Robust Uncertainty Quantification and Scalable Computation for Computer Models with Massive Output

by

Mengyang Gu

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
James O. Berger, Supervisor

_____
Robert L. Wolpert

_____
Surya Tokdar

_____
Barbara Engelhardt

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2016

Abstract

Robust Uncertainty Quantification and Scalable Computation for Computer Models with Massive Output

by

Mengyang Gu

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
James O. Berger, Supervisor

_____
Robert L. Wolpert

_____
Surya Tokdar

_____
Barbara Engelhardt

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2016

Abstract

Computer models have been widely used to reproduce the behavior of engineering, physical, biological and human processes. The rapid development of technology has empowered scientists and engineers to implement large-scale simulations via computer models and to collect real data from different kinds of resources, with the ultimate goal of predicting real processes through modeling. To achieve this typically requires a host of interactions with data and statistics, a process that has come to be called *uncertainty quantification* (UQ). This thesis develops statistical models focusing on two aspects of the problem in UQ: computational feasibility for huge functional data and robust parameter estimation in fitting models. To achieve these two goals, new techniques, theories and numerical procedures are developed and studied.

Chapter 1 frames the issue of modeling data from computer experiments and introduces the Gaussian stochastic process (GaSP) emulator as a crucial step in UQ. Chapter 2 provides a practical approach for simultaneously emulating/approximating computer models that produce massive output; for instance, output over a huge range of space-time coordinates (as necessary for the discussed application of hazard quantification for the Soufrière Hills volcano in Montserrat island). Chapter 3 and Chapter 4 are both about the parameter estimation problem for the GaSP emulator. Chapter 3 provides new criteria for parameter estimation, called 'robustness parameter estimation criteria'. Properties of the reference prior are studied in a general

setting and a new robust estimation procedure is proposed based on the estimation from marginal posterior modes with optimal parameterizations. Chapter 4 introduces a new class of priors – called jointly robust priors – for the GaSP model where inert inputs (inputs that barely affect the computer model output) are present. This prior has many of the good properties of the reference prior and is more computationally tractable. Chapter 5 discusses another problem with big functional data, in which the number of observations in a function is large. An exact algorithm that can compute the likelihood of the GaSP model linearly in time is introduced, and regression, separable GaSP models and nonseparable GaSP models are discussed in a unified framework. Finally, Chapter 6 provides some concluding remarks and describes possible future work.

To my family

Contents

List of Tables

List of Figures

Acknowledgements

I am grateful to all people who have helped during my PhD study at Duke. I would like to express my deepest gratitude to my advisor Dr. James O. Berger for his continued support and encouragement during my PhD study. Not only I learnt from his wisdom in doing research, but also from his unique way that explains the ideas so well to different audience. Our meeting has been an integral part of my study, which led to the accomplishment of this thesis. Further thanks go to my thesis committee members, Dr. Robert Wolpert, Dr. Surya Tokdar and Dr. Barbara Engelhardt, for their help and discussion along the way.

I would also like to thank all faculty members at the department of statistical science at Duke University, where I have many wonderful courses and seminars. I sincerely thank Dr. Li Ma, who was my first year advisor. He gave me lots of help every time I encountered a barrier at the first year. Further thanks go to several staff members in our department, Karen Whitesell, Lori Rauch, Nicole Scott and Larry Hall. Without you, I could not finish my PhD study and research smoothly.

I also thank Dr. Mike West, who introduced me to an interesting research project and a great summer internship program in IBM Thomas J Waston research center. Further thanks go to two mentors in IBM Watson, Dr. Dharmashankar Subramanian and Dr. Debarun Bhattacharjya, who bring lots of interesting discussion about applications of various research topics.

I am much obliged to the National Science Foundation. My research at Duke

is mostly supported by NSF grants DMS-1007773, DMS-1228317, EAR-1331353, and DMS-1407775. I am also grateful to the International Society for Bayesian Analysis, the American Statistical Association and Society for Industrial and Applied Mathematics for several travel awards to conferences.

Thanks to all of my colleagues and friends at Duke. With you, life is more colorful here.

Finally, I dedicate this work to my family. I want to express my heartfelt thanks to my parents, without whom I cannot accomplish this long journey. I also dedicate this thesis to my grandfather, who passed away during my PhD study. He is the best friend of me when I was a kid. His encouragement gave me extra support and motivation to continue my research.

1

Introduction

Uncertainty quantification is both an old and a new concept. Measurement errors, for instance, arise with almost every experiment in most scientific fields; one of the main issues is then to characterize and reduce the uncertainties through the development of technologies and statistical modeling. A more recent focus of uncertainty quantification is in the interactions and synthesis of mathematical models, statistics, field/real experiments, and probability theory, with a particular emphasize on large-scale simulations from computer experiments. The challenges not only come from the need to combine mathematical and statistical modeling, but also from the size of the information that is often available. The focus in this thesis is to provide statistical models for uncertainty quantification that are scalable to massive data produced in computer experiments and real experiments, through fast and robust statistical inference.

Computer models (also called simulators), are used to generate data to reproduce physical, engineering and human processes. They are particularly important when the data is limited or expensive to obtain. However, there are several roadblocks in using computer models to explain and predict the underlying processes. One impor-

tant issue with simulators is that they can be very computationally expensive to run. For instance, many of our illustrations will concern the TITAN2D computer model (Pitman et al. (2003); Patra et al. (2005)), which simulates the volcanic pyroclastic flow that surges down a volcano after an eruption; this model can require up to 2 hours for a single run. A second important issue is that there are usually many sources of uncertainty involved in utilizing the computer model for prediction, e.g. the discrepancy between the computer model and a real process, the uncertainty in the parameters of the computer model, errors in real data concerning the underlying process, etc. Uncertainty quantification is the process of combining all this information and uncertainty to make reliable predictions.

For a computationally expensive simulator, one crucial aspect of uncertainty quantification is the development of an *emulator* (an approximation) to the simulator that is accurate and which can be run very quickly; the uncertainty quantification tasks are then carried out with the emulator. The emulator is used to approximate what the simulator would have produced at input values for which it was not possible to run the simulator. A key feature of statistical emulators is that they have an internal assessment of their approximation accuracy, which makes possible a realistic assessment of uncertainty in predictions.

A Gaussian Stochastic Process (GaSP) is a common tool for analyzing spatially correlated data. For example, in geostatistics, it has long been used to model various types of data with complicated patterns (Gelfand et al. (2010); Banerjee et al. (2014)). The GaSP model can take different sources of uncertainty into account, and make predictions in a complete probabilistic way. (Early uses of GaSPs are known as the Kriging method (Cressie and Cassie (1993))). In recent years, GaSPs have also been used to build emulators of computer models (Sacks et al. (1989); Kennedy and O'Hagan (2001); Oakley and O'Hagan (2002); Bayarri et al. (2007b); Spiller et al. (2014)).

Data from a computer model is typically rather different than spatial data. First, the input space of the computer model (e.g. the space of model parameters, initial conditions, boundary conditions, etc.) often has high dimension, while the maximum dimension for spatial data is typically three. Second, the inputs are typically variables on completely different scales, so the effect of the inputs on the correlations will be highly variable. Consequently, the assumption of isotropy (defined more formally in Section 1.1), which is often adopted in spatial processes, usually does not hold for modeling data from computer models. Indeed, for computer models, it is common to use a product correlation function (Sacks et al. (1989); Paulo (2005); Bayarri et al. (2009)), with typically very different correlation parameters for each input; the product form also keeps computations tractable, and this choice will be followed herein. A third difference is that many computer models are deterministic, or close to being deterministic, while noise in data from spatial processes can be large. A fourth difference is that, by design, data from computer models is typically taken at widely dispersed values of the inputs, whereas this may well not be so for spatial data.

Combining all these distinct features, the GaSP emulator has gradually become one of the formal paradigms for uncertainty quantification arising from the inputs of the simulators, statistics, and field data. In the next section, we briefly review the GaSP emulator.

1.1 Gaussian Stochastic Process Emulator

To set notation, let $\mathbf{x} \in \mathscr{X}$ denote the p-dimensional vector of inputs to the simulator, and let $y(\mathbf{x})$ denote the resulting simulator output, assumed in this section to be real-valued. The simulator $y(\mathbf{x})$ is viewed as an unknown function (because the simulator is expensive to run, we will at most be able to evaluate $y(\mathbf{x})$ at a few points), modeled

via a *stationary* Gaussian Process,

$$y(\cdot) \sim GaSP(\mu(\cdot),\, \sigma^2 c(\cdot,\cdot))\,, \tag{1.1}$$

having mean function $\mu(\cdot)$ and stationary covariance $\sigma^2 c(\cdot,\cdot)$ with variance σ^2, and correlation function $c(\cdot,\cdot)$. The stationarity means that the covariance between two points $\mathbf{x}_a, \mathbf{x}_b \in \mathscr{X}$ only depends on their separation $\mathbf{x}_a - \mathbf{x}_b$. Stationarity is a common assumption, although some nonstationary GaSP models have been proposed in the recent literature (see, e.g., Higdon et al. (1999); Gramacy and Lee (2012); Ba and Joseph (2012)).

For any inputs $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ from $\mathscr{X}$, the likelihood from the GaSP is a multivariate normal distribution,

$$(y(\mathbf{x}_1), \ldots, y(\mathbf{x}_n))^T \mid \boldsymbol{\mu},\, \sigma^2,\, \mathbf{R} \sim \mathcal{MN}((\mu(\mathbf{x}_1), \ldots, \mu(\mathbf{x}_n))^T, \sigma^2\mathbf{R})\,, \tag{1.2}$$

where σ^2 is the unknown variance and $\mathbf{R}$ is the correlation matrix or Gram matrix (Rasmussen (2006)) with (i,j) element $c\,(\mathbf{x}_i, \mathbf{x}_j)$. It is common to model the mean function via regression,

$$\mu(\mathbf{x}) = \mathbf{h}(\mathbf{x})\boldsymbol{\theta} = \sum_{t=1}^{q} h_t(\mathbf{x})\theta_t\,, \tag{1.3}$$

where $\mathbf{h}(\mathbf{x}) = (h_1(\mathbf{x}), h_2(\mathbf{x}), ..., h_q(\mathbf{x}))$ is a vector of specified basis functions and θ_t is the unknown regression parameter for basis function h_t.

A process is called isotropic when the correlation function is only a function of $d_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||$, for any $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^T \in \mathscr{X}$ and $\mathbf{x}_j = (x_{j1}, \cdots, x_{jp})^T \in \mathscr{X}$, where $||\cdot||$ is the Euclidean distance. As mentioned earlier, the isotropic assumption is restrictive for emulating complicated functions and the following product correlation function is often assumed instead:

$$c(\mathbf{x}_i,\, \mathbf{x}_j) = \prod_{l=1}^{p} c_l(x_{il}, x_{ji}), \tag{1.4}$$

where $c_l(\cdot, \cdot)$ is a one-dimensional correlation function for the l^{th} coordinate of the input vector.

The simulator is run at a set of n chosen inputs $\mathbf{x}^{\mathscr{D}} = \{\mathbf{x}_1^{\mathscr{D}}, \ldots, \mathbf{x}_n^{\mathscr{D}}\}$, often selected using some "space filling" technique over the input domain $\mathscr{X}$, e.g., using a Latin Hypercube Design (Sacks et al. (1989)); let $\mathbf{y}^{\mathscr{D}} = (y(\mathbf{x}_1^{\mathscr{D}}), \ldots, y(\mathbf{x}_n^{\mathscr{D}}))^T$ denote the corresponding simulator outputs. Given the product correlation function in (1.4), the correlation matrix of these inputs is thus

$$\mathbf{R} = \mathbf{R}_1 \circ \mathbf{R}_2 \circ ... \circ \mathbf{R}_p. \tag{1.5}$$

Although maximum likelihood estimation of $\boldsymbol{\theta}$ and σ^2 can be undertaken, it is better to utilize Bayesian inference to account for the uncertainty in these parameters. We follow common practice and utilize the objective Bayesian approach, which specifies the standard reference prior for these parameters (Berger et al. (2001); Paulo (2005)):

$$\pi^R(\boldsymbol{\theta}, \sigma^2) \propto \frac{1}{\sigma^2}\,.$$

Estimation of parameters in the correlation function is an even more crucial issue, discussion of which will be delayed until Chapter 3 and Chapter 4; thus, for now, $\mathbf{R}$ will simply be assumed to be known.

With the above setup, the emulator can be defined. It is a prediction, at a new input value $\mathbf{x}^*$, of the corresponding simulator output $y(\mathbf{x}^*)$. Indeed, the predictive distribution of $y(\mathbf{x}^*)$, given $\mathbf{y}^{\mathscr{D}}$ and $\mathbf{R}$, is a t-distribution

$$y(\mathbf{x}^*) \mid \mathbf{y}^{\mathscr{D}}, \mathbf{R} \sim \mathcal{T}(\hat{y}(\mathbf{x}^*), \hat{\sigma}^2 c^{**}, n - q)\,, \tag{1.6}$$

with $n-q$ degrees of freedom, where

$$\begin{aligned}
\hat{y}(\mathbf{x}^*) &= \mathbf{h}(\mathbf{x}^*)\hat{\boldsymbol{\theta}} + \mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1}\left(\mathbf{y}^{\mathscr{D}} - \mathbf{h}(\mathbf{x}^{\mathscr{D}})\hat{\boldsymbol{\theta}}\right), \\
\hat{\sigma}^2 &= (n-q)^{-1}\left(\mathbf{y}^{\mathscr{D}} - \mathbf{h}(\mathbf{x}^{\mathscr{D}})\hat{\boldsymbol{\theta}}\right)^T\mathbf{R}^{-1}\left(\mathbf{y}^{\mathscr{D}} - \mathbf{h}(\mathbf{x}^{\mathscr{D}})\hat{\boldsymbol{\theta}}\right), \\
c^{**} &= c(\mathbf{x}^*,\mathbf{x}^*) - \mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}^*) + \left(\mathbf{h}(\mathbf{x}^*) - \mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}^*)\right)^T \\
&\quad \times \left(\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})\right)^{-1}\left(\mathbf{h}(\mathbf{x}^*) - \mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}^*)\right), \qquad (1.7)
\end{aligned}$$

with $\hat{\boldsymbol{\theta}} = \left(\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\,\mathbf{h}(\mathbf{x}^{\mathscr{D}})\right)^{-1}\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{y}^{\mathscr{D}}$ being the generalized least squares estimator for $\boldsymbol{\theta}$, $\mathbf{h}(\mathbf{x}^{\mathscr{D}})$ being the $n \times q$ basis design matrix with (i,j) element $h_j(\mathbf{x}_i)$, and $\mathbf{r}(\mathbf{x}^*) = (c(\mathbf{x}^*,\mathbf{x}_1^{\mathscr{D}}),\ldots,c(\mathbf{x}^*,\mathbf{x}_n^{\mathscr{D}}))^T$.

Note that, at the design points $\mathbf{x}_i^{\mathscr{D}}$, $1 \leqslant i \leqslant n$, the emulator is an interpolator of the simulator because, when $\mathbf{x}^* = \mathbf{x}_i^{\mathscr{D}}$, $\mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1} = \mathbf{e}_i^T$, where $\mathbf{e}_i$ is the n dimensional vector with the i^{th} entry being 1 and the others being 0. At other inputs, it not only provides a prediction of the simulator (i.e., $\hat{y}(\mathbf{x}^*)$) but also an assessment of the accuracy of the prediction; since this was developed from a Bayesian perspective, it also incorporates the uncertainty arising from estimating $\boldsymbol{\theta}$ and σ^2.

Once the emulator is constructed, the uncertainty quantification tasks are then performed using the emulator. Advantages of the GaSP emulator are that the prediction of unobserved runs is fast (linear in terms of number of points for prediction), and it is fully probabilistic, meaning that different sources of uncertainty can be in integrated coherently. Challenges and research topics related to this emulator are described in the next section.

1.2 Challenges and research questions

The challenges in the development of GaSP emulators that are discussed in this thesis arise from three issues: computation when the simulator output k is massive,

computation when the number of runs n is large, and difficulties in fitting the GaSP emulators to simulator output. These are briefly summarized here.

The first problem is illustrated by the TITAN2D computer model. The actual output from a single run is pyroclastic flow information at up to $k = 10^9$ space-time coordinates, and trying to emulate such massive output is a challenge. There have been some approaches that develop GaSP emulators at a small number of coordinates k; for instance, one could simply build independent GaSP emulators at each coordinate and estimate the parameters of each emulator using the outputs at that coordinate. Another popular approach is to define a *separable* GaSP model, where the covariance matrix between coordinates and the correlation matrix between inputs are modeled separately in the model (Conti and O'Hagan (2010)). Other approaches include specifying (or estimating) a basis function (e.g. through principle component decomposition (Higdon et al. (2008)) or wavelets (Bayarri et al. (2007b))), with the weights of the basis then being modeled as the independent Gaussian Processes to achieve the correlation structure over the input space; the resulting covariance structures are generally "nonseparable", and belong to a class of models called linear models of coregionalization (LMC) (Gelfand et al. (2004); Banerjee et al. (2014)). More general nonseparable covariance structures include process convolutions, created by convolving a base process with a smoothing correlation function (Higdon et al. (2002); Alvarez and Lawrence (2011)). See Álvarez et al. (2011) and Fricker et al. (2013) for a general overview of these methods.

None of these methods are directly implementable when k is huge. Indeed, the Conti and O'Hagan (2010) method requires that $k < n$, where n is the number of observed computer model runs. The general separable GaSP model requires direct modeling of the covariance between the space-time coordinates, necessitating $O(k^2)$ storage and $O(k^3)$ computations for a needed inversion of the covariance matrix, both untenable for TITAN2D. The other approaches mentioned are similarly constrained

to apply to only modestly large k.

Chapter 2 discusses a method of emulation that is actually linear in k, and thus can be implemented for k as large as 10^9. Numerical results show that the performance of this emulator is surprisingly excellent, and theoretical results are presented to explain why this is so. It is also shown in the chapter how the resulting emulator can be used to achieve the ultimate scientific goal involving TITAN2D, which is to determine hazard probabilities for future volcanic eruptions at all locations of interest.

The second computational problem that is addressed in the thesis is computation when the number of runs n of the computer model is large. Computing the GaSP typically requires inversion of the resulting covariance matrix, which is a computation of order $O(n^3)$ operations. Many approximation methods have been proposed and discussed in the literature, e.g. low rank approximation (Cressie and Johannesson (2008); Banerjee et al. (2008)), covariance tapering and compactly support covariance (Furrer et al. (2012); Kaufman et al. (2008, 2011)), use of Gaussian Markov Random field representations (Rue et al. (2009); Lindgren et al. (2011)), and likelihood approximation (Eidsvik et al. (2013)); see, e.g., Chapter 3 in Sun et al. (2012) for an overview of these methods.

Interestingly, under certain circumstances, computation of the likelihood and prediction using the GaSP can be done linearly in n, even if neither the covariance matrix and precision matrix (the inverse of the covariance matrix) are sparse. Situations under which this is so, as well as connections between regression modeling, separable GaSP modeling and nonseparable GaSP modeling are discussed and explored in Chapter 5.

The main roadblock in emulation is that the correlation matrix $\mathbf{R}$ in the GaSP emulator virtually always depends on unknown correlation parameters (Kaufman and Shaby (2013)). These parameters are known to be "notoriously difficult" to deal with

(Kennedy and O'Hagan (2001)). Usually no closed form estimator is available and many numerical problems arise in using standard estimates, e.g, the MLE (Oakley (1999); Lopes (2011); Spiller et al. (2014)).

This difficulty is tackled herein by utilization of posterior modes as the estimates of the correlation parameters. The posterior distributions arise from use of reference priors and some other default objective priors, proposed in Berger et al. (2001) and extended in Paulo (2005); Ren et al. (2012, 2013). It is found, herein, that choice of the parameterization of the correlation parameters is crucial and the properties of various posterior modes are given. It is found that common parameterizations – e.g., those used in Bayarri et al. (2009); Spiller et al. (2014)) – are not optimal, in that lack an important "robustness" property. This is discussed and illustrated in Chapter 3.

A final challenge related to this roadblock is that of being able to reduce the number of inputs for the emulator, by finding so called inert inputs (Linkletter et al. (2006)), i.e. the inputs that barely affect the outputs of the computer model. In order to do this, a new class of priors, that has the desirable properties of the reference prior and is computationally more convenient for determining inert inputs, is introduced in Chapter 5.

1.3 Outline

Chapter 2 provides a practical approach for emulating computer models that produce massive output. The TITAN2D simulator is introduced and used to illustrate all results, including the major scientific goal of prediction of hazard probabilities over a wide region. Theoretical justification for this emulator is given, and its performance compared with competitors.

Chapter 3 discusses the problem of parameter estimation in the GaSP setting. Although there is a vast literature on GaSP models, there is limited discussion of

the difficulties in estimation of the correlation parameters. This problem is discussed from the perspective of a new criterion, called the 'robustness parameter estimation criteria'. A new robust estimation procedure is proposed, based on finding the posterior mode resulting from use of a reference prior and a certain parameterization of the correlation parameters. As a byproduct, the posterior propriety of the resulting posterior is shown for a product covariance with general random designs. The proposed method is compared theoretically and numerically with other methods, including maximum likelihood estimation.

Chapter 4 proposes a new class of priors, called the jointly robust prior, which maintain the good features of the reference prior, while overcoming issues that arise when trying to eliminate weak inputs from the emulator. A new R Package (Gu et al. (2016)) that implements the robust parameter estimation with the reference prior and the jointly robust prior, is also introduced and compared with other available approaches and packages for various tasks.

Chapter 5 discusses the problem when the number of runs, n, of the computer model is large (of size 10^6 in the application considered therin). A general class of models is introduced for which the resulting computation can be done linearly in n. This class unifies several different models, including linear regression and separable GaSP models, through a nonseparable GaSP framework. A method is also introduced for combining different sources of correlation to boost the accuracy of the prediction and uncertainty quantification. An exact algorithm that can compute the likelihood linearly in terms of number of sites is provided, along with Bayesian inference for uncertainty quantification. As an application, results on the interpolation problem of methylation levels in epigenetics are given.

2

Parallel Partial Gaussian Process Emulation for Computer Models with Massive Output

In this chapter, we consider the problem of emulating (approximating) computer models (simulators) that produce massive output. The specific simulator we study is a computer model of volcanic pyroclastic flow, a single run of which produces up to 10^9 outputs over a space-time grid of coordinates. An emulator that is computationally suitable for such massive output is developed, and studied from practical and theoretical perspectives. On the practical side, the emulator does unexpectedly well in predicting what the simulator would produce, even better than much more flexible and computationally intensive alternatives. This allows the attainment of the scientific goal of this work, accurate assessment of the hazards from pyroclastic flows over wide spatial domains. Theoretical results are also developed that provide insight into the unexpected success of the massive emulator. Generalizations of the emulator are introduced that allow for a nugget, which is useful for the application to hazard assessment.

2.1 Literature Review and Motivations

Computer models – henceforth *simulators* – are used to generate data to reproduce the behavior of physical, engineering or human processes. We will be working with the testbed simulator TITAN2D, (see Pitman et al. (2003); Patra et al. (2005) for details), which simulates the volcanic pyroclastic flow that surges down a volcano after an eruption, based on inputs such as the initiating volume of flow. A key issue with such simulators is that they are typically very computationally expensive to run; TITAN2D requires up to 2 hours for a single run.

As introduced earlier, GaSP model is a very well studied paradigm for emulation (see e.g. Sacks et al. (1989); Bayarri et al. (2007b, 2009)). This chapter focuses on a challenging aspect of the problem, namely emulating a simulator that produces massive output over a coordinate space. For instance, TITAN2D produces flow information at approximately 10^9 space-time coordinates. While there is a vast body of research concerning emulating the simulator at one of a small number of simulator outputs, simultaneously emulation of the output over many coordinates is less studied. Some papers that do so are Higdon et al. (2008); Rougier (2008); Rougier et al. (2009); Xiao et al. (2010); Marrel et al. (2011); these are further discussed in Section 2.5.2, and representative methods will later be compared with the methodology introduced here.

The scientific motivation for this work is to determine hazard probabilities for future volcanic eruptions. In previous studies of volcanic hazard (see, e.g. Spiller et al. (2014)), the hazard probability, at a specific location, is the probability of a catastrophic event happening at least once during next T years; a catastrophic event is typically characterized by a maximum flow height larger than 1 meter during the flow event. In Bayarri et al. (2009), the estimation of the hazard probability at two locations (Plymouth and Bramble Airport) on Montserrat island were given. In

Lopes (2011); Spiller et al. (2014), this was extended to a number of locations in Belham Valley, an at risk area on Montserrat. One of the main scientific goals of this work is to enable computation of these hazard probabilities, not at individual locations, but simultaneously over a large spatial region. Furthermore, policymakers might be interested in events other than just maximum flow height exceeding a meter; they could want to use a lower threshold, or some other measure entirely, such as damage to structures by the force of the flow. To achieve the flexibility to answer any such posed question, an emulator is needed that can quickly predict the entirety of the output of TITAN2D.

The inputs to the simulator will be denoted $(\mathbf{x}, \mathbf{s})$, where $\mathbf{x}$ describes the driving inputs for the simulator (e.g. the volume of the pyroclastic flow) and $\mathbf{s}$ denotes a coordinate (e.g. the space-time coordinate) at which the simulator evaluates pyroclastic flow; this notation is not convenient for the later technical development, but is useful in this introduction.

The main idea of this development is that $\mathbf{x}$ *must* be accurately involved in the emulation (there is no chance in predicting a pyroclastic flow without adjusting for the volume of the flow, and we must consider volumes that have not yet been observed), but it is often just fine to perform the predictions of flow on just the space of $\mathbf{s}$, which will typically be a fixed grid of space-time coordinates; it is not typically necessary to interpolate into new space-time locations, if the original grid is detailed enough.

The straightforward approach to emulating the simulator simultaneously at many locations is discussed in Conti and O'Hagan (2010); Lee et al. (2011, 2012) and utilized for TITAN2D in Spiller et al. (2014). This approach, which is called the Many Single (MS) emulator approach, is simply to fit separate emulators at each coordinate. We will be using Gaussian stochastic process (GaSP) emulators, which are characterized by an unknown mean function, an unknown variance, and unknown

correlation parameters. In the MS emulator approach, these are all determined separately at each location, resulting in a highly computationally intensive process.

This chapter provides a computationally feasible alternative to the MS approach, which we call the *parallel partial* Gaussian stochastic process (PP GaSP) emulator approach. This approach has the following features:

- There are independent emulators at each of k coordinates $\mathbf{s}_1, \ldots, \mathbf{s}_k$ (with k being up to 10^9 for TITAN2D).
- Each coordinate emulator is allowed a different mean function and variance because pyroclastic flows behave very differently at different locations on the mountain (e.g., the height of the flow at locations near the initiation of the flow event will be much larger than at locations far from this point).
- All coordinates share common Gaussian process correlation parameters, and these are estimated from the joint likelihood of all emulators.

The name "parallel partial" is used to reflect the fact that the locations have probabilistically independent parallel emulators, but they are only partially independent, in the sense that they share common correlation parameters, estimated from the overall likelihood (as will be discussed in Section 2.3 and 2.7). The PP GaSP emulator is computationally feasible because it is is linear in k; more precisely, after some pre-computation steps, computation of the emulator predictions for a new input $\mathbf{x}^*$ at all k locations requires $O(n^2 + nk)$ numerical operations, where n is the number of simulator runs upon which the emulator is based. Such computational details are discussed in Section 2.5.1.

One natural concern with this approach is that the simulator is (usually) very tightly constrained at nearby locations, while the PP GaSP emulator provides independent predictions at each location. A related concern is the use of common

Gaussian process correlation parameters at all locations, as opposed to the more flexible modeling of the MS emulator. The surprising reality is that the PP emulator is not only accurate in emulation over the k coordinate points, but usually seems to be substantially better than alternatives such as the MS emulator, which do allow for differing correlation parameters. Both theoretical reasons and numerical evidence for this will be presented.

2.2 Background

2.2.1 The TITAN2D testbed

We introduce TITAN2D testbed in this section. The four inputs to the TITAN2D simulator are the initial flow volume V, initial angle of the flow ϕ, basal friction angle δ_{bed}, and internal friction angle δ_{int}. TITAN2D produces numerous outputs, one of them being the pyroclastic flow height at every space-time grid point. The PP emulator developed herein is perfectly capable of handling the entire space-time grid, but the time component is not of particular practical interest, for the simple fact that damage from pyroclastic flow is primarily due to the largest flow that hits a given spatial location. Therefore, the simulator output of particular interest, at a given location on the island, is

$$y(V, \phi, \delta_{bed}, \delta_{int}) = \textit{maximum flow height over time}\,,$$

this being a good surrogate for the damage inflicted at the location. We will thus work with this simulator output in our illustrations and evaluations, including the ultimate goal of producing probabilistic hazard maps for the region. Actually, for reasons discussed in Bayarri et al. (2009), we fit the emulator to $\log(y + 1)$ and then transform the predictions back. Means and variances do not transform directly through this transformation, but posterior medians and quantiles do, and are what we use in actual computations; we will suppress this detail in our notation.

The design input space $\mathscr{D}$ consisted of 2048 points chosen according to a maximin Latin hypercube design (Stein (1987)) over the relevant region $[10^5, 10^{9.5}] \times [0, 2\pi) \times [5.45, 18.45] \times [15, 35]$ for the four inputs. TITAN2D was run at these 2048 inputs, and the resulting vectors of maximum flow heights over the spatial grid of the island were recorded.

The GaSP emulator discussed in Section 1.1 has been introduced in Bayarri et al. (2009); Spiller et al. (2014) for emulating TITAN2D computer model at the specific interested location. Because the flow height is positively correlated with the initial volume of the flows, the basis functions is typically specified as $\mathbf{h}(\mathbf{x}) = (1, V)$, so that the mean function will simply be the regression $\theta_1 + \theta_2 V$. A commonly used correlation function for inputs $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ and $\mathbf{x}_j = (x_{j1}, \ldots, x_{jp})$ is the exponential family correlation of the form (Rasmussen (2006)),

$$c(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{ -\sum_{t=1}^{p} \left(\frac{|x_{it} - x_{jt}|}{\gamma_t} \right)^{\alpha_t} \right\} . \tag{2.1}$$

with $\gamma_t \in (0, \infty)$ and $\alpha_t \in [1, 2]$. The emulator is developed from runs of the simulator at a set of n chosen inputs $\mathbf{x}^{\mathscr{D}} = \{\mathbf{x}_1^{\mathscr{D}}, \ldots, \mathbf{x}_n^{\mathscr{D}}\}$, selected using a Latin Hypercube Design (LHD) over the input space $\mathscr{X}$ (Sacks et al. (1989); Forrester et al. (2008)). The points in $\mathbf{x}^{\mathscr{D}}$ are typically chosen as far apart as possible, in order to sample the simulator at as many diverse points as possible. This means that the parameters α_t are not highly influential and typically have quite flat likelihood surfaces. They also are typically highly confounded with the γ_t and σ^2, causing computational and inferential difficulties if left in the model (see e.g. Zhang (2004); Gelfand et al. (2010)). It is thus common to fix them to a constant value – often 1.9 (which we adopt herein), to reflect a typical desire for smoothness of the emulator, yet avoiding numerical problems that can arise with the choice $\alpha_t = 2$.

The γ_t will be estimated as the modes of their marginal posterior densities arising

from first integrating out $\boldsymbol{\theta}$ and σ^2, with respect to $\pi^R(\cdot)$, and then multiplying this marginal likelihood by the reference prior for the γ_t. There are several technical issues involved in the implementation, the details of which we delay until Section 2.7; for now we just assume the availability of estimates $\hat{\gamma}_t$.

A complication that arises in TITAN2D is that the second input, ϕ, is periodic, ranging from 0 to 2π, so that a correlation function that respects periodicity is needed. Spiller et al. (2014) introduces a formal way to deal with such an input, using a circular correlation function with the "periodic folding" form. Details are described in Appendix A.1. Note that, while we focus here on prediction of hazard probabilities for the Soufrière Hill Volcano (SHV) on Montserrat island using TITAN2D computer model, the methodology can be used for hazard prediction for any volcanic pyroclastic flows.

2.2.2 Integrating different sources of information

For the goal of quantifying the volcanic hazard rate at Montserrat Island, several pieces of information are available for this problem, summarized below.

i. TITAN2D computer model that generates $y_j(\mathbf{x}_i)$, the pyroclastic flow heights for all spatial-temporal coordinates j at a given input $\mathbf{x}_i$.

ii. Some prior information for $\mathbf{x}$.

iii. Some field data of inputs, denoted as $\mathbf{x}_i^F$ (11 flows) (Bayarri et al. (2015)).

iv. Occurrence of the pyroclastic flow is modeled as a stationary poisson process with the rate $\lambda \approx 22/yr$.

The integration of this information requires a statistical model. There is obviously computational issue since the number of spatial coordinates are large and the hazard quantification needs to be all interested locations simultaneously. We first extend

the GaSP model to this massive output problem and then discuss the scientific goal in Section 2.4.

2.3 Parallel partial emulation

2.3.1 The PP GaSP emulator

As discussed before, TITAN2D will generate massive data over many coordinates during each simulator run. Let k denote the total number of space-time grid points that are considered for each simulator run; with TITAN2D, k can be as big as 10^9 but, for the reasons discussed in Section 2.2.2, we will herein restrict consideration to only the spatial grid. Let $y_j(\mathbf{x})$ denote the simulator output at the j^{th} coordinate, so that $\mathbf{y}(\mathbf{x}) = (y_1(\mathbf{x}), \ldots, y_k(\mathbf{x}))$ is the entire simulator output arising from input $\mathbf{x}$. In this section we develop a computationally efficient and accurate emulator of that entire output.

As discussed in Section 2.1, we assume that an *independent* GaSP of the form (1.1) is assigned to each coordinate, with prior mean functions of the regression form $\mathbf{h}(\mathbf{x})\boldsymbol{\theta}_j$, where $\mathbf{h}(\mathbf{x})$ is a *common* q-vector of given basis functions and the $\boldsymbol{\theta}_j$ are *differing* unknown regression coefficients; *differing* unknown prior variances σ_j^2; and *common* estimated correlation parameters $\hat{\boldsymbol{\gamma}}$. Assuming common basis functions and estimated correlation parameters is the key to the computational simplification.

Let $\mathbf{y}_j^{\mathscr{D}}$ denote the column vector of simulator output values at the j^{th} coordinate when run over the design input values, as discussed in Section 1.1. We also utilize the same standard objective prior for the mean and variance parameters

$$\pi^R(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k, \sigma_1^2, \ldots, \sigma_k^2) \propto \frac{1}{\prod_{j=1}^{k} \sigma_j^2}. \tag{2.2}$$

Since the GaSPs at each coordinate are independent given the range parameters $\boldsymbol{\gamma}$, the prior is of a product form in the parameters of the different coordinate GaSPs,

and $\hat{\boldsymbol{\gamma}}$ is common across coordinates, it is immediate that the overall GaSP, at a new input $\mathbf{x}^*$, is the product of k independent t-distributions, with that for the j^{th} coordinate being

$$y_j(\mathbf{x}^*) \mid \mathbf{y}_j^{\mathscr{D}}, \hat{\boldsymbol{\gamma}} \sim \mathcal{T}(\hat{y}_j(\mathbf{x}^*), \hat{\sigma}_j^2 c^{**}, n-q)\,, \tag{2.3}$$

with $n-q$ degrees of freedom, where

$$\hat{y}_j(\mathbf{x}^*) = \mathbf{h}(\mathbf{x}^*)\hat{\boldsymbol{\theta}}_j + \mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1}\left(\mathbf{y}_j^{\mathscr{D}} - \mathbf{h}(\mathbf{x}^{\mathscr{D}})\hat{\boldsymbol{\theta}}_j\right), \tag{2.4}$$

$$\hat{\sigma}_j^2 = (n-q)^{-1}\left(\mathbf{y}_j^{\mathscr{D}} - \mathbf{h}(\mathbf{x}^{\mathscr{D}})\hat{\boldsymbol{\theta}}_j\right)^T \mathbf{R}^{-1}\left(\mathbf{y}_j^{\mathscr{D}} - \mathbf{h}(\mathbf{x}^{\mathscr{D}})\hat{\boldsymbol{\theta}}_j\right), \tag{2.5}$$

with $\hat{\boldsymbol{\theta}}_j = \left(\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\ \mathbf{h}(\mathbf{x}^{\mathscr{D}})\right)^{-1}\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{y}_j^{\mathscr{D}}$ being the generalized least squares estimator for $\boldsymbol{\theta}_j$, and $\mathbf{R}$, $\mathbf{h}(\mathbf{x}^{\mathscr{D}})$, $\mathbf{r}(\mathbf{x}^*)$ and c^{**} being defined in Section 1.1. From algebraic rearrangement of (2.4), the following lemma is immediate.

Lemma 2.3.1. *Letting* $\mathbf{y}^{\mathscr{D}} = (\mathbf{y}_1^{\mathscr{D}}, \mathbf{y}_2^{\mathscr{D}}, ..., \mathbf{y}_k^{\mathscr{D}})$ *denote the* $n \times k$ *matrix of all the simulator output at the design points, the predictive mean of the PP GaSP at new input* $\mathbf{x}^*$*, namely* $\hat{\mathbf{y}}(\mathbf{x}^*) = (\hat{y}_1(\mathbf{x}^*), \hat{y}_2(\mathbf{x}^*), ..., \hat{y}_k(\mathbf{x}^*))$*, can be expressed as*

$$\hat{\mathbf{y}}(\mathbf{x}^*) = \boldsymbol{\omega}(\mathbf{x}^*)\mathbf{y}^{\mathscr{D}}\,, \tag{2.6}$$

where

$$\boldsymbol{\omega}(\mathbf{x}^*) = \left(\mathbf{h}(\mathbf{x}^*) - \mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})\right)\left(\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\ \mathbf{h}(\mathbf{x}^{\mathscr{D}})\right)^{-1} \times \mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1} + \mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1}\,.$$

The weights $\boldsymbol{\omega}(\mathbf{x}^*)$ are usually called Kriging weights (Cressie and Cassie (1993)). There are two immediate important consequences of (2.6). First, the PP GaSP emulator is not only an interpolator of the simulator at the design inputs, but it is also a weighted sum of the simulator runs (each row of $\mathbf{y}^{\mathscr{D}}$ being the simulator

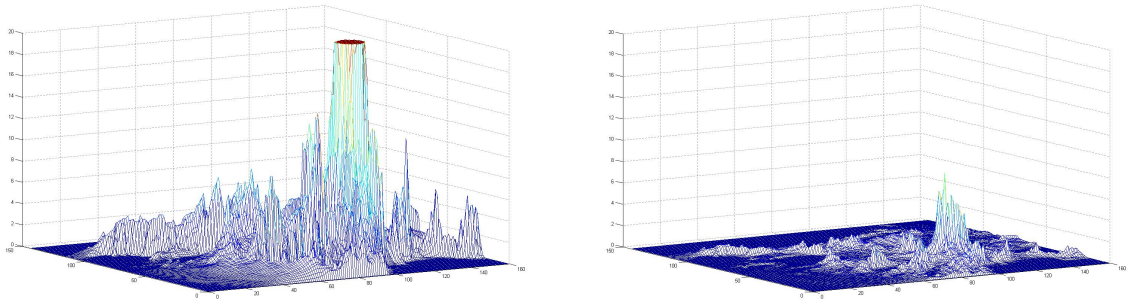

FIGURE 2.1: Median (truncated at 20 meters at the volcanic center region) and interquartile range of the GaSP emulator of 'maximum flow height over time' for TITAN2D, at 23,040 spatial locations over Montserrat Island and for new input values $V^* = 10^{6.9984}$, $\varphi^* = 3.3487$, $\delta^*_{bed} = 10.8790$, and $\delta^*_{int} = 31.0300$.

output – at one of the n input values – over all k coordinates, and $\omega(\mathbf{x}^*)$ being an n-vector). This ensures that the emulator inherits the smoothness of the simulator and probably some of the dynamics. Note, in contrast, that developing a separate emulator at each coordinate would not have this property, in that this would result in different weights for the simulator output at each coordinate.

Second, the weights, $\boldsymbol{\omega}(\mathbf{x}^*)$, depend only on computation of the q-vector $\mathbf{h}(\mathbf{x}^*)$ and the n-vector $\mathbf{r}(\mathbf{x}^*)$, together with pre-computable matrices and vectors. The entire computation of the emulator is thus linear in k, the key to the computational simplification. (More details of the computation are given in Section 2.5.1.)

Note that it is crucial that the outputs over all coordinates share the same correlation parameters $\hat{\boldsymbol{\gamma}}$. If not, each coordinate would have a different design correlation matrix $\mathbf{R}$, requiring the inversion of an $n \times n$ matrix at each coordinate; the computational situation is actually then even worse, as shown in Section 2.5.1, because of the need to separately estimate the $\hat{\boldsymbol{\gamma}}_j$. As shown in Section 2.5.1, there is also a considerable penalty for not having the same basis elements at each coordinate, although the penalty is not nearly as severe as that for allowing differing correlation parameters.

Figure 2.1 shows the median (truncated at 20 meters at the volcanic center region) and interquartile range of the PP GaSP emulator of TITAN2D for a new input, based on $n = 50$ simulator design runs; only 50 runs are used in this illustration because even this small number of runs seems to capture the main features of the model output. Note that the GaSP assessment of accuracy suggests small uncertainty at most of the locations. We will see in Section 2.5 that these internal emulator uncertainties do accurately reflect the real accuracy in emulation of TITAN2D.

2.3.2 Adding a nugget to the PP GaSP emulator

In TITAN2D, the output flow height is almost constant (for fixed values of the other inputs) as the internal friction angle δ_{int} varies over its range $[15^\circ, 30^\circ]$ based on Bayesian model selection for GaSP (Linkletter et al. (2006); Savitsky et al. (2011)) and sensitivity analysis (Iooss and Lemaître (2014)), and confirmed by the simulation study in Subsection 2.5.2. Using a weak input in emulation has the same drawbacks as using a weak covariate in regression – the inaccuracies introduced by incorporating the weak input or covariate into the model can lead to worse predictions than omitting them. However, if a simulator input is omitted in the emulator (Andrianakis and Challenor (2012)), the emulator can no longer be an interpolator, so that the GaSP model is then inappropriate. The standard solution is to add a nugget (a noise term) to the GaSP model, such as $\tilde{y}(\cdot) = y(\cdot) + \varepsilon$, where $y(\cdot)$ is the earlier noise-free GaSP and ε is i.i.d. mean-zero Gaussian white noise. In particular, we assume that the covariance function for the new process $\tilde{y}_j(\cdot)$ at coordinate j is

$$\sigma_j^2 \tilde{c}(\mathbf{x}_l, \mathbf{x}_m) = \sigma_j^2 \{c(\mathbf{x}_l, \mathbf{x}_m) + \eta 1_{l=m}\}\,; \tag{2.7}$$

note that we assume the nugget parameter η is common across all coordinates (needed for the same reasons we required common correlation parameters $\boldsymbol{\gamma}$). We parameterize the nugget in this way to allow for marginalizing out over σ_j^2 (Ren et al. (2012);

Kazianka and Pilz (2012)). After adding the nugget, the covariance matrix for the design input space $\mathscr{D}$ at coordinate j is

$$\sigma_j^2 \tilde{\mathbf{R}} = \sigma_j^2 (\mathbf{R} + \eta \mathbf{I}) \,. \tag{2.8}$$

For a new input, $\boldsymbol{x}^*$, the joint distribution of the new and design outputs at coordinate j is

$$\begin{pmatrix} y_j(\mathbf{x}^*) \\ \mathbf{y}_j^{\mathscr{D}} \end{pmatrix} | \, \boldsymbol{\theta}_j, \sigma_j^2, \boldsymbol{\gamma}, \eta \sim \mathcal{N} \left(\begin{pmatrix} \mathbf{h}(\mathbf{x}^*)\boldsymbol{\theta}_j \\ \mathbf{h}(\mathbf{x}^{\mathscr{D}})\boldsymbol{\theta}_j \end{pmatrix}, \sigma_j^2 \begin{pmatrix} \tilde{c}(\mathbf{x}^*, \mathbf{x}^*) & \mathbf{r}^T(\mathbf{x}^*) \\ \mathbf{r}(\mathbf{x}^*) & \tilde{\mathbf{R}} \end{pmatrix} \right). \tag{2.9}$$

for $1 \leqslant j \leqslant k$. The nugget parameter η will be estimated along with the input correlation parameters, as discussed in Section 2.7.2, leading to $\hat{\boldsymbol{\gamma}}$ and $\hat{\eta}$ that will be used to develop the emulator. Indeed the resulting PP GaSP with nugget, is defined exactly as in (2.3), with the simple change of replacing $\mathbf{R}$ by $\tilde{\mathbf{R}}$ (computed using $\hat{\boldsymbol{\gamma}}$ and $\hat{\eta}$). One of the advantages of the original PP emulator is that it is an interpolator of the simulator at the input design points; this will no longer be true of the PP emulator with a nugget, but in Appendix A.2, we show that it is close to being an interpolator.

The improvement, for TITAN2D, in going from a four-input emulator to a three-input emulator with nugget, is indicated in Table 2.1 and Table 2.2 in Section 2.5.2. For an indication as to the overall accuracy of the PP emulator with nugget, we consider a crucial feature of the TITAN2D output, namely the contour on the island at which the maximum flow height is 1m; as discussed in Bayarri et al. (2009), the interior of this contour defines the region in which the pyroclastic flow is viewed as being catastrophic. The PP emulator with nugget of TITAN2D was developed using only $n = 50$ runs of the simulator (the small number in order to hopefully see some differences between the emulator and the simulator). The 1m contours on the island

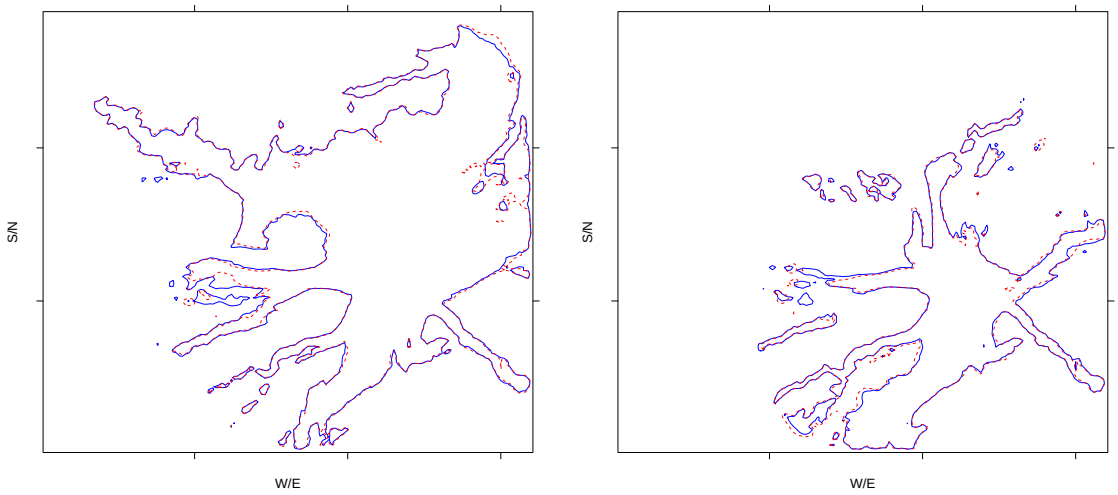

FIGURE 2.2: 1m spatial contours of maximum pyroclastic flow height on Montserrat Island for two held-out values of the inputs. The red dashed contour is from the actual simulator run, while the blue solid contour is the prediction from the PP GaSP emulator with 3 inputs (V, δ_{bed}, ϕ) and an estimated nugget. The held out testing inputs for the left figure are $V^* = 10^{7.1368}$, $\varphi^* = 1.8484$, $\delta^*_{bed} = 12.2940$, and $\delta^*_{int} = 24.2140$. Those for the right figure are $V^* = 10^{6.8292}$, $\varphi^* = 4.5360$, $\delta^*_{bed} = 12.7880$, and $\delta^*_{int} = 27.3000$.

were then computed for a large number of held-out design inputs using the emulator and then the simulator runs.

Two typical results are presented in Figure 2.2; the red curves are actual contours from the TITAN2D simulator, while the blue curves are the contours from the emulator. The contours match surprisingly well, especially considering the challenging topography (the 'holes' in the contours reflect topographical features such as hills, known to the simulator but not directly known to the emulator) and the use of only 50 training runs.

2.4 Flexible hazard quantification

As discussed in Section 2.1, the scientific goal for this work was to enable flexible assessments of hazard from pyroclastic flow over a wide region. In particular, we

focus here on developing contour plots of probabilities that maximum flow heights from SHV will exceed any threshold H of interest, over a time period T and over the entire at risk part of Montserrat island. Using the PP GaSP emulator, the entire distribution of flow heights over the island (as inputs vary) can be estimated, and this, in turn, can be used to answer a very wide range of hazard questions. We specifically develop the whole island hazard maps for $T = 2.5$ years and H equal 0.5, 1.0, or 2.0.

2.4.1 Uncertainty in the inputs and the occurrence of pyroclastic flows

We first need to account for the uncertainty in the inputs $\mathbf{x}^* = (V, \phi, \delta_{bed})$. The distribution of these input is studied in Bayarri et al. (2009); Spiller et al. (2014), and we follow their analysis. The distribution of (V, ϕ) is assumed to be of the form

$$p(V, \phi \mid V_m) \propto \alpha V_m^{\alpha} V^{-\alpha-1} 1_{V>V_m} 1_{0 \leqslant \phi < 2\pi},$$

i.e., a uniform distribution on $[0, 2\pi)$ for ϕ and (independently) a Pareto distribution for the initial volume V. V_m was chosen to be 5×10^4, since flows smaller than this value have no impact on hazard assessments of interest. Based on data giving the volumes of observed pyroclastic flows from SHV, a full Bayesian analysis was conducted in Bayarri et al. (2009) and Spiller et al. (2014) for the Pareto shape parameter α, but the variance of the posterior was so small that we simply utilize $\alpha = 0.64$ (the posterior mean and MLE) in the ensuing analysis.

The basal friction angle, δ_{bed} is assumed to be independent of V and ϕ, and is known to be decreasing in V. Based on available data relating V to δ_{bed}, we follow Bayarri et al. (2015); Spiller et al. (2014) and fit a linear model to the following transformed V and δ_{bed}:

$$\log_{10}(\tan^{-1}(\delta_{\text{bed}})) = a + b \log_{10} V + \epsilon\,, \tag{2.10}$$

where $\epsilon \sim N(0, \sigma^2_{\text{bed}})$. Eleven observations of (V, δ_{bed}) at Montserrat island (Bayarri et al. (2015)) are available to fit the model and, utilizing the objective prior $\pi(a, b, \sigma^2_{\text{bed}}) = 1/\sigma^2_{\text{bed}}$, the posterior predictive distribution $\pi(\delta_{\text{bed}} \mid V)$ is found and utilized in the ensuing analysis. (we will call the above the posterior distribution of $(V, \phi, \delta_{\text{bed}})$, although it is only an approximate posterior in terms of V.)

Conditional on the occurrence of a pyroclastic flow (PF), denote the density of flow height at location j by $f_j(\cdot)$ and denote the cumulative distribution function as $F_j(\cdot)$. This will be estimated by the distribution of flow heights arising from the PP GaSP emulator, as the inputs $\mathbf{x}^* = (V, \phi, \delta_{bed})$ are drawn from their posterior distribution described above. Actually we need a sample from $f_j(\cdot)$ for each coordinate j in the following, which we will (approximately) obtain by drawing a sample of inputs $(\mathbf{x}^*_1, \ldots, \mathbf{x}^*_{N*})$ from the input posterior distribution, and then computing the PP GaSP with nugget emulator mean (2.6) at each input, simultaneously obtaining a sample at all locations. Strictly speaking we should sample from the PP GaSP posterior t-distributions but, in this application, these distributions are extremely concentrated around their modes since we will be using all 2048 simulator runs to build the emulator; the variation caused by the uncertainty in $\mathbf{x}^* = (V, \phi, \delta_{bed})$ is several orders of magnitude larger than the uncertainty in the PP Gasp.

Hazard prediction, for a period of time T at location j, is based on the distribution of the maximum pyroclastic flow height, $Y_j^{\{T\}}$, that occurs over that period at the location; the following lemma gives the density and α-quantile of this distribution, under the assumption that pyroclastic flows arise from a stationary Poisson process with yearly intensity λ. (At SHV, $\lambda \approx 22$ /year, as found in Bayarri et al. (2009)). We acknowledge that stationarity can be a critical assumption, but it is the most frequently used assumption to provide tractable results (Spiller et al. (2014); Bayarri et al. (2015)).

Lemma 2.4.1. *Under the assumption that pyroclastic flows arise from a stationary Poisson process with yearly intensity λ, the density of $Y_j^{\{T\}}$, the maximum flow height over that period at the location j, is*

$$p_j^{\{T\}}(y) = 1_{\{y=0\}} \exp(-\lambda T) + f_j(y)\lambda T \exp\{\lambda T(F_j(y) - 1)\}\,.$$

The α-quantile of this distribution is

$$y_j^{\alpha} = \begin{cases} 0 & \alpha \leqslant \exp(-\lambda T)\,, \\ F_j^{-1}\left(\dfrac{\log(\alpha)}{\lambda T} + 1\right) & \alpha > \exp(-\lambda T)\,. \end{cases} \tag{2.11}$$

Proof. The random number of occurrences, M, of PF's over time period T follows a Poisson distribution with mean $\tilde{\lambda} = \lambda T$. If $M = m$ were to happen over the next T years, the maximum flow height at coordinate j is then the largest order statistic, having density $mF_j(y)^{m-1}f_j(y)$. If $M = 0$, which happens with probability $\exp(-\tilde{\lambda})$, the maximum flow height is obviously 0. Marginalizing out over M gives

$$\begin{aligned} p_j^{\{T\}}(y) &= 1_{\{y=0\}} \exp(-\tilde{\lambda}) + \sum_{m=1}^{\infty} m f_j(y) F_j(y)^{m-1} \frac{\tilde{\lambda}^m}{m!} \exp(-\tilde{\lambda}) \\ &= 1_{\{y=0\}} \exp(-\tilde{\lambda}) + f_j(y)\tilde{\lambda} exp\{\tilde{\lambda}(F_j(y) - 1)\} \sum_{m=1}^{\infty} \frac{\tilde{\lambda}^{m-1}}{(m-1)!} \exp(-\tilde{\lambda} F_j(y)) \\ &= 1_{\{y=0\}} \exp(-\tilde{\lambda}) + f_j(y)\tilde{\lambda} exp\{\tilde{\lambda}(F_j(y) - 1)\}\,. \end{aligned}$$

Expression (A.3) is an immediate consequence.

