cambridge, UK and New York, NY. Cambridge University Press.
- Jeffreys, H. (1939). *Theory of Probability*. 1st edition. The Clarendon Press, Oxford.
- 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.

- Kass, R. E. and Raftery, R. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**(430), 773–795.
- Kimeldorf, G. and Wahba, G. (1971). Some Results on Tchebycheffian Spline Functions. *Journal of Mathematical Analysis and Applications*, **33**, 82–95.
- Knapik, B. and Salomond, J. B. (2018). A General Approach to Posterior Contraction in Nonparametric Inverse Problems. *Bernoulli*, **24**, 2091–2121.
- Knapik, B. T., van der Vaart, A. W., and van Zanten, J. H. (2011). Bayesian Inverse Problems with Gaussian Priors. *The Annals of Statistics*, **39**, 2626–2657.
- König, H. (1986). *Eigenvalue Distribution of Compact Operators*. Birkhäuser.
- Korhola, A., Vasko, K., Toivonen, H. T. T., and Olander, H. (2002). Holocene Temperature Changes in Northern Fennoscandia Reconstructed from Chironomids Using Bayesian Modelling. *Quaternary Science Reviews*, **21**, 1841–1860.
- Krutchkoff, R. G. (1967). Classical and Inverse Regression Methods of Calibration. *Technometrics*, **9**, 425–435.
- Kundu, S. and Dunson, D. B. (2014). Bayes Variable Selection in Semiparametric Linear Models. *Jounal of American Statistical Association*, **109**, 437–447.
- Lange, K. (2010). *Numerical Analysis for Statisticians*. New York, Springer.
- Lavine, M. (1992). Some Aspects of Polya Tree Distributions for Statistical Modelling. *The Annals of Statistics*, **20**, 1222–1235.
- Lavine, M. (1994). More Aspects of Polya Tree Distributions for Statistical Modelling. *The Annals of Statistics*, **22**, 1161–1176.
- Lenk, P. J. (1988). The Logistic Normal Distribution for Bayesian, Nonparametric, Predictive Densities. *Journal of the American Statistical Association*, **83**, 509–516.

- Lenk, P. J. (1991). Towards a Practicable Bayesian Nonparametric Density Estimator. *Biometrika*, **78**, 531–543.
- Lenk, P. J. (2003). Bayesian Semiparametric Density Estimation and Model Verification Using a Logistic-Gaussian Process. *Journal of Computational and Graphical Statistics*, **12**, 548–565.
- Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g Priors for Bayesian Variable Selection. *Journal of American Statistical Association*, **103**(481), 410–423.
- Lindley, D. (1957). A Statistical Paradox. *Biometrika*, **44**, 187–192.
- Lorentz, G. G. (1966). Metric Entropy and Approximation. *Bulletin of the American Mathematical Society*, **72**, 903–937.
- Lupo, A., Kininmonth, W., Armstrong, J. S., and Green, K. (2013). Global Climate Models and Their Limitations. In C. D. Isdo, R. M. Carter, and S. F. Singer, editors, *Climate Change Reconsidered II: Physical Science*, pages 7–148, Chicago, IL: The Heartland Institute.
- Macon, N. and Spitzbart, A. (1958). Inverses of Vandermonde Matrices. *The American Mathematical Monthly*, **65**(2), 95–100.
- Maitra, T. and Bhattacharya, S. (2016a). Asymptotic Theory of Bayes Factor in Stochastic Differential Equations: Part I. Available at <https://arxiv.org/pdf/1503.09011.pdf>.
- Maitra, T. and Bhattacharya, S. (2016b). On Convergence of Bayes Factor in Stochastic Differential Equations: Part II. Available at <https://arxiv.org/pdf/1504.00002.pdf>.
- Manabe, S. (1970). The Dependence of Atmospheric Temperature on the Concentration of Carbon Dioxide. In S. F. Singer, editor, *Global Effects of Environmental Pollution*, pages 25–29, Dordrecht. Springer.

- Manabe, S. and Stouffer, R. J. (1993). Century-scale Effects of Increased Atmospheric CO₂ on the Ocean-atmosphere System. *Nature*, **364**, 215–218.
- Massart, P. (2003). Concentration Inequalities and Model Selection. Volume 1896 of Lecture Notes in Mathematics. Springer-Verlag. Lectures given at the 33rd Probability Summer School in Saint-Flour.
- Masson-Delmotte, V., Schulz, M., Abe-Ouchi, A., Beer, J., *et al.* (2013). Chapter 5: Information from Paleoclimate Archives. IPCC AR5 WG1 2013.
- Meinguet, J. (1979). Multivariate Interpolation at Arbitrary Points Made Simple. *Journal of the Applied Mathematics and Physics*, **30**, 292–304.
- Meng, X. L. and Wong, W. H. (1996). Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration. *Statistica Sinica*, **6**, 831–860.
- Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, London.
- Mitchell, J. M. (1970). A Preliminary Evaluation of Atmospheric Pollution as a Cause of the Global Temperature Fluctuation of the Past Century. In S. F. Singer, editor, *Global Effects of Environmental Pollution*, pages 139–155, Dordrecht. Springer.
- Mukhopadhyay, S. and Bhattacharya, S. (2013). Cross-Validation Based Assessment of a New Bayesian Palaeoclimate Model. *Environmetrics*, **24**, 550–568.
- Mukhopadhyay, S., Bhattacharya, S., and Dihidar, K. (2011). On Bayesian “Central Clustering”: Application to Landscape Classification of Western Ghats. *Annals of Applied Statistics*, **5**, 1948–1977.
- Mukhopadhyay, S., Roy, S., and Bhattacharya, S. (2012). Fast and Efficient Bayesian Semi-parametric Curve-fitting and Clustering in Massive Data. *Sankhya. Series B*, **71**, 77–106.

- Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004). Optimal Sample Size for Multiple Testing: the Case of Gene Expression Microarrays. *Journal of the American Statistical Association*, **99**(468), 990–1001.
- Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*. Springer, New York, USA.
- Muth, R. F. (1960). The Demand for Non-Farm Housing. In A. C. Harberger, editor, *The Demand for Durable Goods*. The University of Chicago.
- Newey, W. K. (1991). Uniform Convergence in Probability and Stochastic Equicontinuity. *Econometrica*, **59**, 1161–1167.
- Newton, M. A., Czado, C., and Chappell, R. (1996). Bayesian Inference for Semi-parametric Binary Regression. *Journal of the American Statistical Association*, **91**, 142–153.
- Nordhaus, W. (1977). Strategies for the Control of Carbon Dioxide (cowles foundation discussion papers). Cowles Foundation for Research in Economics, Yale University. Retrieved from <https://econpapers.repec.org/RePEc:cwl:cwldpp:443>.
- Olivier, J. G. and Peters, J. A. H. W. (2019). Trends in Global CO₂ and Total Greenhouse Gas Emissions. The Hague: PBL Netherlands Environmental Assessment Agency.
- Orbanz, P. (2014). Lecture Notes on Bayesian Nonparametrics. Available at http://stat.columbia.edu/~porbanz/papers/porbanz_BNP_draft.pdf.
- O’Sullivan, F. (1986). A Statistical Perspective on Ill-Posed Inverse Problems. *Statistical Science*, **1**, 502–512.

- O'Sullivan, F., Yandell, B. S., and Raynor, W. J. (1986). Automatic Smoothing of Regression Functions in Generalized Linear Models. *Journal of the American Statistical Association*, **81**, 96–103.
- Pillai, N., Wolpert, R. L., and Clyde, M. A. (2007). A Note on Posterior Consistency of Nonparametric Poisson Regression Models. Available at <https://pdfs.semanticscholar.org/27f5/af4d00cef092c8b19662951cc316c2e222b7.pdf>.
- Press, S. J. and Scott, A. (1975). Missing Variables in Bayesian Regression. In S. E. Fienberg and A. Zellner, editors, *Studies in Bayesian Econometrics and Statistics*, North-Holland, Amsterdam.
- Rahmstorf, S., Cazenave, A., Church, J. A., Hansen, J. E., Keeling, R. F., Parker, D. E., and Somerville, R. C. J. (2007). Recent Climate Observations Compared to Projections. *Science*, **316**, 709–709.
- Rahmstorf, S., Foster, G., and Cazenave, A. (2012). Comparing Climate Projections to Observations up to 2011. *Environmental Research Letters*, **7**, 44035.
- Rasch, D., Enderlein, G., and Herrendörfer, G. (1973). Biometrie. Deutscher Landwirtschaftsverlag, Berlin.
- Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, Massachusetts.
- Rasool, S. L. and Schneider, S. H. (1971). Atmospheric Carbon Dioxide and Aerosols: Effects of Large Increases on Global Climate. *Science*, **173**, 138–141.
- Reifen, C. and Toumi, R. (2009). Climate Projections: Past Performance No Guarantee of Future Skill? DOI: 10.1029/2009GL038082.
- Robert, C. P. (1993). A Note on Jeffreys-Lindley Paradox. *Statistica Sinica*, **3**, 601–608.

- Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer-Verlag, New York.
- Roy, S. and Bhattacharya, S. (2020). Function Optimization with Posterior Gaussian Derivative Process. arXiv:2010.13591.
- Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. Chapman & Hall/CRC, Boca Raton.
- Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer Series in Statistics. Springer-Verlag, New York, Inc.
- Sarkar, S. K., Zhou, T., and Ghosh, D. (2008). A General Decision Theoretic Formulation of Procedures Controlling FDR and FNR from a Bayesian Perspective. *Statistica Sinica*, **18**(3), 925–945.
- Sawyer, J. S. (1972). Man-made Carbon Dioxide and the “greenhouse” Effect. *Nature*, **239**, 23–26.
- Schervish, M. J. (1995). *Theory of Statistics*. Springer-Verlag, New York.
- Schimek, M. J. (2013). *Smoothing and Regression: Approaches, Computation, and Application*. John Wiley and Sons, New Jersey, USA.
- Schneider, S. H. and Thompson, S. L. (1981). Atmospheric CO₂ and Climate: Importance of the Transient Response. *Journal of Geophysical Research*, **86**, 3135–3147.
- Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels*. MIT Press, USA.
- Sethuraman, J. (1994). A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, **4**, 639–650.
- Shalizi, C. R. (2009). Dynamics of Bayesian Updating With Dependent Data and Misspecified Models. *Electronic Journal of Statistics*, **3**, 1039–1074.