□

That we have a closed from expression for the quantiles of $p_j^{\{T\}}$ is key to being able to efficiently employ the PP GaSP emulator to simultaneously compute hazard probabilities over all relevant locations at SHV. Simulation of M would not allow for such efficient use of the emulator.

Quantiles of $p_j^{\{T\}}$ typically transform into quantiles of F_j in the far right tails. For instance, suppose we are interested in quantiles of $p_j^{\{T\}}$ at levels $\alpha = (0.01, 0.1, 0.6, 0.95)$ when $T = 2.5$ years and $\lambda \approx 22$ times/year. Then (A.3) implies that we need the corresponding $(0.9163, 0.9581, 0.9907, 0.9990)$ quantiles of F_j. These latter quantiles will be found, at each location j, as the corresponding empirical quantiles from the sample of N^* draws from F_j that were discussed above. The point here, is that typically it will suffice to only retain the largest 10% of these draws in order to find the desired quantiles of $p_j^{\{T\}}$; this is a significant saving, since one must store these draws over all locations j.

2.4.2 Quantification of the hazard at SHV

We first fit the PP GaSP emulator (i.e., obtain estimates of $(\hat{\boldsymbol{\gamma}}, \hat{\eta})$), using all simulator runs $\mathbf{y}^{\mathscr{D}}$, utilizing the composite likelihood method discussed in the next section. Then we sample $N^* = 10^5$ inputs $(\mathbf{x}_1^*, \ldots, \mathbf{x}_{N*}^*)$ from their posterior distribution discussed in Section 2.4.1. At each input we compute the PP GaSP posterior predictive mean, and collect the samples at each location j (i.e., the j^{th} coordinates of the predictive means) to provide an (approximate) sample from $F_j(.)$ at each location j (possibly saving only the larges 10% of samples at each location). For any threshold H, we compute the estimate of $F_j(H)$ as the proportion of samples from this distribution smaller than H, and we marginalize out the occurrence of PF to get the estimated probability that the maximum flow heights exceed H at each location using Lemma 2.4.1. This is summarized in the Algorithm 1.

Figure 2.3 gives contours of the probabilities that the maximum flow heights exceed 0.5, 1 and 2 meters over the next $T = 2.5$ years over Montserrat Island. The upper row in Figure 2.3 shows the hazard probabilities produced by PP GaSP with only $N = 50$ runs from TITAN2D, uniformly sampled from the whole $2,048$ runs and the lower rows shows the results using all $2,048$ runs. We can see two rows of figures

Algorithm 1 Flexible full hazard map

(1) Run TITAN2D N times for each design $\mathbf{x}_i^{\mathscr{D}}$, $i = 1, \ldots, N$, and record the output pyroclastic flow $\mathbf{y}^{\mathscr{D}}$.
(2) Build the PP GaSP emulator discussed in Section 2.3, based on all N runs on the design points.
(3) Sample $\mathbf{x}_i^*$, for $i = 1, \ldots, N^*$, from the posterior distribution of inputs discussed in Section 2.4.1.
(4) Compute the PP GaSP posterior predictive mean (2.6) for each sample $\mathbf{x}_i^*$, and collect the samples at each coordinate j to provide the sample from the flow height distribution at location j.
(5) For any threshold H that is of interest, use the proportion of the predictions from the samples $\mathbf{x}_i^*$ in step (4) that are smaller than H as the estimate of $F_j(H)$.
(6) Use the estimate of $F_j(H)$ to obtain the probability of maximum flow heights over the next T years larger than H at location j, by use of (A.3).

are similarly yet some distinguishable difference can be identified, especially at the small hazard probabilities area. This is because TITAN2D outputs have many zeros at these locations, so that it can easily happen that all $N = 50$ runs simply report zero at a location where there is hazard. An interesting problem (outside the scope of this paper) is that of determining the minimum number of runs of TITAN2D for an accurate hazard assessment.

Belham Valley is a small region in the northwest part of Montserrat, plotted as shaded area in Figure 2.3. The coastal area to the north of Belham Valley is still inhabited and so is of primary interest for risk assessment. The upper (uninhabited) parts of the valley have a large probability (≈ 0.9) to have more than 1 meter flows within the next $T = 2.5$ years, while the lower parts of the valley have comparatively small hazard probability. The borders of some inhabited regions have probability larger than 0.1 of being hit by pyroclastic flows higher than 0.5 meter, which is a significant concern.

2.5 Validation and numerical comparisons

In this section, we study the performance of the PP emulator in the context of TITAN2D output, and compare it with the MS GaSP emulator (Conti and O'Hagan

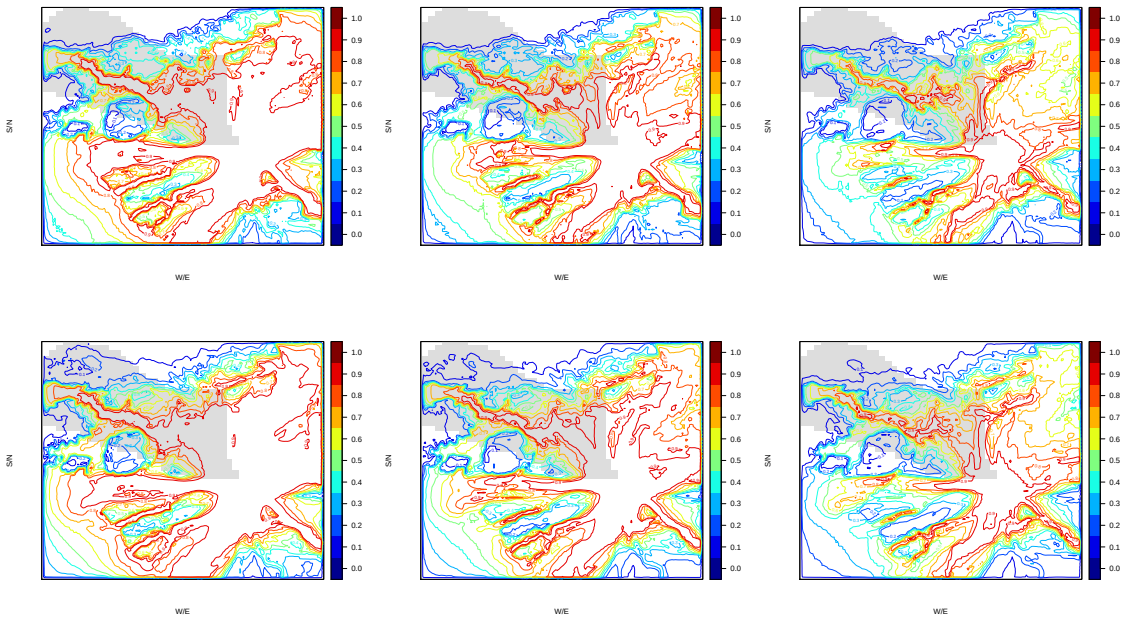

FIGURE 2.3: For SHV, contours of the probabilities that the maximum flow heights exceed 0.5 (left), 1 (center) and 2 (right) meters over the next $T = 2.5$ years at each location on SHV. The shaded area is Belham Valley, which is still inhabited. The results of the upper row utilizes $N = 50$ runs to build PP GaSP and the lower row utilizes the whole $N = 2048$ runs.

(2010)) and other emulators defined in Section 2.5.2. Initially, we thought that the MS GaSP emulator would be the gold standard, since it allows for adaptation of the correlation parameters to the particular coordinate; in contrast the PP emulator insists on the same correlation parameters across all coordinates. Quite surprisingly, we did not find this to be so. The comparisons between emulators will be in terms of computational cost and out of sample prediction.

2.5.1 Computational cost

It is useful to divide the computational cost of the PP emulator into three phases.

- The first (see (2.6)) is the one-time costs of computing $\tilde{\mathbf{R}}^{-1}$, $\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\tilde{\mathbf{R}}^{-1}$, and $\left(\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\tilde{\mathbf{R}}^{-1}\,\mathbf{h}(\mathbf{x}^{\mathscr{D}})\right)^{-1}$ and estimating $(\hat{\boldsymbol{\gamma}}, \hat{\eta})$. The first three have maximum cost $O(n^3)$; we consider the cost of estimating $(\hat{\boldsymbol{\gamma}}, \hat{\eta})$ later.

- The second phase is computation of $\boldsymbol{\omega}(\mathbf{x}^*)$ in (2.6) at a new input $\mathbf{x}^*$, a computation of order $O(n^2)$ utilizing the phase one pre-computations. This cost may seem minor compared to the phase one cost of $O(n^3)$ but, for many uses of the emulator, such as performing an MCMC analysis, this may have to be repeated many thousands of times whereas the phase one computation is not repeated.

- Finally, the computation of the emulator mean in (2.6) is then $O(nk)$, which can be much larger that the phase one and two costs in our situation, since k can be so much larger than n. (In the TITAN2D testbed, n is a maximum of 2048, while k can be as large as 10^9.) Similarly, it can be shown that the computational cost for computing all the variances of the PP GaSP is $O(n^2k)$; this is substantially more expensive than computing the PP emulator mean, but often it will only be necessary to compute the variances at some of the coordinates, to obtain a feel for the accuracy of the emulator.

The MS emulator has different $(\hat{\boldsymbol{\gamma}}, \hat{\eta})$ and, hence, different $\tilde{\mathbf{R}}$ at each coordinate. The inversions of $\tilde{\mathbf{R}}$ thus have to be done k times in the pre-computation stage, leading to a pre-computational cost of order $O(n^3k)$. Even the MS emulator mean, after this pre-computation, is an expensive $O(n^2k)$, essentially because a new $\boldsymbol{\omega}(\mathbf{x}^*)$ must be computed at each coordinate. Basically, when k is huge and n is large, use of the MS emulator is not computationally feasible.

Actually, the pre-computation of $(\hat{\boldsymbol{\gamma}}, \hat{\eta})$ in the PP GaSP is the severest computational challenge, if one attempts to use the full likelihood to estimate $(\hat{\boldsymbol{\gamma}}, \hat{\eta})$. The reason is that, for each candidate $(\hat{\boldsymbol{\gamma}}, \hat{\eta})$ used in trying to fit the full likelihood, a new inversion of $\tilde{\mathbf{R}}^{-1}$ is needed, and the subsequent computation of the likelihood (see Section 2.7.2) is of order $O(n^2k)$ for the PP GaSP emulator. Hence the full cost of estimating $\hat{\boldsymbol{\gamma}}$ is $O(tn^2k) + O(tn^3)$, where t is the number of iterations needed in the

estimation process. (In the testbed examples considered here, t is roughly 200.) For the MS emulator, the total computational cost involved in estimating the differing $\hat{\boldsymbol{\gamma}}$ at the coordinates is of order $O(tn^3k)$, which is prohibitive in settings such as here. The computational time in seconds for the PP GaSP and MS GaSP emulators, are given in Table 2.1 and Table 2.2 for two computational scenarios, these actual times reflect the extreme theoretical disparity discussed above.

The expense of estimating $(\boldsymbol{\gamma}, \eta)$ suggests that various approximation strategies be utilized. In Section 2.7.3, we will consider two such strategies, basing the estimation on only subsets of the designed inputs $\mathbf{x}^{\mathscr{D}}$, and use of composite likelihoods.

2.5.2 Out of sample prediction

Here we compare the performance of the PP GaSP and MS GaSP emulators in out-of-sample prediction. We also include emulators from the next section in the comparison; these have been considered for situations similar to ours in recent literature.

Coregionalization emulators

Another approach to emulation of multiple outputs is the Linear Model of Coregionalization (LMC) emulator (see, e.g. Fricker et al. (2013)). In this approach, the output $\mathbf{Y}(\mathbf{x})_{[k\times 1]}$ is modeled as

$$\mathbf{Y}(\mathbf{x}) = \boldsymbol{\mu}(\mathbf{x}) + \mathbf{A}\mathbf{v}(\mathbf{x}) + \boldsymbol{\epsilon}\,, \tag{2.12}$$

where $\mathbf{A}$ is a $k \times k_0$ matrix, $k_0 < k$, and $\mathbf{v}(\mathbf{x}) = (\mathbf{v}_1(\mathbf{x}), \ldots, \mathbf{v}_{k0}(\mathbf{x}))^T$, with the $\mathbf{v}_i(\mathbf{x})$ being zero mean independent GaSP emulators and $\boldsymbol{\epsilon}$ is independent noise. Denote the observed output matrix as $\mathbf{Y}_{[k\times n]} = (\mathbf{Y}(\mathbf{x}_1), \ldots, \mathbf{Y}(\mathbf{x}_n))$. In Higdon et al. (2008), the output is normalized and then represented by a singular value decomposition (SVD), $\mathbf{Y} - \bar{\mathbf{Y}}_{\mathbf{row}} = \mathbf{U}\mathbf{D}\mathbf{V}^T$,where $\bar{\mathbf{Y}}_{\mathbf{row}}$ is the row mean of $\mathbf{Y}$. $\mathbf{A}$ is then estimated as the first k_0 columns of $\mathbf{U}\mathbf{D}/\sqrt{n}$. In Paulo et al. (2012), $\mathbf{A}$ is estimated as the column

in the eigen-decomposition of the observed variance matrix of $\mathbf{Y}$, and $\boldsymbol{\epsilon}$ is omitted because no dimension reduction is used. In Rougier (2008), dimension reduction with varying $\mathbf{h}(\mathbf{x})$ in each coordinate is also discussed.

Note that, when $k > n$, non-zero singular values of $\mathbf{Y}$ by SVD and eigenvalues by eigen-decomposition are equal to or smaller than n. In our situation, $k \gg n$, while the rank of the estimated $\hat{\mathbf{A}}$ is at most n, using either of these two approaches. In the numerical comparisons we will include the LMC emulator, with $\mathbf{A}$ being estimated by the eigenvectors of the observed covariance matrix of the output (Paulo et al. (2012)). We will also compare the method of estimating the correlation parameters and nugget that is developed in Section 2.7, which we call *robust estimation*, with the standard method in DiceKriging package (Roustant et al. (2012)).

Design of the numerical study

To evaluate the accuracy of various variants of the PP emulator and alternative emulators, we divide the simulator runs into two parts, those used for development of the emulator and those used for out-of-sample assessment of accuracy. We utilized only $n = 50$ runs to design the emulator because of the extreme computational difficulty of working with the MS emulator (the primary emulator for comparison) for larger n, as discussed in Section 2.5.1. Also, we surprisingly found that the PP-emulator based on only 50 runs is quite accurate, and using a large number of runs to build the emulators would likely have made it more difficult to see differences or problems.

We consider two evaluation scenarios. The first encompasses the entire island except the crater, but is limited to flow volumes $6 < \log_{10}V < 7.5$; 683 runs are available in this region. The second scenario focuses on regions of the island with moderate to small expected flows, since these regions are the subject of current major risk assessment. We omit the crater region from the analysis because there

is no interest in hazard prediction there, and the flows are so large that they could adversely affect the estimation of the GaSP correlation parameters. Locations where all 50 simulator runs had maximum flow heights of zero were also eliminated from the analysis. The total number of remaining spatial coordinates was $23,040$; $k = 17,311$ coordinates for the first case and $k = 14,911$ for the second. Of course, utilizing a single emulator over such a large domain might well not work, and a natural strategy to consider is to divide the domain into more homogeneous regions and develop separate emulators over each region; luckily, this did not seem to be necessary for our scientific application.

Before proceeding with the complex emulators, it is useful to check that simple methods, such as linear regression, are not adequate for the problem. Thus Figure 2.4 compares the use of simple linear regression of the output versus V at a specific location, based on the 50 training runs of the simulator, with the PP GaSP built on the same 50 runs, for predicting the 633 other simulator runs. Clearly the linear regression estimates are far less accurate.

Prediction criteria

Diagnostics for GaSP emulation have been discussed in Bastos and O'Hagan (2009). The criteria that we focus on are out-of-sample prediction and accuracy in uncertainty quantification. In the following, we denote x_i^*, $1 \leqslant i \leqslant n^*$, as the held-out runs to verify the performance of the emulators; we have $n^* = 633$. The specific

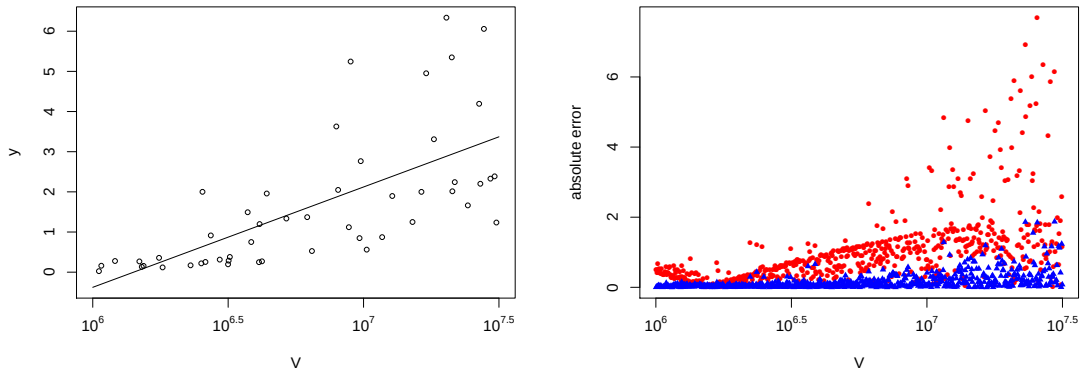

FIGURE 2.4: The left figure is the least squares fit of simulator output to volume, for 50 simulator runs at a specific location. The right figure compares use of this least squares fit to estimate the outputs of 633 other simulator runs (the red dots), corresponding to other input values at the same location, with use of the PP GaSP (developed from the same 50 simulator runs) to estimate the 633 outputs (the blue triangles). Accuracy is measured by the absolute error of the prediction $|y(x_i^*) - \hat{y}(x_i^*)|$.

criteria employed are the following:

$$\begin{aligned} MSE &= \frac{\sum_{j=1}^{k}\sum_{i=1}^{n^*}\left(y_j(\mathbf{x}_i^*) - \hat{y}_j(\mathbf{x}_i^*)\right)^2}{kn^*}, \\ P_{CI}(95\%) &= \frac{1}{kn^*}\sum_{j=1}^{k}\sum_{i=1}^{n^*} 1\{y_j(\mathbf{x}_i^*) \in CI_{ij}(95\%)\}, \\ L_{CI}(95\%) &= \frac{1}{kn^*}\sum_{j=1}^{k}\sum_{i=1}^{n^*} \text{length}\{CI_{ij}(95\%)\}, \end{aligned}$$

where $\hat{y}_j(\mathbf{x}_i^*)$ is the prediction of the output of the ith held-out run, $\mathbf{x}_i^*$, at the jth spatial coordinate; $CI_{ij}(95\%)$ is the 95% posterior credible interval based on (2.3); and length$\{CI_{ij}(95\%)\}$ is the length of the 95% posterior credible interval. An ideal emulator would have relatively low Mean Square Error (MSE), $P_{CI}(95\%)$ close to the 95% nominal level, and short average credible interval lengths.

Table 2.1: Performance of various emulators of max flow height over spatial grids in all locations except the crater area and non-flow areas. The first emulator uses all 4 inputs while the remaining four emulators use 3 inputs (V, δ_{bed}, ϕ) and nugget(s), all with the same regressor $\mathbf{h}(\mathbf{x}) = (1, V)$. The emulators are evaluated based on $n^* = 633$ held-out inputs over $k = 17,311$ locations. The last row shows the computational time needed to estimate the correlation parameters and nuggets in the emulators (the dominant part of the computational cost), using R and [C++].

	PP GaSP robust est.	PP GaSP robust est.	MS GaSP robust est.	MS GaSP DiceKriging	LMC GaSP robust est.
	4 inputs	3 inputs and estimated nugget(s)			
MSE	0.109	0.097	0.103	0.114	0.123
$P_{CI}(95\%)$	0.926	0.950	0.924	0.900	0.903
$L_{CI}(95\%)$	0.521	0.536	0.491	0.462	0.449
time (s)	50.0	28.1 [2.0]	31337.7	4493.2	83.6

Emulation over the non-crater region with constrained flow volumes

Table 2.1 presents the results for the $k = 17,311$ non-crater locations with constrained flow volumes.

First, note that the computation times for the emulators reflect what was discussed in Section 2.5.1; the PP GaSP emulator is roughly *three orders of magnitude* faster than the MS emulator. The MS emulator with parameters estimated by DiceKriging was faster, because it incorporated certain optimization techniques and the underlying codes were written in C, but it was still two orders of magnitude slower than PP GaSP using R codes. The speed of LMC emulator was similarly to PP GaSP because it projects the k dimensional space onto a n dimensional subspace (as discussed in Section 2.5.2).

The PP GaSP emulator based on three inputs and the nugget outperformed the PP GaSP emulator based on four inputs. It had better MSE and more accurate coverage, with only slightly longer credible intervals. This was also true for the second test situation, as evidenced in Table 2.2.

The PP GaSP emulator had the lowest out-of-sample MSE result among the four

emulators based on three inputs and a nugget and, as importantly, produced 95% credible intervals that actually covered approximately 95% of the held-out outputs. In contrast, the other emulators were over-confident in their accuracy assessments. This is not surprising for the LMC GaSP emulator, since its projection onto a n dimensional subspace is too restricted, but it is surprising for the MS emulator, which we had entertained as being the gold standard because of its increased flexibility. The average length of the credible intervals for the PP emulator with 3 inputs and a nugget was slightly longer than for the other emulators but, again, that is very likely due to the other emulators being over-confident.

The MSE's of all of the emulators are rather impressive, especially when realizing that the output values they are predicting ranged from 0 to 40 in the non crater area. Likewise the small average size of the credible intervals is impressive for predicting outputs over that range.

The reason for the comparatively poor performance of the MS emulator is that fairly often (i.e. at some significant fraction of the coordinates), the estimates of the correlation parameters are bad, because (i) only limited number of computer runs are used ($n = 50$); (ii) each location will have many simulator runs with zero flow heights, which can cause problems for Gaussian processes. The first issue could be dealt with by using more simulator runs to develop the emulators, but this drastically increases the computational cost. The second issue, however, is generic in emulating the TITAN2D computer model; each pyroclastic flow will only hit some of the locations on the island, with the others receiving zero flow. In contrast, while the PP emulator may not have the optimal correlation parameters at any coordinate, the stability of their estimation ensures good average prediction.

MS emulator implemented in DiceKriging package optimizes both the range parameters $\boldsymbol{\gamma}$ and smoothness parameters α of the power exponential correlation function at the same time, and typically results in smaller out of sample MSE than the

Table 2.2: Performance of various emulators of max flow height over the $k = 14,911$ locations in the moderate to small flow area. The first emulator uses 4 inputs while the remaining four emulators use 3 inputs (V, δ_{bed}, ϕ) and nugget(s), all with the same regressor $\mathbf{h}(\mathbf{x}) = (1, V)$. The emulators are evaluated based on $n^* = 633$ held-out inputs. The last row shows the computational time needed to estimate the correlation parameters and nuggets in the emulators (the dominant part of the computational cost), using R and [C++].

	PP GaSP robust est.	PP GaSP robust est.	MS GaSP robust est.	MS GaSP DiceKriging	LMC GaSP robust est.
	4 inputs	3 inputs and estimated nugget(s)			
MSE	0.057	0.050	0.055	0.061	0.062
$P_{CI}(95\%)$	0.930	0.950	0.924	0.900	0.900
$L_{CI}(95\%)$	0.350	0.358	0.319	0.299	0.298
time (s)	42.0	38.8 [2.1]	27150.9	3835.4	81.2

one with fixed smoothness parameters α and estimated range parameters $\boldsymbol{\gamma}$, when the moderate number of runs are used. However, it is not performing as good as MS emulators with a robust estimation partly because of the periodic folding adjustment (Spiller et al. (2014)) for the initial flow angle ϕ, but the major improvement arises from the robust estimation of the correlation and nugget parameters that is given in Section 2.7. Chapter 3 discusses the estimation of GaSP correlation parameters in details.

The LMC GaSP emulator using eigen-decomposition to estimate the orthogonal basis matrix $\mathbf{A}$ (Paulo et al. (2012)) performs the worst among 4 emulators. Using an orthogonal basis with 50 dimensions does not seem to be flexible enough to capture the variations among the $k = 17,911$ locations. Remember that this is the case when k is very large, in Chapter 5, we will show an example that when $n \gg k$, the strategy should then change accordingly.

Emulation over the region having only moderate to small flows

Table 2.2 presents the MSE results for the $14,911$ locations in the small to moderate flow region. The PP GaSP outperforms the MS GaSP by more than 10% in terms of

MSE and again has considerably better confidence properties. The degraded performance of the MS GaSP here is probably due to the fact that the small flow regions have numerous 0 max flow heights, which can cause problems in the estimation of the range and nugget parameters. And, of course, the computational advantage of the PP emulator was enormous.

2.6 The near irrelevance of spatial correlation in emulator construction

2.6.1 The identical predictive mean and variance by PP GaSP emulator

A seemingly natural extension of the PP GaSP emulator is to introduce spatial correlation into the model, as in Conti and O'Hagan (2010), in recognition of the fact that there is typically strong spatial dependence between simulator outputs at nearby inputs. (Recall that the PP emulator assumes each output is independent.) To keep the computation manageable, the spatial correlations and model input correlations are typically presumed to be separable, i.e., the covariance function for $\mathbf{y}^{\mathscr{D}}$, conditional on $\mathbf{\Theta}$, is assumed to be a Kronecker product of a $k \times k$ spatial correlation matrix $\mathbf{\Sigma}$ and the $n \times n$ input correlation matrix $\mathbf{R}$, leading to the following matrix-normal density for the Gaussian process:

$$p(\mathbf{y}^{\mathscr{D}} \mid \mathbf{\Sigma}, \mathbf{\Theta}, \boldsymbol{\gamma}) = \frac{\exp(-\frac{1}{2} tr[\mathbf{\Sigma}^{-1}(\mathbf{y}^{\mathscr{D}} - \mathbf{h}(\mathbf{x}^{\mathscr{D}})\mathbf{\Theta})^T \mathbf{R}^{-1}(\mathbf{y}^{\mathscr{D}} - \mathbf{h}(\mathbf{x}^{\mathscr{D}})\mathbf{\Theta})])}{(2\pi)^{nk/2} \mid \mathbf{\Sigma}|^{n/2} |\mathbf{R}|^{k/2}}, \quad (2.13)$$

where $\mathbf{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_k)$ is the $q \times k$ matrix of parameters of the mean function for the k spatial coordinates.

In Conti and O'Hagan (2010), a Jeffreys-type non-informative prior,

$$\pi(\mathbf{\Theta}, \mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-(k+1)/2}, \quad (2.14)$$

was considered, since one can then exactly marginalize out $\mathbf{\Theta}$ and $\mathbf{\Sigma}$ when k is small.

This does not work, however, if $k > n - q$, the situation we are considering, since there is then a non-integrable singularity in the posterior at $\mathbf{\Sigma} = \mathbf{0}$.

A wide variety of other prior distributions on $\mathbf{\Sigma}$ can be considered, including priors that effectively give $\mathbf{\Sigma}$ a lower dimensional structure. Indeed we propose one such prior in Appendix A.3, which is effective in smoothing random draws from the PP emulator; recall that, because of the independence assumption at each coordinate, draws directly from the PP GaSP emulator will be quite rough, although the median, mean and quantiles of the PP GaSP are smooth.

Surprisingly, however, for *any* prior on $\mathbf{\Sigma}$ the resulting emulator mean will simply be the PP emulator mean (assuming the usual constant prior is used for the parameters of the mean function), and the resulting emulator variance function will almost equal the PP emulator variance function. Thus there is no need to introduce spatial correlation structure into the emulator with regard to the response space if only the mean and pointwise variance functions are concerned. This delightful simplification is established in the next theorem.

Theorem 2.6.1. *For the GaSP with separable covariance structure in (2.13), given correlation parameters $\boldsymbol{\gamma}$ and the objective prior*

$$\pi(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k \mid \mathbf{\Sigma}, \boldsymbol{\gamma}) \propto 1 \tag{2.15}$$

for the parameters of the mean function, the following hold:

1. *The posterior mean of the GaSP, for an unobserved $\mathbf{x}^*$ and at coordinate j, is identical to the PP emulator posterior mean in (2.4).*

2. *The posterior variance of the GaSP, for an unobserved $\mathbf{x}^*$ and at coordinate j, depends on $\mathbf{\Sigma}$ only through the posterior mean of the j^{th} diagonal term, $\mathrm{E}[\mathbf{\Sigma}_{jj} \mid \mathbf{y}^{\mathscr{D}}, \boldsymbol{\gamma}]$; it is identical to the PP emulator posterior variance if $\mathrm{E}[\mathbf{\Sigma}_{jj} \mid \mathbf{y}^{\mathscr{D}}, \boldsymbol{\gamma}] = \frac{(n-q)\hat{\sigma}_j^2}{n-q-2}$, with $\hat{\sigma}_j^2$ defined in (2.5), under the new prior for $\mathbf{\Sigma}$.*

Proof. The joint distribution of $\mathbf{y}(\mathbf{x}^*) = (y_1(\mathbf{x}^*), y_2(\mathbf{x}^*), \ldots, y_k(\mathbf{x}^*))$ and $\mathbf{y}^{\mathscr{D}}$ is a matrix normal distribution,

$$\begin{pmatrix} \mathbf{y}(\mathbf{x}^*) \\ \mathbf{y}^{\mathscr{D}} \end{pmatrix} \mid \boldsymbol{\Theta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma} \sim \mathcal{N}_{(n+1),k} \left(\begin{pmatrix} \mathbf{h}(\mathbf{x}^*)\boldsymbol{\Theta} \\ \mathbf{h}(\mathbf{x})\boldsymbol{\Theta} \end{pmatrix}, \begin{pmatrix} c(\mathbf{x}^*, \mathbf{x}^*) & \mathbf{r}^T(\mathbf{x}^*) \\ \mathbf{r}(\mathbf{x}^*) & \mathbf{R} \end{pmatrix}, \Sigma \right),$$

where $N_{(n+1),k}(\cdot, \cdot, \cdot)$ is a $(n+1) \times k$ matrix normal distribution. From Gupta and Nagar (1999), it follows that

$$\mathrm{E}[\mathbf{y}(\mathbf{x}^*) \mid \mathbf{y}^{\mathscr{D}}, \Theta, \boldsymbol{\gamma}, \boldsymbol{\Sigma}] = \mathbf{h}(\mathbf{x}^*)\boldsymbol{\Theta} + \mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1}(\mathbf{y}^{\mathscr{D}} - \mathbf{h}(\mathbf{x}^{\mathscr{D}})\boldsymbol{\Theta}),$$

and, for the j^{th} coordinate,

$$\mathrm{E}[\mathbf{y}_j(\mathbf{x}^*) \mid \mathbf{y}^{\mathscr{D}}, \boldsymbol{\theta}_j, \boldsymbol{\gamma}, \boldsymbol{\Sigma}] = \mathbf{h}(\mathbf{x}^*)\boldsymbol{\theta}_j + \mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1}(\mathbf{y}_j^{\mathscr{D}} - \mathbf{h}(\mathbf{x}^{\mathscr{D}})\boldsymbol{\theta}_j). \tag{2.16}$$

Using the objective prior in Equation (2.15) results in

$$\boldsymbol{\theta}_j \mid \mathbf{y}^{\mathscr{D}}, \boldsymbol{\gamma}, \boldsymbol{\Sigma} \sim \mathcal{N}\left(\hat{\boldsymbol{\theta}}_j, \Sigma_{jj}[(\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}}))]^{-1}\right),$$

where $\hat{\boldsymbol{\theta}}_j = \{\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})\}^{-1}\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{y}_j^{\mathscr{D}}$. Taking the expectation over $\boldsymbol{\theta}_j$ in (2.16), results in the expression for the posterior mean in (2.4), as was to be established.

For the posterior variance, after marginalizing out the parameters of the mean function with the objective prior in Equation (2.15), it is shown in Conti and O'Hagan (2010) that

$$\begin{aligned} &\mathrm{Var}[y_j(\mathbf{x}^*) \mid \mathbf{y}^{\mathscr{D}}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}] = \sigma_j^2\{(c(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}^*)) \\ &+ [\mathbf{h}(\mathbf{x}^*) - \mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}^*)]^T[\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})]^{-1}[\mathbf{h}(\mathbf{x}^*) - \mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}^*)]\}, \end{aligned}$$

where $\sigma_j^2 = \Sigma_{jj}$. Now

$$\begin{aligned} &\mathrm{Var}[y_j(\mathbf{x}^*) \mid \mathbf{y}^{\mathscr{D}}, \boldsymbol{\gamma}] \\ &= \mathrm{Var}_{\boldsymbol{\Sigma}|\mathbf{y}^{\mathscr{D}}, \boldsymbol{\gamma}}\left[\mathrm{E}[y_j(\mathbf{x}^*) \mid \mathbf{y}^{\mathscr{D}}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}]\right] + \mathrm{E}_{\boldsymbol{\Sigma}|\mathbf{y}^{\mathscr{D}}, \boldsymbol{\gamma}}\left[\mathrm{Var}[y_j(\mathbf{x}^*) \mid \mathbf{y}^{\mathscr{D}}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}]\right], \end{aligned}$$

but the first term is zero, since the posterior mean does not depend on $\boldsymbol{\Sigma}$. Noting that $\mathrm{Var}[y_j(\mathbf{x}^*) \mid \mathbf{y}^{\mathscr{D}}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}] = \sigma_j^2 \times c^{**}$, where c^{**} is as in (1.7), it is immediate that

$$\mathrm{Var}[y_j(\mathbf{x}^*) \mid \mathbf{y}^{\mathscr{D}}, \boldsymbol{\gamma}] = \mathrm{E}[\sigma_j^2 \mid \mathbf{y}^{\mathscr{D}}, \boldsymbol{\gamma}]\ c^{**}\,,$$

as was to be established. □

Note that, when $n-q$ is moderately large, as is usually the case, (2.5) will approximately equal the new posterior expectation of σ_j^2, since almost all the information about σ_j^2 is contained in the likelihood, not the prior. Thus, in practice, one can just use the PP emulator mean and variance, and ignore the spatial structure, unless draws from the emulator are required.

2.7 Estimating the correlation parameters

In this section, we discuss estimation of the correlation parameter $\boldsymbol{\gamma}$. First the reference prior for $\boldsymbol{\gamma}$ in the situation of vector-valued outputs (as considered in this paper) is derived. Next, an estimation strategy is proposed, utilizing the posterior mode. Finally, to overcome the computational challenge, a composite likelihood approach is considered.

2.7.1 The reference priors for vector output

When dealing with a Gaussian process with a single real output, the reference prior under an isotropic kernel was derived in Berger et al. (2001) and under a product correlation matrix in Bayarri et al. (2009). As recommended in Berger et al. (2001); Paulo (2005), we follow the strategy of first marginalizing out the parameters of the mean function (with respect to a constant prior) and then deriving the reference prior for $\boldsymbol{\gamma}$ from the marginal likelihood; the result is given in the following theorem.

Theorem 2.7.1. *(Reference prior for PP GaSP without a nugget). The reference*

prior of the PP GaSP for vector output has the form

$$\pi^R(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k,\sigma_1^2,\ldots,\sigma_k^2,\boldsymbol{\gamma}) \propto \frac{\pi^R(\boldsymbol{\gamma})}{\prod_{i=1}^{k}\sigma_i^2},$$

with $\pi^R(\boldsymbol{\gamma}) \propto |\mathbf{I}^*(\boldsymbol{\gamma})|^{1/2}$*, where* $\mathbf{I}^*(\boldsymbol{\gamma})$ *is the expected Fisher Information matrix*

$$\mathbf{I}^*(\boldsymbol{\gamma}) = \begin{pmatrix} n-q & tr(\mathbf{W}_1) & tr(\mathbf{W}_2) & \ldots & tr(\mathbf{W}_p) \\ & tr(\mathbf{W}_1^2) & tr(\mathbf{W}_1\mathbf{W}_2) & \ldots & tr(\mathbf{W}_1\mathbf{W}_p) \\ & & tr(\mathbf{W}_2^2) & \ldots & tr(\mathbf{W}_2\mathbf{W}_p) \\ & & & \ddots & \vdots \\ & & & & tr(\mathbf{W}_p^2) \end{pmatrix}_{(p+1)\times(p+1)}, \tag{2.17}$$

with $\mathbf{W}_t = \dot{\mathbf{R}}_t\mathbf{Q}$*, for* $1 \leqslant t \leqslant p$*, where* p *is the number of range parameters in the correlation matrix* $\mathbf{R}$*,* $\dot{\mathbf{R}}_t$ *is the derivative of the correlation matrix* $\mathbf{R}$ *with respect to the* t^{th} *parameter, and* $\mathbf{Q} = \mathbf{R}^{-1}\mathbf{P}$ *with* $\mathbf{P} = \mathbf{I} - \mathbf{h}(\mathbf{x}^{\mathscr{D}})\{\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})\}^{-1}\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}$*.*

Proof. As in Berger et al. (2001), we derive the reference prior based on the marginal likelihood after integrating out the parameters of the mean function $\boldsymbol{\theta}$ with a constant prior. The log marginal likelihood, conditional on $(\boldsymbol{\gamma}, \boldsymbol{\sigma}^2)$, is

$$\log(\mathcal{L}(\mathbf{y}|\boldsymbol{\gamma},\boldsymbol{\sigma}^2)) \propto -\frac{n-q}{2}\sum_{i=1}^{k}\log(\sigma_i^2) - \frac{k}{2}\log(|\mathbf{R}|) \tag{2.18}$$

$$-\frac{k}{2}\log(|\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})|) - \frac{(n-q)}{2}\sum_{i=1}^{k}\log(S_i^2),$$

with

$$S_i^2 = (\mathbf{y}_i^{\mathscr{D}})^T\mathbf{Q}\mathbf{y}_i^{\mathscr{D}}. \tag{2.19}$$

As in the proof of Theorem 2 in Berger et al. (2001), direct computation yields

$$\begin{aligned}
\mathrm{E}\left(\frac{\partial \log(\mathcal{L}(\mathbf{y}|\boldsymbol{\gamma},\boldsymbol{\sigma}^2))}{\partial \sigma_i^2}\right)^2 &= \frac{n-q}{2\sigma_i^4}\,,\\
\mathrm{E}\left(\frac{\partial \log(\mathcal{L}(\mathbf{y}|\boldsymbol{\gamma},\boldsymbol{\sigma}^2))}{\partial \sigma_i^2}\frac{\partial \log(\mathcal{L}(\mathbf{y}|\boldsymbol{\gamma},\boldsymbol{\sigma}^2))}{\partial \sigma_j^2}\right) &= 0\,,\\
\mathrm{E}\left(\frac{\partial \log(\mathcal{L}(\mathbf{y}|\boldsymbol{\gamma},\boldsymbol{\sigma}^2))}{\partial \gamma_l}\right)^2 &= \frac{k}{2}\mathrm{tr}(\mathbf{W}_l^2)\,,\\
\mathrm{E}\left(\frac{\partial \log(\mathcal{L}(\mathbf{y}|\boldsymbol{\gamma},\boldsymbol{\sigma}^2))}{\partial \gamma_l}\frac{\partial \log(\mathcal{L}(\mathbf{y}|\boldsymbol{\gamma},\boldsymbol{\sigma}^2))}{\partial \gamma_m}\right) &= \frac{k}{2}\mathrm{tr}(\mathbf{W}_l\mathbf{W}_m)\,,\\
\mathrm{E}\left(\frac{\partial \log(\mathcal{L}(\mathbf{y}|\boldsymbol{\gamma},\boldsymbol{\sigma}^2))}{\partial \sigma_i^2}\frac{\partial \log(\mathcal{L}(\mathbf{y}|\boldsymbol{\gamma},\boldsymbol{\sigma}^2))}{\partial \gamma_l}\right) &= \frac{1}{2\sigma_i^2}\mathrm{tr}(\mathbf{W}_l)\,,
\end{aligned}$$

where $1 \leqslant i \neq j \leqslant k$ and $1 \leqslant l \neq m \leqslant p$. The Fisher information matrix is

$$|\mathbf{I}^*(\boldsymbol{\gamma},\boldsymbol{\sigma}^2)| \propto \left| \begin{array}{cc} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{array} \right|_{(k+p)\times(k+p)}\,,$$

with

$$\mathbf{A} = \begin{pmatrix} \frac{n-q}{2\sigma_1^4} & 0 & 0 & 0 \\ 0 & \frac{n-q}{2\sigma_2^4} & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{n-q}{2\sigma_k^4} \end{pmatrix}_{k\times k}, \quad \mathbf{B} = \begin{pmatrix} \frac{\mathrm{tr}(\mathbf{W}_1)}{2\sigma_1^2} & \frac{\mathrm{tr}(\mathbf{W}_2)}{2\sigma_1^2} & \cdots & \frac{\mathrm{tr}(\mathbf{W}_p)}{2\sigma_1^2} \\ \frac{\mathrm{tr}(\mathbf{W}_1)}{2\sigma_2^2} & \frac{\mathrm{tr}(\mathbf{W}_2)}{2\sigma_2^2} & \cdots & \frac{\mathrm{tr}(\mathbf{W}_p)}{2\sigma_2^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\mathrm{tr}(\mathbf{W}_1)}{2\sigma_k^2} & \frac{\mathrm{tr}(\mathbf{W}_2)}{2\sigma_k^2} & & \frac{\mathrm{tr}(\mathbf{W}_p)}{2\sigma_k^2} \end{pmatrix}_{k\times p},$$

$$\mathbf{C} = \begin{pmatrix} \frac{k\mathrm{tr}(\mathbf{W}_1^2)}{2} & \frac{k\mathrm{tr}(\mathbf{W}_1\mathbf{W}_2)}{2} & \cdots & \frac{k\mathrm{tr}(\mathbf{W}_1\mathbf{W}_p)}{2} \\ \frac{k\mathrm{tr}(\mathbf{W}_1\mathbf{W}_2)}{2} & \frac{k\mathrm{tr}(\mathbf{W}_2^2)}{2} & \cdots & \frac{k\mathrm{tr}(\mathbf{W}_2\mathbf{W}_p)}{2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{k\mathrm{tr}(\mathbf{W}_1\mathbf{W}_p)}{2} & \frac{k\mathrm{tr}(\mathbf{W}_2\mathbf{W}_p)}{2} & \cdots & \frac{k\mathrm{tr}(\mathbf{W}_p^2)}{2} \end{pmatrix}_{p\times p}.$$

The result soon follows from $\pi^R(\boldsymbol{\gamma},\boldsymbol{\sigma}^2) \propto |\mathbf{I}^*(\boldsymbol{\gamma},\boldsymbol{\sigma}^2)|^{1/2}$.

□

Theorem 2.7.2. *(Reference prior for PP GaSP with a nugget). The reference prior of the PP GaSP with a nugget, for vector output, has the form*

$$\pi^{\tilde{R}}(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k,\sigma_1^2,\ldots,\sigma_k^2,\boldsymbol{\gamma},\eta) \propto \frac{\pi^{\tilde{R}}(\boldsymbol{\gamma},\eta)}{\prod_{i=1}^{k}\sigma_i^2},$$

with $\pi^{\tilde{R}}(\boldsymbol{\gamma},\eta) \propto |\tilde{\mathbf{I}}^*(\boldsymbol{\gamma},\eta)|^{1/2}$*, where* $\tilde{\mathbf{I}}^*(\boldsymbol{\gamma},\eta)$ *is the expected fisher information matrix*

$$\tilde{\mathbf{I}}^*(\boldsymbol{\gamma},\eta) = \begin{pmatrix} n-q & tr(\tilde{\boldsymbol{W}}_1) & tr(\tilde{\mathbf{W}}_2) & \ldots & tr(\tilde{\mathbf{W}}_{p+1}) \\ & tr(\tilde{\mathbf{W}}_1^2) & tr(\tilde{\mathbf{W}}_1\tilde{\mathbf{W}}_2) & \ldots & tr(\tilde{\mathbf{W}}_1\tilde{\mathbf{W}}_{p+1}) \\ & & tr(\tilde{\mathbf{W}}_2^2) & \ldots & tr(\tilde{\mathbf{W}}_2\tilde{\mathbf{W}}_{p+1}) \\ & & & \ddots & \vdots \\ & & & & tr(\tilde{\mathbf{W}}_{p+1}^2) \end{pmatrix}_{(p+2)\times(p+2)}, \quad (2.20)$$

with $\tilde{\mathbf{W}}_t = \dot{\tilde{\mathbf{R}}}_t\tilde{\mathbf{Q}}$*, for* $1 \leqslant t \leqslant p$*, where* p *is the number of range parameters in the correlation matrix* $\tilde{\mathbf{R}}$*,* $\dot{\tilde{\mathbf{R}}}_t$ *is the derivative of the correlation matrix* $\tilde{\mathbf{R}}$ *with respect to the* t^{th} *parameter, and* $\tilde{\mathbf{Q}} = \tilde{\mathbf{R}}^{-1}\tilde{\mathbf{P}}$ *with* $\tilde{\mathbf{P}} = \mathbf{I} - \mathbf{h}(\mathbf{x}^{\mathscr{D}})\{\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\tilde{\mathbf{R}}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})\}^{-1}\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\tilde{\mathbf{R}}^{-1}$*.*

Proof. The proof is a direct generalization of Ren et al. (2012), essentially following the same steps as the proof of Theorem 2.7.1. □

2.7.2 Marginal Posterior

We will utilize the marginal posterior density of $\boldsymbol{\gamma}$ and η to perform the estimation for these parameters. Starting with the full likelihood, multiplying by the reference prior, and integrating out the parameters of the mean function, $(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k)$, and variance parameters, $(\sigma_1^2,\ldots,\sigma_k^2)$, results in

$$p(\boldsymbol{\gamma},\eta \mid \mathbf{y}^{\mathscr{D}}) \propto \mathcal{L}(\mathbf{y}^{\mathscr{D}} \mid \boldsymbol{\gamma},\eta)|\tilde{\mathbf{I}}^*(\boldsymbol{\gamma},\eta)|^{1/2}, \quad (2.21)$$

with

$$\mathcal{L}(\mathbf{y}^{\mathscr{D}} \mid \boldsymbol{\gamma},\eta) \propto |\tilde{\mathbf{R}}|^{-k/2}|\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\tilde{\mathbf{R}}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})|^{-k/2}\prod_{i=1}^{k}[(\mathbf{y}_i^{\mathscr{D}})^T\tilde{\mathbf{Q}}\mathbf{y}_i^{\mathscr{D}}]^{-(n-q)/2}. \quad (2.22)$$

We reparameterize the parameters by

$$(\xi_1, \ldots, \xi_p, \tau) = (\log(1/\gamma_1^{\alpha_1}), \ldots, \log(1/\gamma_p^{\alpha_p}), \log(\eta)).$$

We then estimate the parameters $(\boldsymbol{\xi}, \tau)$ as the mode of this marginal posterior, namely

$$(\hat{\xi}_1, \ldots \hat{\xi}_p, \hat{\tau}) = \underset{\xi_1,\ldots,\xi_p,\tau}{\operatorname{argmax}} \mathcal{L}(\mathbf{y}^{\mathscr{D}} \mid \xi_1, \ldots, \xi_p, \tau)\pi^{\tilde{R}}(\xi_1, \ldots, \xi_p, \tau). \tag{2.23}$$

The difference between the posterior mode of $(\boldsymbol{\gamma}, \eta)$ and $(\boldsymbol{\xi}, \tau)$ arises because of the Jacobian of the transformation. The marginal likelihood alone can have bad behavior, such as being maximized as parameters go to infinity. Using the marginal posterior, with respect to the reference prior, will substantially eliminate such bad behavior and achieve comparatively better results, if a correct parameterization is used.

Note that there have been a variety of parameterizations for GaSP's that have been used in the past literature, other than the $\boldsymbol{\gamma}$ parameterization in (2.1). The parameterization $\beta_j = 1/\gamma_j^{\alpha_j}$ was discussed in Paulo (2005) and the parameterization $\xi_j = \log(1/\gamma_j^{\alpha_j})$ was introduced in Spiller et al. (2014). The effectiveness of the different parameterizations, when combined with use of the posterior mode, is extensively discussed in Chapter 3, where a robustness argument in favor of the above parameterization is given. For the MS GaSP with a nugget, the same strategy is used: form the marginal posterior distribution of $(\boldsymbol{\xi}, \tau)$ and utilize the posterior mode of these parameters in the MS GaSP.

One concern with using (2.22) is that the assumption of independence of coordinates is clearly wrong, so that this likelihood is almost certainly much too concentrated. This, by itself, would not be a problem, since we am simply using it to obtain estimates of the correlation parameters, but it is possible that the likelihood

Table 2.3: MSE comparison between PP GaSP and Oracle PP GaSP with $n = 50$ and $n^* = 633$.

	PP GaSP	Oracle PP GaSP
MSE at non crater area	0.09726675	0.09464283
MSE at small flow area	0.04964399	0.04837848

would also be biased in some way. To investigate this, we define the following "oracle estimator," which views the posterior predictive mean in Lemma 2.3.1 as a function of the range parameters and nugget $(\boldsymbol{\xi}, \tau)$ and optimizes over the choice of these parameters:

$$(\xi_1^{oracle}, \ldots, \xi_p^{oracle}, \tau^{oracle}) = \underset{\xi_1,\ldots,\xi_p,\tau}{argmin} \frac{\sum_{j=1}^{k}\sum_{i=1}^{n^*}\left(y_j(\mathbf{x}_i^*) - \hat{y}_j^{oracle}(\mathbf{x}_i^*)\right)^2}{kn^*}, \tag{2.24}$$

where

$$\begin{aligned} \hat{y}_j^{oracle}(\mathbf{x}_i^*) &= \boldsymbol{\omega}(\xi_1, \ldots, \xi_p, \tau)\mathbf{y}_j^{\mathscr{D}} \\ \boldsymbol{\omega}(\xi_1, \ldots, \xi_p, \tau) &= \left(\mathbf{h}(\mathbf{x}^*) - \mathbf{r}^T(\mathbf{x}^*)\tilde{\mathbf{R}}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})\right)\left(\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\tilde{\mathbf{R}}^{-1}\ \mathbf{h}(\mathbf{x}^{\mathscr{D}})\right)^{-1} \times \\ &\quad \mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\tilde{\mathbf{R}}^{-1} + \mathbf{r}^T(\mathbf{x}^*)\tilde{\mathbf{R}}^{-1}. \end{aligned}$$

Table 2.3 compares the MSE of the oracle PP GaSP and the PP GaSP. For both of the regions under consideration, the PP GaSP has almost the same MSE as the oracle, so that the use of (2.22) in estimating the correlation parameters and nugget seems justified.

2.7.3 Using composite likelihood

As discussed in Section 2.5.1, the major computational challenge in developing the PP emulator is estimating the parameters $(\boldsymbol{\gamma}, \eta)$, the computation being of order $O(tn^2k) + O(tn^3)$, with t being the number of iterations (typically about 200) needed

to find a good approximation to the marginal posterior mode discussed in the previous section. A variety of strategies have been proposed to reduce this computational burden. Use of covariance tapering and compactly supported correlation functions were studied and successfully applied to large spatial datasets in Kaufman et al. (2008, 2011). Other possibilities include estimating the parameters using only some subsets of the input design points (i.e., significantly reduce n) or using only some coordinates of the simulator output (i.e, significantly reduce k).

Another approach, and that which we will adopt here, is to use marginal composite likelihood for the input parameter estimation, in that this can be done in a way that guarantees accurate parameter estimation (at least asymptotically). The idea of composite likelihood can be traced back to pseudo-likelihood (Besag (1974)) and partial likelihood (Cox (1975)). It has been studied intensively in recent years, (see e.g. Lindsay et al. (2011) and Varin et al. (2011) for recent developments).

We will utilize the Independent Marginal Composite Likelihood (ICML) approach, replacing $\mathcal{L}(\mathbf{y}^{\mathscr{D}} \mid \boldsymbol{\gamma}, \eta)$ in (2.21) by the following product of sub-likelihoods:

$$\mathcal{L}_C(\mathbf{y}^{\mathscr{D}} \mid \boldsymbol{\gamma}, \eta) = \prod_{t=1}^{T} \left\{ \mathcal{L}_t(\mathbf{y}_t^{\mathscr{D}} \mid \boldsymbol{\gamma}, \eta) \right\} , \tag{2.25}$$

where the $\mathcal{L}_t(\mathbf{y}_t^{\mathscr{D}} \mid \boldsymbol{\gamma}, \eta)$ are "parts" of the full likelihood, formed from batches of sub rows of the $n \times k$ matrix $\mathbf{y}^{\mathscr{D}}$, and it is assumed that there is no correlation between the different batches. Specifically, we form $T = n/n_0$ batches of the design inputs, each batch being of size n_0, by simple random sampling of the inputs. Imposing independence of the batches results in the correlation matrix over the input space

$$\tilde{\mathbf{R}}_C = \begin{pmatrix} \tilde{\mathbf{R}}_1 & & & \\ & \tilde{\mathbf{R}}_2 & & \\ & & \ddots & \\ & & & \tilde{\mathbf{R}}_T \end{pmatrix},$$

where each $\tilde{\mathbf{R}}_t$, $1 \leqslant t \leqslant T$ is a batch with m inputs and the other elements of the correlation matrix are 0. The composite marginal posterior for the parameters $(\boldsymbol{\gamma}, \eta)$ is then

$$p_C(\boldsymbol{\gamma}, \eta \mid \mathbf{y}^{\mathscr{D}}) \propto \mathcal{L}_C(\mathbf{y}^{\mathscr{D}} \mid \boldsymbol{\gamma}, \eta) \pi^{\tilde{R}}(\boldsymbol{\gamma}, \eta)\,.$$

Defining $(\hat{\boldsymbol{\gamma}}_C, \hat{\eta}_C)$ as the composite maximum likelihood estimator, under the regular conditions (Lindsay (1988); Severini (2000)), we have that, as $n \to \infty$,

$$\sqrt{n}[(\hat{\boldsymbol{\gamma}}_C, \hat{\eta}_C) - (\boldsymbol{\gamma}, \eta)] \xrightarrow{d} \mathcal{N}(0, \mathbf{G}^{-1})\,,$$

where

$$\mathbf{G} = \mathbf{G}(\boldsymbol{\gamma}, \eta) = \mathbf{H}(\boldsymbol{\gamma}, \eta)\mathbf{J}^{-1}(\boldsymbol{\gamma}, \eta)\mathbf{H}(\boldsymbol{\gamma}, \eta)\,,$$

$$\mathbf{H}(\boldsymbol{\gamma}, \eta) = -\,\mathrm{E}\Big(\frac{\partial^2 \mathcal{L}_C(\mathbf{y}^{\mathscr{D}} \mid \boldsymbol{\gamma}, \eta)}{\partial(\boldsymbol{\gamma}, \eta)^2}\Big)\,, \quad \mathbf{J}(\boldsymbol{\gamma}, \eta) = \mathrm{Var}\Big(\frac{\partial \mathcal{L}_C(\mathbf{y}^{\mathscr{D}} \mid \boldsymbol{\gamma}, \eta)}{\partial(\boldsymbol{\gamma}, \eta)}\Big)\,.$$

Because of these asymptotic results, the use of composite likelihood to estimate $(\boldsymbol{\gamma}, \eta)$ is reasonable when n is large.

In choosing n_0, we make use of the 'folklore' notion that the number of design points necessary to effectively estimate p correlation parameters is $10p$. For TITAN2D, there are either 4 correlation parameters or 3 with a nugget, so n_0 should be at least 40. We then utilize $n_0 = 50$ to form each batch to see the performance.

Table 4 presents the MSE in prediction for three different ways of estimating the range and nugget parameters. The first two use the ICML approach with 4 blocks each and with $n = 50$; the first one utilizes averages of 4 blocks for prediction, while the second utilizes the full $n \times n$ matrix $\tilde{\mathbf{R}}$ for prediction, as it only requires one inversion. The third method is the full PP GaSP, using the full correlation matrix to do both estimation and prediction. Clearly using the full correlation matrix for prediction is much better than merely using blocks for the prediction. Using the full likelihood for estimation of the range and nugget parameters is slightly better than

Table 2.4: The MSE and computational time in seconds using R at the non-crater area and the small flow area based on $n = 200$ inputs. The first column uses ICML with block size $n_0 = 50$ to do estimation of the range and nugget parameters, and also uses composite likelihood to do prediction. The second column uses composite likelihood to do the parameter estimation, but uses the full likelihood for prediction. The third column shows the results for the full PP GaSP. The number of held-out runs for the evaluation is $n^* = 483$.

MSE (time)	PP GaSP ICML block est, block pred,	PP GaSP ICML block est, full pred	PP GaSP full lik full pred
non-crater area	0.088 (103.6s)	0.063 (111.9s)	0.062 (534.4s)
small flow area	0.050 (113.4s)	0.034 (133.7s)	0.033 (573.5s)

using the ICML approach, but the difference is modest. And the second method needs only $O(tmnk + tm^2n)$ flops, as compared to $O(tn^2k + tn^3)$ flops for the full PP GaSP method, so the ICML approach can be very attractive.

3

Robust Gaussian Process Emulation

In this chapter, We consider estimation of the parameters of a Gaussian Stochastic Process (GaSP), in the context of emulation (approximation) of computer models. Objective priors for these parameters, such as the Jeffreys-rule prior and reference prior, have been studied in the literature for the situations of an isotropic covariance function for the GaSP or separability in the design of model runs used in the GaSP construction. In this paper, we consider a general multi-dimensional design setting (e.g., a Latin Hypercude Design (LHD)) with commonly used anisotropic covariance functions for the GaSP emulator, and discuss properties of the objective priors and marginal likelihoods for the parameters of the GaSP. Propriety of the resulting posterior of the GaSP parameters is established, but the main focus is on estimation of the GaSP parameters through various generalized maximum likelihood methods, mostly involving finding posterior modes; this is because full Bayesian analysis in problems involving computer model emulation is typically prohibitively expensive. With the poorly behaved likelihood, finding posterior modes is a difficult estimation problem, so the study of the robustness of the estimators is crucial. We demonstrate that certain parameterizations result in more robust estimators than others, and that

some parameterizations which are in common use should clearly be avoided. These results are applicable to many frequently used covariance functions, e.g., power exponential, Matérn, rational quadratic and spherical covariance; we also generalize the results to GaSP with a nugget parameter. Both theoretical and numerical evidence is presented concerning the performance of the studied procedures.

3.1 Literature Review and Motivations

A GaSP model typically consists of mean parameters, a variance parameter (if stationarity is assumed), and the parameters in the correlation functions, such as range and roughness parameters (introduced in more detail in the next section). Although the mean parameters and variance parameter are relatively easy to deal with, estimation of parameters in the correlation functions is "notoriously" difficult (Kennedy and O'Hagan (2001)). For instance, maximum likelihood estimation (MLE) of these parameters has been widely recognized to be unstable Li and Sudjianto (2005); Lopes (2011). This instability is partially caused by needing the inverse of covariance matrices that are often close to singular, when evaluating the likelihood. This can often be overcome by adding a nugget to stabilize the computation, but studies have found that the features of the emulator can significantly change when a nugget is added Andrianakis and Challenor (2012). Another difficulty that will be discussed herein is that serious problems can arise when the covariance matrix is estimated to be near-diagonal, and this can easily happen when a product correlation structure is used because, if even when one of the terms in the product is close to zero, the correlation will be close to zero. Two R packages, DiceKriging and DiceOptim use several different ways to avoid the unstable results, such as using expected improvement criteria and lower and upper bounds for range parameters Roustant et al. (2012). Although those methods can yield stabilized computations, they produce larger predictive errors, as shown in Section 3.5.

In this work, we seek to stabilize the computation, without significantly affecting emulator predictive accuracy, by utilizing formal objective prior distributions (namely reference priors) and then finding posterior modes (for the correlation parameters). The first use of reference priors in modeling spatially correlated data was Berger et al. (2001); this paper was restricted to consideration of an isotropic covariance function, with only one range/scale parameter. Reference priors for an anisotropic process were introduced in Paulo (2005), and there properties were studied in the context of product correlation functions and separable designs (e.g., a lattice) for the input values over which the computer model is run. These results were extended in Ren et al. (2013), although still under the assumption of separability of the design. Most designs used for creating emulators of computer models – such as the very popular Latin Hypercube Design (LHD) – are, however, non-separable, and so we need to extend the analysis of the reference priors (and likelihoods) to cover non-separable situations and to include the possibility of a nugget parameter. (Objective priors for isotropic GaSPs with a nugget were discussed recently in Ren et al. (2012) and Kazianka and Pilz (2012).)

Posterior modes of the correlation parameters depend on the parameterization used for the correlation parameters and it was surprisingly found that this choice of parameterization can make a major difference in the "robustness" of the posterior mode Bayarri et al. (2009, 2015). The word "robust" in this context was first used in Spiller et al. (2014) and will be formally defined in Section 3.3 but, informally, a robust procedure avoids the numerical issues discussed above while producing an emulator with good predictive performance. In this investigation it was also found that robustness is considerably more difficult to obtain for the anisotropic case with product correlation function than for the isotropic case. As an example, the posterior density of the range parameters goes to infinity when the correlation matrix, for a product correlation function, is close to a matrix of ones, under one frequently

used parameterization, while this does not happen in the isotropic case. One of the major contributions of this work is in making the study of robustness of the parameterization rigorous by determining such "tail rates" of posterior distributions for the various parameterizations.

3.2 Parameter estimation in Gaussian Stochastic Processes

3.2.1 Background and correlation function

We assume stationary GaSP model, defined in Equation (1.1), with the mean structure $\mu(\cdot)$ defined in Equation (1.3) and product correlation defined in Equation (1.4) in Chapter 1. The computer model is run at n design points $\mathbf{x}^{\mathscr{D}} = (\mathbf{x}_1^{\mathscr{D}}, ..., \mathbf{x}_n^{\mathscr{D}})$ and the outputs are denoted as $\mathbf{y}^{\mathscr{D}} = (\mathbf{y}_1^{\mathscr{D}}, ..., \mathbf{y}_n^{\mathscr{D}})$.

Some frequently chosen correlation functions are listed in Table 3.1 (dropping the subscript l). The correlation functions $c_l(\cdot, \cdot)$ typically have a range parameter $\gamma_l > 0$ and a roughness parameter $\alpha_l > 0$. As mentioned earlier, the points in $\mathbf{x}^{\mathscr{D}}$ are typically chosen as far apart as possible, in order to sample the computer model output at as many diverse points as possible. Consequently, the roughness parameters α_l, $1 \leqslant l \leqslant p$, are not highly influential and typically have quite flat likelihood surfaces. They also are typically highly confounded with the γ_l and σ^2, causing computational and inferential difficulties if left in the model (Zhang (2004); Gelfand et al. (2010)). It is thus common (and herein adopted) to fix the roughness parameters as certain prespecified values and focus only on estimation of the range parameters.

One of most frequently used correlation functions is the Gaussian correlation, which is the special case of $\alpha_l = 2$ in the power exponential correlation function. The sample paths of the resulting GaSP process are infinitely differentiable, which is sometimes desirable in applications. However, the choice of $\alpha_l = 2$ has been criticized since it often yields too smooth sample paths for many applications (Stein (2012))

Table 3.1: Popular choices of correlation functions, when $c_l(x_{il}, x_{jl}) \equiv c(d)$, with $d = |x_{il} - x_{jl}|$. Here α is the roughness parameter, γ is the range parameter, $\Gamma(\cdot)$ is the gamma function and $\mathcal{K}_\alpha(\cdot)$ is the modified Bessel function of second kind of order α. $\nu(\gamma)$ and $\omega(\gamma)$ are terms in the Taylor expansion of the correlation functions, as $\gamma \to \infty$, that will be needed later.

	$c(d)$	$\nu(\gamma)$	$\omega(\gamma)$
Power Exponential	$\exp\{-(d/\gamma)^\alpha\}$, $\alpha \in (0, 2]$	$\gamma^{-\alpha}$	$\gamma^{-\alpha}$
Spherical	$\left(1 - \frac{3}{2}\left(\frac{d}{\gamma}\right) + \frac{1}{2}\left(\frac{d}{\gamma}\right)^3\right)\mathbf{1}_{[d/\gamma \leqslant 1]}$	γ^{-1}	γ^{-2}
Rational Quadratic	$\left(1 + \left(\frac{d}{\gamma}\right)^2\right)^{-\alpha}$, $\alpha \in (0, +\infty)$	γ^{-2}	γ^{-2}
Matérn	$\frac{1}{2^{\alpha-1}\Gamma(\alpha)}\left(\frac{d}{\gamma}\right)^\alpha \mathcal{K}_\alpha\left(\frac{d}{\gamma}\right)$, $0 < \alpha < 1$	$\gamma^{-2\alpha}$	$\gamma^{-2+2\alpha}$
	$\alpha = 1$	$\frac{\log(\gamma)}{\gamma^2}$	$\frac{1}{\log(\gamma)}$
	$1 < \alpha < 2$	γ^{-2}	$\gamma^{2-2\alpha}$
	$\alpha = 2$	γ^{-2}	$\frac{\log(\gamma)}{\gamma^2}$
	$\alpha > 2$	γ^{-2}	γ^{-2}

and because computational difficulties can arise with this choice (see Appendix B.1). Thus $1 < \alpha_l < 2$ is typically chosen in the power exponential family (Bayarri et al. (2009)), although the process is then not even once differentiable, sometimes not ideal for applications.

Another popular choice of the correlation function is the Matérn class. When $\alpha_l = (2k + 1)/2$ for $k \in \mathbb{N}$, the Matérn correlation has a closed form expression. For example, when $\alpha_l = 1/2$, the Matérn correlation reduces to the power exponential correlation with $\alpha_l = 1$. And when $\alpha_l \to \infty$, it reduces to Gaussian correlation. One nice feature of Matérn correlation is that its sample path are $\lfloor \alpha_l - 1 \rfloor$ times differentiable, so the smoothness of the process can be directly controlled by the roughness parameters. Hence, it has become the recommended choice for the correlation function in spatial modeling (Stein (2012)). One of the most frequently used Matérn

correlation functions is $\alpha_l = 5/2$, which has the form

$$c_l(d_l) = \left(1 + \frac{\sqrt{5}d_l}{\gamma_l} + \frac{5d_l^2}{3\gamma_l^2}\right) \exp\left(-\frac{\sqrt{5}d_l}{\gamma_l}\right), \tag{3.1}$$

where d_l stands for any of the $|x_{il} - x_{jl}|$.

Use of Matérn correlation functions has been less popular in the computer model emulation literature. Here is an argument as to why (5.12) should be seriously considered for emulation, noting first that it is computationally very tractable. Denoting $\tilde{d}_l = d_l/\gamma_l$, the following are easy to establish for (5.12).

- When $\tilde{d}_l \to 0$, $c_l(\tilde{d}_l) \approx 1 - C\tilde{d}_l^2$ with $C > 0$ being a constant. This thus behaves similarly to $\exp(-\tilde{d}_l^2) \approx 1 - \tilde{d}_l^2/2$, which corresponds to the power exponential correlation with $\alpha_l = 2$ (i.e., Gaussian correlation). This suggests that the Matérn correlation in (5.12) will maintain the smoothness induced by Gaussian correlation for nearby inputs.

- When $\tilde{d}_l \to \infty$, the dominant part of $c_l(\tilde{d}_l)$ is $\exp\left(-\sqrt{5}\tilde{d}_l\right)$ which matches the power exponential correlation with $\alpha_l = 1$. Thus the Matérn correlation in (5.12) prevents the correlation from decreasing quickly with distance, as does the Gaussian correlation. This can be of benefit in computer model emulation since some inputs may have almost no effect on the computer model, which would correspond to near constant correlations even for distant inputs.

We have also found that the Matérn correlation function with $\alpha_l = 5/2$ yields very good empirical results compared with other correlation functions in emulation. It is also the default correlation function in the DiceKriging package (Roustant et al. (2012)). For these reasons, it will be used as the default correlation function for the numerical study in Section 3.5. (However, the following theoretical results are applicable to the much larger class of correlation functions listed in Table 3.1, as shown in Section 3.3.)

3.2.2 Marginal likelihood and marginal posterior

Marginal likelihood

Although maximum likelihood estimation of all parameters of the covariance function is possible, it has become standard to treat the mean parameters and variance in a fully objective Bayesian fashion, since they can be dealt with in closed form in the Bayesian computations. Thus these parameters are assigned the objective prior

$$\pi(\boldsymbol{\theta}, \sigma^2) \propto \frac{1}{(\sigma^2)^a},$$

with a fixed $a > 0$, where $a = 1$ corresponds to the standard reference prior. (It has become customary to also compare results with other choices of a, so we allow that in what follows.)

Using this prior to marginalize out the mean and variance parameters in the likelihood function, we obtain the marginal likelihood

$$\mathcal{L}(\mathbf{y}^{\mathscr{D}}|\boldsymbol{\gamma}) \propto |\mathbf{R}|^{-\frac{1}{2}} |\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})|^{-\frac{1}{2}} \left(S^2\right)^{-(\frac{n-q}{2}+a-1)}, \tag{3.2}$$

where $\mathbf{h}(\mathbf{x}^{\mathscr{D}})$ is the $n \times q$ basis matrix with (i, j) term $h_j(\mathbf{x}_i^{\mathscr{D}})$; $S^2 = (\mathbf{y}^{\mathscr{D}})^T \mathbf{Q} \mathbf{y}^{\mathscr{D}}$ with $\mathbf{Q} = \mathbf{R}^{-1}\mathbf{P}_{\boldsymbol{R}}$ and $\mathbf{P}_{\boldsymbol{R}} = \mathbf{I}_n - \mathbf{h}(\mathbf{x}^{\mathscr{D}})\{\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})\}^{-1}\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}$, with $\mathbf{I}_n$ being the identity matrix of size n.

Assuming the roughness parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ have been pre-specified, the range parameters of the correlation function can now be estimated by maximizing the marginal likelihood (we denote the resulting estimators the MMLE estimators) in (3.2). While this approach was argued in Bayarri et al. (2009) to be superior to overall maximum likelihood estimation, we will see that it is still non-robust, in the sense that will be defined in Section 3. (The main problem is that the marginal likelihood will often not go to zero in the tails and, indeed, can be increasing.) Thus

it was argued in Lopes (2011) and Spiller et al. (2014) that the marginal likelihood needs to be augmented by the reference prior for the range parameters.

Reference prior and posterior

The reference prior for a separable product correlation function was developed in Paulo (2005) and is given by

$$\pi^R(\boldsymbol{\theta}, \sigma^2, \boldsymbol{\gamma}) \propto \frac{\pi^R(\boldsymbol{\gamma})}{(\sigma^2)^a}\,, \tag{3.3}$$

with $\pi^R(\boldsymbol{\gamma}) \propto |\mathbf{I}^*(\boldsymbol{\gamma})|^{1/2}$, where $\mathbf{I}^*(\cdot)$ is the expected fisher information matrix as below,

$$\mathbf{I}^*(\boldsymbol{\gamma}) = \begin{pmatrix} n-q & tr(\mathbf{W}_1) & tr(\mathbf{W}_2) & \dots & tr(\mathbf{W}_p) \\ & tr(\mathbf{W}_1^2) & tr(\mathbf{W}_1\mathbf{W}_2) & \dots & tr(\mathbf{W}_1\mathbf{W}_p) \\ & & tr(\mathbf{W}_2^2) & \dots & tr(\mathbf{W}_2\mathbf{W}_p) \\ & & & \ddots & \vdots \\ & & & & tr(\mathbf{W}_p^2) \end{pmatrix}_{(p+1)\times(p+1)}, \tag{3.4}$$

where $\mathbf{W}_l = \dot{\mathbf{R}}_l\mathbf{Q}$, for $1 \leqslant l \leqslant p$, $\dot{\mathbf{R}}_l$ is the derivative of the correlation matrix $\mathbf{R}$ with respect to the l^{th} range parameter.

The marginal posterior of $\boldsymbol{\gamma}$ with regard to this reference prior is

$$p(\boldsymbol{\gamma}|\mathbf{y}^{\mathscr{D}}) \propto \mathcal{L}(\mathbf{y}^{\mathscr{D}}|\boldsymbol{\gamma})|\mathbf{I}^*(\boldsymbol{\gamma})|^{1/2}. \tag{3.5}$$

Sampling from this posterior requires a Metropolis-type algorithm and each evaluation of the likelihood typically requires $O(n^3)$ flops for the inverse of the correlation matrix, which is computationally prohibitive for many applications. Moreover, the computation error can be very large when the correlation matrix is close to the matrix of all ones. Because of these computational reasons, it is common (Bayarri et al. (2009); Spiller et al. (2014)) to instead simply estimate $\boldsymbol{\gamma}$ by its marginal posterior mode, using (3.5) ,

$$(\hat{\gamma}_1, \dots \hat{\gamma}_p) = \underset{\gamma_1,\dots,\gamma_p}{argmax}(\mathcal{L}(\mathbf{y}^{\mathscr{D}}|\gamma_1, \dots, \gamma_p)\pi^R(\gamma_1, \dots, \gamma_p)). \tag{3.6}$$

Parameterizations

Maximum likelihood estimation is invariant under the choice of parameterization, but the posterior mode is not invariant because of the presence of the Jacobian for the prior. Quite surprisingly, we have found that the choice of parameterization often makes a big difference in the robustness of the posterior mode. Here are three common ways of parameterizing the range parameters in power exponential correlation function (Berger et al. (2001); Paulo (2005); Bayarri et al. (2007a, 2009); Spiller et al. (2014)):

$$c_{\gamma_l}(|x_{il} - x_{jl}|) = \exp\{-(|x_{il} - x_{jl}|/\gamma_l)^{\alpha_l}\}, \tag{3.7}$$

$$c_{\tilde{\beta}_l}(|x_{il} - x_{jl}|) = \exp\{-\tilde{\beta}_l|x_{il} - x_{jl}|^{\alpha_l}\}, \tag{3.8}$$

$$c_{\tilde{\xi}_l}(|x_{il} - x_{jl}|) = \exp\left\{-\exp(\tilde{\xi}_l)|x_{il} - x_{jl}|^{\alpha_l}\right\}, \tag{3.9}$$

for any $l = 1, \cdots, p$.

Table 3.1 gives various correlation functions in their natural parameterizations, in which the range parameter and roughness parameter are independent; we will call this the α-*free* parameterization of the range parameter. In contrast, in the above parameterizations of the power exponential correlation function, $\tilde{\beta}_l = \gamma_l^{-\alpha_l}$ and $\tilde{\xi}_l = \log(\gamma_l^{-\alpha_l})$ both depend on α_l. We will also consider the following transformations of the α-*free* parameterization (dropping the subscript l for convenience).

Definition 3.2.1. *For the range parameters γ in Table 3.1,*

(i) $\beta = 1/\gamma$ will be called the inverse range parameter;

(ii) $\xi = log(1/\gamma)$ will be called the log inverse range parameter.

Note that $\tilde{\beta} = \beta^{\alpha}$ and $\tilde{\xi} = \alpha\xi$. The mode of the posterior distributions for the $\tilde{\xi}$ and ξ parameterizations will be the same (properly transformed), because the Jacobians of the transformations differ only by the prefixed constant α; thus we need

to consider only the ξ — and not the $\tilde{\xi}$ — parameterization of the power exponential correlation function in what follows. On the other hand, the posterior modes of $\tilde{\beta}$ and β are not the same (when transformed), so we will have to consider both parameterizations in what follows.

3.2.3 Profile likelihood

For comparison purposes, we will also consider the full likelihood approach, which also utilizes maximum likelihood estimators (MLE) for the mean and variance parameters, $\hat{\boldsymbol{\theta}}_{MLE} = \hat{\boldsymbol{\theta}}$, $\hat{\sigma}^2_{MLE} = (n-q)\hat{\sigma}^2/n$, where $\hat{\boldsymbol{\theta}}$ and $\hat{\sigma}^2$ are defined in (1.6). Plugging $\hat{\boldsymbol{\theta}}_{MLE}$ and $\hat{\sigma}^2_{MLE}$ into (1.2) and ignoring the normalizing constant, the likelihood reduces to the profile likelihood

$$\mathcal{L}(\mathbf{y}^{\mathscr{D}}|\hat{\sigma}^2_{MLE}, \hat{\boldsymbol{\theta}}_{MLE}, \boldsymbol{\gamma}) \propto |\mathbf{R}|^{-\frac{1}{2}} (S^2)^{-\frac{n}{2}} . \tag{3.10}$$

where S^2 is defined in Equation (3.2). To complete the MLE analysis, $\boldsymbol{\gamma}$ is estimated by the mode of this profile likelihood.

The predictive distribution of a new input $\mathbf{x}^*$, conditional on the previous outputs and the MLE, are

$$y(\mathbf{x}^*)|\mathbf{y}^{\mathscr{D}}, \hat{\sigma}^2_{MLE}, \hat{\boldsymbol{\theta}}_{MLE}, \hat{\boldsymbol{\gamma}}_{MLE} \sim N(\hat{y}_{MLE}(\mathbf{x}^*), \hat{\sigma}^2_{MLE} c^*_{MLE}), \tag{3.11}$$

where $\hat{y}_{MLE}(\mathbf{x}^*) = \hat{y}(\mathbf{x}^*)$, with $\hat{y}(\mathbf{x}^*)$ defined in (1.6) and $c^*_{MLE} = c(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{r}^T(\mathbf{x}^*)\hat{\mathbf{R}}^{-1}_{MLE}\mathbf{r}(\mathbf{x}^*)$ with $\hat{\mathbf{R}}_{MLE}$ being the correlation matrix by plugging in $\hat{\boldsymbol{\gamma}}_{MLE}$.

The profile likelihood sometimes is very flat in the tails, resulting in $\hat{\boldsymbol{\gamma}}$ being near zero and $\hat{\mathbf{R}}_{MLE}$ being near $\mathbf{I}_n$ (see the details in Section 3.3). This can be shown to result in the predicted mean $\hat{y}_{MLE}(\mathbf{x}^*)$ being essentially an impulse function at each of the observations, while following the GaSP mean elsewhere. Figure 3.1 gives an example of this scenario, where the GaSP mean is assumed to be a constant. In the left panel of the figure, the roughness parameter was $\alpha = 1$ for a power exponential

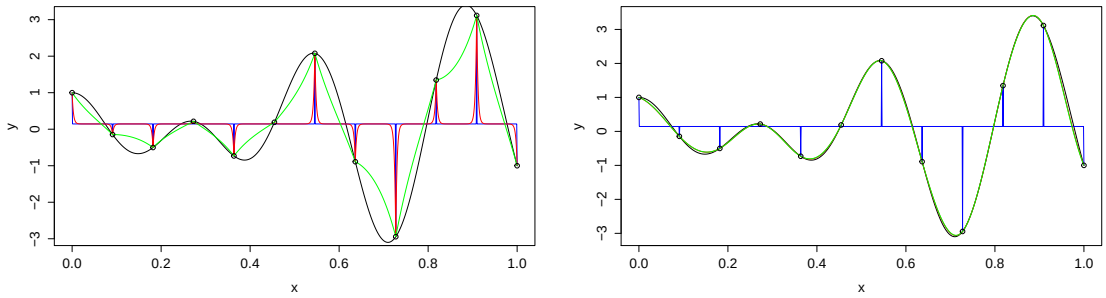

FIGURE 3.1: Emulation of the function $y = 3sin(5\pi x)x + cos(7\pi x)$, graphed as the black solid curves (overlapping the green curves in the right panel). The design for the input x is equally spaced from $[0, 1]$ with $n = 12$, with the resulting function values indicated by the black circles. A constant GaSP mean is used. The left panel is for $\alpha = 1$ and the right panel for $\alpha = 1.9$, for the power exponential correlation function. The blue curves (which are essentially unit impulse functions at the observations and constant elsewhere) give the emulator mean obtained from the MLE/profile likelihood approach; the red curves give the emulator mean from the MMLE approach; and the green curves give the emulator mean arising from the maximum posterior mode approach with the reference prior. The red curves are overlapping with the green curves in the right panel.

correlation function, and both the MLE and MMLE became essentially degenerate, while the prediction from the posterior mode approach was reasonable (although not quite smooth enough). In the right panel of the figure, the roughness parameter was $\alpha = 1.9$; here both the MMLE and marginal posterior mode approach gave excellent predictions, but the profile likelihood approach still resulted in a degenerate prediction. Such degeneracies are somewhat unusual in one-dimension, but are not particularly unusual with higher dimensional inputs, as shown numerically shown in Section 3.5.

3.3 Robust parameter estimation for GaSP Models

In this section, we explore the ways in which GaSP emulator construction can fail, developing the "robustness criteria" that are needed to avoid such failures. We then examine which estimation methods satisfy the criteria. To begin, it is pedagogically

useful to look at a special case, where the analysis is essentially closed form.

3.3.1 A closed form example for the profile likelihood and marginal likelihood

Suppose the input is one-dimensional and that the design is equally spaced with the design points being d_0 units apart. Consider a constant mean $h(x) = 1$ and power exponential correlation with roughness parameter $\alpha = 1$. Denote $\rho = e^{-d_0/\gamma}$, write $c(x_i, x_j) = \rho^{\Delta_{ij}}$, with $\Delta_{ij} = |x_i - x_j|/d_0$, and write $y(x_i^{\mathscr{D}})$ as y_i to simplify the notation. The logarithm of the profile likelihood can then be given in closed form as follows,

$$\begin{aligned}
&\log\mathcal{L}(\mathbf{y}^{\mathscr{D}}|\hat{\sigma}^2_{MLE}, \hat{\boldsymbol{\theta}}_{MLE}, \rho) \\
&\propto \frac{1}{2}\log(1-\rho^2) - \frac{n}{2}\log\left\{\sum_{i=1}^{n} y_i^2 - 2\rho\left(\sum_{i=1}^{n-1} y_i y_{i+1}\right) + \rho^2\left(\sum_{i=2}^{n-1} y_i^2\right)\right. \\
&\left. -\frac{1-\rho}{n-(n-2)\rho}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j - 2\rho\left(\sum_{i=1}^{n}\sum_{j=2}^{n-1} y_i y_j\right) + \rho^2\left(\sum_{i=2}^{n-1}\sum_{j=2}^{n-1} y_i y_j\right)\right]\right\}.
\end{aligned} \tag{3.12}$$

The proof is given in Appendix B.2. The logarithm of the marginal likelihood (obtained by integrating out the mean and variance parameters using the standard reference prior) can be written as (from Lopes (2011), with some typos corrected)

$$\begin{aligned}
&\log\mathcal{L}(\mathbf{y}^{\mathscr{D}}|\rho) \\
&\propto -\frac{1}{2}\log\left(\frac{n-(n-2)\rho}{1+\rho}\right) - \frac{n-1}{2}\log\left\{\sum_{i=1}^{n} y_i^2 - 2\rho\left(\sum_{i=1}^{n-1} y_i y_{i+1}\right) + \rho^2\left(\sum_{i=2}^{n-1} y_i^2\right)\right. \\
&\left. -\frac{1-\rho}{n-(n-2)\rho}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j - 2\rho\left(\sum_{i=1}^{n}\sum_{j=2}^{n-1} y_i y_j\right) + \rho^2\left(\sum_{i=2}^{n-1}\sum_{j=2}^{n-1} y_i y_j\right)\right]\right\}.
\end{aligned} \tag{3.13}$$

As $\rho \to 0$ and $\rho \to 1$, the limiting values for the log profile likelihood and log marginal likelihood are given in Table 3.2. If the mode for either log likelihood is at

Table 3.2: The tail behaviors of the log-marginal likelihood and log-profile likelihood.

	when $\rho \to 0$	when $\rho \to 1$
$\log\mathcal{L}(\mathbf{y}^{\mathscr{D}}\|\rho)$	$-\frac{n-1}{2}\log\{\sum_{i=1}^{n}(y_i-\bar{y})^2\} - \frac{1}{2}\log(n)$	$-\frac{n-1}{2}\log\{\sum_{i=1}^{n-1}(y_{i+1}-y_i)^2\}$
$\log\mathcal{L}(\mathbf{y}^{\mathscr{D}}\|\hat{\sigma}^2_{MLE},\hat{\boldsymbol{\theta}}_{MLE},\rho)$	$-\frac{n}{2}\log\{\sum_{i=1}^{n}(y_i-\bar{y})^2\}$	$-\infty$

(or near) zero, the emulator will degenerate, as in Figure 3.1. From Table 3.2, it is clear that the log likelihoods will not typically go to zero as $\rho \to 0$. Indeed, we can find a condition under which the mode of the log-profile likelihood occurs at zero as $\rho \to 0$, stated in Lemma 3.3.1.

Lemma 3.3.1. *A necessary and sufficient condition that the mode of the profile likelihood in (3.12) is at* $\rho = 0$ [causing the unwelcome degeneracy]*, is*

$$\sum_{i=1}^{n-1}(y_i-\bar{y})(y_{i+1}-\bar{y}) \leqslant 0, \tag{3.14}$$

where $\bar{y} = \sum_{i=1}^{n} y_i$.

We postpone the proof of Lemma 3.3.1 until Appendix B.2. The intuition behind Lemma 3.3.1 comes from the fact that in this case, the GaSP becomes an autoregressive model of order 1. When the empirical lag-1 autocorrelation is less than zero, the profile likelihood estimate of the correlation ρ will be zero, since the correlation ρ is parameterized to be nonnegative here.

If either likelihood is maximized as $\rho \to 1$, then $\mathbf{R} \to \mathbf{1}_n\mathbf{1}_n^T$, where $\mathbf{1}_n$ is the vector of all ones, so that the correlation matrix becomes ill-conditioned, causing large approximation errors in computation of its inverse. Table 3.2 shows that the profile likelihood goes to zero as $\rho \to 1$, so that this problem does not arise for the profile likelihood. But it can arise for the marginal likelihood, as we have seen in

examples (although we do not have a theoretical condition under which the mode will converge to this solution).

For the general case considered in the remainder of the paper, explicit results such as that in Lemma 1 are not available. However, we can still look at the tail rates (corresponding to ρ going to 0 or 1) for the various likelihoods and posteriors and assess when problems will occur. We formalize these notions in the next section, through our criteria for robust estimation.

3.3.2 Robust estimation

As discussed in the previous section, when $\mathbf{R} \approx \mathbf{I}_n$, the GaSP predictive mean will degenerate to the fitted mean and impulse functions at the observed inputs, as happened in Figure 3.1. When $\mathbf{R} \approx \mathbf{1}_n\mathbf{1}_n^T$, the correlation matrix $\mathbf{R}$ is almost singular, leading to very large computational errors in the GaSP predictive mean. Robust estimation of the correlation parameters is defined as avoidance of these two possible problems.

Definition 3.3.1. *(Robust Estimation.) Estimation of the parameters in the GaSP is called robust, if the following two situations do NOT happen:*

(i) $\hat{\mathbf{R}} = \mathbf{1}_n\mathbf{1}_n^T$,

(ii) $\hat{\mathbf{R}} = \mathbf{I}_n$,

where $\hat{\mathbf{R}}$ *is the estimated correlation matrix.*

Note that the GaSP predictive mean in (1.7) is not well-defined in these two situations when the inputs are at one of the design points, but it can be defined as the limit as $\hat{\mathbf{R}} \to \mathbf{1}_n\mathbf{1}_n^T$, and $\hat{\mathbf{R}} \to \mathbf{I}_n$.

The following basic lemma is immediate from the definition of the correlation matrix.

Lemma 3.3.2. *Robustness is lacking in either of the following two cases.*

Case 1. If for all $1 \leqslant l \leqslant p$, $\hat{\gamma}_l = \infty$ *(or* $\hat{\xi}_l = -\infty$ *or* $\hat{\beta}_l = 0$ *in the other parameterizations), then* $\hat{\mathbf{R}} = \mathbf{1}_n \mathbf{1}_n^T$.

Case 2. If any $\hat{\gamma}_l = 0$ *(equivalent to* $\hat{\xi}_l = \infty$ *or* $\hat{\beta}_l = \infty$*), then* $\hat{\mathbf{R}} = \mathbf{I}_n$.

Note that it is generally fine if some (but not all) of the estimated γ_l are close to ∞, because this will just make $\hat{\mathbf{R}}_l \approx \mathbf{1}_n \mathbf{1}_n^T$ for some l but not $\hat{\mathbf{R}} \approx \mathbf{1}_n \mathbf{1}_n^T$. In such a situation, the inputs associated with the large γ_l can be called *inert* inputs, since they will have only a small effect on the outputs. Indeed, this is a desirable situation, since such inputs could be removed from the emulator, simplifying and improving the approximation. The identifiability of these inert inputs motivates for a new class of prior, discussed in detail in Chapter 4.

The MLE, MMLE and marginal posterior modes (for the various parameterizations) all reduce to mode estimation with regard to a function $G(\boldsymbol{\gamma})$. Thus the following guarantees that one of the problem situations cannot occur.

Corollary 3.3.1. *Estimation of* $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$ *as the mode of a nonnegative function* $G(\boldsymbol{\gamma})$ *is robust if*

$$G(\boldsymbol{\gamma}) \to 0,$$

under the following two situations:

(i) $\forall l$, $1 \leqslant l \leqslant p$, $\gamma_l \to 0$,

(ii) $\gamma_l \to \infty$ *for all* $1 \leqslant l \leqslant p$.

Corollary 3.3.2. *Estimation of any monotonic transformation of the range parameters* $\boldsymbol{\zeta} = \mathbf{f}(\boldsymbol{\gamma}) = (f_1(\boldsymbol{\gamma}), \ldots, f_p(\boldsymbol{\gamma}))^T$*, by the mode of its marginal posterior, is robust if*

$$\mathcal{L}(\mathbf{y}|\mathbf{f}^{-1}(\boldsymbol{\zeta}))\pi^R(\mathbf{f}^{-1}(\boldsymbol{\zeta}))\left|\frac{\partial \mathbf{f}^{-1}(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}}\right| \to 0$$

under the following two situations:

(i) $\forall l,\ 1 \leqslant l \leqslant p,\ \mathbf{f}_l^{-1}(\boldsymbol{\zeta}) \to 0,$

(ii) $\mathbf{f}_l^{-1}(\boldsymbol{\zeta}) \to \infty$ *for all* $1 \leqslant l \leqslant p.$

where $\mathbf{f}^{-1}(\boldsymbol{\zeta}) = (\mathbf{f}_1^{-1}(\boldsymbol{\zeta}),\ ...,\ \mathbf{f}_p^{-1}(\boldsymbol{\zeta}))^T.$

3.3.3 Robustness Results

From the results in the previous section, it is clear that we should compute the tail rates, in terms of $\boldsymbol{\gamma}$, of the marginal likelihood, profile likelihood, and the various posteriors to see if they are robust. Computation of the tail rates of the posteriors requires computation of the tail rates of the reference prior, as well as the tail rates of the marginal likelihood. We need the following two mild assumptions (c.f., Berger et al. (2001); Ren et al. (2012)) to establish the main results concerning these rates. The assumptions hold for all the correlation functions listed in Table 3.1, for which the functions ν_l and ω_l below are also given.

Assumption 3.3.1. *For any* $d_l \geqslant 0$ *and* $1 \leqslant l \leqslant p,\ c_l(d_l) = c_l^0(d_l/\gamma_l)$*, where* $c_l^0(\cdot)$ *is a correlation function that satisfies* $\lim_{u\to\infty} c_l^0(u) = 0.$

Assumption 3.3.2. *For any* $1 \leqslant l \leqslant p,$ *as* $\gamma_l \to \infty,$

$$\mathbf{R}_l(\gamma_l) = \mathbf{1}_n\mathbf{1}_n^T + \nu_l(\gamma_l)\mathbf{D}_l + \nu_l(\gamma_l)\omega_l(\gamma_l)(\mathbf{D}_l^* + \mathbf{B}_l(\gamma_l)),$$

where $\mathbf{D}_l$ *is a nonsingular and symmetric matrix with* $\mathbf{1}_n^T\mathbf{D}_l^{-1}\mathbf{1}_n \neq 0,$ $\mathbf{D}_l^*$ *is a fixed matrix,* $\nu_l(\gamma_l) > 0$ *is a non-increasing and differentiable function,* $\omega_l(\gamma_l)$ *is a differentiable function, and* $\mathbf{B}_l(\gamma_l)$ *is a differentiable matrix (incorporating the higher order terms of the expansion), satisfying*

$$\nu_l(\gamma_l) \to 0, \quad \omega_l(\gamma_l) \to 0, \frac{w_l'(\gamma_l)}{\frac{\partial}{\partial\gamma_l}\log\ \nu_l(\gamma_l)} \to 0, \quad ||\mathbf{B}_l(\gamma_l)||_\infty \to 0, \quad \frac{||\frac{\partial}{\partial\gamma_l}\mathbf{B}_l(\gamma_l)||_\infty}{\frac{\partial}{\partial\gamma_l}\log(\omega_l(\gamma_l))} \to 0.$$

where $\omega_l'(\gamma_l) = \partial\omega_l(\gamma_l)/\partial\gamma_l$, $\|\boldsymbol{B}\|_\infty = \max_{i,j}|a_{ij}|$ *with* a_{ij} *being the* (i,j) *entry of a matrix* $\boldsymbol{B}$.

The first assumption ensures that the correlation function will decrease to zero as the distance between two points goes to infinity. The second assumption follows from a Taylor expansion of the correlation function as $\gamma_l \to \infty$. The following lemma, whose proof is given in Appendix B.3, gives the tail rates for the marginal and profile likelihoods.

Lemma 3.3.3. *(Tail rates of the marginal likelihood and profile likelihood.) If Assumption 1 and Assumption 2 hold for each of the* $\mathbf{R}_l$, $1 \leqslant l \leqslant p$, *the marginal likelihood and profile likelihood have the following tail rates.*

(i) If $\gamma_l \to 0^+$ *for any* l, $1 \leqslant l \leqslant p$, *the marginal likelihood and profile likelihood both exist and are greater than zero.*

(ii) If $\gamma_l \to \infty$ *for all* l, $1 \leqslant l \leqslant p$, *and* $\mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$ *denotes the column space of the mean basis matrix* $\mathbf{h}(\mathbf{x}^{\mathscr{D}})$, *the marginal likelihood satisfies*

$$\mathcal{L}(\mathbf{y}^{\mathscr{D}}|\boldsymbol{\gamma}) = \begin{cases} O\left((\sum_{l=1}^{p}\nu_l(\gamma_l))^{a-1/2}\right), & \mathbf{1}_n \notin \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}})), \\ O\left((\sum_{l=1}^{p}\nu_l(\gamma_l))^{a-1}\right), & \mathbf{1}_n \in \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))\,. \end{cases}$$

The profile likelihood, in this case, satisfies

$$\mathcal{L}(\mathbf{y}^{\mathscr{D}}|\boldsymbol{\gamma}, \hat{\sigma}^2_{MLE}, \hat{\boldsymbol{\theta}}_{MLE}) = O\left((\sum_{l=1}^{p}\nu_l(\gamma_l))^{1/2}\right).$$

Part (i) of this lemma indicates that the marginal likelihood and profile likelihood could have their modes at $\mathbf{R} = \mathbf{I}_n$ and thus could potentially be non-robust; one such case was given in Figure 3.1.

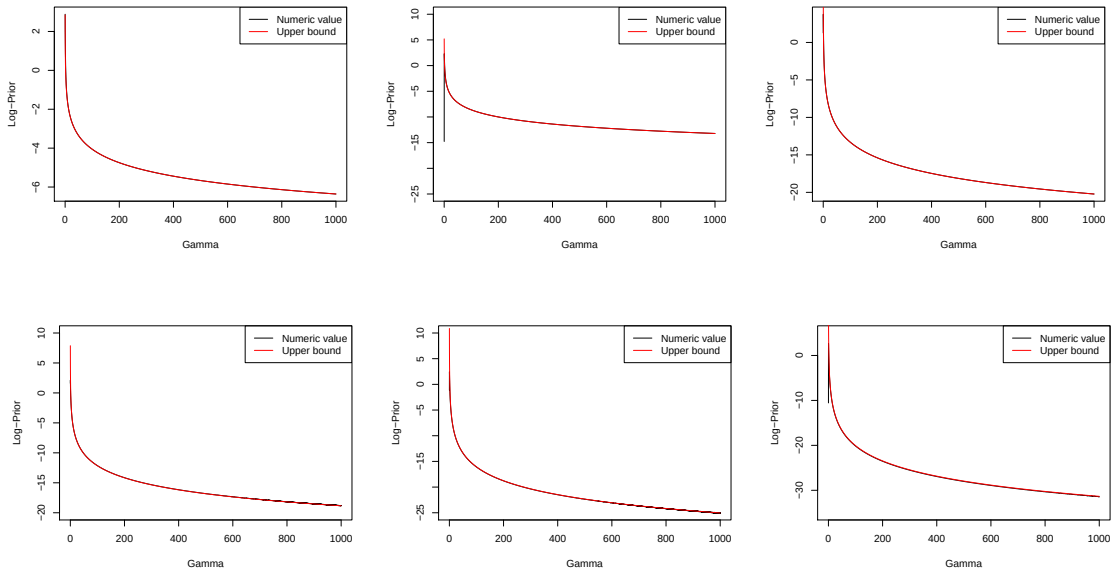

FIGURE 3.2: The tail behavior of the reference prior (black curves), and its upper bound (red curves) from Lemma 3.3.4 part (ii), when $\gamma_1 = \ldots = \gamma_p \to \infty$. The power exponential correlation function is used with fixed $\alpha_l = 1.9$, $1 \leqslant l \leqslant p$. The first row is for the case in which $\mathbf{1} \notin \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$, while the second row is for the case $\mathbf{1} \in \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}})$. From left to right, the dimension of the inputs are $p = 1$, $p = 2$ and $p = 3$. The prior and bounds are evaluated at points uniformly sampled from $\mathbb{R}^p$. The black curves and the red curves overlap when γ_l is large.

Part (ii) of the lemma shows that the mode of the marginal likelihood could be at $\mathbf{R} = \mathbf{1}_n\mathbf{1}_n^T$ for the frequently used setting of $a = 1$ and $\mathbf{1}_n \in \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$. On the other hand, the profile likelihood will decrease to zero at this limit, so it cannot be non-robust in this fashion. A byproduct of Lemma 3.3.3 is that, when $a = 1$ and $\mathbf{1}_n \in \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$, use of a constant prior for $\boldsymbol{\gamma}$ would result in an improper posterior distribution, consistent with the result for isotropic case given in Berger et al. (2001).

The asymptotic behavior of the reference prior for the two limiting cases of interest are given in the following lemma, whose proof is given in Appendix B.3.

Lemma 3.3.4. *(Tail rates of the prior.) If Assumption 1 and Assumption 2 hold for each of the* $\mathbf{R}_l$, $1 \leqslant l \leqslant p$, *then* $\pi^R(\boldsymbol{\gamma})$ *has the following two limiting properties. Here* $\boldsymbol{\gamma}_E$ *denotes the vector of* γ_l *for all* $l \in E$, $E \subset \{1, 2, ..., p\}$, *and* $\boldsymbol{\gamma}_{-E}$ *denotes the*

complementary vector.

(i) As $\boldsymbol{\gamma}_E \to \mathbf{0}$,

$$\pi^R(\boldsymbol{\gamma}) \leqslant C(\boldsymbol{\gamma}_{-E}) \left[\prod_{l \in E} tr \left(\frac{\partial \mathbf{R}}{\partial \gamma_l} \right)^2 \right]^{1/2},$$

where $C(\boldsymbol{\gamma}_{-E})$ *is constant in* $\boldsymbol{\gamma}_E$.

(ii) As $\gamma_l \to \infty$ *for all* l, $1 \leqslant l \leqslant p$,

$$\pi^R(\boldsymbol{\gamma}) \leqslant C_1 \left| \frac{\prod_{l=1}^{p} \nu_l'(\gamma_l)}{(\sum_{l=1}^{p} \nu_l(\gamma_l))^p} \right|. \tag{3.15}$$

where $\nu_l'(\gamma_l) = \partial \nu_l(\gamma_l)/\partial \gamma_l$. *Furthermore, if* $\mathbf{1} \in \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$ *and* $p \geqslant 2$,

$$\pi^R(\boldsymbol{\gamma}) \leqslant C_2 \left| \frac{\prod_{l=1}^{p} \nu_l'(\gamma_l)}{(\sum_{l=1}^{p} \nu_l(\gamma_l))^p} \right| \left| \sum_{l=1}^{p} \frac{\nu_l^2(\gamma_l)\omega_l'(\gamma_l)}{\nu_l'(\gamma_l)\nu_m(\gamma_m)} \right|,$$

for every index m *between* 1 *and* p; *if* $\mathbf{1} \in \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$ *and if* $p = 1$,

$$\pi^R(\boldsymbol{\gamma}) \leqslant C_3 |\omega_1'(\gamma_1)|.$$

where C_1, C_2 *and* C_3 *are all positive and not related to* γ_l.

The bounds for the one-dimensional case in Lemma 3.3.4 (ii) were proved in Berger et al. (2001). These results are a generalization of the p dimensional results in Paulo (2005), which considered only separable designs.

Interestingly, the bounds in part (ii) of Lemma 3.3.4 seem to be almost exact in numerical examples we have studied for the power exponential correlation function. Figure 3.2 presents some of the evidence for this.

The following theorem states that, under the $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ parameterizations and when $a = 1$, the mode of the marginal posterior for the reference prior for the range parameters will typically be robust for the correlation functions listed in Table 3.1. Similar theorems can be stated for other choices of a but, since $a = 1$ is the near universal choice, we restrict the statement of the results to that case.

Theorem 3.3.1. *Under the parameterizations of the range parameter $\boldsymbol{\gamma}$ and log inverse range $\boldsymbol{\xi}$ in Definition 3.2.1, the posterior mode in (3.5) with $a = 1$ is robust for the product form of the power exponential, spherical, and Matérn correlation functions over the range of $\boldsymbol{\alpha}$ listed in Table 3.1. In addition, the posterior mode of $\boldsymbol{\gamma}$ is robust for the rational quadratic correlation function if $\alpha_l > 1/2$, $1 \leqslant l \leqslant p$ and the posterior mode of $\boldsymbol{\xi}$ is robust for the rational quadratic correlation function over the entire domain of $\boldsymbol{\alpha}$.*

Proof. Theorem 3.3.1 can be proved by verifying Corollary 3.3.2 using the results from Lemma 3.3.3 and Lemma 3.3.4. □

While use of the mode of the marginal posterior for the $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ parameterizations is robust, the mode of the marginal posterior under other parameterizations, such as the $\tilde{\boldsymbol{\beta}}$ parameterization in (3.8), can be non-robust. Indeed, directly applying Lemma 3.3.4 and Lemma 3.3.3, the bounds on the tail rates of the marginal posterior under the various parameterizations (and also for the profile and marginal likelihood) are given in Table 3.3. For simplicity, we assume roughness parameters are kept the same, i.e. $\alpha_1 = \alpha_2 = ... = \alpha_p = \alpha$. The blue highlighted entries are those in which the tail rate is constant, so that there is a potential problem of non-robustness.

The red highlighted entries in Table 3.3 are quite surprising, as here the marginal posterior density becomes infinite in the tail, so that the mode will be at the problematical $\mathbf{1}_n\mathbf{1}_n^T$. (The following Corollary 3.3.3 establishes that there is no other infinite mode.) That the posterior mode for the $\tilde{\boldsymbol{\beta}}$ parameterization has this bizarre

behavior has not been previously recognized, and should clearly rule out use of this parameterization (at least when estimating by the marginal posterior mode with the standard reference prior). Figure 3.3 gives numerical evidence of this feature, where we plot the log-marginal posterior as a function of $\tilde{\beta}_1 = \tilde{\beta}_2 = \tilde{\beta}$. Both examples have local modes with a finite marginal posterior, while the real modes with infinite value occur as $\tilde{\beta}_1 = \tilde{\beta}_2 \to 0$.

The following lemma is needed to establish posterior propriety in the next subsection and also to establish Corollary 3. It calculates the tail rates when only some but not all the range parameters are close to zero.

Lemma 3.3.5. *Assume Assumption 1 and Assumption 2 hold for each* $\mathbf{R}_l$, $1 \leqslant l \leqslant p$. *As* $\gamma_{l_1} \to \infty$ *for* $1 \leqslant l_1 \leqslant p_1$ *with* $p_1 < p$, $\gamma_{l_2} \to 0$ *for* $p_1 + 1 \leqslant l_2 \leqslant p_2$ *and* γ_{l_3} *is bounded between* 0 *and* ∞ *for* $p_2 + 1 \leqslant l_3 \leqslant p$, *a bound on the tail rate of the marginal posterior of* $\boldsymbol{\gamma}$ *is*

$$L(\mathbf{y}|\boldsymbol{\gamma})\pi^R(\boldsymbol{\gamma}) \leqslant C_4 \prod_{l_1=1}^{p_1} |\nu'_{l_1}(\gamma_{l_1})| \left[\prod_{l_2=p_1+1}^{p_2} tr\left(\frac{\partial \mathbf{R}}{\partial \gamma_{l_2}}\right)^2 \right]^{1/2},$$

where $C_4 > 0$ *is a positive constant .*

The details of the proof are given in Appendix B.3. The following corollary is a direct consequence of the above lemma and states that, when the power exponential correlation is used, the only possible infinite mode of the marginal posterior of $\boldsymbol{\beta}$ is at $\tilde{\beta}_l \to 0$ for all $1 \leqslant l \leqslant p$.

Corollary 3.3.3. *For the power exponential correlation function, if there is one* l, $1 \leqslant l \leqslant p$, *for which* $\tilde{\beta}_l > K$ *where* K *is a positive constant, then the marginal posterior of* $\tilde{\boldsymbol{\beta}}$ *using the standard reference prior (3.3) with* $a = 1$ *satisfies*

$$p(\tilde{\boldsymbol{\beta}}|\mathbf{y}^{\mathscr{D}}) \leqslant O(1).$$

Table 3.3: Tail behaviors of the profile likelihood, the marginal likelihood and the posterior distributions for different parameterizations of the power exponential correlation function, using the reference prior in (3.3) with $a = 1$. In the 2nd and 4th columns, E is a nonempty set such that for $l \in E$, $\gamma_l \to 0$ (equivalent to $\beta_l \to \infty$ or $\xi_l \to \infty$), and C and C_l are positive numbers depending on $|x_{il} - x_{jl}|$, $1 \leqslant i, j \leqslant n$, $l \in E$. In the 3rd and 5th columns, $\gamma_l \to \infty$ (equivalent to $\beta_l \to 0$ or $\xi_l \to -\infty$), for all $1 \leqslant l \leqslant p$; in the stated tail rates, $\gamma_{(1)}$ is defined as the minimum of the γ_l, $\beta_{(p)}$ is the largest β_l, and $\xi_{(p)}$ is the largest ξ_l, where $1 \leqslant l \leqslant p$. Blue highlights the cases where the tail behavior is constant, so that there is danger of non-robustness. Red highlights the cases where the posterior goes to infinity in the tail, necessarily leading to non-robustness, as this will be shown to be the unique mode.