- Shepp, L. A. (1966). Radon-Nikodym Derivatives of Gaussian Measures. *Annals of Mathematical Statistics*, **37**, 321–354.
- Shimodaira, H. (1998). An Application of Model Comparison Techniques to Model Selection. *Annals of the Institute of Statistical Mathematics*, **50**(1), 1–13.
- Storey, J. D. (2003). The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value. *Ann. Statist.*, **31**(6), 2013–2035.
- Stouffer, R. J. and Manabe, S. (2017). Assessing Temperature Pattern Projections Made in 1989. *Nature Climate Change*, **7**, 163–165.
- Stouffer, R. J., Manabe, S., and Bryan, K. (1989). Interhemispheric Asymmetry in Climate Response to a Gradual Increase of Atmospheric CO₂. *Nature*, **342**, 660–662.
- Takezawa, K. (2006). *Introduction to Nonparametric Regression*. John Wiley & Sons, New Jersey, USA.
- Tikhonov, A. (1963). Solution of Incorrectly Formulated Problems and the Regularization Method. *Soviet Math. Dokl.*, **5**, 1035–1038.
- Tikhonov, A. and Arsenin, V. (1977). *Solution of Ill-Posed Problems*. Wiley, New York.
- Uspensky, J. V. (1937). *Introduction to Mathematical Probability*. McGraw-Hill, New York, USA.
- van der Vaart, A. W. and van Zanten, J. H. (2008). Rates of Contraction of Posterior Distributions Based on Gaussian Process Priors. *The Annals of Statistics*, **36**, 1435–1463.
- van der Vaart, A. W. and van Zanten, J. H. (2009). Adaptive Bayesian Estimation Using a Gaussian Random Field with Inverse Gamma Bandwidth. *The Annals of Statistics*, **37**, 2655–2675.

- van der Vaart, A. W. and van Zanten, J. H. (2011). Information Rates of Nonparametric Gaussian Process Methods. *Journal of Machine Learning Research*, **12**, 2095–2119.
- van Erven, T. and Harremoës, P. (2014). Rényi Divergence and Kullback-Leibler Divergence. *IEEE Transactions on Information Theory*, **60**, 3797–3820.
- Vasko, K., Toivonen, H. T., and Korhola, A. (2000). A Bayesian Multinomial Gaussian Response Model for Organism-based Environmental Reconstruction. *Journal of Paleolimnology*, **24**, 243–250.
- Vehtari, A. and Ojanen, J. (2012). A Survey of Bayesian Predictive Methods for Model Assessment, Selection and Comparison. *Statistics Surveys*, **6**, 142–228.
- Villa, C. and Walker, S. (2015). On the Mathematics of the Jeffreys-Lindley Paradox. Available at arXiv:1503.04098.
- Vollmer, S. (2013). Posterior Consistency for Bayesian Inverse Problems Through Stability and Regression Results. *Inverse Problems*, **29**. Article number 125011.
- Wahba, G. (1978). Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression. *Journal of the Royal Statistical Society B*, **40**, 364–372.
- Wahba, G. (1990). Spline Functions for Observational Data. CBMS-NSF Regional Conference series, SIAM. Philadelphia.
- Walker, S. G. (2004). Modern Bayesian Asymptotics. *Statistical Science*, **19**, 111–117.
- Walker, S. G., Damien, P., and Lenk, P. (2004). On Priors With a Kullback-Leibler Property. *Journal of the American Statistical Association*, **99**, 404–408.
- Williams, E. J. (1969). A Note on Regression Methods in Calibration. *Technometrics*, **11**, 189–192.

- Wong, W. H. and Ma, L. (2010). Optional Polya Tree and Bayesian Inference. *The Annals of Statistics*, **38**, 1433–1459.
- Wu, H. and Zhang, J.-T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches*. John Wiley and Sons, New Jersey, USA.
- Yang, Y., Bhattacharya, A., and D.Pati (2018). Frequentist Coverage and Sup-norm Convergence Rate in Gaussian Process Regression. ArXiv Preprint.
- Ye, X., Wang, K., Zou, Y., and Lord, D. (2018). A Semi-nonparametric Poisson Regression Model for Analyzing Motor Vehicle Crash Data. *PLoS One*, **23**, 15.