	$\mathbf{1}_n \in \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$		$\mathbf{1}_n \notin \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$	
	$l \in E$, $\gamma_l \to 0$	$\gamma_l \to \infty$ for all l	$l \in E$, $\gamma_l \to 0$	$\gamma_l \to \infty$ for all l
Profile Lik	$O(1)$	$O(\gamma_{(1)}^{-\alpha/2})$	$O(1)$	$O(\gamma_{(1)}^{-\alpha/2})$
Marginal Lik	$O(1)$	$O(1)$	$O(1)$	$O((\gamma_{(1)}^{-\alpha/2}))$
Post $\boldsymbol{\gamma}$, $p = 1$	$O(\frac{\exp(-C/\gamma^{\alpha})}{\gamma^{(\alpha+1)}})$	$O(\gamma^{-\alpha-1})$	$O(\frac{\exp(-C/\gamma^{\alpha})}{\gamma^{(\alpha/2+1)}})$	$O(\gamma^{-\alpha/2-1})$
$p \geqslant 2$	$O(\prod_{l\in E} \frac{\exp(-C_l/\gamma_l^{\alpha})}{\gamma_l^{(\alpha+1)}})$	$O(\frac{\prod_{l=1}^{p} \gamma_l^{-\alpha-1}}{\gamma_{(1)}^{(1-p)\alpha}})$	$O(\prod_{l\in E} \frac{\exp(-C_l/\gamma_l^{\alpha})}{\gamma_l^{(\alpha/2+1)}})$	$O(\frac{\prod_{l=1}^{p} \gamma_l^{-\alpha-1}}{\gamma_{(1)}^{(1/2-p)\alpha}})$
Post $\tilde{\boldsymbol{\beta}}$, $p = 1$	$O(\exp(-\tilde{\beta}C))$	$O(1)$	$O(\tilde{\beta}^{\frac{1}{2}} \exp(-\tilde{\beta}C))$	$O(\tilde{\beta}^{-1/2})$
$p \geqslant 2$	$O(\prod_{l\in E} \exp(-\tilde{\beta}_l C_l))$	$O(\tilde{\beta}_{(p)}^{-(p-1)})$	$O((\sum_{l\in E} \tilde{\beta}_l)^{\frac{1}{2}} \prod_{l=1}^{p} \exp(-\tilde{\beta}_l)C_l)$	$O(\tilde{\beta}_{(p)}^{-(p-1/2)})$
Post $\boldsymbol{\xi}$, $p = 1$	$O(\exp(-\exp(\xi)C + \xi))$	$O(\exp(\xi))$	$O(\exp(-\exp(\xi)C + \frac{3}{2}\xi))$	$O(\exp(\xi/2))$
$p \geqslant 2$	$O(\prod_{l\in E} \exp(-\exp(\xi_l)C_l + \xi_l))$	$O(\frac{\exp(\sum_{l=1}^{p-1} \xi_l)}{\exp((p-2)\xi_{(p)})})$	$O(\prod_{l\in E} \exp(-\exp(\xi_l)C_l) + \frac{3}{2}\xi_l)$	$O(\frac{\exp(\sum_{i=1}^{p-1} \xi_l)}{\exp((p-1/2)\xi_{(p)})})$

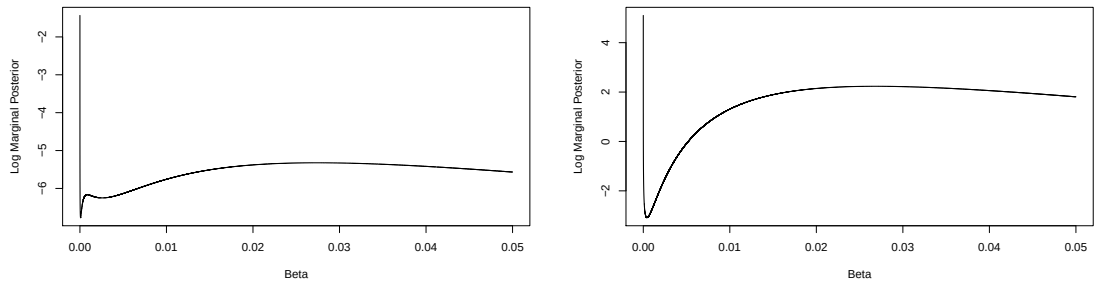

FIGURE 3.3: Examples of the marginal posterior of $\tilde{\beta}$ in the power exponential family with $\alpha = 1.9$, when emulating the modified Branin function (Picheny et al. (2013)), which has $p = 2$ inputs. Two data sets of size $n = 20$ were generated using an LHD design with $\mathbf{1} \in \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$. The black curves are the log marginal posterior of $\tilde{\beta}$ arising from setting $\tilde{\beta}_1 = \tilde{\beta}_2 = \tilde{\beta}$, and both exhibit infinite modes at 0.

3.3.4 Posterior propriety

Propriety of the posterior distribution for $\boldsymbol{\gamma}$ (and, hence, for all other parameterizations) is established in the following theorem for general designs, generalizing the theorems in Berger et al. (2001) under the isotropic assumption and in Paulo (2005) for separable designs. For simplicity, we assume $\alpha_1 = \alpha_2 = ... = \alpha_p = \alpha$.

Theorem 3.3.2. *When $\alpha_1 = \alpha_2 = ... = \alpha_p = \alpha$, the reference prior in (3.3) with $a = 1$ yields posterior propriety for GaSP models with the power exponential, spherical, rational quadratic and Matérn correlation functions, under general p-dimensional designs.*

The proof of Theorem 3.3.2 is in Appendix B.3.

3.4 Robust inference when a nugget is added the GaSP model

3.4.1 Background and parameter estimation

Some inputs have little effect on the output of the computer model. Such inputs are called inert inputs (Linkletter et al. (2006)) and are usually not used in building the emulator (Spiller et al. (2014)). However, when inert inputs are omitted in the

emulator, the emulator can no longer be an interpolator at the design points so that the GaSP model is then inappropriate. The common solution is to add a small noise term – called a *nugget* – to account for the error, such as $\tilde{y}(\cdot) = y(\cdot) + \epsilon$, where $y(\cdot)$ is the noise-free GaSP and ϵ is an i.i.d. mean-zero Gaussian white noise. (Of course, sometimes there is also uncertainty in the GaSP output itself, in which case adding a noise term is appropriate even if there are no inert inputs.) After adding the noise, the covariance function for the new process $\tilde{y}(\cdot)$ can be expressed as

$$\sigma^2\tilde{c}(\mathbf{x}_l, \mathbf{x}_m) \hat{=} \sigma^2\{c(\mathbf{x}_l, \mathbf{x}_m) + \eta\delta_{lm}\}, \tag{3.16}$$

where η is defined to be the nugget-variance ratio and δ_{lm} is a Dirac delta function when $l = m$, $\delta_{lm} = 1$. We parameterize the nugget in this way to allow for marginalizing the likelihood over σ^2 (c.f., Ren et al. (2012)). After adding the nugget, the covariance matrix becomes

$$\sigma^2\tilde{\mathbf{R}} = \sigma^2(\mathbf{R} + \eta\mathbf{I}_n). \tag{3.17}$$

The reference prior for a real-value output and isotropic GaSP model with a nugget has been discussed in Ren et al. (2012); Kazianka and Pilz (2012). Extending it to the GaSP model with multiple range parameters results in the the following form:

$$\pi^{\tilde{R}}(\boldsymbol{\theta}, \sigma^2, \boldsymbol{\gamma}, \eta) = \pi^{\tilde{R}}(\boldsymbol{\theta}, \sigma^2)\pi^{\tilde{R}}(\boldsymbol{\gamma}, \eta \mid \boldsymbol{\theta}, \sigma^2) \propto \frac{\pi^{\tilde{R}}(\boldsymbol{\gamma}, \eta)}{(\sigma^2)^a}, \tag{3.18}$$

with $\pi^{\tilde{R}}(\boldsymbol{\gamma}, \eta) \propto |\tilde{\mathbf{I}}^*(\boldsymbol{\gamma}, \eta)|^{1/2}$, and $\tilde{\mathbf{I}}^*(\cdot)$ is the expected fisher information matrix,

$$\tilde{\mathbf{I}}^*(\boldsymbol{\gamma}, \eta) = \begin{pmatrix} n-q & tr(\tilde{\boldsymbol{W}}_1) & tr(\tilde{\mathbf{W}}_2) & \dots & tr(\tilde{\mathbf{W}}_{p+1}) \\ & tr(\tilde{\mathbf{W}}_1^2) & tr(\tilde{\mathbf{W}}_1\tilde{\mathbf{W}}_2) & \dots & tr(\tilde{\mathbf{W}}_1\tilde{\mathbf{W}}_{p+1}) \\ & & tr(\tilde{\mathbf{W}}_2^2) & \dots & tr(\tilde{\mathbf{W}}_2\tilde{\mathbf{W}}_{p+1}) \\ & & & \ddots & \vdots \\ & & & & tr(\tilde{\mathbf{W}}_{p+1}^2) \end{pmatrix}_{(p+2)\times(p+2)}, \tag{3.19}$$

where $\tilde{\mathbf{W}}_l = \dot{\tilde{\mathbf{R}}}_l \tilde{\mathbf{Q}}$, for $1 \leqslant l \leqslant p$, p is the number of range parameters in the correlation matrix $\tilde{\mathbf{R}}$ and $\dot{\mathbf{R}}_l$ is the partial derivative of $\tilde{\mathbf{R}}$ with respect to the l^{th} range parameters and $\tilde{\mathbf{Q}} = \tilde{\mathbf{R}}^{-1}\mathbf{P}_{\tilde{\mathbf{R}}}$ with $\mathbf{P}_{\tilde{\mathbf{R}}} = \mathbf{I} - \mathbf{h}(\mathbf{x}^{\mathscr{D}})\{\mathbf{h}(\mathbf{x}^{\mathscr{D}})\tilde{\mathbf{R}}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})\}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})\tilde{\mathbf{R}}^{-1}$.

As in the previous sections, one can estimate the nugget and range parameters by their marginal maximum posterior mode,

$$(\hat{\gamma}_1, \ldots \hat{\gamma}_p, \hat{\eta}) = \underset{\gamma_1,\ldots,\gamma_p,\eta}{argmax}\, \mathcal{L}(\mathbf{y}^{\mathscr{D}}|\gamma_1, \ldots, \gamma_p, \eta)\pi^{\tilde{\mathbf{R}}}(\gamma_1, \ldots, \gamma_p, \eta). \tag{3.20}$$

3.4.2 Robustness of the posterior mode

Note that

$$\tilde{\mathbf{R}} = \mathbf{R}_1 \circ \mathbf{R}_2 \circ \ldots \circ \mathbf{R}_p \circ \mathbf{R}_{p+1},$$

where $\tilde{\mathbf{R}}_{p+1} = \mathbf{1}_n\mathbf{1}_n^T + \eta\mathbf{I}_n$. Also, $\mathbf{R}_{p+1}$ follows Assumption 3.3.2 with $\nu_{p+1} = \eta$, $\omega_{p+1} = 0$, and $\gamma_{p+1} = \eta$. Using these facts and in parallel to Lemma 3.3.3 and Lemma 3.3.4, the tail rates of the likelihood and the prior for the GaSP with a nugget are given in the following lemmas. The proof of the lemmas and theorems in this section are given in Appendix B.4.

Lemma 3.4.1. *If Assumption 1 and Assumption 2 hold for each of the* $\mathbf{R}_l$, $1 \leqslant l \leqslant p$ *, the marginal likelihood and profile likelihood have the following tail rates.*

(i) If $\gamma_l \to 0$ *for any* l, $1 \leqslant l \leqslant p$, *the marginal likelihood and profile likelihood are both greater than zero.*

(ii) If $\gamma_l \to \infty$ *for all* l, $1 \leqslant l \leqslant p$,

$$\mathcal{L}(\mathbf{y}^{\mathscr{D}}|\boldsymbol{\gamma}, \eta) = \begin{cases} O\left((\sum_{l=1}^{p} \nu_l(\gamma_l) + \eta)^{a-1/2}\right), & \mathbf{1}_n \notin \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}})), \\ O\left((\sum_{l=1}^{p} \nu_l(\gamma_l) + \eta)^{a-1}\right), & \mathbf{1}_n \in \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}})), \end{cases}$$

and the profile likelihood, in this case, satisfies

$$\mathcal{L}(\mathbf{y}^{\mathscr{D}}|\boldsymbol{\gamma}, \hat{\sigma}^2_{MLE}, \hat{\boldsymbol{\theta}}_{MLE}) = O(\sum_{l=1}^{p} \nu_l(\gamma_l) + \eta)^{1/2}.$$

Lemma 3.4.2. *If Assumption 1 and Assumption 2 hold for each of the R_l, $1 \leqslant l \leqslant p$, then $\pi^{\tilde{R}}(\gamma, \eta)$ has following two limiting properties. Here $\boldsymbol{\gamma}_E$ denotes the vector of γ_l for all $l \in E$, $E \in \{1, 2, ..., p\}$, and $\boldsymbol{\gamma}_{-E}$ denotes the complementary vector.*

(i) When $\boldsymbol{\gamma}_E \to \mathbf{0}$ for all $l \in E$, $E \subset \{1, 2, ..., p\}$, then

$$\pi^R(\boldsymbol{\gamma}) \leqslant \tilde{C}_{\boldsymbol{\gamma}_{-E}} \left[\prod_{l \in E} tr\left(\frac{\partial \tilde{\mathbf{R}}}{\partial \gamma_l} \right) \right]^{1/2}.$$

where $\tilde{C}_{\boldsymbol{\gamma}_{-E}}$ is a constant in $\boldsymbol{\gamma}_E$.

(ii) As $\gamma_l \to \infty$ for all $1 \leqslant l \leqslant p$ and $\eta \to 0$, if $\mathbf{1} \notin \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$, then

$$\pi^R(\boldsymbol{\gamma}) \leqslant \tilde{C}_1 \left| \frac{\prod_{l=1}^{p} \nu_l'(\gamma_l)}{(\sum_{l=1}^{p} \nu_l(\gamma_l) + \eta)^{p+1}} \right|;$$

further, if $\mathbf{1} \in \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$ and if $p \geqslant 2$

$$\pi^R(\boldsymbol{\gamma}) \leqslant \tilde{C}_2 \left| \frac{\prod_{l=1}^{p} \nu_l'(\gamma_l)}{(\sum_{l=1}^{p} \nu_l(\gamma_l) + \eta)^{p+1}} \right| \left| \sum_{l=1}^{p} \frac{\nu_l^2(\gamma_l) \omega_l'(\gamma_l)}{\nu_l'(\gamma_l) \nu_m(\gamma_m)} \right|,$$

for every index m between 1 to p; if $\mathbf{1} \in \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$ and $p = 1$,

$$\pi^R(\boldsymbol{\gamma}) \leqslant \tilde{C}_3 \frac{\nu_1(\gamma_1)|\omega_1'(\gamma_1)|}{(\nu_1(\gamma_1) + \eta)^2}.$$

where $\tilde{C}_1, \tilde{C}_2$ and $\tilde{C}_3$ are positive constants.

Directly applying Lemma 3.4.1 and Lemma 3.4.2 yields the bounds on the tail rates of the marginal posterior under the various parameterizations (and also for the profile and marginal likelihood) in Table 3.4. For simplicity, we assume $\alpha_1 = \alpha_2 = ... = \alpha_p = \alpha$.

Comparing Table 3.3 with Table 3.4, it is clear that addition of the nugget can cause a loss of robustness of the posterior mode for the $(\gamma_1, \gamma_2, ..., \gamma_p, \eta)^T$ and $(\xi_1, \xi_2, ..., \xi_p, \eta)^T$ parameterizations, in certain cases. Luckily, a simple reparameterization of η, to $\tau = log(\eta)$, with estimation by the corresponding posterior mode achieves robustness, as shown in the following theorem.

Theorem 3.4.1. *When $a = 1$, marginal posterior mode estimation of $(\gamma_1, ..., \gamma_p, \tau)^T$, and $(\xi_1, ..., \xi_p, \tau)^T$, where $\tau = log(\eta)$, is robust for the product form of the power exponential family, spherical, and Matérn correlation functions listed in Table 3.1, and for the rational quadratic correlation function when $\alpha > 1/2$. In addition, marginal posterior mode estimation of $(\xi_1, ..., \xi_p, \tau)^T$, for $1 \leqslant l \leqslant p$, is robust for the rational quadratic correlation function for all $\alpha > 0$, $1 \leqslant l \leqslant p$.*

Proof. Theorem 3.4.1 can be proved by verifying Corollary 3.3.2 using the results from Lemma 3.4.1 and Lemma 3.4.2. □

Note that the posterior mode of $(\gamma_1, ..., \gamma_p, \eta)^T$ is also robust for most of the correlation functions.

3.4.3 Posterior propriety for the GaSP model with a nugget

Lemma 3.4.3. *Assume Assumption 1 and Assumption 2 hold for each of the $\tilde{\mathbf{R}}_l$, $1 \leqslant l \leqslant p$. As $\gamma_{l_1} \to \infty$ for $1 \leqslant l_1 \leqslant p_1$ with $p_1 < p$, $\gamma_{l_2} \to 0$ for $p_1 + 1 \leqslant l_2 \leqslant p_2$, and if γ_{l_3} is bounded away from 0 and ∞, for $p_2 + 1 \leqslant l_3 \leqslant p$, the tail rates of the*

Table 3.4: Tail behaviors of the profile likelihood, the marginal likelihood and the posterior distributions for different parameterizations of the power exponential correlation function, using the reference prior in (3.16) with $a = 1$. E is a nonempty set such that for $l \in E$, $\gamma_l \to 0$ (equivalent to $\tilde{\beta}_l \to \infty$ or $\xi_l \to \infty$), and C and C_l are positive numbers not depending on $\gamma_l \in E$ (or $\tilde{\beta}_l \in E$ or $\xi_l \in E$). In the 3rd and 5th columns, $\gamma_l \to \infty$ (equivalent to $\tilde{\beta}_l \to 0$ or $\xi_l \to -\infty$), for all $1 \leqslant l \leqslant p$; in the stated tail rates, $\gamma_{(1)}$ is defined as minimum of the γ_l, $\tilde{\beta}_{(p)}$ is the largest $\tilde{\beta}_l$, and $\xi_{(p)}$ is the largest ξ_l, where $1 \leqslant l \leqslant p$. Blue highlights the cases where the tail behavior is constant; red highlights the cases where the posterior goes to infinity in the tail; and green highlights situations in which the rate might go to zero, a constant or infinity, depending on the speed of η and γ_l to their limits and the choice of the roughness parameter α.

	$\mathbf{1}_n \in \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$		$\mathbf{1}_n \notin \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$	
	$l \in E$, $\gamma_l \to 0$	$\gamma_l \to \infty$ for all l and $\eta \to 0$	$l \in E$, $\gamma_l \to 0$	$\gamma_l \to \infty$ for all l and $\eta \to 0$
Profile Lik	$O(1)$	$O((\gamma_{(1)}^{-\alpha} + \eta)^{\frac{1}{2}})$	$O(1)$	$O((\gamma_{(1)}^{-\alpha} + \eta)^{\frac{1}{2}})$
Marginal Lik	$O(1)$	$O(1)$	$O(1)$	$O((\gamma_{(1)}^{-\alpha} + \eta)^{\frac{1}{2}})$
Post $\boldsymbol{\gamma}$, $p = 1$	$O(\frac{exp(-C/\gamma^{\alpha})}{\gamma^{(\alpha+1)}})$	$O(\frac{\gamma^{-2\alpha-1}}{(\gamma^{-\alpha}+\eta)^2})$	$O(\frac{exp(-C/\gamma^{\alpha})}{\gamma^{(\alpha/2+1)}})$	$O(\frac{\gamma^{-\alpha-1}}{(\gamma^{-\alpha}+\eta)^{3/2}})$
$p \geqslant 2$	$O(\prod_{l\in E} \frac{exp(-C_l/\gamma_l^{\alpha})}{\gamma_l^{(\alpha+1)}})$	$O(\frac{\prod_{l=1}^{p} \gamma_l^{-\alpha-1}}{(\gamma_{(1)}^{-\alpha}+\eta)^p})$	$O(\prod_{l\in E} \frac{exp(-C_l/\gamma_l^{\alpha})}{\gamma_l^{(\alpha/2+1)}})$	$O(\frac{\prod_{l=1}^{p} \gamma_l^{-\alpha-1}}{(\gamma_{(1)}^{-\alpha}+\eta)^{p+1/2}})$
Post $\tilde{\boldsymbol{\beta}}$, $p = 1$	$O(exp(-\tilde{\beta}C))$	$O(\frac{\tilde{\beta}}{(\tilde{\beta}+\eta)^2})$	$O(\tilde{\beta}^{\frac{1}{2}} exp(-\tilde{\beta}C))$	$O((\tilde{\beta}+\eta)^{-3/2})$
$p \geqslant 2$	$O(\prod_{l\in E} exp(-\tilde{\beta}_l C_l))$	$O((\tilde{\beta}_{(p)}+\eta)^{-p})$	$O((\sum_{l\in E} \tilde{\beta}_l)^{\frac{1}{2}} \prod_{l=1}^{p} exp(-\tilde{\beta}_l)C_l)$	$O((\tilde{\beta}_{(p)}+\eta)^{-p-1/2})$
Post $\boldsymbol{\xi}$, $p = 1$	$O(exp(-exp(\xi)C + \xi))$	$O(\frac{exp(2\xi)}{(exp(\xi)+\eta)^2})$	$O(exp(-exp(\xi)C + \frac{3}{2}\xi))$	$O(\frac{exp(\xi)}{(exp(\xi)+\eta)^{3/2}})$
$p \geqslant 2$	$O(\prod_{l\in E} exp(-exp(\xi_l)C_l + \xi_l))$	$O(\frac{exp(\sum_{l=1}^{p} \xi_l)}{(exp(\xi_{(p)})+\eta)^p})$	$O(\prod_{l\in E} exp(-exp(\xi_l)C_l) + \frac{3}{2}\xi_l)$	$O(\frac{exp(\sum_{l=1}^{p} \xi_l)}{(exp(\xi_{(p)})+\eta)^{p+1/2}})$

marginal posterior of $\boldsymbol{\gamma}$ *are*

$$\mathcal{L}(\mathbf{y}^{\mathscr{D}}|\boldsymbol{\gamma},\eta)\pi^R(\gamma,\eta) \leqslant \tilde{C}_4 \prod_{l_1=1}^{p_1} \nu'_{l_1}(\gamma_{l_1}) \left[\prod_{l_2=p_1+1}^{p_2} tr\left(\frac{\partial \tilde{\mathbf{R}}}{\partial \gamma_{l_2}}\right)^2 \right]^{1/2},$$

where $\tilde{C}_4 > 0$ *is a constant.*

Propriety of the posterior distribution for $\boldsymbol{\gamma}$ and η (and, hence, for all other parameterizations) is established in the following theorem, generalizing the theorems in Ren et al. (2012); Kazianka and Pilz (2012) under the isotropic assumption with a nugget. For simplicity, we assume $\alpha_1 = \alpha_2 = ... = \alpha_p = \alpha$.

Theorem 3.4.2. *When* $\alpha_1 = \alpha_2 = \ldots = \alpha_p = \alpha$*, the reference prior in (3.18) with* $a = 1$ *yields posterior propriety for the GaSP models with a nugget under the power exponential, spherical, rational quadratic and Matérn correlation functions, for general p-dimensional designs.*

3.5 Numerical results

3.5.1 Comparison criteria

In this section, we numerically compare the performance of several of the methods discussed above, including the MLE and marginal posterior mode estimation with parameterizations $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ (the log inverse of $\boldsymbol{\gamma}$). We do not include the MLLE method or results for the $\tilde{\boldsymbol{\beta}}$ parameterization because of the robustness problems these methods have, as indicated in Table 3.3 with Table 3.4. A constant GaSP mean is assumed for all cases, i.e. $\mathbf{h}(\mathbf{x}) = \mathbf{1}$ and we use the Matérn correlation with $\alpha = 5/2$ in (5.12) for all methods. Also included are the results produced by the DiceKriging package (Roustant et al. (2012)), where the Matérn correlation is also the default setting.

We mainly compare the out of sample prediction evaluated by Mean Square Error (MSE). In each simulation, we use n runs (n is small, typically chosen to be $n \approx 10p$ to $n \approx 20p$, where p is the number of inputs) to build the GaSP emulator and then record the out-of-sample MSE of $n^* = 10,000$ held-out outputs. This is repeated for $N = 500$ random designs, with the resulting average MSE being reported. The criteria are thus

$$\begin{aligned} \text{MSE}_j &= \frac{1}{n^*}\sum_{i=1}^{n^*}(y(\mathbf{x}_{ij}^*) - \hat{y}(\mathbf{x}_{ij}^*))^2, \\ \text{AvgMSE} &= \sum_{j=1}^{N}\text{MSE}_j/N, \end{aligned}$$

where $\mathbf{x}_{ij}^*$ is the i^{th} held-out input in the j^{th} design and $\hat{y}(\mathbf{x}_{ij}^*)$ is its prediction. To provide a better visual comparison between the methods, we also study the out-of-sample Normalized-RMSE

$$\text{Normalized-RMSE}_j = \sqrt{\sum_{i=1}^{n^*}(y(\mathbf{x}_{ij}^*) - \hat{y}(\mathbf{x}_{ij}^*))^2 / \sum_{i=1}^{n^*}(y(\mathbf{x}_{ij}^*) - \bar{\mathbf{y}}_j)^2},$$

where $\bar{\mathbf{y}}_j$ is the mean of the observed output for the j^{th} experiment, $j = 1, .., N$. The Normalized-RMSE of an effective method should range from 0 to 1.

3.5.2 GaSP model without a nugget

We test the following five functions:

i. 1 dimensional Higdon function from Higdon et al. (2002),

$Y = \sin(2\pi X/10) + 0.2\sin(2\pi X/2.5)$, where $X \in [0, 10]$.

ii. 2 dimensional Lim function from Lim et al. (2002),

Table 3.5: Average MSE of the four estimation procedures for the five experimental functions. The sample size is $n = 20p$ for the Higdon function and $n = 10p$ for the others. Designs are generated by maxmin LHD. The baseline MSE is 0.52, 3.8, 52, 0.71, and 24 for these five functions if only the mean of the training output is used for the predictions.

	Robust GaSP $\boldsymbol{\xi}$	Robust GaSP $\boldsymbol{\gamma}$	MLE	DiceKriging
1-dim Higdon	.00011	.00012	.00013	.00013
2-dim Lim	.0064	.0080	.021	.0083
3-dim Pepelyshev	.083	.15	3.5	.79
4-dim Park	.00011	.00011	.033	.00063
5-dim Friedman	.026	.038	4.7	.44

$Y = \frac{1}{6}[(30 + 5X_1 \sin(5X_1))(4 + \exp(-5X_2)) - 100] + \epsilon$, where $X_i \in [0, 1]$ for $i = 1, 2$.

iii. 3 dimensional Pepelyshev function from Dette and Pepelyshev (2010),

$Y = 4(X_1 - 2 + 8X_2 - 8X_2^2)^2 + (3 - 4X_2)^2 + 16\sqrt{X_3 + 1}(2X_3 - 1)^2$, where $X_i \in [0, 1]$ for $i = 1, 2, 3$.

iv. 4 dimensional Park function from Park (1991),

$Y = \frac{2}{3}\exp(X_1 + X_2) - X_4 \sin(X_3) + X_3$, where $X_i \in [0, 1)$ for $i = 1, 2, 3, 4$.

v. 5 dimensional Friedman function from Friedman (1991),

$Y = 10 \sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5$, where $X_i \in [0, 1]$ for $i = 1, 2, 3, 4, 5$.

The average MSEs of the four estimation methods for the five functions are shown in Table 3.5. The Robust GaSP methods clearly outperformed MLE and DiceKriging, with the $\boldsymbol{\xi}$ parameterization yielding the best performance. Note that all methods used the same GaSP prediction equations; the only difference was in the estimates of the correlation parameters.

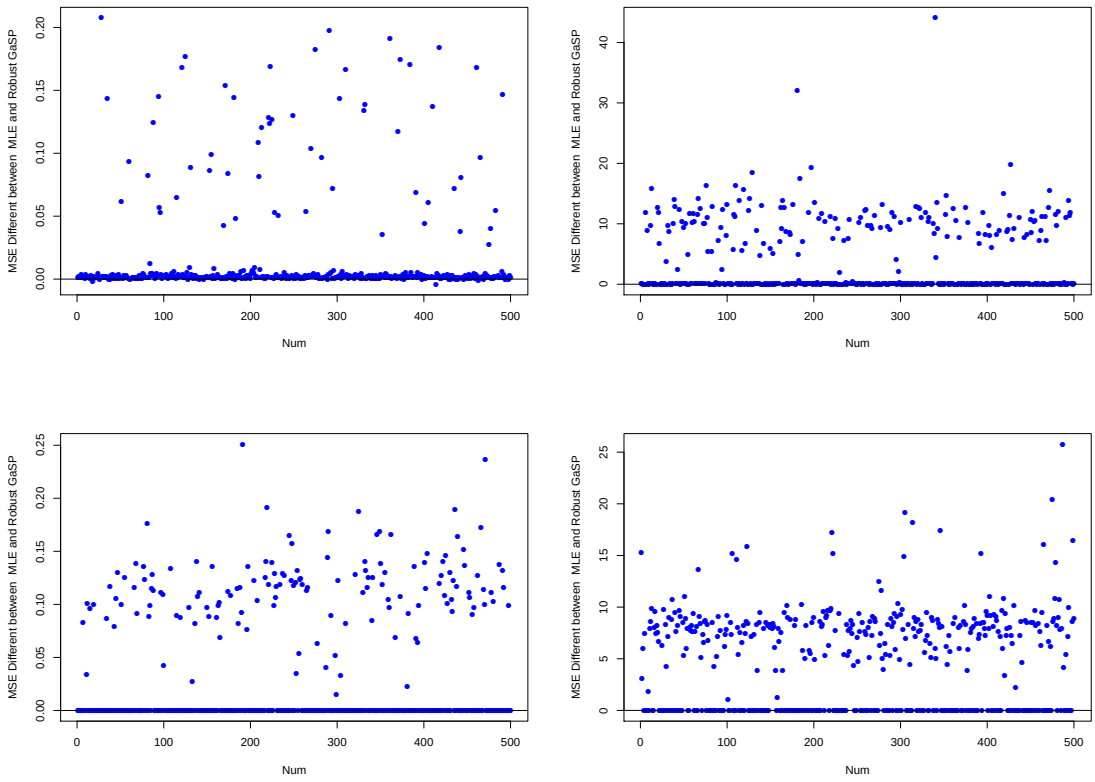

FIGURE 3.4: Plot of MSE for the MLE GaSP minus MSE for the robust GaSP under the $\boldsymbol{\xi}$ parameterization, for each of $N = 500$ designs for the Lim function (upper left), Pepelyshev function (upper right), Park function (lower left)) and Friedman function (lower right).

Figure 3.4 gives the difference of MSE_j of prediction, for each of 500 designs j (for functions ii, iii, iv and v), between the MLE GaSP and the robust GaSP under the $\boldsymbol{\xi}$ parameterization. Note that, for a significant proportion of the designs, the MLE GaSP is much worse than the robust GaSP. In these cases, the MLE GaSP estimate yields a covariance matrix that is close to $\hat{\mathbf{R}} \approx \mathbf{I}_n$, in which the prediction degenerates to the fitted mean with impulse functions at the observed values of the inputs.

Figure 3.5 gives the difference of MSE_j of prediction, for each of 500 designs j, between the DiceKriging GaSP and the robust GaSP under the $\boldsymbol{\xi}$ parameterization.

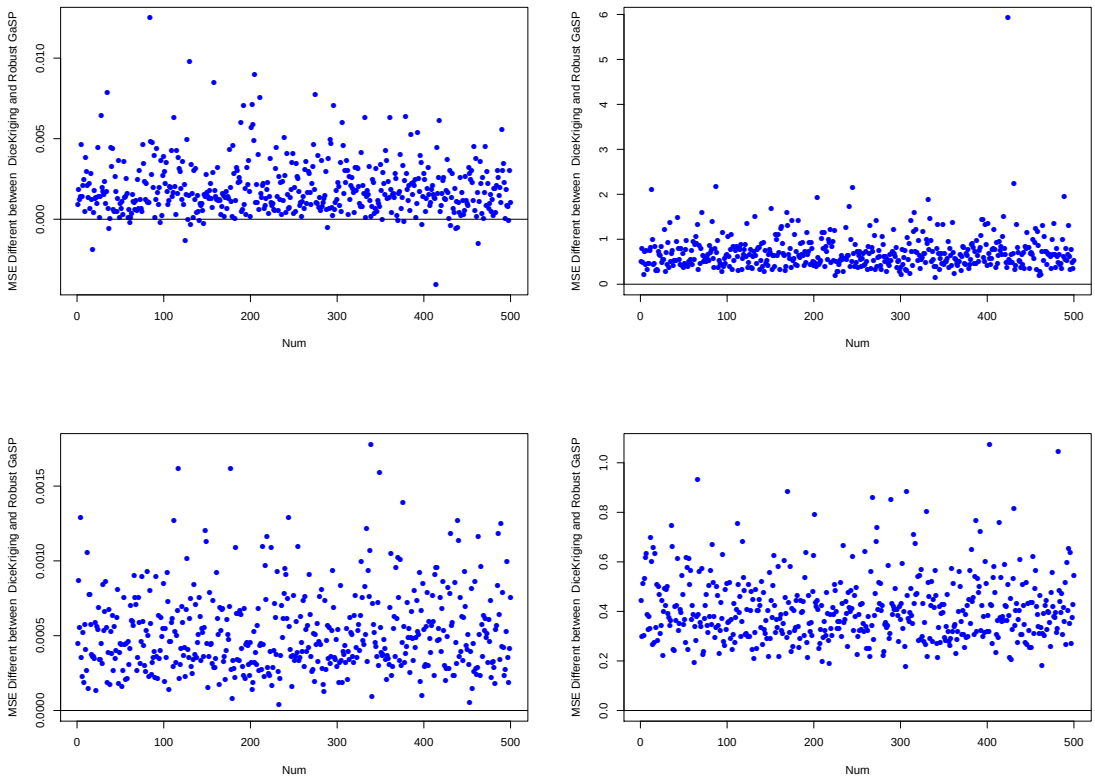

FIGURE 3.5: Plot of MSE for DiceKriging minus MSE for the robust GaSP under the $\boldsymbol{\xi}$ parameterization, for each of $N = 500$ designs for the Lim function (upper left), Pepelyshev function (upper right), Park function (lower left)) and Friedman function (lower right)..

The DiceKriging package uses a number of techniques to avoid unstable prediction of the correlation parameters (Roustant et al. (2012); Li and Sudjianto (2005)), and does so much better than the MLE GaSP, as a comparison of Figure 3.4 and Figure 3.5 indicates (the y-axis scales are considerably smaller for DiceKriging). Clearly, however, DiceKriging produces inferior correlation parameter estimates than does the robust GaSP in virtually all of the design cases for the four functions in Figure 3.5; indeed, only for a few design choices for the Lim function does DiceKriging produce better predictions than the robust GaSP.

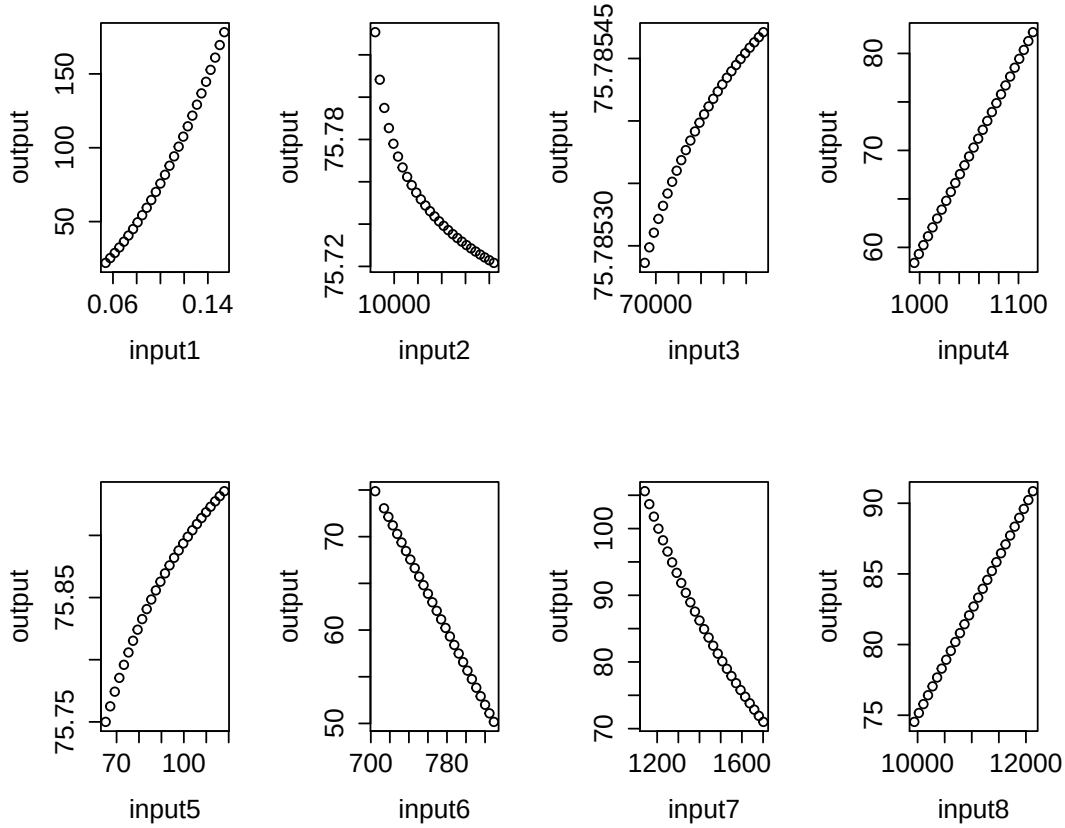

FIGURE 3.6: Values of Borehole function by varying one input at a time.

3.5.3 GaSP model with a nugget

The borehole function models water flow through a borehole and is given by

$$Y = \frac{2\pi T_u(H_u - H_l)}{ln(r/r_\omega)(1 + \frac{2LT_u}{ln(r/r_\omega r_\omega^2 K_\omega)} + \frac{T_u}{T_l})},$$

where $r_\omega, r, T_u, H_u, T_I, H_l L, K_\omega$ are the 8 inputs defined in a rectangle (Morris et al. (1993); An and Owen (2001)). Figure Figure 3.6 presents plots of the borehole function made by fixing seven of the inputs and varying one. It seems that the 2^{nd}, 3^{rd} and 5^{th} inputs barely affect the output, and this can be shown to hold globally over the input space. Thus in this section, we only use the five influential inputs to build the GaSP model and add a nugget to account for the error.

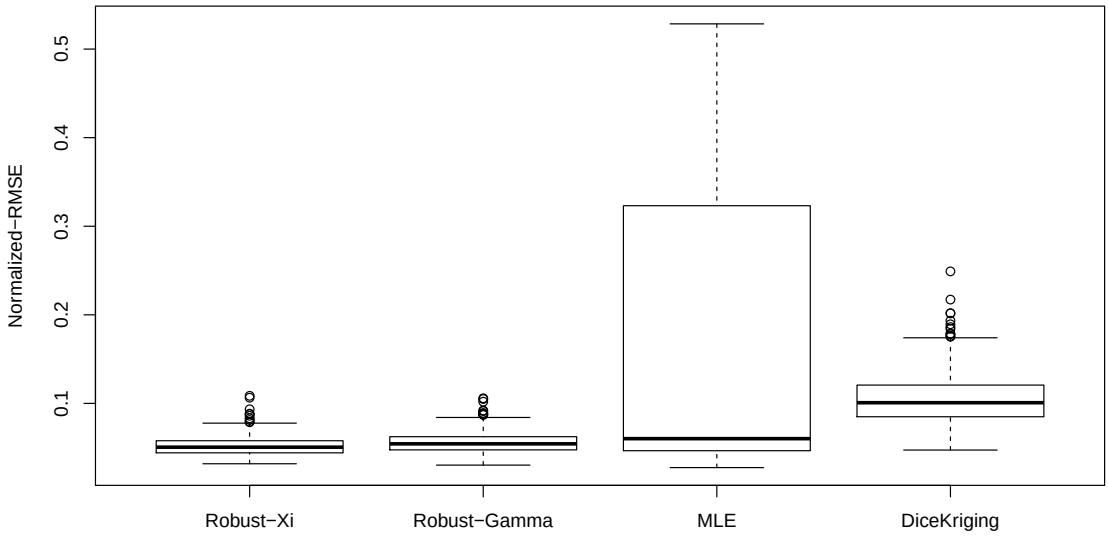

FIGURE 3.7: Boxplots, for the four estimation methods, of the Normalized RMSE for prediction of the Borehole function, based on $n = 30$ design points to build the emulator and averaging over $N = 500$ different designs generated from a Maximin LHD design. The average baseline MSE is 2080.087, using only the mean for prediction. The Average MSE for the 4 methods (from the left to the right) are 5.932, 6.786, 92.84 and 24.55.

The results of Normalized-RMSE for the borehole function are shown in Figure 3.7. The Normalized-RMSE of the GaSP with parameters estimated by MLE is much higher than the others, because it degenerates to the fitted basis mean, corresponding to $\tilde{\mathbf{R}} \to (1 + \hat{\eta})\mathbf{I}_n$ in many cases. Although the nugget might stabilize the computation when $\mathbf{R} \approx \mathbf{1}_n\mathbf{1}_n$, it cannot help when $\mathbf{R}$ becomes nearly proportional to $\mathbf{I_n}$. In contrast, the robust GaSP, with a correct parameterization, prevents these bad cases from materializing.

DiceKriging produces more stable results and yields a smaller RMSE than does MLE. But it is worse than the robust GaSP in almost all cases, in terms of Normalized-RMSE. The largest MSE for robust GaSP, with either the $\boldsymbol{\gamma}$ or $\boldsymbol{\xi}$ parameterization,

is around 0.1, which is quite small, considering that only $n = 30$ observations were utilized to build the emulators.

4

Jointly robust prior for the GaSP model

This chapter introduces a new prior for the GaSP model, as a substitute to the reference prior. In the last chapter, marginal posterior mode estimation is proposed using the reference prior with certain parameterization, with a focus on robustness of parameter estimation. Estimation with the robustness property stabilizes the computation, and out-of-sample prediction results suggests the superiority of robust methods.

Robust parameter estimation with the marginal posterior modes, however, depends on the parameterizations of the reference prior. Some properties of the reference prior that contribute to the good performance in prediction are free of parameterizations. For instance, the reference prior automatically scales to the number of observations (n), the number of dimensions of the input space (p), and is invariant to the scale of inputs. These properties make the reference prior a suitable default choice of the prior in most situations.

On the other hand, the reference prior has limitations in certain scenarios. One challenge with the reference prior arises on the computational side. The derivative of the prior is computationally expensive to compute. And for some correlation

functions, e.g. the Matérn correlation with roughness parameter $\alpha = 5/2$ defined in Equation (5.12), there are substantial prior mass in a region such that $\mathbf{R} \approx \mathbf{1}_n \mathbf{1}_n^T$, making it impossible to compute the prior precisely. Another challenge presents when the marginal posterior mode is used to estimate parameters for computer models with inert inputs (inputs which barely affect the outputs). In such case, posterior mode estimation with the reference prior may not be able to identify such inputs; see Section 4.1.1 for details.

A new class of priors – called jointly robust priors – is introduced in this chapter for the GaSP model. This prior is developed to have all good properties of the reference prior that were considered, including the robustness property, scalability to the number of observations, the number of dimensions and the range of inputs. In additional, it has a closed form expression, which is computationally cheaper. Numerical analysis also suggests that it can identify inert inputs with no additional computational cost.

4.1 Literature Review and Motivation

4.1.1 The Reference prior

Throughout this chapter, we consider the GaSP model discussed in Section 1.1. The GaSP model is defined in (1.1), with the mean function in (1.3), and the correlation function function with the product form in (1.4). The parameters in the model are the mean parameters, variance parameters and range parameters (or their transformations), while the roughness parameters in the correlation function are fixed. Assume the computer model has been avaluated at n design points $\mathbf{x}^{\mathscr{D}} = (\mathbf{x}_1^{\mathscr{D}}, ..., \mathbf{x}_n^{\mathscr{D}})$ and the outputs are $\mathbf{y}^{\mathscr{D}} = (y_1^{\mathscr{D}}, ..., y_n^{\mathscr{D}})$. The correlation matrix follows a product form $\mathbf{R} = \mathbf{R}_1 \circ ... \circ \mathbf{R}_p$.

Features of the reference prior

Some common properties of the reference prior $\pi^R(\cdot)$ defined in (3.3) for range parameters (and their transformations) are discussed here. The first three properties do not depend on the parameterization used, although we will illustrate the properties with the inverse range parameter $\beta_l = 1/\gamma_l$ (thus also holding for the range parameters γ_l and $\xi_l = \log(1/\gamma_l)$). The estimation by marginal posterior mode with β_l is typically not robust, but one could obtain the estimate of the inverse range parameter $\hat{\beta}_l$ by first estimating $\hat{\gamma}_l$ (or $\hat{\xi}_l$) by its mode (as it is a robust estimate) and transforming to $\hat{\beta}_l$.

First, when the dimension of the inputs increases, the reference prior mass moves from large values of β_l to small values of β_l marginally, for each $l = 1, ..., p$. This is an important property since, as any of $\hat{\beta}_l \to \infty$, $\hat{\mathbf{R}} = \mathbf{I}_n$, a degenerate case that should be avoided, which is discussed extensively in Chapter 3. When p increases, the chance that at least one β_l is estimated to be large increases, if the prior mass does not change along with p and, consequently, the chance that $\hat{\mathbf{R}} = \mathbf{I}_n$ also increases. The reference prior adapts to the increase of dimension by concentrating more prior mass at smaller β_l, avoiding $\hat{\mathbf{R}} \approx \mathbf{I}_n$, when p increases.

Second, when a denser design is used in a fixed domain of the input space, the prior mass of the reference prior parameterized by β_l moves to larger values slowly, and the variance of the prior also increases. This is helpful for computation in practice, because as points fill with a fixed domain of the input space, the covariance matrix becomes singular if $\hat{\beta}_l$ does not change. Increasing $\hat{\beta}_l$ along with the design points thus helps solve the covariance matrix inversion problem. From a theoretical point of view, the posterior mode is biased, due to the penalty induced by the prior, so inflating the prior variance when increasing the number of the observations is important for maintaining the efficiency of the posterior mode estimation. Increasing

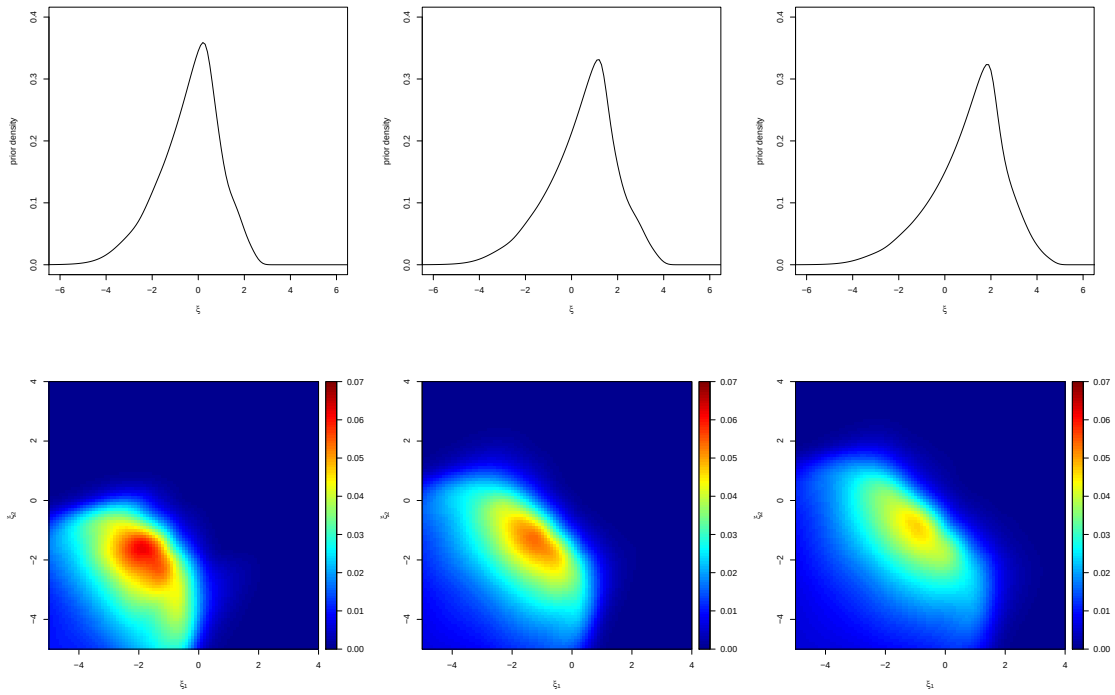

FIGURE 4.1: The reference prior $\pi^R(\boldsymbol{\xi})$ for the power exponential correlation function with roughness parameters $\alpha_l = 1.9$, $1 \leqslant l \leqslant p$. The dimensions of the inputs are $p = 1$ in the first row and $p = 2$ in the second row. From left to right, the number of design points are $n = 20$, $n = 50$ and $n = 100$. The designs are generated from a maximin Latin Hypercube on $[0, 1]$ for the first row and on $[0, 1] \times [0, 1]$ for the second row.

the inverse range parameters along with design points has also been found to be needed to have L_2 consistency in calibration of the computer model, while MLE is not L_2 consistent in such scenarios (Tuo and Wu (2015); Tuo et al. (2015)).

The third property of interest is that the reference prior is invariant to location-scale transformations of the inputs. When we apply a location-scale transformation, such that $\tilde{x}_l = \frac{x_l - c_0}{c_1}$ with two real constants c_0 and c_1, the new reference prior is $\tilde{\pi}^R(\beta_1, ..., \beta_l, ..., \beta_p) = \pi^R(\beta_1, ..., \beta_l/c_1, ..., \beta_p)$. This makes the prior scale naturally to the range of the inputs; as a consequence, we do not need to normalize the inputs.

In addition, the posterior modes for two robust parameterizations $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ in (3.2.1) have similar tail rates, at the limits when $\mathbf{R} = \mathbf{I}_n$ and $\mathbf{R} = \mathbf{1}_n \mathbf{1}_n^T$. When

$\gamma_l \to 0$ (equivalent to $\xi_l \to \infty$) for any $l = 1, ..., p$, the tail rates of the of reference prior decrease at approximately an exponential rate; when $\gamma_l \to \infty$ (equivalent to $\xi_l \to -\infty$) for all $l = 1, ..., p$, the tail rates of the reference prior are polynomial with a slightly different ratio, depending on the parameterization. The first part of the tail rates induces an exponential penalty to the likelihood at the limit of $\mathbf{R} = \mathbf{I}_n$, which is shared through all three parameterizations defined in (3.2.1). The second part depends on the parameterization. The robustness property under range parameters $\boldsymbol{\gamma}$ and log inverse range parameters $\boldsymbol{\xi}$ both have slow polynomial decreasing rates (when γ_l increases or ξ decreases), allowing the marginal likelihood to come in play at this limit.

Reference prior with a different number of design points and different input dimensions are shown in Figure 4.1. They are plotted for the $\boldsymbol{\xi}$ parameterization because of better visualization. Comparing the first row of figures (in which the input has 1 dimension) and the second row of figures (where the input has 2 dimension), the prior mass concentrates on regions with smaller values (in $\boldsymbol{\xi}$ parameterization), if the inputs have higher dimensions. When the number of design points increases in a fixed domain, the prior mass is more spread out and the mode of the prior moves to the region with larger values of $\boldsymbol{\xi}$.

Problems of the reference prior with mode estimation

Estimation by the posterior mode when utilizing the reference prior does have some drawbacks. In particular, the presence of inert inputs can cause difficulties in the computation. Having inert inputs is a fairly common scenario with computer models, as we saw in Chapter 2 for TITAN2D, where the internal friction angle, δ_{int}, had negligible effect on the output. When using product correlation functions, the hope is that, for an inert input l, $\hat{\mathbf{R}}_l = \mathbf{1}_n \mathbf{1}_n^T$, in which it will not affect the correlation matrix $\mathbf{R}$. Using posterior mode estimation with reference prior, this would happen if

$\hat{\gamma}_l \to \infty$ (or $\hat{\xi}_l \to -\infty$). However, as shown in the following corollary, robust marginal posterior mode estimation utilizing the reference prior cannot identify inert inputs.

Corollary 4.1.1. *The marginal posterior of range parameters $\boldsymbol{\gamma}$ (or the logarithm of inverse range parameters $\boldsymbol{\xi}$) goes to 0 if some, but not all, $\gamma_l \to \infty$ (or $\xi_l \to 0$), $l = 1, \cdots, p$, for the power exponential, spherical, rational quadratic and Matérn correlation functions, when the standard reference prior (3.3) is used, with $a = 1$.*

Proof. It is a direct consequence of Lemma 3.3.5. □

According to Corollary 4.1.1, the marginal posterior mode with two parameterizations will never appear at $\hat{\mathbf{R}}_l = \mathbf{1}_n \mathbf{1}_n^T$ for any l, as the posterior density is 0 if $\hat{\mathbf{R}}_l = \mathbf{1}_n \mathbf{1}_n^T$ for some but not all l. The identifiability of inert inputs with the posterior mode estimation, however, requires the posterior density is positive when $\mathbf{R}_l = \mathbf{1}_n \mathbf{1}_n^T$ for some but not all l (otherwise it is not a robust parameter estimation). Other transformation of the reference prior is also less likely to both maintain the robustness parametrization and can identify of inert inputs. Such difficulties could lead to inferior prediction results when some inert inputs are presented in computer models.

Besides, the computational challenges still persist with the use of the reference prior, even if posterior modes estimation is used in replace of MCMC posterior samples. The closed form derivatives of the reference prior are very computationally intensive, while frequently used optimizing algorithms typically rely on the derivatives information. The numerical derivatives, require more evaluations of the likelihood and are thus very time consuming. In addition, the reference prior could also induce some extra local modes, making the optimization algorithm harder to converge to global modes.

Finally, even though the reference prior is proper, the normalizing constant, the first moment and the higher moments of the reference prior do not have a close form

expression. Sometimes the numerical values are also hard to obtain, because some non-negligible prior mass are at $\mathbf{R} \approx \mathbf{1}_n \mathbf{1}_n^T$, in which the computation of the inverse of the correlation matrix becomes very unstable because of the singularity, while the reference prior requires such computation. All these features, to some extents, prohibit the study of the reference prior in the GaSP model.

4.1.2 Other priors

Several different priors have been proposed for the GaSP model, often with a product form, i.e. $\pi(\beta) = \pi(\beta_1)\pi(\beta_2)...\pi(\beta_p)$. For the forms of β_l, other than the Jefferys-type of prior (Berger et al. (2001); Paulo (2005); Ren et al. (2012); Kazianka and Pilz (2012)), many vague proper/improper priors were previously studied. For example, $\pi(\beta_l) \propto 1/\beta_l$ was utilized in Kennedy and O'Hagan (2001); $\pi(\beta_l) = \frac{1}{1+\beta_l^2}$ was assumed in Conti and O'Hagan (2010), and the widely used gamma prior for the inverse range parameter of β was also studied extensively in the isotropic case (c.f. van der Vaart and van Zanten (2009)). Priors were also defined in other parameterization, e.g. an independent beta prior for transformation $\rho_l = 1/exp(\beta_l)$, $l = 1, ..., p$, is utilized in Higdon et al. (2008). Spike and Slab prior for the same parameterization is proposed for model selection in the GaSP model (Savitsky et al. (2011)).

Though eliciting the prior information has been discussed in Oakley (1999, 2002), it is rather hard to faithfully transform subjective prior knowledge in Gaussian Process model with the product correlation matrix $\mathbf{R} = \mathbf{R}_1 \circ \mathbf{R}_2 \circ ... \circ \mathbf{R}_p$, as the parameters in $\mathbf{R}$ typically do not have real interpretation. Moreover, if the posterior modes should be used for quick computation, these subjective priors are usually either non-robust (and thus two unwelcome limiting cases discussed in Section 3.3.2 could happen frequently) or incapable to identify the inert inputs.

4.1.3 Sensitivity analysis

Sensitivity analysis in computer model concerns with the problem of learning how change of inputs affects the outputs (Oakley and O'Hagan (2004)). The inputs, in the computer model, typically associate with a distribution $\pi(\mathbf{x})$. The uncertainty in $\pi(\mathbf{x})$ can be propagated through the computer model by generating $\mathbf{x}_i \sim \pi(\mathbf{x})$ and running the computer model on $\mathbf{x}_i$, $1 \leqslant i \leqslant N$. The change of outputs y can be numerically analyzed through the simulated $y(\mathbf{x}_i)$, when the computer model is allowed to run many times.

The change of outputs at a specific input point $\mathbf{x} = \mathbf{x}_0$ can be studied using the derivative of the computer model at this point. Such study is called local sensitivity analysis. An related and more interested topic is the global sensitivity analysis, which studies the change of outputs associated with the whole input ranges (see e.g. Iooss and Lemaître (2014) for a review of these studies).

One of the main goal associated with the global sensitivity analysis is to identify how much a set of inputs influence the variability of outputs, which is studied through the decomposition of the variance of a function, or the functional analysis of the variance (functional ANOVA), introduced below.

Functional ANOVA

For a function $f(\mathbf{x})$, it is possible to decompose the function as following (Hoeffding (1948)),

$$f(\mathbf{x}) = z_0 + \sum_{i=1}^{p} z_i(x_i) + \sum_{i<j}^{p} z_{ij}(\mathbf{x}_{i,j}) + ... + z_{12...p}(\mathbf{x}),$$

where $\mathbf{x}_{i,j} = (x_i, x_j)$ and $\mathbf{x} = (x_1, ..., x_p)$. One can obtain these element functions by the conditional expectation,

$$
\begin{aligned}
z_0 &= \mathrm{E}[f(\mathbf{x})], \\
z_i(x_i) &= \mathrm{E}[f(\mathbf{x})|x_i] - z_0, \\
z_{ij}(\mathbf{x}_{i,j}) &= \mathrm{E}[f(\mathbf{x})|\mathbf{x}_{i,j}] - z_0 - z_i - z_j, \\
&\dots
\end{aligned}
$$

$z_i(x_i)$ is called the main effect; $z_{i,j}(\mathbf{x}_{i,j})$ is called the second order effect and so on. $\mathbf{x}$ is assumed to have a random distribution $\pi(\mathbf{x})$, reflecting the belief of the input values (and thus the expectation is on $\mathbf{x}$).

The decomposition of the variance can be defined as (Efron and Stein (1981)),

$$
\mathrm{Var}[f(\mathbf{x})] = \sum_{i=1}^{p} W_i + \sum_{i<j}^{p} W_{ij} + ... + W_{12...p},
$$

where $W_i = \mathrm{Var}[\mathrm{E}[f(\mathbf{x})|x_i]] = \mathrm{Var}[z_i(x_i)]$, $W_{ij} = \mathrm{Var}[\mathrm{E}[f(\mathbf{x})|\mathbf{x}_{i,j}]] - W_i - W_j$. Further define $V_i = \mathrm{Var}[E[f(\mathbf{x})|x_i]]$. Two principal measures called main effect index and total effect index were defined as (Sobol' (1990); Homma and Saltelli (1996); Saltelli et al. (2000); Sobol' (2001)),

$$
\begin{aligned}
S_i &= V_i / \mathrm{Var}[f(\mathbf{x})], \\
S_{T_i} &= V_{T_i} / \mathrm{Var}[f(\mathbf{x})] = 1 - S_{-i},
\end{aligned}
$$

where $V_{T_i} = \mathrm{Var}[f(\mathbf{x})] - \mathrm{Var}[\mathrm{E}(f(\mathbf{x})|\mathbf{x}_{-i})]$. S_i is referred as the main effect index of x_i and S_{T_i} is referred as the total effect index of x_i.

As pointed out in Oakley and O'Hagan (2004), S_i has a very clear practical meaning. If we were to know the real value of the i^{th} input, $x_i = x_i^r$, the uncertainty left is thus $\mathrm{Var}[f(\mathbf{x})|x_i^r]$, and the decrease of the uncertainty is $\mathrm{Var}[f(\mathbf{x})] - \mathrm{Var}[f(\mathbf{x})|x_i^r]$. Since we do not know x_i, it is common to take the expectation. Consequently, the

decrease of the variance is then $\mathrm{Var}[\mathrm{E}[f(\mathbf{x})|x_i]] = V_i$. This means if we were able to select one input to explore its true value, we will select x_i that maximizes V_i.

However, if one were able to select two inputs to explore, the answer is NOT to select the largest main effect index, but to select the largest

$$V_{i,j} = \mathrm{Var}[E(Y|\mathbf{x}_{i,j})] = \mathrm{Var}[z_i(x_i) + z_j(x_j) + z_{ij}(\mathbf{x}_{ij})].$$

Thus, many higher order indices are needed to compute if one are interested in exploring the first few largest influential inputs; however, the total number of indices is 2^p, which is computationally intensive. Main effect indices only serves as an approximation, though they are frequently used in reality due to the computational reason.

When $f(\mathbf{x})$ and $\pi(\mathbf{x})$ have simple forms, the above main effect indices and higher order indices may be computed explicitly. But in a general scenario, these indices do not have a closed form expression, thus the estimation of these indices with numerical evaluation becomes important. Monte Carlo methods are proposed to evaluate these indices (Sobol' (1990, 2001)).

The shortage of the Monte Carlo method is that it typically needs lots of computer model runs for numerically estimation, which may sometimes be unrealistic as the computer model may also be very slow. One approach that significantly reduces the number of evaluation of the functions is discussed in Oakley and O'Hagan (2004). The idea is to use a small number of runs to fit the GaSP emulator and use it to replace the computer model (which allows generating runs via the posterior predictive distribution). The estimation of the indices can be implemented based on the emulator built on only very small number of runs from the computer model. We compare with these methods in Section 4.4.

Bayesian model selection

Although there are vast amount of literature in the sensitivity analysis of complex compute models. Variable selection in GaSP model is less discussed in the literature. Some papers that do so are Schonlau and Welch (2006); Linkletter et al. (2006); Savitsky et al. (2011). In Schonlau and Welch (2006), the variable is selected one by one through a screening algorithm with functional ANOVA, while the number of model that needs to be computed is at the order of p^2. In Linkletter et al. (2006), the size of the range parameters are used as indicators to decide whether the input is influential and a Metropolis Hasting algorithm is used for sampling the full posterior distribution. In Savitsky et al. (2011), a spike and slab prior is used for the transformation of the range parameters. However, the difficulty with model selection strategy comes from the computational burden, as the model space is 2^p, and each evaluation of the likelihood in the GaSP model requires $O(n^3)$ flops.

4.1.4 Goal of the new prior

We introduce a new prior here, for the inverse range parameters $\boldsymbol{\beta}$, that has the good properties of the reference prior discussed in Section 4.1.1 and avoids some of the drawbacks mentioned in Section 4.1.1. Indeed, this prior has following properties:

(i) (Dimension of inputs p.) The prior mass of $\boldsymbol{\beta}$ concentrates on smaller values when the dimension of the input space increases.

(ii) (Number of observations n.) The prior mass of $\boldsymbol{\beta}$ moves to larger values when a denser design is utilized.

(iii) (Scale of inputs.) The prior is invariant to location-scale transformations of inputs $\mathbf{x}$.

(iv) (Exponential tail rates on $\mathbf{R} \approx \mathbf{I}_n$.) When $\beta_l \to \infty$ for any l, the prior density decreases to 0 exponentially fast.

(v) (Polynomial tail rates on $\mathbf{R} \approx \mathbf{1}_n \mathbf{1}_n^T$.) When $\beta_l \to 0$ for all l, the prior density decreases to 0 with a polynomial rate.

(vi) (Identifiability of inert inputs.) When some but not all $\beta_l = 0$, while the others are finite, the prior density is positive.

(vii) (Closed form expression.) The prior is proper and has a closed form expression with explicit derivatives, normalizing constant and moments.

The first three properties are free of parameterization and is parallel to the good properties of reference prior. The fourth and fifth properties, are, indeed related to the parameterization and they match the tail rates of reference prior under $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ parameterization. Note that the reference prior with $\boldsymbol{\beta}$ parametrization is typically not robust. We define such the prior on $\boldsymbol{\beta}$ parameterization mainly because this form is easier to operate than the other parameterizations.

The sixth feature allows the use of the posterior mode estimation if some inert inputs are presented. As discussed before, a formal Bayesian model selection procedure has 2^p models to explore. A screening algorithm may scale the searching path to the order of p^2 but may still be too large to compute, since the computation of the likelihood requires $O(n^3)$ flops. Using this prior, we only need to evaluate one model, and hopefully can identify the inert outputs based upon. Since this prior is not designed for variable selection, we do not expect to replace the Bayesian model selection procedure, but following empirical results suggest that it is a good pre-experimental check of whether there might be inert inputs, before running a further computationally intensive algorithm.

4.2 Jointly robust prior and its properties

The new prior, called the *jointly robust (JR) prior* and denoted as $\pi^{JR}(\boldsymbol{\beta})$, has the following simple form:

$$\pi^{JR}(\beta_1, ..., \beta_p) = c(\sum_{l=1}^{p} C_l\beta_l)^a exp(-b\sum_{l=1}^{p} C_l\beta_l), \tag{4.1}$$

where c is a normalizing constant and a, b and C_l are positive parameters of the prior. Note that when $p = 1$, the jointly robust prior is a gamma distribution. This prior has the desired exponential tail rates and polynomial tail rates at its two limits, and will be seen to match other desirable features of the reference prior, such as scaling for dimension and the number of observations.

The choice of prior parameters is discussed in Section 4.2.2. a and b potentially depend on the dimension of inputs and the number of observations. C_l is a factor depending on the l^{th} dimension of input vectors, which is defined to take care of the scale of inputs in each dimension. The closed form expression and its properties are given in the following section.

4.2.1 Properties of the jointly robust prior

Some properties of the jointly robust prior are shown in this section and the proofs of this section are in Appendix C. First of all, the normalizing constant of the jointly robust prior is computed in the following Lemma 4.2.1.

Lemma 4.2.1. *(Normalizing constant.) The jointly robust prior is proper and has the normalizing constant* $c = \frac{b^{a+p}(p-1)!\prod_{l=1}^{p} C_l}{\Gamma(a+p)}$, *where* $\Gamma(\cdot)$ *is the gamma function.*

The prior mean and the prior variance of β_l, $l = 1, ..., p$ and $\sum_{l=1}^{p} C_l\beta_l$ are given in the following two lemmas.

Lemma 4.2.2. *(Prior mean.) The prior mean is as follows,*

(i.) $\mathrm{E}_{\pi^{JR}}[\beta_l] = \frac{a+p}{pC_l b}$, *for* $l = 1, ..., p$.

(ii.) $\mathrm{E}_{\pi^{JR}}[\sum_{l=1}^{p} C_l \beta_l] = \frac{a+p}{b}$.

Lemma 4.2.3. *(Prior variance.) The prior variance is as follows,*

(i.) $\mathrm{Var}_{\pi^{JR}}[\beta_l] = \frac{(a+p)(p^2+p+ap-a)}{p^2(p+1)C_l^2 b^2}$, *for* $l = 1, ..., p$.

(ii.) $\mathrm{Var}_{\pi^{JR}}[\sum_{l=1}^{p} C_l \beta_l] = \frac{a+p}{b^2}$.

The following lemma gives the tail rates of the jointly robust prior.

Lemma 4.2.4. *(Tail rates.)*

(i.) When any $\beta_l \to \infty$, $l \in E$, $E \in \{1, 2, ..., p\}$, *the logarithm of the jointly robust prior approximately decreases linearly with the rate* $-b\sum_{l \in E} \beta_l$;

(ii.) When $\beta_l \to 0$ *for all* $l = 1, ..., p$, *the logarithm of jointly robust prior decreases at the rate of* $a\log(\sum_{l=1}^{p} C_l \beta_l)$.

(iii.) When $\beta_l \to 0$, $l \in E$, $E \in \{1, 2, ..., p\}$, $\#E < p$, $\pi^{JR}(\beta_1, ..., \beta_p)$ *is positive.*

The first and second part of the reference prior matches the tail rates of the reference prior with two robust parameterizations discussed in Chapter 3. And the third part is an improvement, which allows the identification of inert inputs by the marginal posterior mode with the jointly robust prior. Besides, the closed form derivative of the jointly robust prior is also easy to compute.

4.2.2 On choice of the prior parameters

The parameters of the jointly robust prior in Equation (4.1) consist of the scale parameter a, the rate parameter b and input scaling parameters C_l, for $l = 1, ..., p$. The input scaling parameters C_l can be taken to be the mean of absolute distance of the l^{th} design inputs $|x_{il}^{\mathscr{D}} - x_{jl}^{\mathscr{D}}|$, for $1 \leqslant i, j \leqslant n$ and $i \neq j$. In this choice, under the

Table 4.1: The default choice of parameters in the jointly robust prior.

	form	values
a	a_0	$a_0 = 0.2$
b	$b_0 n^{-1/p}(a+p)$	$b_0 = 1$
C_l	mean of $\lvert x_{il} - x_{jl}\rvert$, for $1 \leqslant i, j \leqslant n$, $i \neq j$	/

transformation $\tilde{x}_{il} = \frac{x_{il}-c_0}{c_1}$, for all $i = 1, ..., n$, the new prior is $\tilde{\pi}^{JR}(\beta_1, ..., \beta_l, ..., \beta_p) = \pi^{JR}(\beta_1, ..., \beta_l/c_1, ..., \beta_p)$.

The choice of a and b can be tricky. a controls the tail rate at $\beta_l \to 0$ for all l. As we see in Table 3.3, reference prior have slightly different tail rates with different parameterization (though they are all polynomial). The rates at this limits are small so $a \in (0, 1]$ might be a reasonable choice. In the following results, we let $a = .2$.

Lemma 4.2.2 sheds some lights in specifying the default normalizing constant. To match the behavior of reference priors to the change of dimensions, we need to have $b \geqslant O(p)$. b should also be related to the number of design points n. When we have n random design points (e.g. from LHD) at a p dimensional input space, the average "effective" sample size for every dimension is $n^{1/p}$. It is helpful to first consider a fixed separable design, i.e. Lattice. If we have n design points and assume the number of design points is the same at each dimension, then every dimension has $n^{1/p}$ design points. The random design should approximately be the same as the lattice in a long run. Taking this into account, we let $b = n^{-1/p}(a + p)$. When n increases, b decreases, and consequently, the variance of the jointly robust prior increases at each dimension marginally and more mass moves to larger values of β_l.

We summarize the default choice of the prior parameters in the following table. $a_0 = 0.2$ and $b_0 = 1$ are specified as the default values, but one may choose other values to reflect their own prior belief at these parameters.

4.2.3 Marginal posterior mode estimation

The joint prior for all parameters in the GaSP model has the following form,

$$\pi(\boldsymbol{\theta}, \sigma^2, \boldsymbol{\beta}) \propto \frac{\pi^{JR}(\boldsymbol{\beta})}{\sigma^2},$$

where $\pi^{JR}(\boldsymbol{\beta})$ is the above jointly robust prior. We first marginalize out the mean and variance parameter, and obtain the marginal likelihood $\mathcal{L}(\mathbf{y}^{\mathscr{D}}|\beta_1, ..., \beta_p)$. We then estimate the inverse-range parameters by marginal posterior modes,

$$(\hat{\beta}_1, ..., \hat{\beta}_p) = \underset{\beta_1,...,\beta_p}{\operatorname{argmax}} \mathcal{L}(\mathbf{y}^{\mathscr{D}}|\beta_1, ..., \beta_p)\pi^{JR}(\beta_1, ..., \beta_p). \tag{4.2}$$

The following results show the marginal posterior mode estimation is robust for all correlation functions in Table 3.1.

Corollary 4.2.1. *The marginal posterior mode estimation of $(\beta_1, ..., \beta_p)^T$ in Equation (4.2) is robust for the product form of the power exponential, Matérn, spherical, and rational quadratic correlation functions listed in Table 3.1, if the prior parameters are specified in Table 4.1.*

The above results are a direct consequence of the tail rates of the marginal likelihood in Lemma 3.3.3 and the tail rates of the jointly robust prior in Lemma 4.2.4. In Chapter 3, we show the marginal posterior mode estimator with the reference prior by $\boldsymbol{\xi}$ yields the best performance of all methods we compare. Such estimation is,

$$(\hat{\xi}_1, ..., \hat{\xi}_p) = \underset{\xi_1,...,\xi_p}{\operatorname{argmax}} \mathcal{L}(\mathbf{y}^{\mathscr{D}}|\xi_1, ..., \xi_p)\pi^{R}(\xi_1, ..., \xi_p). \tag{4.3}$$

where $\pi^R(\xi_1, ..., \xi_p)$ is the reference prior parameterized by log inverse range parameters $\boldsymbol{\xi}$.

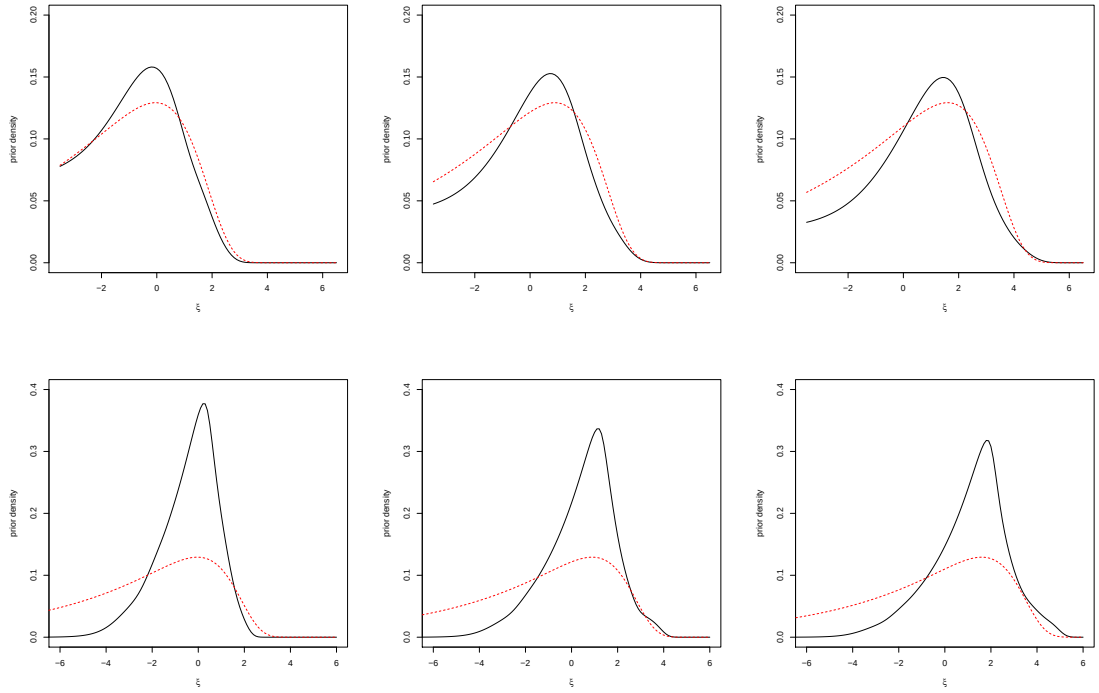

FIGURE 4.2: The reference prior $\pi^R(\boldsymbol{\xi})$ (black curves) and the jointly robust prior $\pi^{JR}(\boldsymbol{\beta})$ in the ξ space (red curves). Matérn correlation with $\alpha = 5/2$ is assumed in the first row and power exponential correlation with $\alpha = 1.9$ is assumed in the second row. From the left panel to the right panel, the number of design points is $n = 20$, $n = 50$ and $n = 100$. Designs are sampled from maximin Latin Hypercube at $[0, 1]$.

The difference in estimation in Equation (4.2) and Equation (4.3) is the prior $\pi^{JR}(\beta_1, ..., \beta_p)$ and $\pi^R(\xi_1, ..., \xi_p)$. We plots them in Figure 4.2 in the space $\boldsymbol{\xi}$ (again, because of better visual effect). Note that since we propose to estimate $\boldsymbol{\beta}$ with its mode in Equation 4.2, not a transformation defined in other parameterizations, the plot of $\pi^{JR}(\boldsymbol{\beta})$ is in the space of $\boldsymbol{\xi}$ without the Jacobian term. Such plot of $\pi^{JR}(\boldsymbol{\beta})$, could lead to an improper prior in the space of $\boldsymbol{\xi}$, and thus the plots in Figure 4.2 are up to a normalizing constant. But the point here, is to compare the difference between two marginal posterior mode estimation in Equation (4.2) and Equation (4.3).

As we can see in Figure 4.2, the spread of the prior is similar, especially the mode

between two priors is similar for different number of sample size. This relies on our choice of default prior parameters in Section 4.2.2.

4.2.4 Size of the inverse range parameters

Assume the inverse parameters $(\hat{\beta}_1, ..., \hat{\beta}_p)$ are estimated in Equation (4.2). When $\beta_l = 0$, the l^{th} input is not in the correlation function in GaSP model. In the estimation, exact zero estimation is less often to obtain, we propose to use the *estimated normalized inverse range parameters*

$$P_l = \frac{C_l \hat{\beta}_l}{\sum_{i=1}^{p} C_i \hat{\beta}_i}, \tag{4.4}$$

as an indicator of how importance the l^{th} input is. The involvement of C_l is to take into account the scale of different inputs. The part $\sum_{i=1}^{p} C_i \hat{\beta}_i$ in the denominator is the overall size of the estimation and the $C_l \hat{\beta}_l$ is the contribution by the l^{th} inputs. The sum is of P_l is 1 and the average P_l is $1/p$.

The size of the inverse range parameters has been discussed to indicate whether the input is influential or inert, but with a different prior (Linkletter et al. (2006)). However, the jointly robust prior yields much better results, as compared in the Section 4.4.

One may use a certain threshold of the estimated normalized inverse range parameters to predict whether the input is inert or not,

$$P_l \leqslant \frac{p_0}{p},$$

where p_0 may be chosen as a constant between 0 to 1. Such values could also depend on the number of observations, dimension of the inputs and the expected number of inputs to be chosen. We do not try to present a method for model selection. The point, here, is that the computation of the P_l does not take any

extra computation (as the posterior modes are typically needed for building a GaSP model), and can serve as a indicator to tell an input is inert or not.

4.3 Jointly robust prior with a noise

When a nugget is present in the GaSP model, such that $\tilde{\mathbf{R}} = \mathbf{R} + \eta\mathbf{I}$, the jointly robust prior can be extended for the case with a nugget in a natural way,

$$\pi^{JR}(\beta_1, ..., \beta_p, \eta) = \tilde{c}(\sum_{l=1}^{p} C_l\beta_l + \eta)^a exp(-b(\sum_{l=1}^{p} C_l\beta_l + \eta)), \tag{4.5}$$

where $\tilde{c} = \frac{b^{a+p+1}p!\prod_{l=1}^{p} C_l}{\Gamma(a+p+1)}$. The properties of the jointly robust prior can be extended to the one with a noise by letting $\beta_{p+1} = \eta$ and $C_{p+1} = 1$. We also propose to use a local-scale prior to first marginalize out $(\boldsymbol{\theta}, \sigma^2)$ and estimate the inverse range parameters and noise-variance ratio parameter by

$$(\hat{\beta}_1, ..., \hat{\beta}_p, \hat{\eta}) = \underset{\beta_1,...,\beta_p,\eta}{\operatorname{argmax}} \mathcal{L}(\mathbf{y}^{\mathscr{D}}|\beta_1, ..., \beta_p, \eta)\pi^{JR}(\beta_1, ..., \beta_p, \eta). \tag{4.6}$$

Similarly, such mode estimation is also robust, stated in the following Corollary.

Corollary 4.3.1. *The marginal posterior mode estimation of $(\hat{\beta}_1, ..., \hat{\beta}_p, \hat{\eta})^T$ in Equation (4.6) is robust for the product form of the power exponential, Matérn, spherical, and rational quadratic correlation functions listed in Table 3.1, if the prior parameters are specified in Table 4.1.*

The above result is a direct consequence of tail rates of the marginal likelihood of the GaSP model with a noise in Lemma 3.4.1 and the tail rates of the jointly robust prior in Lemma 4.2.4.

Table 4.2: Average MSE of the three estimation procedures for the five experimental functions. The sample size is $n = 20p$ for the Higdon function and $n = 10p$ for the others. Designs are generated by maximin LHD. The baseline MSE is 0.52, 3.8, 52, 0.71, and 24 for these five functions if only the mean of the training output is used for the predictions.

	Robust GaSP $\boldsymbol{\xi}$	Jointly robust prior	DiceKriging
1-dim Higdon	1.09×10^{-4}	1.09×10^{-4}	1.43×10^{-4}
2-dim Lim	6.48×10^{-3}	6.72×10^{-3}	8.47×10^{-3}
3-dim Pepelyshev	8.40×10^{-2}	8.44×10^{-2}	7.93×10^{-1}
4-dim Park	1.10×10^{-4}	1.08×10^{-4}	6.58×10^{-4}
5-dim Friedman	2.65×10^{-2}	2.60×10^{-2}	4.51×10^{-1}

4.4 Numerical results

4.4.1 Predictive results

Example 4.4.1. *(Many test functions continued.) We first redo the test of 5 different functions shown in Section 3.5 in Chapter 3. The setting are all kept the same.*

The MSE table are shown in Table 4.2. The MSE results of the jointly robust prior are almost the same as the reference prior with $\boldsymbol{\xi}$, the parameterization that leads to the least MSE for this set of testing functions so far.

The similar predictive results come from the similar estimation of inverse range parameters. We plot such estimation of 3-dim Pepelyshev function, 4-dim Park function and 5-dim Friedman function in Figure 4.3. The posterior estimation of $C_l\hat{\beta}_l$ plotted in the first row (by reference prior with $\boldsymbol{\xi}$ parameterization) and the second row (by jointly robust prior) are very similar. This is because the posterior modes are used in both estimation, and the penalty term induced from two different priors has similar spread out of prior mass across different dimension and different number of the design. It is not a surprise that we obtain similar estimation of the inverse range parameters, and it further leads to similar predictive errors, as shown

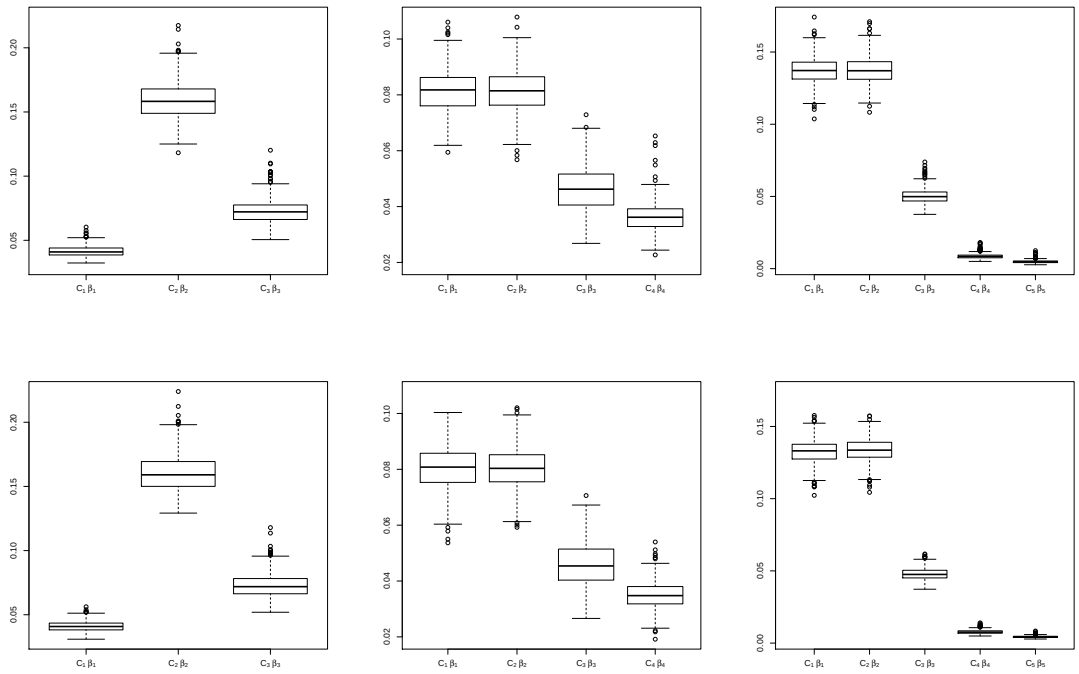

FIGURE 4.3: Box plot of $C_l\hat{\beta}_l$ of each experiment (out of $N = 500$ random design) by the posterior modes of the reference prior with $\boldsymbol{\xi}$ parameterization (the first low) and by the jointly robust prior (the second row). From the left panel to the right panel, The test functions are 3-dim Pepelyshev function, 4-dim Park function and 5-dim Friedman function.

in Table 4.2.

The computational time between two posterior modes estimations, however, is different. As shown in Figure 4.4, the computational time of posterior mode estimation with the jointly robust prior is a lot smaller. This is because the jointly robust prior has a explicit form of derivatives, while the derivative of the reference prior is computationally intensive. The derivatives of the reference prior are computed numerically, which relies on evaluations of the posterior much more times.

Example 4.4.2. *(Borehole function continued.) Borehole function is discussed in Section 3.5 in Chapter 3. This function is recently discussed in Chen et al. (2016) for the topic of whether the basis mean function should be a constant or should include*

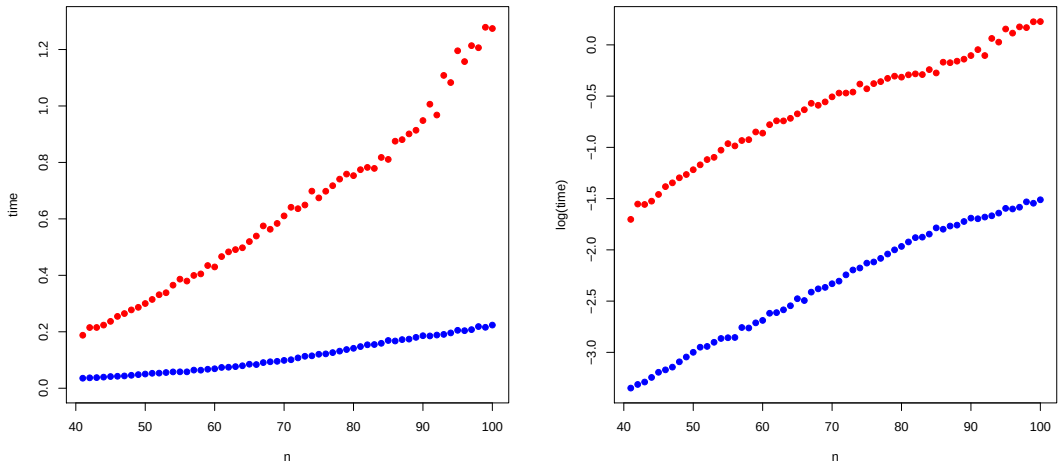

FIGURE 4.4: Time in second and log-time in log(second) between posterior mode estimation with reference prior (red) and with jointly robust prior (blue) for 5-dim Friedman function with different number of observations.

linear terms of inputs. The dataset given by Jerome Sacks and Hao Chen (through personal communication). 27 maximin LHD design points at 8 dimensional input space are generated and 25 random permutation of these design points are used to construct the GaSP model in each experiment, with $n^* = 5000$ *held-out points for testing.*

In Section 3.5 in Chapter 3, we only use 5 influential inputs with a small noise to build the GaSP model, while three inert inputs are not used. The identification of the inert inputs in many applications, however, could be hard. Indeed, in Chen et al. (2016), all inputs are incorporated in the correlation function.

Here we examine the performance of different methods using all inputs in the correlation function shown in the first row of the Figure 4.5, and using 5 influential inputs with a nugget, shown in the second row of the Figure 4.5. The same dataset reported in Chen et al. (2016) is used. In the first row, the left panel and right panel compare the predictive errors evaluated by normalized-RMSE with constant mean function (i.e. $h(\mathbf{x}) = 1$) and with full linear terms in the mean function (i.e.

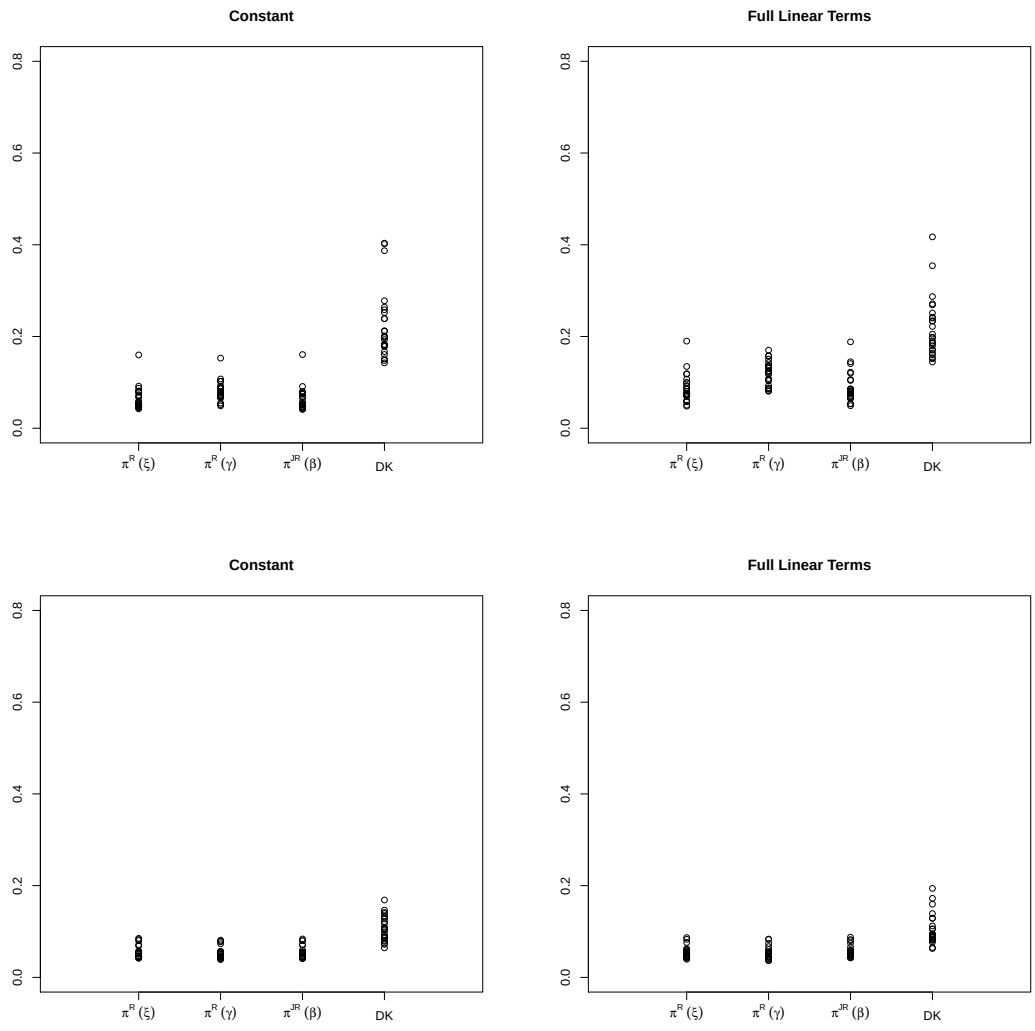

FIGURE 4.5: Plotsof Normalized-RMSE of prediction for the held-out data for each of 25 permutation of maximin random design. The first row is for the case where 8 inputs are used to build the GaSP model, while the second row is for the case that 5 influential inputs are used with a noise. The left panel column shows the results with a constant mean function ($h(\mathbf{x}) = 1$), while the second panel shows results with full linear terms as mean function ($\mathbf{h}(\mathbf{x}) = (1, \mathbf{x})$). The methods from the left to the right are from the marginal posterior mode estimations by the reference prior with $\boldsymbol{\xi}$ parameterization, $\boldsymbol{\gamma}$ parameterization, the jointly robust prior and the DiceKriging package. The number of design points is $n = 27$ in each experiment and Matérn correlation function with $\alpha = 2.5$ are used for all methods.

$\mathbf{h}(\mathbf{x}) = (1, \mathbf{x})$) in the GaSP model. The results by the posterior mode estimation with the jointly robust prior $\pi^{JR}(\beta)$ is shown as the third one in each figure. Compared

with the results reported in Figure 2 in Chen et al. (2016) (in which MLE estimation of the parameters is used), the results with posterior mode estimation with the jointly robust prior seem to outperform, especially in the case where full linear terms are utilized in the mean basis function. Chen et al. (2016) reported normalized-RMSE larger than 0.2 for every experiment, when the mean basis function includes full linear terms, while we don't have any normalized-RMSE larger than 0.2 using posterior mode estimation with the jointly robust prior. This is partly caused by the robustness argument we explored in Chapter 3. MLE estimation is not robust, while the first three methods here are all robust.

Indeed, from the first row of the figure, using constant mean function seems to be better than the one using full linear terms in general. This is because only very limited number of design points ($n = 27$) is used, while there are 8 range parameters, 9 mean parameters (if full linear terms are used as the mean function) and 1 variance parameter in the model, which is for estimation using any method. As a result, we notice that the computation is unstable when a linear term is used, as local modes appear in the posterior. But even in such difficult scenario, the posterior mode estimation with the JR prior seems to be fine. In most of experiments, normalized-RMSE are less than 0.1.

The second row of Figure 4.5 shows results that only 5 influential inputs are used with a noise term. Note the results are now better than the ones at the first row. When very limited number of design points are used, the GaSP model with only influential inputs and a noise term performs better than the one with all inputs in this scenario. This is not a particularly unlikely scenario, as is also the case in PP GaSP emulator for TITAN2D computer model reported.

The estimated normalized inverse range parameters of the first three methods are shown in Figure 4.6. P_l is estimated to be very small by the posterior mode with the jointly robust prior for 3 inert inputs, while they are clearly estimated to be too

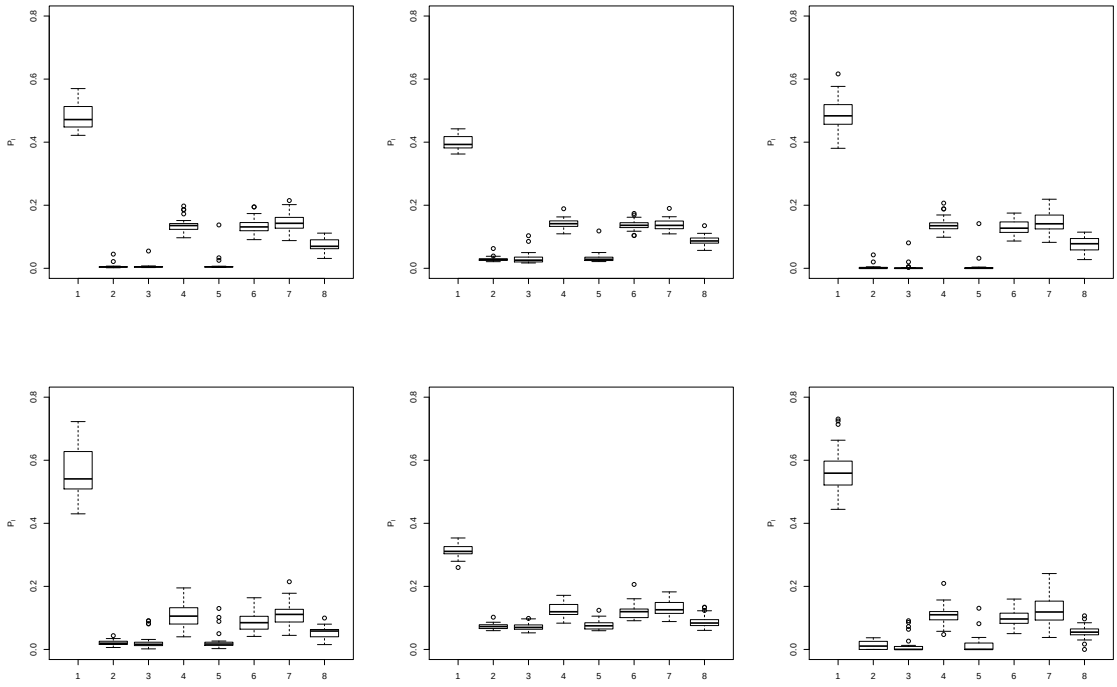

FIGURE 4.6: Box plot of the estimated normalized inverse range parameters P_l of each experiment of the Borehole function by the posterior modes of the reference prior with $\boldsymbol{\xi}$ parameterization (left), with $\boldsymbol{\gamma}$ parameterization (middle) and by the jointly robust prior (right), when all inputs are used in the correlation function. The first row is the case with a constant mean basis and the second is the case with full linear terms.

large by the posterior with the reference prior under $\boldsymbol{\gamma}$ parameterization, as the prior density forbids $\beta_l \to 0$. The reference prior with $\boldsymbol{\xi}$ seems to estimate P_l small enough to identify the inert inputs, when $h(\mathbf{x}) = 1$. Actually, the estimated P_2, P_3 and P_5 by the posterior mode of the reference prior with $\boldsymbol{\xi}$ parameters are around 3955, 4699 and 3615 times the values of with posterior mode estimation with the jointly robust prior on average with $h(\mathbf{x}) = 1$. The separate estimation of different types of inputs by the posterior mode with the jointly robust prior allows the identifiability of the inert inputs explored below.

4.4.2 Identification of inert inputs

We first test the performance of two examples reported in Linkletter et al. (2006).

Example 4.4.3.

$$Y = 0.2X_1 + 0.2X_2 + 0.2X_3 + 0.2X_4 + \epsilon,$$

where $\epsilon \sim N(0, 0.05^2)$ *and* $X_l \in [0, 1]$, $l = 1, .., 4$. *6 completely noise input variables are also added in this example.*

Example 4.4.4.

$$\begin{aligned} Y = {} & 0.2X_1 + 0.2/2X_2 + 0.2/4X_3 + 0.2/8X_4 \\ & + 0.2/16X_5 + 0.2/32X_6 + 0.2/64X_7 + 0.2/128X_8 + \epsilon, \end{aligned}$$

where $\epsilon \sim N(0, 0.05^2)$ *and* $X_l \in [0, 1]$, $l = 1, .., 4$. *2 completely noise input variables are also added in this example.*

In both Example 4.4.3 and Example 4.4.4, the number of design points is $n = 54$ and the correlation function is assumed to be Gaussian (same as in Linkletter et al. (2006)). In both examples, the function is linear, however, in the GaSP model we only use the constant mean function $h(\mathbf{x}) = 1$, pretending that we do not know the linear trend of the real function. The purpose of Example 4.4.3 is to see whether GaSP can identify the signal, and the purpose of Example 4.4.4 is designed to see how small the signal that GaSP can identify. A small noise is added to represent the error induced by the numerical solution of the computer model.

$N = 1,000$ random designs from the maximin LHD design are generated and the estimated normalized inverse range parameters P_l are shown in Figure 4.7. In the left figure, it is clear that the first four inputs are much more important than the rest of input. Indeed these are 4 signals while the other are noises. The second

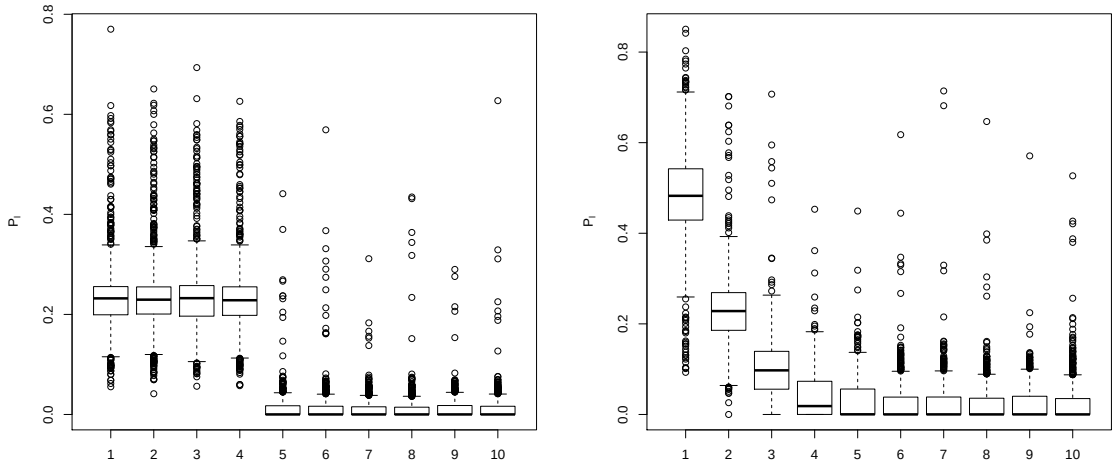

FIGURE 4.7: Estimated normalized inverse range parameters P_l in Example 4.4.3 (left) and Example 4.4.4 (right).

Table 4.3: Proportion of times each input is identified as important in Example 4.4.3 and Example 4.4.4. JR prior denotes the posterior mode estimation with the jointly robust prior and P_l with different p_0 is used to identify the inert inputs. RDVS selection is a method introduced in Linkletter et al. (2006) with different choice of percentile (PT).

Example 4.4.3	1	2	3	4	5	6	7	8	9	10
JR prior, $p_0 = 1$	.979	..985	.987	.982	.011	.013	.006	.007	.005	.008
JR prior, $p_0 = .75$	.996	.996	.999	.997	.014	.014	.007	.008	.006	.010
JR prior, $p_0 = .5$	1	.999	1	1	.042	.045	.039	.034	.036	.038
RDVS, 5^{th} PT	.619	.618	.717	.631	.030	.034	.021	.074	.051	.051
RDVS, 10^{th} PT	.852	.855	.910	.880	.061	.064	.053	.137	.076	.102
RDVS, 15^{th} PT	.947	.954	.973	.955	.079	.091	.080	.173	.108	.135
Example 4.4.4	1	2	3	4	5	6	7	8	9	10
JR prior, $p_0 = 1$	.999	.955	.482	.142	.086	.052	.046	.040	.035	.058
JR prior, $p_0 = .75$	1	.981	.638	.241	163	.107	.103	.100	.109	.116
JR prior, $p_0 = .5$	1	.996	.775	.379	.276	.208	.193	.188	.208	.185
RDVS, 5^{th} PT	.679	.180	.062	.025	.016	.023	.017	.031	.009	.036
RDVS, 10^{th} PT	.889	.379	.133	.058	.034	.051	.035	.067	.030	.094
RDVS, 15^{th} PT	.959	.540	.217	.092	.061	.098	.065	.107	.063	.149

figure shows that estimated normalized inverse range parameters P_l can successfully identify the largest 3 to 4 signals.

A detailed comparison with results in Linkletter et al. (2006) are shown in Ta-

ble 4.3. In both case, using P_l under the posterior mode with the jointly robust prior seems to have smaller false positives and false negatives in both examples, compared with the RDVS selection method in Linkletter et al. (2006).

The cut-off value p_0 can be hard to define. However, one major feature the variable selection in the computer model is that typically all inputs are real signals. The task is then not to identify the true signals, but to identify what set of inputs are more important than the others. This seems successful in Example 4.4.3 and Example 4.4.4, as importance of the factors are correctly ordered. This means to have the correct order of importance is the key issue. In the following Example 4.4.5, we test whether the method can have a correct order of signals to noises.

Example 4.4.5. *We test on these functions.*

i. $Y = \frac{1}{6}[(30 + 5X_1 sin(5X_1))(4 + exp(-5X_2)) - 100] + \epsilon$, $X_i \in unif(0, 1)$, $i = 1, ..., 7$, $\epsilon \in N(0, 0.3^2)$.

ii. $Y = 4(X_1 - 2 + 8X_2 - 8X_2^2)^2 + (3 - 4X_2)^2 + 16\sqrt{X_3 + 1}(2X_3 - 1)^2 + \epsilon$, $X_i \in unif(0, 1)$, $i = 1, ..., 6$, $\epsilon \in N(0, 0.05^2)$.

iii. $Y = \frac{2}{3}\exp(X_1 + X_2) - X_4 \sin(X_3) + X_3 + \epsilon$, $X_i \in unif(0, 1)$, $i = 1, ..., 8$, $\epsilon \in N(0, 0.15^2)$.

iv. $Y = 10sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \epsilon$, $X_i \in unif(0, 1)$, $i = 1, ..., 10$, $\epsilon \in N(0, 0.2^2)$.

The first function is a slightly modified version of Lim et al. (2002) by adding 5 noisy inputs and a small Gaussian noise, the second one being a modified version in Dette and Pepelyshev (2010), the third one modifies Park (1991) and the fourth one modifies Friedman (1991).

The results are recorded in Table 4.4. In this table, it records how many sample is needed to have correct signal-noise order larger than 85%. From the left to

Table 4.4: Minimal sample size to have correct signal-noise order proportion larger than 85%. The number is the sample size that is tested and the correct signal-noise order proportion is recorded in the bracket tested with this sample size. $N = 200$ experiments are implemented.

	JR prior	Sobol GaSP	Sobol	Sobol2007 S	Sobol2007 T
Eg (i)	$< 20(.960)$	$< 20(.970)$	$> 130(.835)$	$> 130(.825)$	$> 130(.820)$
Eg (ii)	$< 35(.965)$	$> 40(.290)$	$> 20,000(.800)$	$> 20,000(.780)$	$> 600(.820)$
Eg (iii)	$< 35(.905)$	$> 35(.795)$	$> 4,000(.695)$	$> 4,000(.675)$	$> 4,000(.765)$
Eg (iv)	$< 35(.920)$	$> 35(.780)$	$> 1,000(.810)$	$> 1,000(.805)$	$> 1,000(.790)$

Table 4.5: Average Computational time in seconds by the posterior mode with the jointly robust prior and by Sobol GP for one experiment. The inputs are 5 dimensional with the different number of noisy inputs. $n = 35$ is used for each comparison.

	5 noises	6 noises	7 noises	8 noises
JR prior	0.108	0.113	0.128	0.136
Sobol GaSP	117.29	143.53	175.02	215.74

right, it records the methods using P_l to rank the importance of the inputs, Sobol GaSP (Oakley and O'Hagan (2004); Le Gratiet et al. (2014)), Sobol (Sobol' (1990)), Sobol2007 S and Sobol2007 T (Sobol' et al. (2007)). All Sobol methods are coded in 'Sensitivity' R package (Pujol et al. (2007)).

As shown in Table 4.4, Sobol method (and its variants) using Monte Carlo method needs much more computer model runs to identify the signals, while Sobol GaSP only needs much less runs, consistent with the previous study in Oakley and O'Hagan (2004).

The result using P_l is similar (or slightly better) to Sobol GaSP methods. This might be caused by two reasons. For one thing, as the Sobol GP first fits the GaSP model using the DiceKriging Package, which leads to the inferior estimation of parameters as shown in these two Chapters. For another thing, as now we have couple of signals, so higher order indices might be needed as discussed in Section 4.1.3.

The computation time by the posterior mode estimation with the jointly robust

prior and the Sobol GaSP method is shown in Table 4.5. Sobol GaSP is computationally intensive when the number of (noisy) inputs increases, while the posterior modes evaluation with the jointly robust prior is a lot faster, because only the posterior mode of the full model is needed.

5

Nonseparable GaSP: a unified view and its computational strategy

In Chapter 2, we discussed the PP GaSP emulator for the TITAN2D computer model, where the output at each coordinate is a function over the input space. The output is a $k \times n$ matrix, in which k is the number of space-time coordinates and n is the number of observed runs of the computer model. Since $k \gg n$, the major computational achievement by PP GaSP emulator is that the order of whole computational flops is linear in k so the emulator can be implemented efficiently. This chapter focuses on another scenario, in which the number of functions (k) is relatively small, compared to the number of inputs (n). In a standard GaSP model discussed in Chapter 1, the computational flops are at the order of $O(n^3)$ due to the computation of the inverse of the covariance matrix in the likelihood, which is a barrier when n is large (here $n \approx 10^6$). This chapter introduces a computational strategy that is linear with regard to n using a data augmentation strategy, instead of computing the inverse directly. Moreover, a unified view between the linear regression model, the separable GaSP model and the nonseparable GaSP model will be introduced. An application

to the methylation levels interpolation problem will be given by synthesizing different sources of information for prediction, using the nonseparable GaSP model with the fast computing strategy.

5.1 Literature Review and Motivations

To begin with, let us consider the following matrix data $\mathbf{Y}$, partitioned into 4 blocks,

$$\mathbf{Y}_{[K\times N]} = \begin{pmatrix} \mathbf{y}(\mathbf{x}^{\mathscr{D}}) & \mathbf{y}(\mathbf{x}^*) \\ \mathbf{y}^*(\mathbf{x}^{\mathscr{D}}) & \mathbf{y}^*(\mathbf{x}^*) \end{pmatrix}_{K\times N} \tag{5.1}$$

where the dimension of $\mathbf{y}(\mathbf{x}^{\mathscr{D}})$ is $k \times n$, $\mathbf{y}(\mathbf{x}^*)$ being $k \times n^*$, $\mathbf{y}^*(\mathbf{x}^{\mathscr{D}})$ being $k^* \times n$ and $\mathbf{y}^*(\mathbf{x}^*)$ being $k^* \times n^*$, with $K = k + k^*$ and $N = n + n^*$. In emulating TITAN2D computer model discussed in Chapter 2, the task is to predict the computer model at some new runs $\mathbf{x}^*$, based on some previously evaluated runs, i.e. $(\mathbf{y}(\mathbf{x}^*)^T, \mathbf{y}^*(\mathbf{x}^*)^T)^T$ are not known while $(\mathbf{y}(\mathbf{x}^{\mathscr{D}})^T, \mathbf{y}^*(\mathbf{x}^{\mathscr{D}})^T)^T$ are observed.

There are also many other scenarios, the task of which is to make prediction only on the block $\mathbf{y}^*(\mathbf{x}^*)$, while $\mathbf{y}(\mathbf{x}^{\mathscr{D}})$, $\mathbf{y}^*(\mathbf{x}^{\mathscr{D}})$ and $\mathbf{y}(\mathbf{x}^*)$ are observed, (see e.g. Conti and O'Hagan (2010); Farah et al. (2014) for applications in computer models).

In this study, we discuss another application, which is to interpolate DNA methylation levels at unobserved sites. DNA methylation is an epigenetic modification of DNA, which is shown to be closely involved in the process of DNA replication, gene transcription with strong association with development, aging, and cancer (Das and Singal (2004); Scarano et al. (2005); Cedar and Bergman (2012)). The methylation levels can now be quantified at each CpG site, a region of DNA where a cytosine nucleotide is followed by a guanine nucleotide in the linear sequence of bases along its 5 to 3 direction. Single-site DNA methylation levels can be quantified by whole-genome bisulfite sequencing (WGBS), in which approximately 26 million (out of 28 millions) sites are evaluated in the human genome. However, WGBS is expensive,

limited to conversion bias and hard to perform in certain genomic regions (Zhang et al. (2015)). This motivates the study of other sequencing methods, one of which is Illumina HumanMethylation450 (henceforth Methylation450) BeadChip, measuring the DNA methylation at around $482,000$ CpG sites, less than 2% of total number of CpG sites.

The goal of this study, is to interpolate the WGBS data of a participant, using the Methylation450 data (which is much cheaper and easier to obtain than the WGBS data) and the previously assayed WGBS data of other participants. Let $\mathbf{x}^{\mathscr{D}}$ be the CpG sites of the Methylation450 data and $(\mathbf{x}^{\mathscr{D}}, \mathbf{x}^*)$ be the CpG sites of the whole WGBS data. This draws a clear connection to the matrix data discussed above, as the WGBS data of the first k participants (the block $\mathbf{y}(\mathbf{x}^{\mathscr{D}})$, $\mathbf{y}(\mathbf{x}^*)$) is observed and only the Methylation450 data ($\mathbf{y}(\mathbf{x}^*)$), is examined for the rest of k^* participants. The task is then to predict $\mathbf{y}^*(\mathbf{x}^*)$, the methylation levels that have not be examined in the Methylation450 data.

Denote the methylation levels as $y_i(s_j)$ (a $[0,1]$ real valued output, which is the proportion of probes) that are methylated of the i^{th} participant at the j^{th} CpG site. In this denotation, s_j is treated as an input. In the previous literature, the methylation levels at nearby CpG sites are shown to be correlated to each other (sometimes called co-methylation (Zhang et al. (2015))), particularly when the distance is within 2 kilobase (kb), while such correlation gradually decays when two CpG sites become far away (Eckhardt et al. (2006)). Hence it might be reasonable to model such correlation between CpG sites under the formal GaSP paradigm.

On the other hand, the methylation levels at a CpG site are also similar across different participants. Note that in this scenario, the covariance matrix between each row of the matrix in Equation (5.1) potentially matters, as the prediction can now be conditional on $\mathbf{y}(\mathbf{x}^*)$. In comparison, Theorem 2.6.1 indicates that that off-diagonal terms of the row-wise covariance matrix barely affects the prediction, if one

does not observe $\mathbf{y}(\mathbf{x}^*)$, which supports the independence assumption between each spatial coordinate in the PP GaSP model. This additional block of observations $\mathbf{y}(\mathbf{x}^*)$ motivates the study of modeling the correlation between participants; here we use a nonseparable GaSP model, discussed in Section 5.2.

Before moving on, we want to emphasize the potential computational challenge in this problem. Denote $K = k + k^*$ as the total number of people involved in this study, and $N = n + n^*$ as the total number of interested CpG sites. In the whole WBGS dataset, there are about 2.8×10^7 CpG sites and even in the Methylation450K data, there are roughly 4.5×10^5 CpG sites, both of which are really large. In comparison, the number of participants is relatively small, currently we only have $k = 21$ observations for the WBGS dataset and $k^* = 100$ observations for the Methylation450K dataset. It is the key to scale down the computation linearly in terms of the number of CpG sites in this application.

In the following, we start from two possible linear regression models in Section 5.1.1, and we discuss a separable GaSP model in Section 5.1.2 for the possibility of including correlation structure into the model. In Section 5.2, we introduce a nonseparable GaSP model that unifies these models, with a fast computational algorithm highlighted in Section 5.3.

5.1.1 Two linear regression strategies

We begin by discussing two multiple regression ways for data in Equation (5.1). Denote $\mathbf{Y}(\mathbf{s}^{\mathscr{D}}) := (\mathbf{y}(\mathbf{s}^{\mathscr{D}})^T; \mathbf{y}^*(\mathbf{s}^{\mathscr{D}})^T)^T$ and $\mathbf{Y}(\mathbf{s}^*) := (\mathbf{y}(\mathbf{s}^*)^T; \mathbf{y}^*(\mathbf{s}^*)^T)^T$ as $K \times n$ and $K \times n^*$ blocks of methylation levels respectively; One simple way is to model the data as a linear regression model at each CpG site j,

$$\mathbf{Y}(s_j^*) = \mathbf{H}(s_j^*)\boldsymbol{\beta}_j + \boldsymbol{\epsilon}_j, \tag{5.2}$$

where $\mathbf{Y}(s_j^*) = (\mathbf{y}(s_j^*)^T; \mathbf{y}^*(s_j^*)^T)^T$ is a $K \times 1$ vector, the methylation levels of K people at the j^{th} CpG site and $\epsilon_j \sim N(0, \sigma_j^2)$ as an independent noise. $\mathbf{H}(s_j^*)$ is the covariate matrix at site s_j. In Zhang et al. (2015), 124 site-specific features including methylation levels at nearby CpG sites $\mathbf{Y}(\mathbf{s}_{ne(j)})$ are used as covariates; some machine learning strategies for estimating the regression parameters are compared and the random forest (Liaw and Wiener (2002)) is suggested for the prediction of methylation levels at unobserved CpG sites. We call this strategy as (linear) regression model by site in general. This strategy treats the outcomes at each site independently, while in fact methylation levels of different people at one CpG site is usually similar to each other. The similarity can be interpreted as the correlation between samples at each CpG site due to the shared biological nature. However, the independent assumption between people of Equation (5.2) does not take such information into the model and can thus be imprecise in inference and prediction, as we shall see in Section 5.4 .

Another way is to form the methylation levels of each of k^* participants as an $N \times 1$ vector $\mathbf{y}_i^*(\cdot)^T := (\mathbf{y}_i^*(\mathbf{s}^{\mathscr{D}}); \mathbf{y}_i^*(\mathbf{s}^*))^T$ and use a linear regression model for each participant,

$$\mathbf{y}_i^*(\cdot)^T = \mathbf{H}_i(\cdot)^T \boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i,\ i = 1, ..., k^*, \tag{5.3}$$

where $\mathbf{H}_i(\cdot)^T$ is the covariate matrix for the i^{th} person, $i = 1, .., k^*$. If we only consider using the methylation levels of the first k people (without other feature data), whose WBGS methylation levels are fully observed, we have $\mathbf{H}_i(\cdot)^T = (\mathbf{y}(\mathbf{s}^{\mathscr{D}}); \mathbf{y}(\mathbf{s}^*))^T$, an $N \times k$ matrix. The least squares (LS) estimator of $\boldsymbol{\beta}_i$ is

$$\hat{\boldsymbol{\beta}}_i = \left\{\mathbf{y}(\mathbf{s}^{\mathscr{D}})\mathbf{y}(\mathbf{s}^{\mathscr{D}})^T\right\}^{-1} \mathbf{y}(\mathbf{s}^{\mathscr{D}})\mathbf{y}_i^*(\mathbf{s}^{\mathscr{D}})^T,$$

and thus the prediction of the methylation levels at unobserved CpG sites of the i^{th} person, $i = 1, .., k^*$, is

$$\hat{\mathbf{y}}_i^*(\mathbf{s}^*)^T = \mathbf{y}(\mathbf{s}^*)^T \left\{\mathbf{y}(\mathbf{s}^{\mathscr{D}})\mathbf{y}(\mathbf{s}^{\mathscr{D}})^T\right\}^{-1} \mathbf{y}(\mathbf{s}^{\mathscr{D}})\mathbf{y}_i^*(\mathbf{s}^{\mathscr{D}})^T, \tag{5.4}$$

or equivalently,

$$\hat{\mathbf{y}}_i^*(\mathbf{s}^*) = \mathbf{y}_i^*(\mathbf{s}^{\mathscr{D}})\mathbf{y}^T(\mathbf{s}^{\mathscr{D}}) \left\{\mathbf{y}(\mathbf{s}^{\mathscr{D}})\mathbf{y}(\mathbf{s}^{\mathscr{D}})^T\right\}^{-1} \mathbf{y}(\mathbf{s}^*). \tag{5.5}$$

We call this model as the regression strategy by participant. This treats sites as observations, thus potentially one can utilize n observations, which is big enough to have small variance of the estimator of $\boldsymbol{\beta}_i$. For simplicity of the comparison, only methylation levels of the first k people are used as regressors for now, but feature data can be incorporated into the regressor $\mathbf{H}_i(\cdot)^T$, as discussed in Section 5.2.3.

As opposed to the first regression strategy, the correlation between people at each CpG sites is modeled through the covariates, while sites of the methylation levels are treated as independent observations. This could also be a potential problem for inference because as mentioned before, methylation levels at nearby sites are correlated to each others (Eckhardt et al. (2006); Zhang et al. (2015)). The message here, is that both regression strategies only model a part of correlation, while treating the other part as independent observations, which is possibly inferior in making prediction.

5.1.2 From linear regression to separable GaSP model

The linear regression model (5.2) and (5.3) discussed above are both special cases of modeling the data by the matrix normal distribution. To see this, let

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}(\mathbf{s}^{\mathscr{D}}) & \mathbf{y}(\mathbf{s}^*) \\ \mathbf{y}^*(\mathbf{s}^{\mathscr{D}}) & \mathbf{y}^*(\mathbf{s}^*) \end{pmatrix} \sim \mathcal{N}_{K,N}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}) \tag{5.6}$$

where $\mathcal{N}_{K,N}$ is the matrix normal distribution with $K \times N$ dimension and $\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{00} & \mathbf{\Sigma}_{0*} \\ \mathbf{\Sigma}_{*0} & \mathbf{\Sigma}_{**} \end{pmatrix}$ is a $K \times K$ covariance matrix, with $\mathbf{\Sigma}_{00} = Cov(\mathbf{y}(s), \mathbf{y}(s))$ a $k \times k$ covariance matrix, $\mathbf{\Sigma}_{0*} = Cov(\mathbf{y}(s), \mathbf{y}^*(s))$ a $k \times k^*$ covariance matrix and $\mathbf{\Sigma}_{**} = Cov(\mathbf{y}^*(s), \mathbf{y}^*(s))$ an $k^* \times k^*$ covariance matrix, for every site s.

By properties of the multivariate normal distribution, it can easily be shown that the conditional mean of $\mathbf{y}^*(\mathbf{s}^*)$ given the other three blocks data is $\hat{\mathbf{y}}^*(\mathbf{s}^*) = \mathbf{\Sigma}_{*0}\mathbf{\Sigma}_{00}^{-1}\mathbf{y}(\mathbf{s}^*)$. Take the maximum likelihood estimation (MLE) applying to $\mathbf{Y}(\mathbf{s}^{\mathscr{D}})$, it follows $\hat{\mathbf{\Sigma}}_{00} = (\mathbf{y}(\mathbf{s}^{\mathscr{D}})\mathbf{y}(\mathbf{s}^{\mathscr{D}})^T)/n$ and $\hat{\mathbf{\Sigma}}_{*0} = (\mathbf{y}^*(\mathbf{s}^{\mathscr{D}})\mathbf{y}(\mathbf{s}^{\mathscr{D}})^T)/n$. Plugging these estimates, the conditional mean of $\mathbf{y}^*(\mathbf{s}^*)$ is

$$\hat{\mathbf{y}}^*(\mathbf{s}^*) = \hat{\mathbf{\Sigma}}_{*0}\hat{\mathbf{\Sigma}}_{00}^{-1}\mathbf{y}(\mathbf{s}^*) = \mathbf{y}^*(\mathbf{s}^{\mathscr{D}})\mathbf{y}^T(\mathbf{s}^{\mathscr{D}})\left\{\mathbf{y}(\mathbf{s}^{\mathscr{D}})\mathbf{y}(\mathbf{s}^{\mathscr{D}})^T\right\}^{-1}\mathbf{y}(\mathbf{s}^*), \tag{5.7}$$

the same as the prediction for model(5.3), if $\mathbf{H}_i(\cdot)^T = (\mathbf{y}(\mathbf{s}^{\mathscr{D}}); \mathbf{y}(\mathbf{s}^*))^T$. The prediction of each participant in model(5.3) is thus equivalently given as the i^{th} row of the Equation (5.7). The prediction of Model(5.2) can also be formulated as the matrix normal likelihood with the MLE estimation of the covariance matrix, but we do not go into details.

To incorporate the correlation between sites into model(5.6), it is natural to define the model

$$\mathbf{Y} \sim \mathcal{N}_{K,N}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{\Lambda}), \tag{5.8}$$

where $\mathbf{\Sigma}$ is the covariance matrix in model(5.6) and $\mathbf{\Lambda} = \begin{pmatrix} \mathbf{R} & \mathbf{r} \\ \mathbf{r}^T & \mathbf{R}^* \end{pmatrix}$, with $\mathbf{R} = Corr(\mathbf{Y}_i(\mathbf{s}^{\mathscr{D}}), \mathbf{Y}_i(\mathbf{s}^{\mathscr{D}}))$, $\mathbf{r} = Corr(\mathbf{Y}_i(\mathbf{s}^{\mathscr{D}}), \mathbf{Y}_i(\mathbf{s}^*))$, and $\mathbf{R}^* = Corr(\mathbf{Y}_i(\mathbf{s}^*), \mathbf{Y}_i(\mathbf{s}^*))$, where $Corr(\cdot)$ denotes the correlation. Model (5.6) is a special case that treats $\mathbf{\Lambda} = \mathbf{I}$. Parallel to the results above, the predictive mean of $\mathbf{y}^*(\mathbf{s}^*)$ in model(5.8), conditional on all other parameters is

$$\mathrm{E}[\mathbf{y}^*(\mathbf{s}^*)|\mathbf{y}(\mathbf{s}^*), \mathbf{y}(\mathbf{s}^{\mathscr{D}}), \mathbf{y}^*(\mathbf{s}^*), \boldsymbol{\Sigma}, \mathbf{R}] = \mathbf{y}^*(\mathbf{s}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{r} + \Sigma_{*0}\Sigma_{00}^{-1}\left\{\mathbf{y}(\mathbf{s}^*) - \mathbf{y}(\mathbf{s}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{r}\right\}. \tag{5.9}$$

The connection between the regression model (5.3) and matrix normal model (5.8) is stated in the following remark.

Remark 5.1.1. *The prediction in Equation (5.5) of the linear regression model by participant is a special case of the posterior mean prediction in Equation (5.9) under the matrix normal model (5.8) if* $\mathbf{R} = \mathbf{I}$, $\boldsymbol{\Sigma}_{00}$ *and* $\boldsymbol{\Sigma}_{0*}$ *are estimated by the plug-in MLE estimation, i.e.* $\hat{\boldsymbol{\Sigma}}_{00} = (\mathbf{y}(\mathbf{s}^{\mathscr{D}})\mathbf{y}(\mathbf{s}^{\mathscr{D}})^T)/n$ *and* $\hat{\boldsymbol{\Sigma}}_{0*} = (\mathbf{y}^*(\mathbf{s}^{\mathscr{D}})\mathbf{y}(\mathbf{s}^{\mathscr{D}})^T)/n$.

The matrix normal model has been studied extensively in the recent literature (see e.g. Wang and West (2009); Conti and O'Hagan (2010)). Since the methylation levels at nearby CpG sites are correlated to each other, it might be reasonable to specify the correlation between two CpG sites, say e.g. s_a and s_b, via a correlation function $c(s_a, s_b)$, depending on the distance between s_a and s_b. This leads to the separable GaSP model defined in Equation (2.13). As discussed in Conti and O'Hagan (2010), the covariance matrix between different people can be handled with a fully Bayesian way by assuming a conjugate prior (i.e. a Wishart distribution) and correlation between sites can be modeled as a correlation function, i.e.$\Lambda_{i,j} = c(s_i, s_j)$. Note that in the separable GaSP model, $c(\cdot, \cdot)$ does not depend on i. This is a key assumption in the separable GaSP model, which generally holds for a model with the matrix normal likelihood. It means that the correlation between sites is assumed to be the same for each participant, and the correlation between participants is also the same for each site. The nonseparable GaSP model goes beyond such assumption, discussed in the following Section 5.2.

More challenges come from the computational part of the above separable GaSP model. As discussed before, to evaluate the likelihood in a straightforward way

involves the inverse the covariance matrix Λ, requiring $O(n^3)$ flops and n is very large. Since we need to compute the likelihood many times, the computation remains to be a potential barrier. In this work, a nonseparable GaSP model will be introduced with following features.

- Gaussian Stochastic processes with the Matérn covariance are used to model correlation between sites and the level of correlation is allowed to vary across people;
- Only $O(N)$ flops are needed for computing the likelihood and making prediction of unobserved methylation levels without approximation;
- Prediction with LS estimator in the regression model (5.3) and the separable GaSP model (5.8) correspond to special cases of this nonseparable GaSP model.

5.2 Nonseparable GaSP model

5.2.1 Nonseparable GaSP model with a sharing noise parameter

Let us think about the model previously mentioned in Chapter 2.5,

$$\mathbf{Y}(s) = \mathbf{A}\mathbf{v}(s) + \boldsymbol{\epsilon}, \tag{5.10}$$

for every site $s \in \mathscr{S}$, where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_K)$ is an independent noise; $\mathbf{A} = (\mathbf{a}_1; ...; \mathbf{a}_K)$ is a $K \times K$ matrix where $\mathbf{a}_i$ is the i^{th} basis function ($K \times 1$ vector) specified later; $\mathbf{v}(\cdot) = (v_1(\cdot), ..., v_K(\cdot))^T$, where each weight function $v_i(\cdot)$ is modeled (independently with regard to i) as a zero mean Gaussian Stochastic Process (GaSP),

$$v_i(\cdot) \sim GaSP(0, \sigma_i^2 c_i(\cdot, \cdot)), \tag{5.11}$$

for $i = 1, ..., K$, where σ_i^2 is the variance and $c_i(s_a, s_b)$ is the correlation function between site s_a and s_b, which is specified as Matérn correlation function with the

roughness parameter equal to 5/2,

$$c_i(d) = \left(1 + \frac{\sqrt{5}d}{\gamma_i} + \frac{5d^2}{3\gamma_i^2}\right) exp\left(-\frac{\sqrt{5}d}{\gamma_i}\right), \tag{5.12}$$

with $d = |s_a - s_b|$ and γ_i being the unknown range parameter. Some of advantages to use above Matérn correlation function were discussed in Section 3.2.1. The computational benefit of choosing the Matérn class of covariance is further explored in Section 5.3. For the purpose of illustration, we do not model the mean/regressor at the current stage for simplicity, but such issue will be addressed in Section 5.2.3.

The separable matrix normal model defined in Equation (5.8) is a special case of model (5.10). To see this, Note that the matrix normal model can be written as $\mathbf{Y} = \mathbf{AZ}$ where $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I} \otimes \mathbf{R})$ and define $\mathbf{\Sigma} = \mathbf{AA}^T$. This corresponds to model (5.10) where all v_i share the same covariance function and $\sigma_0^2 = 0$. A formal statement about the connection between these models in terms prediction will be given after we introduce a way of inference for this model.

For $\mathbf{A}$, we follow Higdon et al. (2008) by first applying SVD to $\mathbf{Y}(\mathbf{s}^{\mathscr{D}}) = \mathbf{UDV}$, and estimating $\mathbf{A}$ as $\mathbf{A} = \mathbf{UD}/\sqrt{n}$ for three main reasons. First, the computation of estimation on $\mathbf{A}$ is linear to n, which is computationally cheap. This is essential because n is at the size of 10^6. Second, $\mathbf{AA}^T = \mathbf{Y}(\mathbf{s}^{\mathscr{D}})\mathbf{Y}(\mathbf{s}^{\mathscr{D}})^T/n = \hat{\Sigma}$, as the MLE of the covariance matrix between people when there is no correlation between sites. This allows us to build a unified view between the joint model and the conditional model as regression, discussed more formally later. Third, we have $\mathbf{a}_i^T\mathbf{a}_j = 0$ if $i \neq j$, and hence $\mathbf{A}^T\mathbf{A}$ is a diagonal matrix, which is also a substantial simplification for computation, as we will see soon. As mentioned in Higdon et al. (2008), the data is typically normalized by its variance and row mean. We will model the variance through the scales of weights functions and postpone the discussion of the mean structure in Section 5.2.3. We suppress the details for now.

Unlike Higdon et al. (2008), we do not reduce the dimension k by only using the largest several basis functions because k is comparatively small compared to the number of sites. Potentially one can further scale down the computation by approximating the full likelihood with the first several largest principle components, as discussed in Higdon et al. (2008), but it is not the focus in this work, as the major computational burden comes from the number of sites.

For each $i = 1, .., K$, $v_i(\cdot)$ follows a GaSP prior independently, one can marginalize out $\mathbf{v}(\mathbf{s}^{\mathscr{D}})$ explicitly. To see this, first vectorize the output to get $\mathbf{Y}_v(\mathbf{s}^{\mathscr{D}}) = vec(\mathbf{Y}(\mathbf{s}^{\mathscr{D}}))$, a $K \times n$ vector. Denote $\mathbf{A}_v = [\mathbf{I}_n \otimes \mathbf{a}_1; ...; \mathbf{I}_n \otimes \mathbf{a}_k]$ and $\mathbf{v}_v(\mathbf{s}^{\mathscr{D}}) = vec(\mathbf{v}(\mathbf{s}^{\mathscr{D}})^T)$, where $\mathbf{v}(\mathbf{s}^{\mathscr{D}}) = (\mathbf{v}_1(\mathbf{s}^{\mathscr{D}}); ...; \mathbf{v}_k(\mathbf{s}^{\mathscr{D}}))^T$ is a $K \times n$ weight matrix, with $\mathbf{v}_i(\mathbf{s}^{\mathscr{D}})$ being a $K \times 1$ vector. It is not hard to see, for $\mathbf{s}^{\mathscr{D}} = (s_1^{\mathscr{D}}, ..., s_n^{\mathscr{D}})$, model(5.10) can be written as,

$$\mathbf{Y}_v(\mathbf{s}^{\mathscr{D}}) = \mathbf{A}_v \mathbf{v}_v(\mathbf{s}^{\mathscr{D}}) + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma_0^2 \mathbf{I}_{nK})$. Note that

$$\mathbf{v}_v(\mathbf{s}^{\mathscr{D}}) \sim \mathcal{MN}(\mathbf{0}, \boldsymbol{\Sigma}_v),$$

where $\boldsymbol{\Sigma}_v = blkdiag(\sigma_1^2 \mathbf{R}_1; ...; \sigma_K^2 \mathbf{R}_K)$ with $\mathbf{R}_l$ being the l^{th} correlation matrix, $l = 1, ..., K$ and $blkdiag(.)$ means the block diagonal matrix. Directly marginalizing out $\mathbf{v}_v(\mathbf{s})$, it is not hard to show the sampling model for $\mathbf{Y}_v(\mathbf{s})$ is (Higdon et al. (2008))

$$\mathbf{Y}_v(\mathbf{s}^{\mathscr{D}}) \sim \mathcal{MN}(\mathbf{0}, \mathbf{A}_v \boldsymbol{\Sigma}_v \mathbf{A}_v^T + \sigma_0^2 \mathbf{I}_{Kn}).$$

The straightforward computation of the likelihood in the above sampling model needs to evaluate the inverse of of $nK \times nK$ covariance matrix $\mathbf{A}_v \boldsymbol{\Sigma}_v \mathbf{A}_v^T + \sigma_0^2 \mathbf{I}_{Kn}$, which is computationally infeasible. We first have the following simplification of the marginal model stated as follows.

Lemma 5.2.1. *If* $\mathbf{A} = \mathbf{UD}/\sqrt{n}$ *and* $\mathbf{Y}(\mathbf{s}^{\mathscr{D}}) = \mathbf{UDV}$ *as SVD decomposition, after integrating out* $\mathbf{v}(\mathbf{s}^{\mathscr{D}})$*, the marginal likelihood of* $\mathbf{Y}(\mathbf{s}^{\mathscr{D}})$ *in model (5.10) follows a product of* K *independent multivariate normal distributions,*

$$L(\mathbf{Y}(\mathbf{s}^{\mathscr{D}})|\sigma_0^2, \sigma_1^2, .., \sigma_K^2, \mathbf{R}_1, ..., \mathbf{R}_K) = \prod_{i=1}^{K} p_{\mathcal{MN}}(\hat{\mathbf{v}}_i(\mathbf{s}^{\mathscr{D}}); \mathbf{0}, \sigma_i^2\mathbf{R}_i + \sigma_0^2(\mathbf{a}_i^T\mathbf{a}_i)^{-1}\mathbf{I}_n),$$

where $p_{\mathcal{MN}}(.\,; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ *is the multivariate normal density with mean* $\boldsymbol{\mu}$*, covariance* $\boldsymbol{\Sigma}$ *and* $\hat{\mathbf{v}}_i(\mathbf{s}^{\mathscr{D}})$ *being the transpose of the* i^{th} *row of* $\hat{\mathbf{v}}(\mathbf{s}^{\mathscr{D}}) = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y}(\mathbf{s}^{\mathscr{D}})$.

Proof. Because $\mathbf{A} = \mathbf{UD}/\sqrt{n}$, $\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T = \mathbf{I}_K$. The likelihood is

$$\begin{aligned}
&\mathcal{L}(\mathbf{Y}_v(\mathbf{s}^{\mathscr{D}})|\mathbf{v}_v(\mathbf{s}^{\mathscr{D}}), \sigma_0^2) \\
=&(2\pi\sigma_0^2)^{-nK/2} exp\left(-\frac{\left(\mathbf{Y}_v(\mathbf{s}^{\mathscr{D}}) - \mathbf{A}_v\mathbf{v}_v(\mathbf{s}^{\mathscr{D}})\right)^T\left(\mathbf{Y}_v(\mathbf{s}^{\mathscr{D}}) - \mathbf{A}_v\mathbf{v}_v(\mathbf{s}^{\mathscr{D}})\right)}{2\sigma_0^2}\right) \\
=&(2\pi\sigma_0^2)^{-nK/2} exp\left(-\frac{\left(\mathbf{v}_v(\mathbf{s}^{\mathscr{D}}) - \hat{\mathbf{v}}_v(\mathbf{s}^{\mathscr{D}})\right)^T\mathbf{A}_v^T\mathbf{A}_v\left(\mathbf{v}_v(\mathbf{s}^{\mathscr{D}}) - \hat{\mathbf{v}}_v(\mathbf{s}^{\mathscr{D}})\right)}{2\sigma_0^2}\right),
\end{aligned}$$

where $\hat{\mathbf{v}}_v(\mathbf{s}^{\mathscr{D}}) = (\mathbf{A}_v^T\mathbf{A}_v)^{-1}\mathbf{A}_v^T\mathbf{Y}_v(\mathbf{s}^{\mathscr{D}})$. $\mathbf{A}_v^T\mathbf{A}_v$ is a diagonal matrix because

$$\mathbf{A}_v^T\mathbf{A}_v = \begin{pmatrix}
(\mathbf{I}_n \otimes \mathbf{a}_1)^T(\mathbf{I}_n \otimes \mathbf{a}_1) & (\mathbf{I}_n \otimes \mathbf{a}_1)^T(\mathbf{I}_n \otimes \mathbf{a}_2) & ... & (\mathbf{I}_n \otimes \mathbf{a}_1)^T(\mathbf{I}_n \otimes \mathbf{a}_K) \\
(\mathbf{I}_n \otimes \mathbf{a}_2)^T(\mathbf{I}_n \otimes \mathbf{a}_1) & (\mathbf{I}_n \otimes \mathbf{a}_2)^T(\mathbf{I}_n \otimes \mathbf{a}_2) & ... & (\mathbf{I}_n \otimes \mathbf{a}_2)^T(\mathbf{I}_n \otimes \mathbf{a}_K) \\
... & ... & ... & ... \\
(\mathbf{I}_n \otimes \mathbf{a}_K)^T(\mathbf{I}_n \otimes \mathbf{a}_1) & (\mathbf{I}_n \otimes \mathbf{a}_K)^T(\mathbf{I}_n \otimes \mathbf{a}_2) & ... & (\mathbf{I}_n \otimes \mathbf{a}_K)^T(\mathbf{I}_n \otimes \mathbf{a}_K)
\end{pmatrix},$$

with $(\mathbf{I}_n \otimes \mathbf{a}_i)^T(\mathbf{I}_n \otimes \mathbf{a}_i) = (\mathbf{I}_n^T \otimes \mathbf{a}_i^T)(\mathbf{I}_n \otimes \mathbf{a}_i) = (\mathbf{I}_n^T\mathbf{I}_n) \otimes (\mathbf{a}_i^T\mathbf{a}_i)$ and $(\mathbf{I}_n \otimes \mathbf{a}_i)^T(\mathbf{I}_n \otimes \mathbf{a}_j) = (\mathbf{I}_n^T\mathbf{I}_n) \otimes (\mathbf{a}_i^T\mathbf{a}_j) = \mathbf{O}$, where $\mathbf{O}$ is a matrix with each element being 0. Marginalizing out $\mathbf{v}_v(\mathbf{s}^{\mathscr{D}})$, one has

$$
\begin{aligned}
&\mathcal{L}(\mathbf{Y}_v(\mathbf{s}^{\mathscr{D}})|\sigma_0^2, \sigma_1^2, .., \sigma_K^2, \mathbf{R}_1, ..., \mathbf{R}_K) \\
=&\int \mathcal{L}(\mathbf{Y}_v(\mathbf{s}^{\mathscr{D}})|\mathbf{v}_v(\mathbf{s}^{\mathscr{D}}), \sigma_0^2) p(\mathbf{v}_v(\mathbf{s}^{\mathscr{D}})|\sigma_1^2, .., \sigma_K^2, \mathbf{R}_1, ..., \mathbf{R}_K) d\mathbf{v}_v(\mathbf{s}^{\mathscr{D}}) \\
=&|\mathbf{\Sigma}_v + \sigma_0^2(\mathbf{A}_v^T\mathbf{A}_v)^{-1}|^{-1/2} exp\left(-\frac{1}{2}\hat{\mathbf{v}}_v(\mathbf{s}^{\mathscr{D}})^T\left(\mathbf{\Sigma}_v + \sigma_0^2(\mathbf{A}_v^T\mathbf{A}_v)^{-1}\right)^{-1}\hat{\mathbf{v}}_v(\mathbf{s}^{\mathscr{D}})\right) \\
=&\prod_{i=1}^{K}|\sigma_i^2\mathbf{R}_i + \sigma_0^2(\mathbf{a}_i^T\mathbf{a}_i)^{-1}\mathbf{I}_n|^{-1/2} exp\left(-\frac{1}{2}\hat{\mathbf{v}}_i(\mathbf{s}^{\mathscr{D}})^T\left(\sigma_i^2\mathbf{R}_i + \sigma_0^2(\mathbf{a}_i^T\mathbf{a}_i)^{-1}\mathbf{I}_n\right)^{-1}\hat{\mathbf{v}}_i(\mathbf{s}^{\mathscr{D}})\right).
\end{aligned}
$$

The last row follows from the fact that $\hat{\mathbf{v}}_i(\mathbf{s}^{\mathscr{D}})^T$ is the i^{th} row of the matrix $\hat{\mathbf{v}}(\mathbf{s}^{\mathscr{D}})$. □

Lemma 5.2.1 means the likelihood implied in model(5.10) can be formulated as a product of independent rows in the transformed matrix, with a sharing noise parameter σ_0^2. The simplification allows us to decompose the likelihood of a multivariate normal distribution with a $Kn \times Kn$ covariance matrix into K products of multivariate normal distributions, each with $n \times n$ covariance. This, of course, relies on the features of our choices of the SVD basis mentioned above and it might not hold in general for other basis functions. After these simplification, the computation of the inverse of an $n \times n$ covariance might still be hard, as n is at the size of 10^6. An algorithm linearly in number of sites is provided in Section 5.3 without approximation.

Lemma 5.2.2. *The posterior predictive distributions for model(5.10) are defined as follows.*

1. *For every* s_j^*,

$$
\mathbf{Y}(s_j^*)|\mathbf{Y}(\mathbf{s}^{\mathscr{D}}), \boldsymbol{\sigma}_{0:K}^2, \boldsymbol{\gamma}_{1:K} \sim \mathcal{MN}\left(\hat{\boldsymbol{\mu}}(s_j^*), \hat{\mathbf{\Sigma}}(s_j^*)\right),
$$

where $\hat{\boldsymbol{\mu}}(s_j^*) = \mathbf{A}\hat{\mathbf{v}}(s_j^*)$ *and* $\hat{\boldsymbol{\Sigma}}(s_j^*) = \sigma_0^2\mathbf{I}_K + \mathbf{A}\mathbf{D}(s_j^*)\mathbf{A}^T$, *with* $\hat{\mathbf{v}}(s_j^*) = \mathbf{r}_i^T(s_j^*)\mathbf{R}_i^{-1}\hat{\mathbf{v}}_i(\mathbf{s}^{\mathscr{D}})$, $\mathbf{D}(s_j^*) = diag(\sigma_1^2 c_1^*(s_j^*), ..., \sigma_k^2 c_K^*(s_j^*))$, $c_i^*(s_j^*) = c_i(s_j^*, s_j^*) - \mathbf{r}^T(s_j^*)\mathbf{R}^{-1}\mathbf{r}(s_j^*)$ *and* $\hat{\mathbf{v}}_i(\mathbf{s}^{\mathscr{D}})$ *is the transpose of the* i^{th} *row of* $\hat{\mathbf{v}}(\mathbf{s}^{\mathscr{D}}) = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y}(\mathbf{s}^{\mathscr{D}})$.

2. *Denote* $\hat{\boldsymbol{\mu}}(s_j^*) = \left(\hat{\boldsymbol{\mu}}_0^T(s_j^*), \hat{\boldsymbol{\mu}}_*^T(s_j^*)\right)^T$ *and* $\hat{\boldsymbol{\Sigma}}(s_j^*) = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{00}(s_j^*) & \hat{\boldsymbol{\Sigma}}_{0*}(s_j^*) \\ \hat{\boldsymbol{\Sigma}}_{*0}(s_j^*) & \hat{\boldsymbol{\Sigma}}_{**}(s_j^*) \end{pmatrix}$. *For every* s_j^*, *we have*

$$\mathbf{y}^*(s_j^*)|\mathbf{y}(\mathbf{s}^{\mathscr{D}}), \mathbf{y}(\mathbf{s}^*), \mathbf{y}^*(\mathbf{s}^{\mathscr{D}}), \sigma_{0:K}^2, \mathbf{R}_{1:K} \sim \mathcal{MN}\left(\hat{\boldsymbol{\mu}}_{*|0}(s_j^*), \hat{\boldsymbol{\Sigma}}_{*|0}(s_j^*)\right),$$

where $\hat{\boldsymbol{\mu}}_{*|0}(s_j^*) = \hat{\boldsymbol{\mu}}_*(s_j^*) + \hat{\boldsymbol{\Sigma}}_{*0}(s_j^*)\hat{\boldsymbol{\Sigma}}_{00}^{-1}(s_j^*)\left(\mathbf{y}(s_j^*) - \hat{\boldsymbol{\mu}}_0(s_j^*)\right)$,

and $\hat{\boldsymbol{\Sigma}}_{*|0}(s_j^*) = \boldsymbol{\Sigma}_{**} - \hat{\boldsymbol{\Sigma}}_{*0}(s_j^*)\hat{\boldsymbol{\Sigma}}_{00}^{-1}(s_j^*)\hat{\boldsymbol{\Sigma}}_{0*}(s_j^*)$.

Note when only $\mathbf{Y}(\mathbf{s}^{\mathscr{D}})$ are observed, $\hat{\boldsymbol{\mu}}(s_j^*)$ can be used to predict $\mathbf{Y}(s_j^*)$. When three blocks of outputs $\mathbf{y}(\mathbf{s}^{\mathscr{D}})$, $\mathbf{y}(\mathbf{s}^*)$ and $\mathbf{y}^*(\mathbf{s}^{\mathscr{D}})$ are observed, $\hat{\boldsymbol{\mu}}_{*|0}(s_j^*)$ can be used as prediction for $\mathbf{y}(s_j^*)$ for any site s_j^*, by properly conditional on all observations. The inference of parameters $\sigma_{0:K}^2$ and $\boldsymbol{\gamma}_{1:K}$ along with the computation strategy will be discussed in Section 5.3.

The prediction by linear regression in Equation (5.3) and separable GaSP model in Equation (5.8) are both special cases of the prediction by $\hat{\boldsymbol{\mu}}_{*|0}(s_j^*)$ with certain choices of parameters, as denoted in the following Remark 5.2.1.

Remark 5.2.1. *If* $\boldsymbol{\Sigma}_{00}$ *and* $\boldsymbol{\Sigma}_{0*}$ *in the separable GaSP model in Equation (5.8) are estimated by* $\hat{\boldsymbol{\Sigma}}_{00} = (\mathbf{y}(\mathbf{s}^{\mathscr{D}})\mathbf{y}(\mathbf{s}^{\mathscr{D}})^T)/n$ *and* $\hat{\boldsymbol{\Sigma}}_{0*} = (\mathbf{y}^*(\mathbf{s}^{\mathscr{D}})\mathbf{y}(\mathbf{s}^{\mathscr{D}})^T)/n$, *the predictive mean of* $\mathbf{y}^*(s_j^*)$ *given in Equation (5.9) is a special case of posterior predictive mean* $\hat{\boldsymbol{\mu}}_{*|0}(s_j^*)$ *in Lemma 5.2.2 by taking* $\sigma_0^2 = 0$, $\sigma_1^2 = ... = \sigma_K^2 = 1$, $\mathbf{R}_1 = ... = \mathbf{R}_K = \mathbf{R}$. *The predictive mean by regression model (5.3) is further taking* $\mathbf{R} = \mathbf{I}_n$.

Note the above results also rely on our choices of basis functions such that $\mathbf{A}\mathbf{A}^T = \mathbf{Y}(\mathbf{s}^{\mathscr{D}})\mathbf{Y}(\mathbf{s}^{\mathscr{D}})^T/n$. It is delightful to have the connection between the regression model, the separable GaSP model and the nonseparable GaSP model in terms of prediction.

Model (5.10) utilizes a sharing noise parameter σ_0^2 and we generalize by assuming different noisy parameters in following Section 5.2.2.

5.2.2 Nonseparable GaSP model with different noise parameters

Consider the model,

$$\begin{aligned} &\mathbf{Y}(s) = \mathbf{A}\tilde{\mathbf{v}}(s) + \epsilon, \\ &\tilde{v}_i(\cdot) \sim GaSP(0, \sigma_i^2\tilde{c}_i(\cdot,\cdot)),\ i = 1, ..., K, \end{aligned} \tag{5.13}$$

for any $s \in \mathscr{S}$, where $\epsilon \sim N(0, \sigma_0^2\mathbf{I}_{nk})$ and $\mathbf{A} = \mathbf{U}\mathbf{D}/\sqrt{n}$ is a $K \times K$ basis matrix specified the same as the previous model(5.10); $\tilde{\mathbf{v}}(\cdot) = (\tilde{v}_1(\cdot), ..., \tilde{v}_K(\cdot))^T$ where each $\tilde{v}_i(\cdot)$ is now modeled (independently with regard to i) as a zero mean $n$oisy Gaussian Stochastic Process (GaSP),

$$\tilde{v}_i(\cdot) = v_i(\cdot) + \varepsilon_i, \tag{5.14}$$

for $i = 1, ..., K$, where each weight function $v_i(\cdot)$ is modeled the same as the previous independent Gaussian Process in Equation (5.11) and ε_i is an independent zero mean white noise with variance τ_i. Denote $\eta_i = \tau_i\sigma_i^2$, the covariance function of the $v_i(\cdot)$ between s_a and s_b can be written as

$$\sigma_i^2\tilde{c}_i(s_a, s_b) = \sigma_i^2(c_i(s_a, s_b) + \eta_i 1_{a=b}),$$

where η_i is the nugget-variance ratio and $c_i(\cdot,\cdot)$ the correlation function. The marginal likelihood of the above model is given in the following Lemma.

Lemma 5.2.3. *The likelihood of $\mathbf{Y}(\mathbf{s}^{\mathscr{D}})$ in model (5.13) follows a product of K independent multivariate normal distributions,*

$$L(\mathbf{Y}(\mathbf{s}^{\mathscr{D}})|\sigma^2_{1:K},\eta_{1:K},\boldsymbol{\gamma}_{1:K}) = \prod_{i=1}^{K} p_{\mathcal{MN}}(\hat{\mathbf{v}}_i(\mathbf{s}^{\mathscr{D}});\mathbf{0},\sigma_i^2(\mathbf{R}_i+\eta_i\mathbf{I}_n)+\sigma_0^2(\mathbf{a}_i^T\mathbf{a}_i)^{-1}\mathbf{I}_n)),$$

where $p_{\mathcal{MN}}(.\,;\mu,\Sigma)$ *is the multivariate normal density with mean* $\boldsymbol{\mu}$ *and covariance* $\boldsymbol{\Sigma}$*;* $\hat{\mathbf{v}}_i(\mathbf{s}^{\mathscr{D}})^T$ *is the* i^{th} *row of the* $\hat{\mathbf{v}}(\mathbf{s}^{\mathscr{D}}) = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y}(\mathbf{s}^{\mathscr{D}})$*.*

It is not hard to see the marginal likelihood in model (5.10) is a special case of the marginal likelihood in model (5.13). Note that the above model might not be identifiable. It is an interesting research topic by incorporating appropriate prior information for σ_0. In the following, we simply constrain $\sigma_0 = 0$, to avoid potential identifiability issues. This leads to $\tilde{\mathbf{v}}(\mathbf{s}^{\mathscr{D}}) = \hat{\mathbf{v}}(\mathbf{s}^{\mathscr{D}}) = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y}(\mathbf{s}^{\mathscr{D}})$. The following lemma give the predictive distribution of model (5.13) with such constraint.

Lemma 5.2.4. *The following conditional distributions for model (5.13) with* $\sigma_0 = 0$ *are defined as follows.*

1. *For every* s_j^**,*

$$\mathbf{Y}(s_j^*)|\mathbf{Y}(\mathbf{s}^{\mathscr{D}}),\boldsymbol{\sigma}^2_{1:K},\boldsymbol{\eta}_{1:K},\boldsymbol{\gamma}_{1:K} \sim \mathcal{MN}\left(\tilde{\boldsymbol{\mu}}(s_j^*),\tilde{\boldsymbol{\Sigma}}(s_j^*)\right),$$

where $\tilde{\boldsymbol{\mu}}(s_j^*) = \mathbf{A}\tilde{\mathbf{v}}(s_j^*)$ *and* $\tilde{\boldsymbol{\Sigma}}(s_j^*) = \mathbf{A}\tilde{\mathbf{D}}(s_j^*)\mathbf{A}^T$*, with* $\tilde{\mathbf{v}}(s_j^*) = \mathbf{r}_i^T(s_j^*)\tilde{\mathbf{R}}_i^{-1}\hat{\mathbf{v}}_i(\mathbf{s}^{\mathscr{D}})$*,* $\tilde{\mathbf{D}}(s_j^*) = diag(\sigma_1^2\tilde{c}_1^*(s_j^*),...,\sigma_K^2\tilde{c}_K^*(s_j^*))$*,* $\tilde{c}_i^*(s_j^*) = \tilde{c}_i(s_j^*,s_j^*) - \mathbf{r}^T(s_j^*)\mathbf{R}^{-1}\mathbf{r}(s_j^*)$ *and* $\tilde{\mathbf{R}}_i = \mathbf{R}_i + \eta_i\mathbf{I}_n$*.*

2. *Denote* $\tilde{\boldsymbol{\mu}}(s_j^*) = \left(\tilde{\boldsymbol{\mu}}_0^T(s_j^*),\tilde{\boldsymbol{\mu}}_*^T(s_j^*)\right)^T$ *and* $\tilde{\boldsymbol{\Sigma}}(s_j^*) = \begin{pmatrix} \tilde{\boldsymbol{\Sigma}}_{00}(s_j^*) & \tilde{\boldsymbol{\Sigma}}_{0*}(s_j^*) \\ \tilde{\boldsymbol{\Sigma}}_{*0}(s_j^*) & \tilde{\boldsymbol{\Sigma}}_{**}(s_j^*) \end{pmatrix}$*. For every* s_j^**,*

$$\mathbf{y}^*(s_j^*)|\mathbf{y}(\mathbf{s}^{\mathscr{D}}),\mathbf{y}(\mathbf{s}^*),\mathbf{y}^*(\mathbf{s}^{\mathscr{D}}),\boldsymbol{\sigma}^2_{1:K},\boldsymbol{\eta}_{1:K},\boldsymbol{\gamma}_{1:K} \sim \mathcal{MN}\left(\tilde{\boldsymbol{\mu}}_{*|0}(s_j^*),\tilde{\boldsymbol{\Sigma}}_{*|0}(s_j^*)\right),$$

where $\tilde{\boldsymbol{\mu}}_{*|0}(s_j^*) = \hat{\boldsymbol{\mu}}_*(s_j^*) + \tilde{\boldsymbol{\Sigma}}_{*0}(s_j^*)\tilde{\boldsymbol{\Sigma}}_{00}^{-1}(s_j^*)\left(\mathbf{y}(s_j^*) - \tilde{\boldsymbol{\mu}}_0(s_j^*)\right)$,

and $\tilde{\boldsymbol{\Sigma}}_{*|0}(s_j^*) = \tilde{\boldsymbol{\Sigma}}_{**} - \tilde{\boldsymbol{\Sigma}}_{*0}(s_j^*)\tilde{\boldsymbol{\Sigma}}_{00}^{-1}(s_j^*)\tilde{\boldsymbol{\Sigma}}_{0*}(s_j^*)$.

Similar to Lemma 5.2.2, $\tilde{\boldsymbol{\mu}}_{*|0}(s_j^*)$ in the above lemma can be utilized to make prediction for $\mathbf{y}^*(s_j^*)$ when the other three blocks of data $\mathbf{y}(\mathbf{s}^{\mathscr{D}})$, $\mathbf{y}(\mathbf{s}^*)$ and $\mathbf{y}^*(\mathbf{s}^{\mathscr{D}})$ are observed.

5.2.3 Combing feature data with a joint model

In methylation levels interpolation problem, some site-specific feature data such as genomic position, DNA sequence properties, cis-regulatory element, are used as covariates in a regression model correlation between sites, which helps predict the unobserved methylation levels. Modeling Regressor/covariates is less studied in the nonseparable GaSP model. In Higdon et al. (2008), data is normalized first with the row means and variance before applying the SVD and in Paulo et al. (2012), only intercept is considered with the MLE estimation. In this section, we introduce a way to model them jointly and show that the previous connection between regression model, the separable GaSP model and the nonseparable GaSP model still holds.

Denote the feature data as a $q \times N$ matrix $\mathbf{H} = \left(\mathbf{h}(\mathbf{s}^{\mathscr{D}}); \mathbf{h}(\mathbf{s}^*)\right)$, where q is the number of features and N is the number of CpG sites. Consider an extended matrix $\mathbf{Y}^e = (\mathbf{H}^T; \mathbf{Y}^T)^T$, modeled as

$$\mathbf{Y}^e(s) = \mathbf{A}^e\tilde{\mathbf{v}}(s) + \epsilon, \tag{5.15}$$

for every $s \in \mathscr{S}$, where the weights $\tilde{\mathbf{v}}(\cdot)$ are defined as the independent GaSPs as in Equation (5.13); ϵ is an independent mean zero white noise with σ_0^2; $\mathbf{Y}^e(\mathbf{s}^{\mathscr{D}}) = \mathbf{U}^e\mathbf{D}^e\mathbf{V}^e$ with $\mathbf{A}^e = \mathbf{U}^e\mathbf{D}^e/n$. The posterior predictive distribution of the model (5.15) can be similarly formulated as lemma 5.2.4.

It is obvious that separable GaSP model can be extended with more feature data similarly by modeling $\mathbf{Y}^e$ as (jointly) a matrix normal distribution with the

covariance $\mathbf{\Sigma}^e \otimes \mathbf{\Lambda}$, where $\Sigma^e = \begin{pmatrix} \mathbf{\Sigma}^e_{00} & \mathbf{\Sigma}^e_{0*} \\ \mathbf{\Sigma}^e_{*0} & \mathbf{\Sigma}^e_{**} \end{pmatrix}$ is now a $(q+K) \times (q+K)$ covariance matrix. When $\hat{\Sigma}^e_{00} = \mathbf{y}^e(\mathbf{s}^{\mathscr{D}})y^e(\mathbf{s}^{\mathscr{D}})^T/n$ and $y^*(\mathbf{s}^{\mathscr{D}})y^e(\mathbf{s}^{\mathscr{D}})^T/n$, where $y^e(\mathbf{s}^{\mathscr{D}}) = (\mathbf{h}(\mathbf{s}^{\mathscr{D}})^T, y(\mathbf{s}^{\mathscr{D}})^T)^T$, the separable GaSP model becomes a special case of the nonseparable GaSP model (5.15) with the similar specification in Remark 5.2.1.

Note the previous regression model (5.3) can be shown to be a special case in which covariate matrix is specified as

$$\mathbf{H}_i(\cdot) = \begin{pmatrix} \mathbf{h}(\mathbf{s}^{\mathscr{D}}) & \mathbf{h}(\mathbf{s}^*) \\ \mathbf{y}(\mathbf{s}^{\mathscr{D}}) & \mathbf{y}(\mathbf{s}^*) \end{pmatrix}_{(q+k)\times N}.$$

The results follow similarly as denoted in Remark 5.2.1.

5.3 Computation strategy of nonseparable models

The straightforward way to compute the conditional distributions in Lemma 5.2.2 and Lemma 5.2.4 requires the inverse of $\mathbf{R}_i$, each with $O(n^3)$ flops, and the storage of $\mathbf{R}_i$ requires $O(n^2)$, for each $i = 1, ..., K$. This is impractical as the likelihood needs to be evaluated many times.

We introduced a less used computational strategy here, based on the connection between Gaussian random field (GMF) and Gaussian Markovian random field (GMRF). The idea can be traced back to Whittle (1954, 1963), where the Matérn covariance is shown to have a Markov structure, and more recently discussed in Hartikainen and Sarkka (2010); Särkkä and Hartikainen (2012). We quickly review these ideas.

Let us consider a continuous time AR(p) process defined by a stochastic differential equation (SDE),

$$c_p f^{(p)}(s) + c_{p-1} f^{(p-1)}(s) + ... + c_0 f(s) = b_0 z(s), \tag{5.16}$$

where $f^{(j)}(s)$ is the j^{th} derivative of $f(s)$ and $z(s)$ is the standard Gaussian white noise process defined on $s \in \mathbb{R}$ and further let $c_p = 1$ to avoid the identifiability issue. The spectral density of Equation (5.16) is

$$S_{\mathbb{R}}(s) = \frac{b_0^2}{|C(2\pi i s)|^2}, \tag{5.17}$$

where i is the imaginary number and the operator $C(\cdot)$ is defined by $C(z) = \sum_{t=0}^{p} c_t z^t$. The form of the above spectral density is

$$S_{\mathbb{R}}(s) = \frac{constant}{polynomial\ in\ s^2}, \tag{5.18}$$

which is a rational functional form. It has been shown that the spectral density of GaSP with the Matérn covariance is (Whittle (1954, 1963))

$$S_{Mat}(s) \propto \frac{1}{(\lambda^2 + s^2)^{(\alpha+1/2)}}, \tag{5.19}$$

where $\lambda = \frac{\sqrt{2\alpha}}{\gamma}$ with the range parameter γ and the roughness parameter α. It is not hard to see that the spectral density in (5.18) also follows a rational functional form. Wiener-Khinchin theorem (Rasmussen (2006)) states that the stationary covariance function of the process is given by the inverse Fourier transform of the spectral density, meaning that the covariance function is then uniquely determined.

The benefits of using the GMRF representation is that the posterior can be computed linearly in time, requiring only $O(N)$ flops for computing the likelihood and making predictions, without doing any approximation. The delightful simplification is discussed in the following subsection.

5.3.1 The computation by continuous time stochastic process

Note both nonseparable GaSP model (5.10) and model (5.13) can be written as a Gaussian Stochastic Process with a noise. For instance, the model (5.13) with $\sigma_0^2 = 0$

can be represented as

$$\begin{aligned}\tilde{v}_i(\cdot) &= f_i(\cdot) + \epsilon_i,\\ f_i(\cdot) &\sim GaSP(0, \sigma_i^2 c_i(\cdot,\cdot)),\\ \epsilon_i &\sim N(0, \tau_i),\end{aligned} \tag{5.20}$$

where $\tau_i = \sigma_i^2 \eta_i$, for $1 \leqslant i \leqslant K$. Denote $\boldsymbol{\theta}_i(s) = (f_i(s), f_i^{(1)}(s), f_i^{(2)}(s))^T$, where $f_i^{(j)}(s)$ is the j^{th} derivative of $f_i(s)$ with regard to s. As discussed above, for each $i = 1, ..., K$, Gaussian Process of Matérn covariance with the roughness parameter $\alpha_i = 5/2$ can be written as a SDE,

$$\frac{d\boldsymbol{\theta}_i(s)}{ds} = \mathbf{H}_i \boldsymbol{\theta}_i(s) + \mathbf{L} z_i(s),$$

or in a matrix form

$$\frac{d}{ds}\begin{pmatrix} f_i(s) \\ f_i^{(1)}(s) \\ f_i^{(2)}(s) \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -\lambda_i^3 & -\lambda_i^2 & -3\lambda_i \end{pmatrix}\begin{pmatrix} f_i(s) \\ f_i^{(1)}(s) \\ f_i^{(2)}(s) \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} z_i(s),$$

where $f_i^{(j)}(s)$ is the j^{th} derivative of the i^{th} GaSP with regard to s; $z_i(s)$ is a zero-mean Gaussian white noise process with variance σ_i^2 and $\lambda_i = \sqrt{2\alpha_i}/\gamma_i$. Denote $q_i = \frac{16}{3}\sigma_i^2 \lambda_i^5$ and $\mathbf{F} = (1, 0, 0)$. The solution of above SDE can be represented explicitly as a continuous time state space model,

$$\begin{aligned}\tilde{v}_i(s_{j+1}) &= \mathbf{F}\boldsymbol{\theta}_i(s_{j+1}) + \epsilon_i,\\ \boldsymbol{\theta}_i(s_{j+1}) &= \mathbf{G}_i(s_j)\boldsymbol{\theta}_i(s_j) + \mathbf{W}_i(s_j)\\ \mathbf{G}_i(s_j) &= e^{\mathbf{H}_i(s_{j+1}-s_j)}\\ \mathbf{W}_i(s_j) &\sim N(0, \mathbf{Q}_i(s_j))\\ \mathbf{Q}_i(s_j) &= \int_0^{s_{j+1}-s_j} e^{\mathbf{H}_i t}\mathbf{L} q_i \mathbf{L}^T e^{\mathbf{H}_i^T t} dt,\end{aligned} \tag{5.21}$$

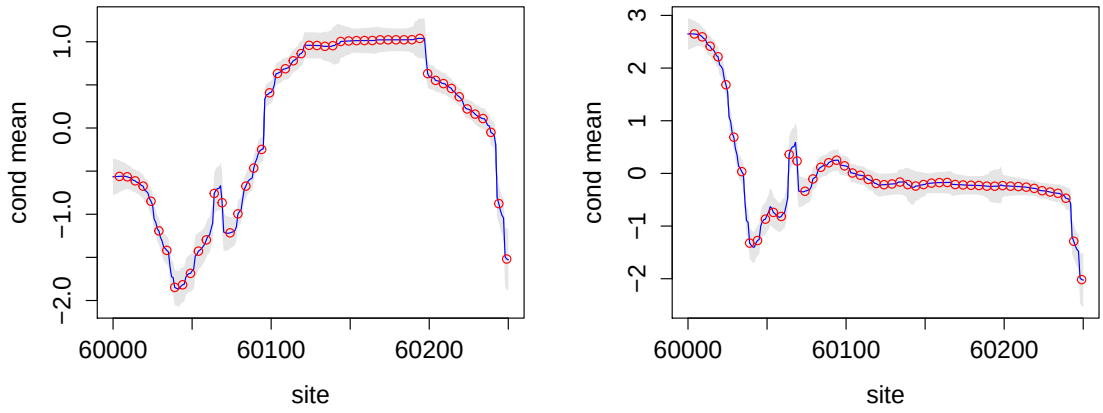

FIGURE 5.1: Comparison of the predictive mean by GF and GMRF. The blue curves are posterior mean of $\tilde{\mathbf{v}}_i(\mathbf{s}^*)|\tilde{\mathbf{v}}_i(\mathbf{s}^{\mathscr{D}})$ by Equation (5.21) and the grey shades are the 95% posterior predictive interval of the mean at a small region for $i = 1$ (left) and $i = 2$(right), for a given set of parameters $(\sigma_i^2, \tau_i, \gamma_i)$. The red dots are posterior mean $\tilde{\mathbf{v}}_i(\mathbf{s}_j^*)|\tilde{\mathbf{v}}_i(\mathbf{s}^{\mathscr{D}})$ of 50 s_j^* at the same region with the same set of parameters by the straightforward computation for the GaSP model. The root of mean square errors (RMSE) between them are 2.04×10^{-13} and 2.41×10^{-12} for the left panel and the right panel respectively.

for $j = 1, ..., n-1$, the stationary distribution of which, is

$$\boldsymbol{\theta}_i(s_0) \sim \mathcal{MN}(0, \mathbf{Q}_i(s_0)),$$

with $\mathbf{Q}_i(s_0) = \int_0^\infty e^{\mathbf{H}_i t}\mathbf{L}q_i\mathbf{L}^T e^{\mathbf{H}_i^T t}dt$. All $\mathbf{G}_i(s_j)$, $\mathbf{Q}_i(s_j)$ and $\mathbf{Q}_i(s_0)$ have closed form expressions and the joint likelihood of $\boldsymbol{\theta}_i$ is a tridiagonal matrix, shown in Appendix D. The above expression (5.21) and expression (5.20) are equivalent in terms of the marginal likelihood of v_i after integrating $\boldsymbol{\theta}_i$ out.

With the above setup, the posterior sampling for all latent states can be computed by a forward filtering and backward sampling/prediction (FFBS) algorithm, which only requires $O(n)$ flops, a lot smaller than $O(n^3)$ for computation of the likelihood of GaSP model directly. The prediction of $\mathbf{s}^*$ also only requires linear flops to the number of size n^*, so the computation is only $O(N)$ altogether. Furthermore,

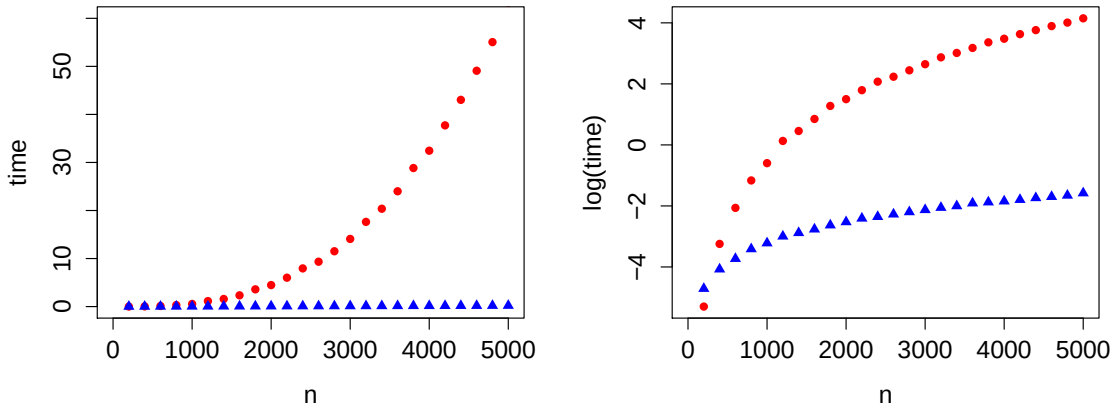

FIGURE 5.2: Comparison of computational time (in seconds) by GF and GMRF for one evaluation of the likelihood and one FFBS step respectively with normal scale (left) and log scale (right). The red dot is by direct evaluation of the likelihood and blue solid triangle is by equation (5.21).

$\boldsymbol{\theta}_i(s)$ can be explicitly marginalized out for all s instead of from posterior sampling, discussed in Section 5.3.2.

Figure 5.1 shows the posterior mean of $\tilde{v}_i(s_j^*)|v_i(\mathbf{s}^{\mathscr{D}})$ (i.e. $\hat{\boldsymbol{\mu}}(s_j^*)$ in Lemma 5.2.4) computed in Equation (5.21) and straightforward computation of GaSP in Equation (5.20), with a given set of parameters. Since they are basically the same quantities computed in two different ways, the difference only depends on machine precision, which is extremely small.

The computational time between them, however, is quite different. As shown in Figure 5.2, since the computational time by FFBS algorithm is linear, it is a lot more efficient than directly evaluation of the likelihood, which requires $O(n^3)$ for matrix inversion. To compute $n = 5,000$ sites, evaluating the likelihood by one FFBS step only takes less than 0.2 second in a desktop, while the direct evaluation takes around 60 seconds as it involves computing the inverse of the covariance matrix.

The fast computation allows us to accomplish our goal of predicting methylation levels at unobserved CpG sites. We conclude this section by the specification of the

prior and a detailed computational strategy, discussed in the following Section 5.3.2.

5.3.2 Prior specification and posterior computation

We first discuss the availability for full posterior computation and then we introduce the marginal posterior mode estimation to avoid MCMC sampling. To begin with, note that the most computationally intensive part of above FFBS algorithm is to sample $3Kn$ latent states $\boldsymbol{\theta}_i(s_j)$ for $i = 1, .., K$ and $j = 1, ..., n$. Fortunately we can marginalize $\boldsymbol{\theta}$ out explicitly and obtain the marginal likelihood

$$p(\tilde{\mathbf{v}}(\mathbf{s}^{\mathscr{D}})|\boldsymbol{\sigma}^2_{1:K}, \boldsymbol{\tau}_{1:K}, \boldsymbol{\gamma}_{1:K}) = \prod_{i=1}^{K} \left\{ p(\tilde{v}_i(s_1)|\boldsymbol{\sigma}^2_i, \boldsymbol{\tau}_i, \boldsymbol{\gamma}_i) \prod_{j=2}^{n} p(\tilde{v}_i(s_j)|\tilde{v}_i(s_{1:j-1}), \sigma_i, \tau_i, \gamma_i) \right\},$$

each term of which follows a normal distribution. This is also called one-step look ahead prediction in a dynamic linear model (West and Harrison (1997); Petris et al. (2009)).

The reference prior discussed in Chapter 3 requires the explicit form of the inverse $\mathbf{R}_i^{-1}$, which is again too computationally intensive. Thus we use the jointly robust discussed in Chapter 4, which is computationally cheap. With the above setting, the posterior can be sampled using a Metropolis-Hasting algorithm (Hoff (2009)).

When the number of sites is particularly large, obtaining thousands of MCMC samples might also be computational consuming. As discussed in Chapter 3, it is typical to estimate the range and noise-variance ratio parameters in the GaSP model by the marginal posterior mode. In this approach, one first marginalizes out σ_i^2 with the local-scale prior $\pi(\sigma_i^2) \propto 1/\sigma_i^2$, and estimate the parameters by

$$(\hat{\beta}_i, \hat{\eta}_i) = \underset{\beta_i, \eta_i}{argmax} \left\{ \mathcal{L}(\tilde{\mathbf{v}}(s_{1:n})|\beta_i, \eta_i) \pi^{JR}(\beta_i, \eta_i) \right\}, \tag{5.22}$$

where $\mathcal{L}(\tilde{\mathbf{v}}(s_{1:n})|\beta_i, \eta_i)$ is the marginal likelihood for $i = 1, ..., N$. Since the reference prior for (β_i, η_i) is also computationally intensive due to the inverse of the covariance

matrix, the jointly robust prior $\pi^{JR}(\beta_i, \eta_i)$ is utilize, introduced in Chapter 4. After obtaining the estimates $(\hat{\beta}_i, \hat{\eta}_i)$ for each i, the predictive distribution (after integrating out σ_i^2) follows a student t distribution,

$$\tilde{v}_i(s_j^*)|\tilde{\boldsymbol{v}}_i(\mathbf{s}^{\mathscr{D}}), \hat{\beta}_i, \hat{\eta}_i, \sigma_i^2 \sim \mathcal{T}(\tilde{v}_i(s_j^*), \hat{\sigma}_i^2 \tilde{c}_i^*(s_j^*), n), \tag{5.23}$$

where $\tilde{v}_i(s_j^*)$ is the i^{th} entry of $\tilde{\mathbf{v}}(s_j^*)$, with $\hat{\sigma}_i^2 = \tilde{\boldsymbol{v}}_i^T(\mathbf{s}^{\mathscr{D}})\tilde{\mathbf{R}}_i^{-1}\tilde{\boldsymbol{v}}_i^T(\mathbf{s}^{\mathscr{D}})/n$, the MLE estimate of σ_i^2, $\tilde{\mathbf{v}}(s_j^*)$ and $\tilde{c}_i^*(s_j^*)$ being defined in Lemma 5.2.4. After obtaining the predictive mean and variance, we rely on the second part in Lemma 5.2.4 for making prediction. we specifically point out that the marginal likelihood $p(\tilde{\mathbf{v}}(s_{1:n})|\beta_i, \eta_i)$ and $\hat{\sigma}_i^2 \tilde{c}_i^*$ can also be computed with linear flops in number of sites, allowing us to make inference without any approximation.

5.4 Numerical Comparison

We show the out of sample prediction results by different numerical methods in this section. In the following, we denote $y_i^*(s_j^*)$, $1 \leqslant i \leqslant n^*$ and $1 \leqslant j \leqslant k^*$ as the held-out methylation levels of the i^{th} participant at the j^{th} CpG site. The specific criteria that we employ are the same as introduced in several previous chapters,

$$\begin{aligned}
RMSE &= \sqrt{\frac{\sum_{i=1}^{k^*}\sum_{j=1}^{n^*}(\hat{y}_i^*(s_j^*) - y_i^*(s_j^*))^2}{k^* n^*}}, \\
P_{CI}(95\%) &= \frac{1}{k^* n^*}\sum_{i=1}^{k^*}\sum_{j=1}^{n^*} 1\{y_i^*(s_j^*) \in CI_{ij}(95\%)\}, \\
L_{CI}(95\%) &= \frac{1}{k^* n^*}\sum_{i=1}^{k^*}\sum_{j=1}^{n^*} \text{length}\{CI_{ij}(95\%)\},
\end{aligned}$$

where $\hat{y}_i^*(s_j^*)$ is the prediction of the held-out methylation levels of the i^{th} participant at the j^{th} CpG site; $CI_{ij}(95\%)$ is the 95% posterior credible interval; and

length$\{CI_{ij}(95\%)\}$ is the length of the 95% posterior credible interval. The rate of accuracy in prediction is also suggested in Zhang et al. (2015) by assuming that the CpG is methylated if there are more than 50% of the probes are methylated. We also present such results.

In the WBGS whole sequencing dataset, we show the results where 25%, 50% and 75% of the sites are held out for testing. And in the Methylation450 dataset, we use 10% as testing since this dataset only contains 2% of the total number of CpG sites in the WBGS whole sequencing dataset.

5.4.1 Real dataset 1: WBGS whole sequencing data

In this section, we test different methods based on the WBGS whole sequencing dataset (Hodges et al. (2011); Ziller et al. (2013)). Out-of-sample prediction results are compared using the criteria discussed above for the first 10^6 methylation levels in the first chromosome. There are 24 participants in total and we randomly sample $k^* = 4$ participants, whose methylation levels are only partially observed, while the methylation levels of the rest of participants are fully observed. In the following Table 5.1, 25%, 50% and 75% of the methylation level of these 4 participants are held out as the testing data set.

For the GaSP model, we first normalize the outputs by its row mean and add back for prediction as in Higdon et al. (2008); we estimated the range and nugget parameters by maximum marginal posterior mode in Equation (5.22) and relied on Lemma 5.2.4 for combining different sources of correlation. We also include other methods such as nearest neighborhood method, linear models (LM) discussed in (5.2) and (5.3), both with LS estimators and random forest (RF) method (Liaw and Wiener (2002)) by site as Equation (5.2) and by participant as Equation (5.3). Results are shown in Table 5.1.

First, as shown in Table 5.1, nonseparable GaSP model has the smallest out

Table 5.1: Comparison of different methods in terms of out of sample prediction for WBGS whole sequencing data. From the upper to the lower, 25%, 50%, 75% of the first million methylation levels of $k^* = 4$ people are held out for testing respectively. LM as linear model and RF as random forest.

25% hold-out CpG sites	RMSE	$P_{CI}(95\%)$	$L_{CI}(95\%)$	Accuracy
Nonseparable GaSP	**0.08367626**	**0.955525**	**0.2800564**	**0.9715888**
Nearest neighborhood	0.1499462	/	/	0.9436125
LM in model (5.2)	0.1081322	0.9162358	0.2797713	0.9630448
RF by site	0.09723264	/	/	0.9660536
LM in model (5.3)	0.09960975	0.9406912	0.3062431	0.9626512
RF by participant	0.09542747	/	/	0.96517
50% hold-out CpG sites	RMSE	$P_{CI}(95\%)$	$L_{CI}(95\%)$	Accuracy
Nonseparable GaSP	**0.0847813**	**0.959131**	**0.290115**	**0.970511**
Nearest neighborhood	0.1471054	/	/	0.9428988
LM in model (5.2)	0.09917482	0.9125848	0.261182	0.9662309
RF by site	0.1027271	/	/	0.964267
LM in model (5.3)	0.09998307	0.940976	0.3078393	0.962405
RF by participant	0.1007986			0.9631261
75% hold-out CpG sites	RMSE	$P_{CI}(95\%)$	$L_{CI}(95\%)$	Accuracy
Nonseparable GaSP	**0.08951984**	**0.9571003**	**0.3011213**	**0.9683547**
Nearest neighborhood	0.1663838	/	/	0.934349
LM in Equation (5.2)	0.1059151	0.9097519	0.2737687	0.9630991
RF by site	0.1079663	/	/	0.9617978
LM in Equation (5.3)	0.1000538	0.9408967	0.3077991	0.9623187
RF by participant	0.09721973	/	/	0.9640213

of sample RMSE, meaning that it is the most accurate in out-of-sample prediction among all these methods for different proportion of the held-out data. It at least improves the RMSE by 15% compared with all other methods. A simple reason is that it models both the correlation between sites and between participant through a coherent statistical model, while the other models only utilize a part of the information. For example, methods in row 2 to 4 only utilize the site-wise correlation by assuming the observation is independent at each CpG site, while the methods in row 5 to 6 only utilize the correlation between participants.

When there are more and more data being held-out as testing data, the corre-

lation between nearby observed methylation levels gets smaller, making it harder for prediction. Yet the nonseparable GaSP model's overall RMSE is less than 0.09, when 75% of CpG sites are held out, still much better than the other methods.

As important, the nonseparable GaSP produce 95% credible interval that covers approximately 95% of the hold-out points, with comparatively short length of credible interval. In contrast, the linear regression model seems over-confident. This is not surprising, because the independence assumption of the linear model (either between sites or between participants) makes the likelihood too concentrated, while the real likelihood (or the one that approximates to it well) is more spread out. As a consequence, the confidence bound by the linear regression model is typically too narrow to cover the real observations as it claims, while the nonseparable GaSP model is able to present an adequate interval that is not over confident.

Also note in Table 5.1, nonseparable GaSP has around 97% accuracy in predicting whether a CpG site is methylated or not (equivalently predicting whether more than half of the probes are methylated in held-out CpG site), which has is the highest compared to all other methods. Note the difference between different methods is not large because most of the CpG sites in human chromosome are methylated. In this dataset, around 90% of the CpG sites are methylated, and thus a benchmark estimator could achieve at least 90% accuracy in prediction. However, in certain interested regions, such as CGI shore regions (Zhang et al. (2015)), around half of CpG sites are not methylated and sometimes these regions are particularly interested for certain goals, in which the difference of prediction is much larger.

We need to mention that the computation of GaSP relies heavily on the fast and exact computation of GaSP discussed in Section 5.3, the speed of which is close to the linear regression model by site. Since the number of methylation levels is at the size of million for one participant, direct computation of the GaSP model is prohibited. Among all models, the most time consuming models are random forest.

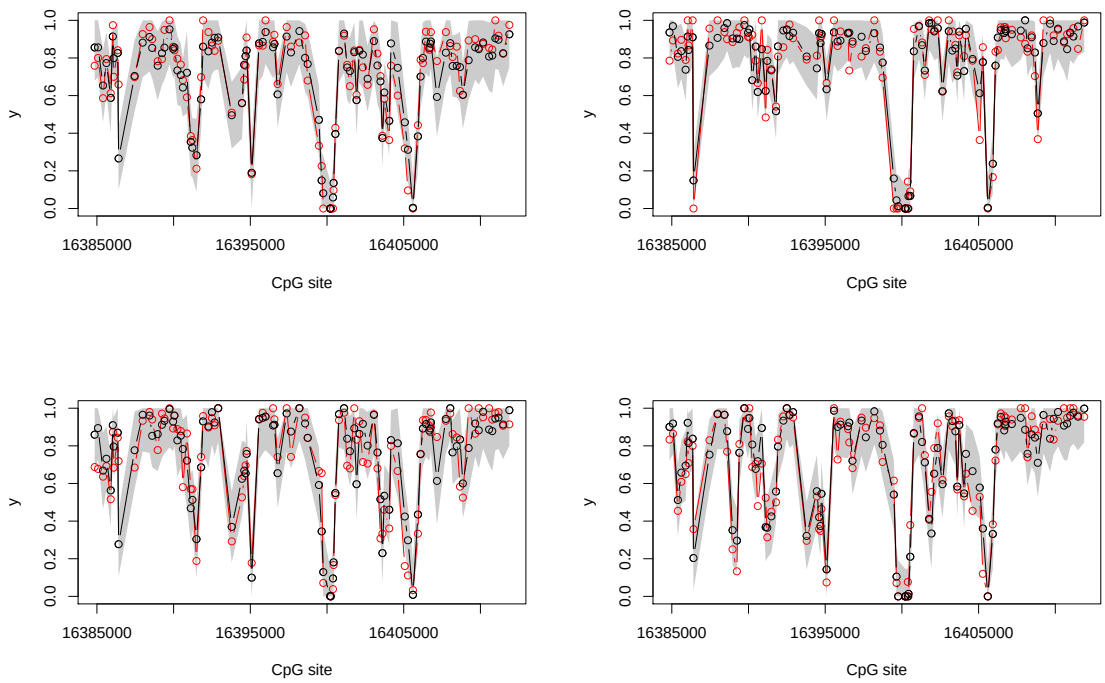

FIGURE 5.3: Prediction of methylation levels (black) by nonseparable GaSP model with estimated parameters and the real methylation levels (red) for 4 testing people at 50 unobserved CpG sites.

We present the results random forest because a version (RF by site) is used and advocated in Zhang et al. (2015) for interpolation of methylation levels. However, because regression models by random forest can only capture a part of correlation patterns, and thus provides inferior RMSE and accuracy results for prediction, than nonseparable GaSP model.

We plot the prediction (posterior mean) and the confidence interval of methylation levels at 50 unobserved CpG sites in Figure 5.3. The plots indeed show large uncertainty across different CpG sites, so only utilizing nearby CpG sites might not be adequate. Similarly, though methylation levels between participants are similar at some CpG sites, large heterogeneity exists at many CpG sites. Incorporating different sources of uncertainty is, thus, a key for modeling these complicated data.

Table 5.2: Comparison of different methods in terms of out of sample prediction for Methylation450K data. 20% CpG sites of the $k^* = 50$ people are held out for testing.

	RMSE	$P_{CI}(95\%)$	$L_{CI}(95\%)$	Accuracy
Nonseparable GaSP	**0.02979088**	**0.9721743**	**0.1218476**	**0.9908924**
Nearest neighborhood	0.1499462	/	/	0.9436125
LM in Equation (5.2)	0.03441514	0.9445401	0.09927303	0.9896839
RF by site	0.03406139	/	/	0.9895047
LM in Equation (5.3)	0.03068592	0.9569643	0.1058513	0.9903573
RF by participant	0.03070545	/	/	0.9901222

The nonseparable GaSP model seems to capture the pattern of methylation levels of each participant effectively, by combining the correlation between sites and between participants through a coherent statistical model. Hence we suggest to use this model to interpolate the methylation levels.

5.4.2 Real dataset 2: Methylation450 data

In this section, we study the numerical results for Methy450K data. There are altogether 100 participants in this study, 20% of the CpG sites of $k^* = 50$ people are held out in the case. We do not want to hold out more sites in this dataset, because this corresponds to only 2% Methylation levels compared to WBGS whole sequencing dataset (Zhang et al. (2015)). Results are recorded in Table 5.2.

As shown in Table 5.2, the nonseparable GaSP model still yields the lowest RMSE. Note the difference now is smaller compared with the linear regression in Equation (5.3) and random forest by participant. Since the data is sparse, the site-wise correlation is not as strong as the correlation by participants. Modeling this part of correlation gives roughly 3% improvement by the nonseparable GaSP model in terms of the RMSE, compared to the ones only utilizing the correlation by participant and more than 10% improvement in RMSE compared to the models only using correlation by site.

6

Concluding remarks and future work

This thesis has considered several different Bayesian models for functional data, with an emphasis on scalable computation and robust parameter estimation. Extensions of the GaSP emulator were introduced to solve a number of problems mainly arising in emulation of computer models.

In Chapter 2, the PP GaSP emulator was introduced for modeling data from computer models over massive space-time coordinates. The PP GaSP emulator was shown to enable fast and accurate prediction of the computer model at new inputs, allowing attainment of the scientific goal of quantifying volcanic hazards over wide spatial domains. The key computational benefits come from the most critical assumption–the independence of the output over different coordinates. Surprisingly, this assumption was shown to actually not be restrictive, if it suffices to have only the emulator mean and variance functions.

For the difficult problem of estimation of the correlation parameters in the GaSP model, a new criterion – called robustness– was introduced; estimation procedures satisfying this property were shown to be considerably superior to those that did not. Marginal posterior mode estimation with the reference prior under some natural pa-

rameterizations were shown to satisfy this criterion. The results depend on the study of "tail rates" of the reference prior and marginal likelihood. Surprisingly, the choice of parameterization makes a major difference in the posterior mode estimation, and previous commonly used parameterizations were shown to be non-robust. Posterior propriety is also shown with and without a nugget. Numerical results were presented to indicate the practical superiority of robust estimation.

Properties of the reference prior were further studied in Chapter 4, including its automatic adaptation to the dimension of the input space, the number of the outputs, and the scale of the inputs. A jointly robust prior was developed as an alternative to the reference prior, for the situation were some of the inputs to the computer model may be (nearly) inert. An R package, called the Robust Gaussian Stochastic Process Emulation, was developed to implement the methods in Chapter 3 and Chapter 4

Data from multiple functions (including the observed and unobserved runs of the computer model) can be viewed as a matrix, where some data are missing in a column or a row. Apart from the problem when the number of functions is large, the computation is also daunting when the number of observations n in a function is large. This is because the standard GaSP model requires $O(n^3)$ flops in each evaluation of the likelihood. This problem was discussed in Chapter 5, where a computational strategy that is linear in the number of observations (which is large) was introduced for certain situations that can be unified under a nonseparable GaSP model.

6.1 Future work on multiple functional outputs

One of the main research goals is to model the data from multiple functions. The TITAN2D data in Chapter 2 at each coordinate is a function over the input space (including the initial volume, initial flow direction, etc.). The space-time coordinate here is fixed; however, the original data from the computer model is sometimes on

dynamic/irregular grids, in order to facilitate the PDE computations. Such simplification, however, is also on the irregular grid. To build the PP emulator on a fixed grid, the irregular grid data was first interpolated to a fixed grid. It would be of considerable interest to be able to incorporate the uncertainty caused by this interpolation into the analysis.

The nonseparable GaSP model introduced in Chapter 5 unifies several models for functional data, including the PP GaSP models. However it also assumes the data is at a regular grid. In other applications, such as longitudinal studies of patients, the data is usually at an irregular grid. One of my research goals is to extend the approach in Chapter 5 to cover such cases.

Further exploration about choices of basis functions, other than the SVD basis discussed in Chapter 5, is also a promising direction for future research. The SVD basis clearly simplifies the computation and the analysis, based on which several models can also be unified. However, there are several limitations in using this basis. For one thing, the basis is estimated from the data and the uncertainty in this estimation are overlooked. Another concern is that the basis would need to be estimated again when there is a small change in the data. One direction is to explore is hierarchical structures for modeling the basis in a "deep-learning" fashion.

6.2 Future work on computation

The computational strategy introduced in Chapter 5 relies on the connection between Gaussian random fields and Gauss Markov random fields. This is clearly not a new topic but it is important, since the inference then requires only $O(n)$ computations rather than $O(n^3)$, even when neither the covariance matrix nor the precision matrix are sparse. Currently the exact algorithm only applies to one-dimensional input spaces. The extension to multi-dimensional input spaces is an important next step. There have been some studies related to this topic in recent years (see e.g.

Lindgren et al. (2011); Särkkä and Hartikainen (2012)), but none of them can be both implemented in linear computational time without an approximation to the likelihood. The GaSP emulator with a product correlation function is separable, so that the spectral density of the process is also separable. This might be a clue for designing a linear algorithm.

Another motivation comes from the applications for interferograms for ground deformation (a key data to forecast earthquakes and volcanic eruptions), e.g. for InSAR data (Lohman and Simons (2005); Montgomery-Brown et al. (2015); Anderson and Poland (2016)). The number of pixels are typically at the 10^5 to 10^6 level, which is often too large to handle. Since the calibration of parameters (the inverse problem) typically rely on posterior sampling, even the algorithm linear in the number of pixels might be too slow. Some downsampling methods have been introduced, e.g. the 'Quadtree' algorithm builds small blocks on places where the scales of pixels have a rapid change and large blocks for places with smooth change. Other algorithms work on a subset of data to approximate posterior samples (Korattikara et al. (2013); Bardenet et al. (2014); Maclaurin and Adams (2014)). It would be interesting to apply the GaSP emulator using only a small number of samples to approximate the likelihood ratio and to predict the posterior samples directly.

6.3 Future work on the GaSP model and its extension

Many interesting research topics are related to inference of the standard GaSP model. Currently, the roughness parameters in the covariance function are fixed, while empirical results indicate that estimating the roughness parameters along with the range parameters can yield better predictions when the number of observations is moderately large. The reference prior clearly can be extended to both the range and roughness parameters. When the posterior mode estimation is utilized in this scenario. Results concerning tail rates, as in Chapter 3, can similarly be found, leading

to more general robust parameter estimation.

Although the posterior density is equal to zero in a robust parameterization when the correlation matrix is an all-one matrix or an identity matrix; however, in reality, the parameters can be estimated to be close to these two cases (though it occurs much less often than using MLE, MMLE or other estimation methods without the robustness property). Other complication comes from another singularity issue. We have found that using Matérn correlation with roughness parameter equal to 2.5, the bounds of prior tail rates in Lemma 3.3.4 are slightly different than the numerical results suggest, because the Equation (B.10) in Lemma B.3.3 will have large computational errors when the correlation matrix is close to be an all-one matrix. These complications have been partially solved in our RobustGaSP R package (Gu et al. (2016)), and we keep working on the rest of the problems.

Robust parameter estimation is also connected with recent developments on "non-local" priors (Johnson and Rossell (2010, 2012)) for model selection. Indeed, the density of the reference prior for two robust parameterizations, γ and $\xi = \log(1/\gamma)$, goes to zero when $\gamma_l \to \infty$ or $\xi_l \to -\infty$ for any l, in which case the correlation matrix is $\mathbf{R} = \mathbf{R}_1 \circ \mathbf{R}_2 \circ ... \circ \mathbf{1}_n\mathbf{1}_n^T \circ ... \circ \mathbf{R}_p$. If the mean basis is specified as a constant, the GaSP model reduces to a smaller model without the l^{th} input. Nonlocal priors are typically discussed for model selection in linear regression models, and argued to balance the different convergence rates of the Bayes factor when the null is true or the alternative is true. Whether similar arguments can be made for the reference priors in the GaSP setting will be interesting to explore.

The jointly robust prior discussed in Chapter 4 can clearly be generalized to other forms. Indeed, when $a = 0$, the jointly robust prior reduces to L_1 penalty on the logarithm of the marginal likelihood, which induces sparsity (i.e., eliminates inert inputs). It would be interesting to study other forms of the priors that have better results in identification of the inert inputs.

The GaSP model is also used in other applications, such as in calibrating inputs in the computer model using data from the real process. Early work in this area includes Kennedy and O'Hagan (2001), where the GaSP model is used to model the discrepancy between the computer model and reality. This approach has been used in many recent applications, but it is not L_2 consistent (Tuo and Wu (2015); Tuo et al. (2015)). The intuition might come from the fact that some usual parameter estimation way (e.g. MLE) for the GaSP model is equivalent to find a solution in the native space (or reproducing kernel Hilbert space), rather than L_p space. However, for calibration of inputs, one might need to focus more on L_p space for explanatory purpose, as the inputs often have practical meanings. The results in Tuo and Wu (2015) suggests that the inverse range parameter should increase along with the number of observations to obtain L_2 consistency, which is the case by the reference prior and the jointly robust prior. And some previous attempts in the modular GaSP approach might also be helpful for the calibration task here.

Finally, GaSP model is needed to be constrained in a certain way for different applications. The outputs from TITAN2D computer model discussed in Chapter 2 might be viewed as censored data, as all outputs (the pyroclastic flow heights) are equal or larger than 0, while there are lots of 0 present in each run of the computer model, as many locations were not hit by the volcanic pyroclastic flows. The $log(y+1)$ transformation does not solve the problems because there are lots of outputs are the same. The methylation levels data introduced Chapter5 is also censored, with the values ranging from $[0, 1]$ and a large proportion of data is either 0 or 1. This clearly violates the multivariate normal likelihood and could be detrimental when one just ignores it in the analysis. Other problems that are related to the constraint problems include modeling/emulating monotonic functions. For all these problems, the GaSP model (or the derivative of the GaSP model) shall be constrained.

Appendix A

Appendix of Chapter 2

A.1 Periodic Folding

The input ϕ in TITAN2D is periodic with values ranging from 0 to 2π. Spiller et al. (2014) introduces a formal way to deal with such an input, using a circular correlation function with the "periodic folding" form

$$c^{cir}(\phi_1, \phi_2) = \sum_{k=-\infty}^{\infty} c(\phi_1 - \phi_2 + 2k\pi). \tag{A.1}$$

Based on Bochner's theorem, it is easy to show that $c^{cir}(,.,)$ is a valid 2π periodic correlation function with spectral representation,

$$c^{cir}(\phi_1, \phi_2) = \frac{c_0}{2\pi} + \frac{1}{\pi} \sum_{n=1}^{\infty} c_n \cos(n(\phi_1 - \phi_2)). \tag{A.2}$$

By using the power exponential correlation $c(\phi_1 - \phi_2) = exp(-\left(\frac{|\phi_1 - \phi_2|}{\gamma_\phi}\right)^{\alpha_\phi})$, it is shown in Spiller et al. (2014) that the coefficient c_n is

$$c_n = \begin{cases} 2\gamma_\phi/(\gamma_\phi^2 + n^2) & \alpha_\phi = 1 \\ e^{-n^2/4\gamma_\phi}\sqrt{\pi/\gamma_\phi} & \alpha_\phi = 2 \\ 2\int_0^\infty e^{-\gamma_\phi x^{\alpha_\phi}} \cos(nx)dx & 1 < \alpha_\phi < 2 \end{cases} . \tag{A.3}$$

We choose $c^{cir}(\cdot,\cdot)$ to be the ten term approximation to (A.2), with $\alpha_\phi = 2$ so that the c_n are available in closed form. This has the advantage of guaranteeing a positive definite covariance function [Spiller et al. (2014)], whereas approximating (A.1) would not guarantee positive definiteness.

A.2 Close to interpolation by PP GaSP with a nugget

One of the advantages of the original PP emulator is that it is an interpolator of the simulator at the input design points; this will no longer be true of the PP emulator with a nugget. It is thus useful to examine extent of lack of interpolation of the latter, using the following measure:

$$Q^{\mathscr{D}} = 1 - \frac{\sum_{j=1}^{k}\sum_{i=1}^{n}\left(y_j(\mathbf{x}_i^{\mathscr{D}}) - \hat{y}_j(\mathbf{x}_i^{\mathscr{D}})\right)^2}{\sum_{j=1}^{k}\sum_{i=1}^{n}\left(y_j(\mathbf{x}_i^{\mathscr{D}}) - \bar{\mathbf{y}}^{\mathscr{D}}\right)^2},$$

where $\hat{y}_j(\mathbf{x}_i^{\mathscr{D}})$ is the posterior mean of the PP GaSP emulator with nugget at design input $\mathbf{x}_i^{\mathscr{D}}$, $y_j(\mathbf{x}_i^{\mathscr{D}})$ is the exact simulator value and the summation is over all design input values and all output coordinates. The measure is scaled by the sum of square deviations of the simulator outputs from the overall output mean $\bar{\mathbf{y}}^{\mathscr{D}}$, so that the measure is unitless; the value of 1 would indicate perfect interpolation.

For the PP-emulator with nugget that was implemented above based on $n = 50$ design simulator runs, the value of the measure, over the 50 runs and $k = 23,040$

output coordinates, is $Q^{\mathscr{D}} \approx 0.998$. Being very close to 1, this indicates that the PP-emulator with nugget is almost an interpolator.

A.3 Smoothing the draws of the PP GaSP emulator

For some uncertainty quantification tasks, actual draws from the PP GaSP are required, e.g., draws of $(y_1(\mathbf{x}^*), \ldots, y_k(\mathbf{x}^*))$. Because of the independence assumption in the PP emulator, nearby coordinates can have quite different values in these draws, which might have a detrimental effect in the uncertainty quantification task. In this section, we introduce an emulator that has a spatial covariance structure $\mathbf{\Sigma}$, which allows for smoothing over nearby coordinates and yet has the exact posterior mean and variance as the PP GaSP Emulator. (That it has the same mean is automatic, as shown in Theorem 1; thus it is only required for the new emulator to yield the same $\hat{\sigma}_j^2$ as the PP emulator.)

The idea is relatively simple: we divide the coordinate space into blocks of small size $s < n - q$ ($s = 4$ and $s = 16$ will be used in the illustrations). The correlation between different blocks will be set to zero while, correlation is allowed within each block. The mean for each block is an $n \times s$ matrix $\boldsymbol{\theta}_b$ and the block covariance matrix, $\mathbf{\Sigma}_b$, is $s \times s$.

For each block b, $1 \leqslant b \leqslant B$, we will mimic the analysis of Conti and O'Hagan (2010), assigning the objective prior

$$\pi(\boldsymbol{\theta}_b, \mathbf{\Sigma}_b \mid \boldsymbol{\gamma}, \nu) \propto |\mathbf{\Sigma}_b|^{-(s+1)/2} \tag{A.4}$$

and marginalizing out $\mathbf{\Sigma}_b$. Then, for each block, the predictive distribution is

$$\mathbf{y}_b(\mathbf{x}^*) \mid \mathbf{y}_b^{\mathscr{D}}, \hat{\boldsymbol{\gamma}} \sim t(\hat{y}_b(\mathbf{x}^*), c^{**}\hat{\mathbf{\Sigma}}_b, n - q), \tag{A.5}$$

where $\mathbf{y}_b^{\mathscr{D}}$ is the $n \times s$ matrix of outputs at the observed inputs and

$$\begin{aligned} \hat{y}_b(\mathbf{x}^*) &= \mathbf{h}(\mathbf{x}^*)\hat{\boldsymbol{\theta}}_b + \mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1}\left(\mathbf{y}_b^{\mathscr{D}} - \mathbf{h}(\mathbf{x}^{\mathscr{D}})\hat{\boldsymbol{\theta}}_b\right), \\ \hat{\boldsymbol{\Sigma}}_b &= \frac{(\mathbf{y}_b^{\mathscr{D}} - \mathbf{h}(\mathbf{x}^{\mathscr{D}})\hat{\boldsymbol{\theta}}_b)^T\mathbf{R}^{-1}(\mathbf{y}_b^{\mathscr{D}} - \mathbf{h}(\mathbf{x}^{\mathscr{D}})\hat{\boldsymbol{\theta}}_b)}{n-q}, \end{aligned} \tag{A.6}$$

with $\hat{\boldsymbol{\theta}}_b = \left(\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\boldsymbol{R}^{-1}\ \mathbf{h}(\mathbf{x}^{\mathscr{D}})\right)^{-1}\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\boldsymbol{R}^{-1}\mathbf{y}_b^{\mathscr{D}}$. Note that the posterior mean and variance are indeed the same as those from the PP emulator at each coordinate, but now there will be smoothing within each block. This sampling algorithm is summarized as Algorithm 1.

Algorithm 2 Predictive sampling from the smoothed PP emulator

(1) Develop the original PP GaSP, computing $\hat{y}_j(\mathbf{x}^*)$ at all coordinates j.
(2) Divide the coordinates into B blocks $b = 1, ..., B$.
(3) Calculate $\hat{\boldsymbol{\Sigma}}_b$ for each block b using (A.6).
(4) Generate predictive samples of the process in each block for new input $\mathbf{x}^*$ using the multivariate t distribution in (A.5).

The only additional computational cost here is computation of the block covariance matrix $\hat{\boldsymbol{\Sigma}}_b$ and generating draws from the multivariate student t distribution, neither of which is very expensive unless s becomes large. But this is being done B times, which will be huge since k is huge; thus we use only moderate values of s, such as 4 and 16.

There is one seemingly incoherent aspect of Algorithm 2, namely the use of the original PP GaSP – not the spatially augmented GaSP – to estimate the correlation parameters $\boldsymbol{\gamma}$. This is intuitively reasonable, because $\boldsymbol{\gamma}$ arises only in the input correlation part of the process, i.e., the original PP GaSP. Indeed, when we tried to use the likelihood from the spatially augmented GaSP to estimate $\boldsymbol{\gamma}$, the results were considerably worse, presumably because the new spatial parameters allowed for additional confounding to arise, without providing sharper inferences about $\boldsymbol{\gamma}$. Finally, note that we also considered a variety of alternate structures and priors for

Table A.1: Percent under smoothing of samples from various emulators for Montserrat Island, for $n^* = 633$ held-out testing points and 23,040 grid points (coordinates).

	PP sample	2×2 Block sample	4×4 Block sample	PP emulator mean
Δ	35.6%	26.8%	19.5%	0.96%

$\mathbf{\Sigma}$, including an autoregressive structure [Farah et al. (2014)], a sparse graphical model through a conditional way [Dobra et al. (2004)] and jointly through the hyper inverse Wishart prior [Carvalho et al. (2007); Wang and West (2009)], none of which improves emulation results of the PP GaSP (evaluated by the criteria discussed in Section 5).

To evaluate the local spatial smoothness of the spatially augmented GaSP, we use the measure

$$\Delta = \left(\frac{\sum_{t=1}^{n^*} \sum_{i \in ne(j), 1 \leqslant i,j \leqslant k} |\tilde{y}_i(\mathbf{x}_t^*) - \tilde{y}_j(\mathbf{x}_t^*)|}{\sum_{t=1}^{n^*} \sum_{i \in ne(j), 1 \leqslant i,j \leqslant k} |y_i(\mathbf{x}_t^*) - y_j(\mathbf{x}_t^*)|} - 1 \right),$$

where $\tilde{y}_i(\mathbf{x}_t^*)$ is the sample output at coordinate i from the new GaSP at held-out input $\mathbf{x}_t^*$, and the outer sums are over the n^* held-out points; and the inner sums are over pairs of neighboring coordinate points i and j, the neighborhood set being defined as the the coordinates that are in the same predefined square block with the size of s. The results of 2×2 and 4×4 block have been shown in Table A.1. This measures the average smoothness, with respect to neighbors, of the sampled GaSP compared with the actual simulator. If Δ is positive, it means there is not enough smoothing while a negative Δ means over smoothing; and its numerical value can be interpreted as the percent of under or over smoothing.

First, note that the PP emulator mean is within 1% of the smoothness of the actual simulator, supporting our earlier statement that the PP emulator would inherit the smoothness of the simulator. In contrast, draws from the PP emulator are 35.6% more variable than the simulator, a clear lack of appropriate smoothness. Using

the block spatially augmented emulators did improve the smoothness with the 4×4 block emulator essentially cutting the under smoothing by half. Clearly much bigger blocks would be needed to drive the smoothness of emulator draws to be closer to the smoothness of the simulator, but that would significantly impact the computational cost.

Appendix B

Appendix of Chapter 3

B.1 Correlation matrix problem caused by the roughness parameters

Example: (Design Singularity.) Consider an one dimension and equally-spaced design on [0,1] with $n = 10$ and $x_i = (i-1)/(n-1)$. Denote the "design correlation" matrix for the l^{th} inputs as $\mathbf{R}_l^0$, $1 \leqslant l \leqslant p$, with the (i, j) entry as $|x_i - x_j|^\alpha$, $1 \leqslant i, j \leqslant n$. If one uses the power exponential correlation functions in Table 3.1 with roughness parameters $\alpha = 2$, the condition number of $\mathbf{R}^0 = \mathbf{R}_1^0 \circ \mathbf{R}_2^0 \circ \cdots \circ \mathbf{R}_p^0$ is larger than 10^{16}. $\mathbf{R}$ in this case is also ill-conditioned with small range parameters γ, e.g., $\gamma = 1$. Although $\mathbf{R}$ is quite far away from $\mathbf{1}_n\mathbf{1}_n^T$, $\mathbf{R}$ will be close to be singular and become almost non-invertible when $n \geqslant 15$.

This type of singularity is reported in Peng and Wu (2014) and other previous literature. When $\mathbf{R}_l^0$ is ill-conditioned for all $1 \leqslant l \leqslant p$, then usually $\mathbf{R}$ is ill-conditioned even if $\mathbf{R}$ is far away from $\mathbf{1}_n\mathbf{1}_n^T$. The second type of matrix singularity is usually related to the choice of roughness parameters α but not related to the estimation of range parameters γ.

One remedy for design singularity is to replace Gaussian covariance by Matérn

covariance, or simply choose the range parameter $\alpha < 2$ in power exponential correlation as in Bayarri et al. (2009); Spiller et al. (2014). Since this type of singularity is a separated problem which can be avoided by a pre-experimental check of the design, we only focused on the singularities caused by the estimated range parameters in this paper.

B.2 Proof for Section 3.3.1

In this section, we first show the validity of Equation (3.12). Note the correlation matrix is,

$$\mathbf{R} = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & ... & \rho^{n-1} \\ \rho & 1 & \rho & \rho^2 & ... & \rho^{n-2} \\ \rho^2 & \rho & 1 & \rho & ... & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho^{n-2} & \rho^{n-3} & \rho^{n-4} & ... & 1 & \rho \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & ... & \rho & 1 \end{pmatrix},$$

and the inverse correlation matrix is

$$\mathbf{R}^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho & 0 & 0 & ... & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & ... & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho & ... & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & ... & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 & ... & 0 & -\rho & 1 \end{pmatrix}.$$

Direct computation yields

$$\mathbf{1}_n^T \mathbf{R}^{-1} \mathbf{1}_n = \frac{n-(n-2)\rho}{1+\rho},$$

$$(\mathbf{y}^{\mathscr{D}})^T \mathbf{R}^{-1} (\mathbf{y}^{\mathscr{D}}) = \frac{\sum_{i=1}^{n} y_i^2 - 2\rho \sum_{i=1}^{n-1} y_i y_{i+1} + \rho^2 \sum_{i=2}^{n-1} y_i^2}{1-\rho^2},$$

$$|\mathbf{R}| = (1-\rho^2)^{n-1},$$

from which the Equation (3.12) follows. To proof Lemma 3.3.1, following quantities are needed

$$a = \sum_{i=1}^{n} y_i^2, \quad b = \sum_{i=1}^{n-1} y_i y_{i+1}, \quad c = \sum_{i=2}^{n-1} y_i^2,$$

$$d = \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j, \quad e = \sum_{i=1}^{n} \sum_{j=2}^{n-1} y_i y_j, \quad f = \sum_{i=2}^{n-1} \sum_{j=2}^{n-1} y_i y_j,$$

and

$$U = a - 2\rho b + \rho^2 c - \frac{1-\rho}{n-(n-2)\rho} \left[d - 2\rho e + \rho^2 f \right].$$

The following lemma is also needed for the proof of Lemma 3.3.1.

Lemma B.2.1. *We have*

$$na \geqslant d, \tag{B.1}$$

$$(n-2)c \geqslant f, \tag{B.2}$$

$$c[n-(n-2)\rho] > (1-\rho)f. \tag{B.3}$$

Proof. The first two inequalities are obvious. As

$$c[n-(n-2)\rho] - (1-\rho)f = (n-2)c\frac{[n-(n-2)\rho]}{n-2} - (1-\rho)f,$$

the third inequality follows from $(n-2)c \geqslant f$ and $\frac{[n-(n-2)\rho]}{n-2} - (1-\rho) = \frac{2}{n-2} > 0$. □

Proof of Lemma 3.3.1. One only needs to prove that log-profile likelihood in Equation (3.12) decreases with ρ for all $0 \leqslant \rho \leqslant 1$. The derivative for Equation (3.12) is

$$\frac{\partial logL(\mathbf{y}^{\mathscr{D}}|\hat{\sigma}^2_{MLE}, \hat{\boldsymbol{\theta}}_{MLE}, \rho)}{\partial \rho} \propto -\frac{\rho}{1-\rho^2} - \frac{n}{2U}\frac{\partial U}{\partial \rho}, \tag{B.4}$$

in which the second term can be written as

$$\begin{aligned}
&\frac{n}{2U}\frac{\partial U}{\partial \rho} \\
=&\frac{n\{(c\rho - b)[n-(n-2)\rho]^2 + (d - 2e\rho + f\rho^2) - (1-\rho)[n-(n-2)\rho](-e+\rho f)\}}{[n-(n-2)\rho]\{(a-2b\rho+c\rho^2)[n-(n-2)\rho] - (1-\rho)(d-2e\rho+f\rho^2)\}}.
\end{aligned}$$

(a) To show necessity, since $\rho \to 1$, $logL(\mathbf{y}^{\mathscr{D}}|\hat{\sigma}^2_{MLE}, \hat{\boldsymbol{\theta}}_{MLE}, \rho) \to -\infty$, a necessary condition is that

$$\lim_{\rho \to 0^+} \frac{\partial logL(\mathbf{y}^{\mathscr{D}}|\hat{\sigma}^2_{MLE}, \hat{\boldsymbol{\theta}}_{MLE}, \rho)}{\partial \rho} \leqslant 0,$$

which implies

$$n^2 b - d - ne \leqslant 0. \tag{B.5}$$

(b) To show sufficiency, we discuss $e \leqslant 0$ and $e > 0$.

(b1) When $e \leqslant 0$, noticing that $U = S^2/(1-\rho^2) > 0$, a sufficient condition is $\frac{\partial U}{\partial \rho} > 0$ or equivalently the numerator of the second term of Equation (B.4) is positive. First, terms related to c and f, by applying Inequality (B.3), are

$$c\rho[n-(n-2)\rho]^2 - (1-\rho)\rho[n-(n-2)\rho]f > 0.$$

Terms that not include c and f are

$$n[n-(n-2)\rho]^2\times$$

$$\left\{-b+\frac{d}{[n-(n-2)\rho]^2}+e\left(\frac{-2\rho}{[n-(n-2)\rho]^2}+\frac{1-\rho}{[n-(n-2)\rho]}\right)\right\}$$

$$\geqslant n[n-(n-2)\rho]^2\{-b+\frac{d}{n^2}+h(\rho)e\},$$

where

$$h(\rho)=\frac{-2\rho}{[n-(n-2)\rho]^2}+\frac{1-\rho}{[n-(n-2)\rho]}.$$

It is easy to show $h(\rho)$ is decreasing monotonically with $h(0)=\frac{1}{n}>0$ and $h(1)=-\frac{1}{2}<0$. since $e\leqslant 0$, we have $h(\rho)e\geqslant h(0)e=\frac{e}{n}$. Thus a sufficient condition is $-b+\frac{d}{n^2}+h(\rho)e\geqslant -b+\frac{d}{n^2}+h(0)e=-b+\frac{d}{n^2}+\frac{e}{n}\geqslant 0$, which is equivalent to

$$b-\frac{d}{n^2}-\frac{e}{n}\leqslant 0.$$

(b2) We show $b-\frac{d}{n^2}-\frac{e}{n}\leqslant 0$ is a sufficient condition for $e>0$ as follows. Inequality (B.6) and inequality (B.6) below are needed for the proof in this part. First, $b-\frac{d}{n^2}-\frac{e}{n}\leqslant 0$ is equivalent to

$$-b\geqslant\frac{-d-ne}{n^2}. \tag{B.6}$$

Second,

$$e=\sqrt{df}\leqslant\frac{\lambda d}{2}+\frac{f}{2\lambda}, \tag{B.7}$$

for any $\lambda>0$. After ignoring the constant n, the numerator of the second

term of Equation (B.4) is

$$
\begin{aligned}
&(c\rho - b)[n-(n-2)\rho]^2 + (d - 2e\rho + f\rho^2) \\
&- (1-\rho)[n-(n-2)\rho](-e+\rho f) \\
=&c\rho\frac{n-2}{n-2}[n-(n-2)\rho]^2 - (1-\rho)[n-(n-2)\rho]\rho f - b[n-(n-2)\rho]^2 \\
&+ (d - 2e\rho + f\rho^2) + (1-\rho)[n-(n-2)\rho]e \\
\geqslant&\rho f\frac{[n-(n-2)\rho]^2}{n-2} - (1-\rho)[n-(n-2)\rho]\rho f - b[n-(n-2)\rho]^2 \\
&+ (d - 2e\rho + f\rho^2) + (1-\rho)[n-(n-2)\rho]e \\
=&\rho f\frac{2}{n-2}[n-(n-2)\rho] - b[n-(n-2)\rho]^2 \\
&+ (d - 2e\rho + f\rho^2) + (1-\rho)[n-(n-2)\rho]e \\
\geqslant&\rho f\frac{2}{n-2}[n-(n-2)\rho] + \frac{-d-ne}{n^2}[n-(n-2)\rho]^2 \\
&+ (d - 2e\rho + f\rho^2) + (1-\rho)[n-(n-2)\rho]e \\
=&\left(-\frac{[n-(n-2)\rho]^2}{n^2} + 1\right)d - \left(2 - \frac{n-2}{n}\rho\right)(2\rho e) \\
&+ \left(\rho\frac{2}{n-2}[n-(n-2)\rho] + \rho^2\right)f \\
\geqslant&\left(-\frac{[n-(n-2)\rho]^2}{n^2} + 1\right)d - \left(2 - \frac{n-2}{n}\rho\right)\rho\left(\lambda d + \frac{f}{\lambda}\right) \\
&+ \left(\rho\frac{2}{n-2}[n-(n-2)\rho] + \rho^2\right)f \\
=&\left(-\frac{[n-(n-2)\rho]^2}{n^2} + 1 - (2 - \frac{n-2}{n}\rho)\rho\lambda\right)d \\
&+ \left(\rho\frac{2}{n-2}[n-(n-2)\rho] + \rho^2 - (2 - \frac{n-2}{n}\rho)\frac{\rho}{\lambda}\right)f
\end{aligned}
$$

The first inequality is from the result in Lemma B.2.1. The second in-

equality follows from Equation (B.6). The third inequality is from Equation (B.7) and the fact $2 - \frac{n-2}{n}\rho \geqslant 0$. Finally, put $\lambda = \frac{n-2}{n}$ into the last equation, then the coefficients of d and f are

* The coefficient of d is $-\frac{[n-(n-2)\rho]^2}{n^2} + 1 - (2 - \frac{n-2}{n}\rho)\rho\frac{n-2}{n} = 0$;
* The coefficient of f is $\rho\frac{2}{n-2}[n-(n-2)\rho] + \rho^2 - (2 - \frac{n-2}{n}\rho)\rho\frac{n}{n-2} = 0$;

from which the proof is complete.

□

B.3 Proof for Section 3.3.3

Here are some needed facts.

Fact 1. (Schur Product Theorem) If $\mathbf{A}_1$ and $\mathbf{A}_2$ are both positive semidefinite matrices (i.o, $\mathbf{A}_1 \geqslant 0$ and $\mathbf{A}_2 \geqslant 0$)

$$\mathbf{A}_1 \circ \mathbf{A}_2 \geqslant 0.$$

Fact 2. For positive semidefinite matrices $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$ and $\mathbf{A}_4$, if $\mathbf{A}_1 \geqslant \mathbf{A}_2$, $\mathbf{A}_3 \geqslant \mathbf{A}_4$,

$$\mathbf{A}_1 \circ \mathbf{A}_3 \geqslant \mathbf{A}_2 \circ \mathbf{A}_4.$$

Fact 3. Let $\mathbf{A}$ be a nonsingular matrix, and $\boldsymbol{u}$ and $\boldsymbol{v}$ be two vectors, then

$$|\mathbf{A} + \boldsymbol{u}\boldsymbol{v}'| = |\mathbf{A}|(1 + \boldsymbol{v}'\mathbf{A}^{-1}\boldsymbol{u}). \tag{B.8}$$

Further, if $\boldsymbol{v}'\mathbf{A}^{-1}\boldsymbol{u} \neq -1$, then $\mathbf{A} + \boldsymbol{u}\boldsymbol{v}'$ is nonsingular and

$$(\mathbf{A} + \boldsymbol{u}\boldsymbol{v}')^{-1} = \mathbf{A}^{-1} - \frac{(\mathbf{A}^{-1}\boldsymbol{u})(\boldsymbol{v}'\mathbf{A}^{-1})}{1 + \boldsymbol{v}'\mathbf{A}^{-1}\boldsymbol{u}}, \tag{B.9}$$

which is called the Sherman-Morrison Formula.

Fact 4. $\mathbf{D}_l$ defined in Assumption 3.3.2 has $n-1$ positive eigenvalues and one negative eigenvalue.

Following three lemmas are needed for the proof. Lemma B.3.1 and Lemma B.3.2 are from Ren et al. (2012); Lemma B.3.3 is from Berger et al. (2001).

Lemma B.3.1. *Suppose* $\mathbf{D}$ *is an* $n \times n$ *symmetric matrix whose eigenvalues are all positive except for one negative, if* $\mathbf{1}_n\mathbf{1}_n^T + \mathbf{D} \geqslant 0$*, there are* $a > 0$*,* $b > 0$ *and* $0 \leqslant s \leqslant 1$ *such that*

$$s\mathbf{1}_n\mathbf{1}_n^T + a\mathbf{I}_n \leqslant \mathbf{1}_n\mathbf{1}_n^T + \mathbf{D} \leqslant \mathbf{1}_n\mathbf{1}_n^T + b\mathbf{I}_n.$$

Lemma B.3.2. *For an* $n \times n$ *matrix* $\mathbf{A} > \mathbf{0}$ *and an* $n \times p$ *full column rank matrix* $\mathbf{h}(\mathbf{x}^{\mathscr{D}})$*,*

$$\mathbf{1}_n^T\{\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})\left(\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{A}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})\right)^{-1}\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{A}^{-1}\}\mathbf{1}_n = 0.$$

if and only if $\mathbf{1}_n \in \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$*.*

Defined $\mathbf{A} = \mathbf{R} - \mathbf{1}_n\mathbf{1}_n^T$ for the following lemma and the rest of proofs in the Appendix.

Lemma B.3.3. *If* $\mathbf{1}_n \in \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$*, then* $\mathbf{A}$ *is nonsingular and*

$$\mathbf{R}^{-1}\mathbf{P}_{\mathbf{R}} = \mathbf{A}^{-1}\mathbf{P}_{\mathbf{A}}, \tag{B.10}$$

where $\mathbf{P}_{\mathbf{R}}$ *is defined in Equation (3.2) and*

$$P_{\mathbf{A}} = \mathbf{I}_{\boldsymbol{\gamma}} - \mathbf{h}(\mathbf{x}^{\mathscr{D}})\{\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{A}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})\}^{-1}\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{A}^{-1}.$$

Proof of Lemma 3.3.3. (i) If $\forall l$, $1 \leqslant l \leqslant p$, $\gamma_l \to 0^+$, one has $\mathbf{R} \to \mathbf{I}_n$, and the marginal likelihood becomes

$$\mathcal{L}(\mathbf{y}^{\mathscr{D}}|\boldsymbol{\gamma}) \propto |\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{h}(\mathbf{x}^{\mathscr{D}})|^{-\frac{1}{2}}(S_0^2)^{-(\frac{n-q}{2}+a-1)},$$

where

$$S_0^2 = (\mathbf{y}^{\mathscr{D}} - \mathbf{h}(\mathbf{x}^{\mathscr{D}})\hat{\boldsymbol{\theta}}_0)^T(\mathbf{y}^{\mathscr{D}} - \mathbf{h}(\mathbf{x}^{\mathscr{D}})\hat{\boldsymbol{\theta}}_0),$$

$$\hat{\boldsymbol{\theta}}_0 = (\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{h}(\mathbf{x}^{\mathscr{D}}))^{-1}\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{y}^{\mathscr{D}}.$$

Similarly, the profile likelihood will be

$$\mathcal{L}(\mathbf{y}^{\mathscr{D}}|\hat{\sigma}^2_{MLE}, \hat{\boldsymbol{\theta}}_{MLE}, \boldsymbol{\gamma}) \propto (S_0^2)^{-n/2}.$$

Hence the marginal likelihood and profile likelihood exist, and their values are positive.

(ii) Because

$$\mathbf{R} = \mathbf{R}_1 \circ \mathbf{R}_2 \circ \cdots \circ \mathbf{R}_p,$$

and for each $\mathbf{R}_l$, $l = 1, \ldots, p$, $\mathbf{R}_l = \mathbf{1}_n\mathbf{1}_n^T + \nu_l(\gamma_l)(\mathbf{D}_l + o(1))$. Applying Lemma B.3.1, it follows that

$$C_{l1}\mathbf{1}_n\mathbf{1}_n^T + C_{l2}\nu_l(\gamma_l)\mathbf{I}_n \leqslant \mathbf{R}_l \leqslant \mathbf{1}_n\mathbf{1}_n^T + C_{l3}\nu_l(\gamma_l)\mathbf{I}_n,$$

and by Fact 2, that

$$b_1\mathbf{1}_n\mathbf{1}_n^T + b_2\mathbf{I}_n \leqslant \mathbf{R} \leqslant \mathbf{1}_n\mathbf{1}_n^T + b_3\mathbf{I}_n, \tag{B.11}$$

where $b_1 = \prod_l^p C_{l1}$, $b_2 = \prod_l^p\{C_{l1} + C_{l2}\nu_l(\gamma_l)\} - \prod_l^p C_{l1}$ and $b_3 = \prod_l^p\{1 + C_{l3}\nu_l(\gamma_l)\} - 1$.

Using Equation (B.8) yields

$$b_2^{n-1}(b_2 + b_1 n) \leqslant \ |\mathbf{R}| \ \leqslant b_3^{n-1}(b_3 + n).$$

Thus

$$|\mathbf{R}| = O(b_2^{n-1}) = O\left((\sum_{l=1}^{p}\nu_l(\gamma_l))^{n-1}\right). \tag{B.12}$$

Using Equation (B.9), it follows that

$$b_3^{-1}(\mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n^T}{b_3+n}) \leqslant \mathbf{R}^{-1} \leqslant b_2^{-1}(\mathbf{I}_n - \frac{b_1\mathbf{1}_n\mathbf{1}_n^T}{b_2+nb_1}), \tag{B.13}$$

and using Equation (B.8), that

$$\begin{aligned} &b_3^{-q}\left|\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{h}(\mathbf{x}^{\mathscr{D}})\right|(1-\frac{\mathbf{1}_n^T\mathbf{P}_x\mathbf{1}_n}{b_3+n}) \\ &\qquad \leqslant \left|\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})\right| \leqslant b_2^{-q}\left|\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{h}(\mathbf{x}^{\mathscr{D}})\right|(1-b_1\frac{\mathbf{1}_n^T\mathbf{P}_x\mathbf{1}_n}{b_2+nb_1}), \end{aligned} \tag{B.14}$$

where $\mathbf{P}_x = \mathbf{h}(\mathbf{x}^{\mathscr{D}})\left(\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{h}(\mathbf{x}^{\mathscr{D}})\right)^{-1}\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})$. Thus if $\mathbf{1}_n \notin \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$,

$$\left|\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})\right| = O\left((\sum_{l=1}^{p}\nu_l(\gamma_l))^{-q}\right). \tag{B.15}$$

If $\mathbf{1}_n \in \mathcal{C}(\mathbf{h}(\mathbf{x}))$, by applying Lemma B.3.2, one additionally has $\mathbf{1}_n^T\mathbf{P}_x\mathbf{1}_n = n$. Applying this fact to Inequality (B.14) yields

$$\begin{aligned} &b_3^{-(q-1)}\left|\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{h}(\mathbf{x}^{\mathscr{D}})\right|\frac{1}{b_3+n} \\ &\qquad \leqslant \left|\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})\right| \leqslant b_2^{-(q-1)}\left|\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{h}(\mathbf{x}^{\mathscr{D}})\right|\frac{1}{b_2+nb_1}, \end{aligned} \tag{B.16}$$

from which it follows that

$$\left|\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})\right| = O\left((\sum_{l}^{p}\nu_l(\gamma_l))^{-(q-1)}\right). \tag{B.17}$$

According to (B.13), $\mathbf{R}^{-1} = O(b_4^{-1}(\mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n^T}{b_5+n}))$, where $b_4^{-1} = O((\sum_l^p(\nu_l(\gamma_l)))) \to 0$ and $b_5^{-1} = O((\sum_l^p(\nu_l(\gamma_l)))) \to 0$, when $\gamma_l \to \infty$ for all $1 \leqslant l \leqslant p$. Plugging

$\mathbf{R}^{-1} = O(b_4^{-1}(\mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n^T}{b_5+n}))$ into $\mathbf{Q}$, if $\mathbf{1}_n \notin \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$, then

$$\begin{aligned}\mathbf{Q} &= \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})\{\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})\}^{-1}\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{R}^{-1} \\ &= O\left(b_4^{-1}(\mathbf{I}_n - \mathbf{P}_x - \frac{(\mathbf{I}_n - \mathbf{P}_x)\mathbf{1}_n\mathbf{1}_n^T(\mathbf{I}_n - \mathbf{P}_x)}{b_5 + n - \mathbf{1}_n^T\mathbf{P}_x\mathbf{1}_n})\right) \\ &= O\left((\sum_{l=1}^{p}\nu_l(\gamma_l))^{-1}\left(\mathbf{I}_n - \mathbf{P}_x - \frac{\mathbf{1}_n\mathbf{1}_n^T}{n}\right)\right). \end{aligned} \tag{B.18}$$

If $\mathbf{1}_n \in \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$, using Equation (B.11) and the fact that $\mathbf{1}_n\mathbf{1}_n^T$ is positive semidefinite, one has

$$b_3^{-1}\mathbf{I}_n \leqslant \mathbf{A}^{-1} \leqslant b_2^{-1}\left(\mathbf{I}_n - \frac{b_1 - 1}{b_2 + n(b_1 - 1)}\mathbf{1}_n\mathbf{1}_n^T\right) \leqslant b_2^{-1}\mathbf{I}_n,$$

where $\mathbf{A}$ is defined before Lemma B.3.3. Define $b_6^{-1} = O((\sum_l^p(\nu_l(\gamma_l)))) \to 0$, when $\gamma_l \to \infty$ for all $1 \leqslant l \leqslant p$. Using Lemma B.3.3, it follows that

$$\begin{aligned}\mathbf{Q} &= \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})\{\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{A}^{-1}\mathbf{h}(\mathbf{x}^{\mathscr{D}})\}^{-1}\mathbf{h}^T(\mathbf{x}^{\mathscr{D}})\mathbf{A}^{-1} \\ &= O\left(b_6^{-1}(\mathbf{I} - \mathbf{P}_x)\right) \\ &= O\left((\sum_{l}^{p}\nu_l(\gamma_l))^{-1}(\mathbf{I} - \mathbf{P}_x)\right). \end{aligned} \tag{B.19}$$

Using Equation (B.18) and Equation (B.19), we have

$$S^2 = (y^{\mathscr{D}})^T\mathbf{Q}\mathbf{y}^{\mathscr{D}} = O\left((\sum_{l}^{p}\nu_l(\gamma_l))^{-1}\right). \tag{B.20}$$

By combining Equation (B.12), (B.15), (B.17) and (B.20), the proof is complete.

□

Proof of Lemma 3.3.4. (i) If $\forall l$, $1 \leqslant l \leqslant p$, $\gamma_l \to 0^+$, one has $\mathbf{R} \to \mathbf{I}_n$ and

$$\mathbf{Q} \to \mathbf{I} - \mathbf{P}_x.$$

For $\forall l$, $1 \leqslant l \leqslant p$, we have

$$tr(\mathbf{W}_l^2) = tr\left[\left(\frac{\partial \mathbf{R}}{\partial \gamma_l}\mathbf{Q}\right)^2\right] \leqslant Ctr\left[\left(\frac{\partial \mathbf{R}}{\partial \gamma_l}\right)^2\right], \tag{B.21}$$

with $C > 0$. Note $\mathbf{I}^*(\boldsymbol{\gamma})$ is the fisher information matrix from the marginal likelihood and is positive semidefinite. According to the Hadamard's inequality, the determinant of positive semidefinite matrix is bounded by its diagonal elements, and thus,

$$\pi^R(\boldsymbol{\gamma}) \propto |\mathbf{I}^*(\boldsymbol{\gamma})|^{1/2} \leqslant \left[(n-q)\prod_{l=1}^{p} tr(\mathbf{W}_l^2)\right]^{1/2} \leqslant C\left[\prod_{l=1}^{p} tr\left(\frac{\partial \mathbf{R}}{\partial \gamma_l}\right)^2\right]^{1/2}. \tag{B.22}$$

(ii) For simplicity, ν_l and ω_l are used to represent $\nu_l(\gamma_l)$ and $\omega_l(\gamma_l)$ in the proof respectively. If $\mathbf{1}_n \notin \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$, Assumption 2 implies for any $1 \leqslant m \leqslant p$,

$$\mathbf{R} = \mathbf{1}_n\mathbf{1}_n^T + \sum_{l=1}^{p} \nu_l \mathbf{D}_l + \mathbf{F}_{-m} + o(\nu_m),$$

where $\mathbf{F}_{-m}$ is an $n \times n$ matrix that does not depend on γ_m. Thus for any $1 \leqslant l \leqslant p$,

$$\left\|\frac{\partial \mathbf{R}}{\partial \gamma_l}\right\|_\infty \leqslant C|\nu_l'|. \tag{B.23}$$

Using Equation (B.18) yields

$$tr(\mathbf{W}_l^2) = tr\left(\frac{\partial \mathbf{R}}{\partial \gamma_l}\mathbf{Q}\right)^2 = O\left(\left(\frac{\nu_l'}{\sum_{l=1}^{p}\nu_l}\right)^2\right), \tag{B.24}$$

and thus Equation (3.15) follows using the fact that the determinant of positive semidefinite matrix is bounded by its diagonal elements.

For the case $\mathbf{1}_n \in \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$, w.l.o.g., we assume $m = 1$. Denote

$$\boldsymbol{\Psi} = \sum_{l=1}^{p} \frac{\nu_l}{\nu_l'}\frac{\partial \mathbf{R}}{\partial \gamma_l} - \mathbf{A}, \tag{B.25}$$

and

$$\frac{\partial \mathbf{R}_1}{\partial \gamma_1} = \frac{\nu_1'}{\nu_1}\left(\mathbf{\Psi} + \mathbf{A} - \sum_{l=2}^{p}\frac{\nu_l}{\nu_l'}\frac{\partial \mathbf{R}}{\partial \gamma_l}\right). \tag{B.26}$$

Using Lemma B.3.3 yields

$$\begin{aligned} tr(\mathbf{W}_1) &= \frac{\nu_1'}{\nu_1} tr\left[\left(\mathbf{\Psi} + \mathbf{A} - \sum_{l=2}^{p}\frac{\nu_l}{\nu_l'}\frac{\partial \mathbf{R}}{\partial \gamma_l}\right)\mathbf{A}^{-1}\mathbf{P_A}\right] \\ &= \frac{\nu_1'}{\nu_1} tr\left[\mathbf{\Psi}\mathbf{A}^{-1}\mathbf{P_A} + \mathbf{P_A} - \sum_{l=2}^{p}\frac{\nu_l}{\nu_l'}\mathbf{W}_l\right], \end{aligned} \tag{B.27}$$

$$tr(\mathbf{W}_1^2) = \left(\frac{\nu_1'}{\nu_1}\right)^2 tr\left[\mathbf{\Psi}\mathbf{A}^{-1}\mathbf{P_A} + \mathbf{P_A} - \sum_{l=2}^{p}\frac{\nu_l}{\nu_l'}\mathbf{W}_l\right]^2, \tag{B.28}$$

and for $2 \leqslant j \leqslant p$,

$$tr(\mathbf{W}_1\mathbf{W}_j) = \frac{\nu_1'}{\nu_1} tr\left[\left(\mathbf{\Psi}\mathbf{A}^{-1}\mathbf{P_A} + \mathbf{P_A} - \sum_{l=2}^{p}\frac{\nu_l}{\nu_l'}\mathbf{W}_l\right)\mathbf{W}_j\right]. \tag{B.29}$$

Note $(\mathbf{P_A})^2 = \mathbf{P_A}$, $tr(\mathbf{W}_l\mathbf{P_A}) = tr(\mathbf{W}_l)$ for all $1 \leqslant l \leqslant p$ and $tr(\mathbf{P_A}) = n-q$. For the reference prior defined in Equation (3.4), first put $\frac{\nu_1'}{\nu_1}$ outside the determinant by dividing $\frac{\nu_1'}{\nu_1}$ on the second column and the second row. Then let the first row multiply by -1 and have it added to the second row. Also let the first column multiply by -1 and have it added to the second column. After the above manipulation of the determinant, it follows that

$$\pi^R(\boldsymbol{\gamma}) \propto \frac{\nu_1'}{\nu_1}\begin{vmatrix} n-q & tr(\mathbf{B}) & tr(\mathbf{W}_2) & \dots & tr(\mathbf{W}_p) \\ & tr(\mathbf{B}^2) & tr(\mathbf{B}\mathbf{W}_2) & \dots & tr(\mathbf{B}\mathbf{W}_p) \\ & & tr(\mathbf{W}_2^2) & \dots & tr(\mathbf{W}_2\mathbf{W}_p) \\ & & & \ddots & \vdots \\ & & & & tr(\mathbf{W}_p^2) \end{vmatrix}^{1/2}. \tag{B.30}$$

where $\mathbf{B} = \mathbf{\Psi}\mathbf{A}^{-1}\mathbf{P_A} - \sum_{l=2}^{p} \frac{\nu_l}{\nu_l'}\mathbf{W}_l$. Further multiple the $(l+1)^{th}$ column by $\frac{\nu_1}{\nu_1'}$, $2 \leqslant l \leqslant p$ and have them added to the second column. Then multiply the $(l+1)^{th}$ row by $\frac{\nu_1}{\nu_1'}$, $2 \leqslant l \leqslant p$ and have them added to the second row. It follows that

$$\pi^R(\boldsymbol{\gamma}) \propto \frac{\nu_1'}{\nu_1} \left| \begin{array}{ccccc} n-p & tr\left(\mathbf{\Psi}\mathbf{A}^{-1}\mathbf{P_A}\right) & tr(\mathbf{W}_2) & ... & tr(\mathbf{W}_p) \\ & tr\left(\mathbf{\Psi}\mathbf{A}^{-1}\mathbf{P_A}\right)^2 & tr\left\{\mathbf{\Psi}\mathbf{A}^{-1}\mathbf{P_A}\mathbf{W}_2\right\} & ... & tr\left\{\mathbf{\Psi}\mathbf{A}^{-1}\mathbf{P_A}\mathbf{W}_p\right\} \\ & & tr(\mathbf{W}_2^2) & ... & tr(\mathbf{W}_2\mathbf{W}_p) \\ & & & \ddots & \vdots \\ & & & & tr(\mathbf{W}_p^2) \end{array} \right|^{1/2}. \tag{B.31}$$

By definition,

$$\mathbf{A} = \sum_{l=1}^{p} \nu_l \mathbf{D}_l + \sum_{l=1}^{p} \nu_l \omega_l \mathbf{D}_l^* + \sum_{l \neq m} \nu_l \nu_m (\mathbf{D}_l \circ \mathbf{D}_m + o(1)), \tag{B.32}$$

and

$$\frac{\partial \mathbf{R}}{\partial \gamma_l} = \nu_l' \mathbf{D}_l + [\nu_l' \omega_l + \nu_l \omega_l'] \mathbf{D}_l^* + \nu_l' \sum_{m \in \{1,\ldots,p\} \setminus l}^{p} \nu_m (\mathbf{D}_l \circ \mathbf{D}_m + o(1)). \tag{B.33}$$

By Equation (B.25), we have

$$\mathbf{\Psi} = O(\sum_{l=1}^{p} \frac{\nu_l^2 \omega_l'}{\nu_l'} \mathbf{D}_l^*). \tag{B.34}$$

Directly applying Cramer's rule to the reference prior in Equation (B.31) yields

$$O\left(\frac{\nu_1'}{\nu_1} \{\prod_{l=2}^{p} tr(\mathbf{W}_l^2) tr\left(\mathbf{\Psi}\mathbf{A}^{-1}\mathbf{P_A}\right)^2\}^{1/2}\right).$$

Using Equation (B.19), (B.24) and (B.34), the result for $p \geqslant 2$ of (ii) in Lemma 3.3.4 follows.

□

Proof of Lemma 3.3.5. For any $\gamma_{l_1} \to \infty$, $1 \leqslant l_1 \leqslant p_1$, $\mathbf{R}_{l_1} \to \mathbf{1}_n \mathbf{1}_n^T$ and the correlation matrix is $\mathbf{R} \to \mathbf{R}_{p_1+1} \circ \mathbf{R}_{p_1+2} \circ ... \circ \mathbf{R}_p$. Note that $\gamma_l < \infty$ with $p_1 + 1 \leqslant l \leqslant p$. The following result is from the first part of Lemma 3.3.3,

$$L(\mathbf{y}^{\mathscr{D}}|\boldsymbol{\gamma}) = O(1) \tag{B.35}$$

Since at least one γ_l does not go to infinity, Equation (B.18) and Equation (B.19) yield,

$$\mathbf{Q} = O(1).$$

Assumption 2 implies

$$\left\| \frac{\partial \mathbf{R}}{\partial \gamma_l} \right\|_\infty \leqslant C|\nu_l'(\gamma_l)|, \tag{B.36}$$

then for any $1 \leqslant l_1 \leqslant p_1$,

$$tr(\mathbf{W}_{l_1}^2) = tr\left(\left(\frac{\partial \mathbf{R}}{\partial \gamma_{l_1}} \mathbf{Q} \right)^2 \right) = O\left(\nu_{l_1}'(\gamma_{l_1})^2 \right). \tag{B.37}$$

Since $p_1 + 1 \leqslant l_2 \leqslant p_2$, $\gamma_l \to 0^+$, the proof of lemma 3.3.4 yields

$$tr(\mathbf{W}_{l_2}^2) \leqslant Ctr\left(\frac{\partial \mathbf{R}}{\partial \gamma_{l_2}} \right)^2.$$

And for $p_1 + 1 \leqslant l_3 \leqslant p$, since γ_{l_3} is finite,

$$tr(\mathbf{W}_{l_3}^2) = tr\left(\frac{\partial \mathbf{R}}{\partial \gamma_{l_3}} \mathbf{Q} \right)^2 = O(1). \tag{B.38}$$

The fact that the determinant of a positive semidefinite matrix is bounded by its diagonal elements yields

$$\pi^R(\boldsymbol{\gamma}) \leqslant C \left| \prod_{l_1=1}^{p_1} \nu_{l_1}'(\gamma_{l_1}) \left[\prod_{l_2=p_1+1}^{p_2} tr\left(\frac{\partial \mathbf{R}}{\partial \gamma_{l_2}} \right)^2 \right]^{1/2} \right|. \tag{B.39}$$

Combining the result of Equation (B.35) and Equation (B.39), the proof is complete.

□

Proof of Theorem 3.3.2. Only Matérn correlation is proved here and the rest of cases can be checked similarly. When $p = 1$, the posterior propriety is shown in Berger et al. (2001). For $p > 1$ case, only the case $\mathbf{1}_n \in \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$ is shown below, the case $\mathbf{1}_n \notin \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}}))$ can be check similarly.

(i) First assume $\gamma_{(1)} \leqslant \gamma_{(2)} \leqslant ... \leqslant \gamma_{(p)}$ and each γ_l goes to ∞ .

(i.1) For the case $0 < \alpha < 1$, the marginal posterior is

$$\pi^R(\boldsymbol{\gamma})L(\mathbf{y}^{\mathscr{D}}|\boldsymbol{\gamma}) \leqslant C\frac{\prod_{l=1}^{p}\gamma_l^{-2\alpha-1}}{\gamma_{(1)}^{-2p\alpha}}\gamma_{(1)}^{-2+2\alpha}, \tag{B.40}$$

with $C > 0$. Noticing that to show that Equation (B.40) is integrable, we only to prove that $\int_M^\infty \int_{\gamma_{(1)}}^\infty ... \int_{\gamma_{(p-1)}}^\infty \pi^R(\boldsymbol{\gamma})L(\mathbf{y}^{\mathscr{D}}|\boldsymbol{\gamma})d\boldsymbol{\gamma}$ is finite, which is easily seen from the following

$$\begin{aligned}
&\int_M^\infty \int_{\gamma_{(1)}}^\infty ... \int_{\gamma_{(p-1)}}^\infty \pi^R(\boldsymbol{\gamma})L(\mathbf{y}^{\mathscr{D}}|\boldsymbol{\gamma})d\boldsymbol{\gamma} \\
\leqslant\ & C\int_M^\infty \int_{\gamma_{(1)}}^\infty ... \int_{\gamma_{(p-1)}}^\infty \prod_{l=2}^{p}\gamma_{(l)}^{-2\alpha-1}\gamma_{(1)}^{2p\alpha-3}d\gamma_{(p)}...d\gamma_{(1)} \\
=\ & C\int_M^\infty \gamma_{(1)}^{2\alpha-3}d\gamma_{(1)} \\
=\ & CM^{2\alpha-2} \\
<\ & \infty,
\end{aligned}$$

for $M > 0$ and $0 < \alpha < 1$. By Fubini theorem, Equation (B.40) is integrable.

(i.2) For the case $\alpha = 1$, as

$$\pi^R(\boldsymbol{\gamma})L(\mathbf{y}^{\mathscr{D}}|\boldsymbol{\gamma}) \leqslant C\frac{\prod_{l=1}^{p}\frac{2\log\gamma_l-1}{\gamma_l^3}}{\left(\frac{\log(\gamma_1)}{\gamma_{(1)}^2}\right)^p}\frac{\frac{1}{\gamma_{(1)}^2(2\log\gamma_{(1)}-1)}}{\frac{\log(\gamma_{(1)})}{\gamma_{(1)}^2}},$$

we have

$$\begin{aligned}
& \int_M^\infty \int_{\gamma_{(1)}}^\infty ... \int_{\gamma_{(p-1)}}^\infty \pi^R(\boldsymbol{\gamma})L(\mathbf{y}^{\mathscr{D}}|\boldsymbol{\gamma})d\boldsymbol{\gamma} \\
\leqslant & \int_M^\infty \int_{\gamma_{(1)}}^\infty ... \int_{\gamma_{(1)}}^\infty \pi^R(\boldsymbol{\gamma})L(\mathbf{y}^{\mathscr{D}}|\boldsymbol{\gamma})d\boldsymbol{\gamma} \\
\leqslant & C\int_M^\infty \frac{\frac{2\log\gamma_{(1)}-1}{\gamma_{(1)}^3}\left(\frac{\log(\gamma_{(1)})}{\gamma_{(1)}^2}\right)^{p-1}}{\left(\frac{\log\gamma_{(1)}}{\gamma_{(1)}^2}\right)^{p+1}\gamma_{(1)}^2(2\log\gamma_{(1)}-1)}d\gamma_{(1)} \\
= & C\int_M^\infty \frac{1}{\gamma_{(1)}\log^2\gamma_{(1)}}d\gamma_{(1)} \\
= & C/\log M \\
< & \infty.
\end{aligned}$$

(i.3) For the case $1 < \alpha < 2$, similarly we have,

$$\begin{aligned}
&\int_M^\infty \int_{\gamma_{(1)}}^\infty ... \int_{\gamma_{(p-1)}}^\infty \pi^R(\boldsymbol{\gamma}) L(\mathbf{y}^{\mathscr{D}}|\boldsymbol{\gamma}) d\boldsymbol{\gamma} \\
\leqslant\ & C \int_M^\infty \int_{\gamma_{(1)}}^\infty ... \int_{\gamma_{(p-1)}}^\infty \frac{\prod_{l=1}^p \gamma_l^{-3}}{\gamma_{(1)}^{-2p}} \gamma_{(1)}^{2-2\alpha} d\gamma_{(p)} ... d\gamma_{(1)} \\
=\ & C \int_M^\infty \gamma_{(1)}^{-2\alpha+1} d\gamma_{(1)} \\
=\ & C M^{-2\alpha+2} \\
<\ & \infty.
\end{aligned}$$

(i.4) For the cases $\alpha = 2$, we have

$$\begin{aligned}
&\int_M^\infty \int_{\gamma_{(1)}}^\infty ... \int_{\gamma_{(p-1)}}^\infty \pi^R(\boldsymbol{\gamma}) L(\mathbf{y}^{\mathscr{D}}|\boldsymbol{\gamma}) d\boldsymbol{\gamma} \\
\leqslant\ & C \int_M^\infty \int_{\gamma_{(1)}}^\infty ... \int_{\gamma_{(p-1)}}^\infty \frac{\prod_{l=1}^p \gamma_l^{-3}}{\gamma_{(1)}^{-2p}} \frac{(2\log(\gamma_{(1)}) - 1)}{\gamma_{(1)}^2} d\gamma_{(p)} ... d\gamma_{(1)} \\
=\ & C \int_M^\infty \gamma_{(1)}^{-3} (2\log(\gamma_{(1)}) - 1) d\gamma_{(1)} \\
=\ & C \frac{\log M}{M^2} \\
<\ & \infty.
\end{aligned}$$

(i.5) For the case $\alpha > 2$,

$$\pi^R(\boldsymbol{\gamma}) L(\boldsymbol{y}^{\mathscr{D}}|\boldsymbol{\gamma}) \leqslant C \frac{\prod_{l=1}^p \gamma_l^{-3}}{\gamma_{(1)}^{-2p}} \gamma_{(1)}^{-2} \leqslant C \prod_{l=1}^p \gamma_l^{-1-2/p}.$$

The right hand side is clearly integrable.

(ii) If there is at least one l, $\gamma_l < \infty$, by applying the result of Lemma 3.3.5, the integral of the product is just the product of the individual integral, so one only needs to check $\nu'(\gamma_l)$ is integrable when $\gamma_l \to \infty$ and $\frac{\partial R}{\partial \gamma_l}$ is integrable when $\gamma_l \to 0$. From Table 3.1, $\nu'(\gamma_l)$ is integrable when $\gamma_l \to \infty$. Besides, when $\gamma_l \to 0$, for Matérn correlation function (Abramowitz et al. (1966)),

$$\mathcal{K}_\alpha(\frac{d}{\gamma}) \to \sqrt{\frac{\pi}{2}} \frac{exp(-\frac{d}{\gamma})}{\sqrt{\frac{d}{\gamma}}}(1 + o(\gamma d)).$$

Thus Matérn correlation function is integrable at $\gamma_l \to 0$.

□

B.4 Proof for Section 3.4.3

Proof of Lemma 3.4.1. Since

$$\tilde{\mathbf{R}} = \mathbf{R} + \eta \mathbf{I}_n,$$

applying similar derivation in the proof of Lemma 3.3.3, it is easy to see that

$$|\tilde{\mathbf{R}}| = O\left((\sum_{l=1}^{p} \nu_l(\gamma_l) + \eta)^{n-1} \right). \tag{B.41}$$

$$\left| \mathbf{h}^T(\mathbf{x}^{\mathscr{D}}) \tilde{\mathbf{R}}^{-1} \mathbf{h}(\mathbf{x}^{\mathscr{D}}) \right| = \begin{cases} O\left((\sum_{l=1}^{p} \nu_l(\gamma_l) + \eta)^{-q} \right), & \mathbf{1}_n \notin \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}})), \\ O\left((\sum_{l=1}^{p} \nu_l(\gamma_l) + \eta)^{-(q-1)} \right), & \mathbf{1}_n \in \mathcal{C}(\mathbf{h}(\mathbf{x}^{\mathscr{D}})). \end{cases} \tag{B.42}$$

$$\tilde{\mathbf{S}}^2 = (\mathbf{y}^{\mathscr{D}})^T \tilde{\mathbf{Q}} \mathbf{y}^{\mathscr{D}} = O\left((\sum_{l=1}^{p} \nu_l(\gamma_i) + \eta)^{-1} \right). \tag{B.43}$$

The result immediately follows. □

Proof of Lemma 3.4.2. The proof of (a) and (b) can be derived similarly as in the Proof of Lemma 3.3.4 by noting $\tilde{\mathbf{R}} = \mathbf{R}_1 \circ \mathbf{R}_2 \circ ... \circ \mathbf{R}_p \circ \mathbf{R}_{p+1}$, where $\mathbf{R}_{p+1} = \eta \mathbf{I}_n + \mathbf{1}_n \mathbf{1}_n^T$, with $\nu_{p+1} = \eta$, $\omega_{p+1} = 0$, and $\gamma_{p+1} = \eta$.

□

Appendix C

Appendix of Chapter 4

Proof of Lemma 4.2.1.

$$
\begin{aligned}
\frac{1}{c} &= \int ... \int (\sum_{l=1}^{p} C_l \beta_l)^a exp(-b(\sum_{l=1}^{p} C_l \beta_l)) d\beta_1 ... d\beta_p \\
&= \int ... \int \frac{(\sum_{l=1}^{p} \tilde{\beta}_l)^a exp(-b(\sum_{l=1}^{p} \tilde{\beta}_l))}{\prod_{l=1}^{p} C_l} d\tilde{\beta}_1 ... d\tilde{\beta}_p, \qquad \text{let } \tilde{\beta}_l = C_l \beta_l, \\
&= \int \frac{z^a \exp(-bz)}{\prod_{l=1}^{p} C_l} \int ... \int_{\tilde{\beta}_2 + ... + \tilde{\beta}_p < z} d\tilde{\beta}_2 ... d\tilde{\beta}_p dz, \qquad \text{let } z = \sum_{l=1}^{p} \tilde{\beta}_l, \\
&= \int \frac{z^a \exp(-bz) z^{p-1}}{\prod_{l=1}^{p} C_l (p-1)!} dz, \\
&= \int \frac{z^{a+p-1} \exp(-bz)}{\prod_{l=1}^{p} C_l (p-1)!} dz, \\
&= \frac{\Gamma(a+p)}{(p-1)! b^{a+p} \prod_{l=1}^{p} C_l}.
\end{aligned}
$$

□

Proof of Lemma 4.2.2.

$$
\begin{aligned}
\mathrm{E}[\beta_j] &= \int \ldots \int \beta_j c(\sum_{l=1}^{p} C_l\beta_l)^a exp(-b(\sum_{l=1}^{p} C_l\beta_l))d\beta_1 \ldots d\beta_p \\
&= \int \ldots \int \frac{\tilde{\beta}_j c(\sum_{l=1}^{p}\tilde{\beta}_l)^a exp(-b(\sum_{l=1}^{p}\tilde{\beta}_l))}{C_j\prod_{l=1}^{p} C_l} d\tilde{\beta}_1 \ldots d\tilde{\beta}_p, \qquad \text{let } \tilde{\beta}_l = C_l\beta_l, \\
&= \int \frac{cz^a \exp(-bz)}{C_j\prod_{l=1}^{p} C_l} \int \ldots \int_{\tilde{\beta}_2+\ldots\tilde{\beta}_p<z} \tilde{\beta}_j d\tilde{\beta}_2 \ldots d\tilde{\beta}_p dz, \qquad \text{let } z = \sum_{l=1}^{p}\tilde{\beta}_l, \\
&= \int \frac{cz^a \exp(-bz)}{C_j\prod_{l=1}^{p} C_l} \int_0^z \frac{(z-\tilde{\beta}_j)^{p-2}\tilde{\beta}_j}{(p-2)!} d\tilde{\beta}_j dz \\
&= \int \frac{c\exp(-bz)z^{a+p}}{C_j\prod_{l=1}^{p} C_l p!} dz, \\
&= \frac{a+p}{pbC_j}.
\end{aligned}
$$

□

Proof of Lemma 4.2.3. Using the similar method as in the above proof for prior mean, we have

$$\mathrm{E}_{\pi^{JR}}[\beta_j^2] = \frac{2(a+p+1)(a+p)}{p(p+1)C_j^2 b^2}.$$

Part (i) follows from $\mathrm{Var}_{\pi^{JR}} = E_{\pi^{JR}}[\beta_j^2] - (E_{\pi^{JR}}[\beta_j])^2$. And using a similar way in calculating the normalizing constant, we have

$$\mathrm{E}[\sum_{l=1}^{p} C_l\beta_l]^2 = \frac{(a+p+1)(a+p)}{b^2},$$

from which the results of Part (ii) follow. □

Appendix D

Appendix of Chapter 5

The quantities of continues time state space model representation of Gaussian Process with Matérn covariance with roughness parameter $\alpha = 2.5$ in Equation (5.21) is shown in this chapter. The following results hold for every subscript i, $1 \leqslant i \leqslant k$, so it is dropped for simplicity. Denote $d_j = |s_j - s_{j-1}|$.

$$e^{\mathbf{H}d_j} = \frac{e^{-\lambda d_j}}{2} \begin{pmatrix} \lambda^2 d_j^2 + 2\lambda + 2 & 2(\lambda d_j^2 + d_j) & d_j^2 \\ -\lambda^3 d_j^2 & -2(\lambda^2 d_j^2 - \lambda d_j - 1) & 2t - \lambda d_j^2 \\ \lambda^4 d_j^2 - 2\lambda^3 d_j & 2(\lambda^3 d_j^2 - 3\lambda^2 d_j) & \lambda^2 d_j^2 - 4\lambda d_j + 2 \end{pmatrix}$$

$$\mathbf{Q}(s_j) = \frac{4\sigma^2\lambda^5}{3} \begin{pmatrix} Q_{1,1}(s_j) & Q_{1,2}(s_j) & Q_{1,3}(s_j) \\ Q_{2,1}(s_j) & Q_{2,2}(s_j) & Q_{2,3}(s_j) \\ Q_{3,1}(s_j) & Q_{3,2}(s_j) & Q_{3,3}(s_j) \end{pmatrix},$$

with

$$Q_{1,1}(s_j) = \frac{e^{-2\lambda d_j}(3 + 6\lambda d_j + 6\lambda^2 d_j^2 + 4\lambda^3 d_j^3 + 2\lambda^4 d_j^4) - 3}{-4\lambda^5},$$

$$Q_{1,2}(s_j) = Q_{2,1}(s_j) = \frac{e^{-2\lambda d_j} t^4}{2},$$

$$Q_{1,3}(s_j) = Q_{3,1}(s_j) = \frac{e^{-2\lambda d_j}(1 + 2\lambda d_j + 2\lambda^2 d_j^2 + 4\lambda^3 d_j^3 - 2\lambda^4 d_j^4) - 1}{4\lambda^3},$$

$$Q_{2,2}(s_j) = \frac{e^{-2\lambda d_j}(1 + 2\lambda d_j + 2\lambda^2 d_j^2 - 4\lambda^3 d_j^3 + 2\lambda^4 d_j^4) - 1}{-4\lambda^3},$$

$$Q_{2,3}(s_j) = Q_{3,2}(s_j) = \frac{e^{-2\lambda d_j} d_j^2 (4 - 4\lambda d_j + \lambda^2 d_j^2)}{2},$$

$$Q_{3,3}(s_j) = \frac{e^{-2\lambda d_j}(-3 + 10\lambda^2 d_j^2 - 22\lambda^2 d_j^2 + 12\lambda^2 d_j^2 - 2\lambda^4 d_j^4) + 3}{4\lambda},$$

and

$$\mathbf{Q}(s_0) = \begin{pmatrix} \sigma^2 & 1 & -\sigma^2\lambda^2/3 \\ 0 & \sigma^2\lambda^2/3 & 1 \\ -\sigma^2\lambda^2/3 & 0 & \sigma^2\lambda^4 \end{pmatrix}.$$

Bibliography

Abramowitz, M., Stegun, I. A., et al. (1966), "Handbook of mathematical functions," *Applied mathematics series*, 55, 62.

Alvarez, M. A. and Lawrence, N. D. (2011), "Computationally Efficient Convolved Multiple Output Gaussian Processes." *Journal of Machine Learning Research*, 12, 1459–1500.

Álvarez, M. A., Rosasco, L., and Lawrence, N. D. (2011), "Kernels for vector-valued functions: a review," *arXiv preprint arXiv:1106.6251*.

An, J. and Owen, A. (2001), "Quasi-regression," *Journal of complexity*, 17, 588–607.

Anderson, K. R. and Poland, M. P. (2016), "Bayesian estimation of magma supply, storage, and eruption rates using a multiphysical volcano model: Kīlauea Volcano, 2000–2012," *Earth and Planetary Science Letters*, 447, 161–171.

Andrianakis, I. and Challenor, P. G. (2012), "The effect of the nugget on Gaussian process emulators of computer models," *Computational Statistics & Data Analysis*, 56, 4215–4228.

Ba, S. and Joseph, V. R. (2012), "Composite Gaussian process models for emulating expensive functions," *The Annals of Applied Statistics*, 6, 1838–1860.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), "Gaussian predictive process models for large spatial data sets," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 825–848.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014), *Hierarchical modeling and analysis for spatial data*, Crc Press.

Bardenet, R., Doucet, A., and Holmes, C. C. (2014), "Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach." in *ICML*, pp. 405–413.

Bastos, L. S. and O'Hagan, A. (2009), "Diagnostics for Gaussian process emulators," *Technometrics*, 51, 425–438.

Bayarri, M., Berger, J., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R., Paulo, R., Sacks, J., and Walsh, D. (2007a), "Computer model validation with functional output," *The Annals of Statistics*, 35, 1874–1906.

Bayarri, M., Berger, J., Calder, E., Patra, A., Pitman, E. B., Spiller, E., and Wolpert, R. (2015), "Probabilistic quantification of hazards: a methodology using small ensembles of physics based simulations and statistical surrogates," *International Journal for Uncertainty Quantification*, 5.

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C.-H., and Tu, J. (2007b), "A framework for validation of computer models," *Technometrics*, 49, 138–154.

Bayarri, M. J., Berger, J. O., Calder, E. S., Dalbey, K., Lunagomez, S., Patra, A. K., Pitman, E. B., Spillerh, E. T., and Wolperti, R. L. (2009), "Using statistical and computer models to quantify volcanic hazards," *Technometrics*, 51, 402–413.

Berger, J. O., De Oliveira, V., and Sansó, B. (2001), "Objective Bayesian analysis of spatially correlated data," *Journal of the American Statistical Association*, 96, 1361–1374.

Besag, J. (1974), "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, 36, 192–236.

Carvalho, C. M., West, M., et al. (2007), "Dynamic matrix-variate graphical models," *Bayesian analysis*, 2, 69–97.

Cedar, H. and Bergman, Y. (2012), "Programming of DNA methylation patterns," *Annual review of biochemistry*, 81, 97–117.

Chen, H., Loeppky, J., Sacks, J., and Welch, W. (2016), "Analysis Methods for Computer Experiments: How to Assess and What Counts?" Tech. Rep. 1.

Conti, S. and O'Hagan, A. (2010), "Bayesian emulation of complex multi-output and dynamic computer models," *Journal of statistical planning and inference*, 140, 640–651.

Cox, D. R. (1975), "Partial likelihood," *Biometrika*, 62, 269–276.

Cressie, N. and Johannesson, G. (2008), "Fixed rank kriging for very large spatial data sets," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 209–226.

Cressie, N. A. and Cassie, N. A. (1993), *Statistics for spatial data*, Wiley, New York.

Das, P. M. and Singal, R. (2004), "DNA methylation and cancer," *Journal of clinical oncology*, 22, 4632–4642.

Dette, H. and Pepelyshev, A. (2010), "Generalized latin hypercube design for computer experiments," *Technometrics*, 52.

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004), "Sparse graphical models for exploring gene expression data," *Journal of Multivariate Analysis*, 90, 196–212.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., et al. (2006), "DNA methylation profiling of human chromosomes 6, 20 and 22," *Nature genetics*, 38, 1378–1385.

Efron, B. and Stein, C. (1981), "The jackknife estimate of variance," *The Annals of Statistics*, pp. 586–596.

Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M., and Niemi, J. (2013), "Estimation and prediction in spatial models with block composite likelihoods," *Journal of Computational and Graphical Statistics*.

Farah, M., Birrell, P., Conti, S., and Angelis, D. D. (2014), "Bayesian Emulation and Calibration of a Dynamic Epidemic Model for A/H1N1 Influenza," *Journal of the American Statistical Association*, 109, 1398–1411.

Forrester, A., Sobester, A., and Keane, A. (2008), *Engineering design via surrogate modelling: a practical guide*, John Wiley & Sons.

Fricker, T. E., Oakley, J. E., and Urban, N. M. (2013), "Multivariate Gaussian process emulators with nonseparable covariance structures," *Technometrics*, 55, 47–56.

Friedman, J. H. (1991), "Multivariate adaptive regression splines," *The annals of statistics*, pp. 1–67.

Furrer, R., Genton, M. G., and Nychka, D. (2012), "Covariance tapering for interpolation of large spatial datasets," *Journal of Computational and Graphical Statistics*.

Gelfand, A. E., Schmidt, A. M., Banerjee, S., and Sirmans, C. (2004), "Nonstationary multivariate process modeling through spatially varying coregionalization," *Test*, 13, 263–312.

Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (2010), *Handbook of spatial statistics*, CRC Press.

Gramacy, R. B. and Lee, H. K. (2012), "Bayesian treed Gaussian process models with an application to computer modeling," *Journal of the American Statistical Association*.

Gu, M., Palomo, J., and Berger, J. (2016), *RobustGaSP: Robust Gaussian Stochastic Process Emulation*, R package version 0.5.

Gupta, A. K. and Nagar, D. K. (1999), *Matrix variate distributions*, CRC Press.

Hartikainen, J. and Sarkka, S. (2010), "Kalman filtering and smoothing solutions to temporal Gaussian process regression models," in *Machine Learning for Signal Processing (MLSP), 2010 IEEE International Workshop on*, pp. 379–384, IEEE.

Higdon, D. et al. (2002), "Space and space-time modeling using process convolutions," *Quantitative methods for current environmental issues*, 3754.

Higdon, D., Swall, J., and Kern, J. (1999), "Non-stationary spatial modeling," *Bayesian statistics*, 6, 761–768.

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008), "Computer model calibration using high-dimensional output," *Journal of the American Statistical Association*, 103, 570–583.

Hodges, E., Molaro, A., Dos Santos, C. O., Thekkat, P., Song, Q., Uren, P. J., Park, J., Butler, J., Rafii, S., McCombie, W. R., et al. (2011), "Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment," *Molecular cell*, 44, 17–28.

Hoeffding, W. (1948), "A class of statistics with asymptotically normal distribution," *The annals of mathematical statistics*, pp. 293–325.

Hoff, P. D. (2009), *A first course in Bayesian statistical methods*, Springer Science & Business Media.

Homma, T. and Saltelli, A. (1996), "Importance measures in global sensitivity analysis of nonlinear models," *Reliability Engineering & System Safety*, 52, 1–17.

Iooss, B. and Lemaître, P. (2014), "A review on global sensitivity analysis methods," *arXiv preprint arXiv:1404.2405*.

Johnson, V. E. and Rossell, D. (2010), "On the use of non-local prior densities in Bayesian hypothesis tests," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 143–170.

Johnson, V. E. and Rossell, D. (2012), "Bayesian model selection in high-dimensional settings," *Journal of the American Statistical Association*, 107, 649–660.

Kaufman, C. and Shaby, B. (2013), "The role of the range parameter for estimation and prediction in geostatistics," *Biometrika*, 100, 473–484.

Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008), "Covariance tapering for likelihood-based estimation in large spatial data sets," *Journal of the American Statistical Association*, 103, 1545–1555.

Kaufman, C. G., Bingham, D., Habib, S., Heitmann, K., and Frieman, J. A. (2011), "Efficient emulators of computer experiments using compactly supported correlation functions, with an application to cosmology," *The Annals of Applied Statistics*, 5, 2470–2492.

Kazianka, H. and Pilz, J. (2012), "Objective Bayesian analysis of spatial data with uncertain nugget and range parameters," *Canadian Journal of Statistics*, 40, 304–327.

Kennedy, M. C. and O'Hagan, A. (2001), "Bayesian calibration of computer models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63, 425–464.

Korattikara, A., Chen, Y., and Welling, M. (2013), "Austerity in MCMC land: Cutting the Metropolis-Hastings budget," *arXiv preprint arXiv:1304.5299*.

Le Gratiet, L., Cannamela, C., and Iooss, B. (2014), "A Bayesian approach for global sensitivity analysis of (multifidelity) computer codes," *SIAM/ASA Journal on Uncertainty Quantification*, 2, 336–363.

Lee, L., Carslaw, K., Pringle, K., Mann, G., and Spracklen, D. (2011), "Emulation of a complex global aerosol model to quantify sensitivity to uncertain parameters," *Atmospheric Chemistry and Physics*, 11, 12253–12273.

Lee, L., Carslaw, K., Pringle, K., and Mann, G. (2012), "Mapping the uncertainty in global CCN using emulation," *Atmospheric Chemistry and Physics*, 12, 9739–9751.

Li, R. and Sudjianto, A. (2005), "Analysis of computer experiments using penalized likelihood in Gaussian Kriging models," *Technometrics*, 47.

Liaw, A. and Wiener, M. (2002), "Classification and regression by randomForest," *R news*, 2, 18–22.

Lim, Y. B., Sacks, J., Studden, W., and Welch, W. J. (2002), "Design and analysis of computer experiments when the output is highly correlated over the input space," *Canadian Journal of Statistics*, 30, 109–126.

Lindgren, F., Rue, H., and Lindström, J. (2011), "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 423–498.

Lindsay, B. G. (1988), "Composite likelihood methods," *Contemporary Mathematics*, 80, 221–39.

Lindsay, B. G., Yi, G. Y., and Sun, J. (2011), "Issues and strategies in the selection of composite likelihoods," *Statistica Sinica*, 21, 71.

Linkletter, C., Bingham, D., Hengartner, N., Higdon, D., and Kenny, Q. Y. (2006), "Variable selection for Gaussian process models in computer experiments," *Technometrics*, 48, 478–490.

Lohman, R. B. and Simons, M. (2005), "Some thoughts on the use of InSAR data to constrain models of surface deformation: Noise structure and data downsampling," *Geochemistry, Geophysics, Geosystems*, 6.

Lopes, D. (2011), "Development and implementation of Bayesian computer model emulators," Ph.D. thesis, Duke University.

Maclaurin, D. and Adams, R. P. (2014), "Firefly Monte Carlo: Exact MCMC with subsets of data," *arXiv preprint arXiv:1403.5693*.

Marrel, A., Iooss, B., Jullien, M., Laurent, B., and Volkova, E. (2011), "Global sensitivity analysis for models with spatially dependent outputs," *Environmetrics*, 22, 383–397.

Montgomery-Brown, E., Wicks, C., Cervelli, P. F., Langbein, J. O., Svarc, J. L., Shelly, D. R., Hill, D. P., and Lisowski, M. (2015), "Renewed inflation of Long Valley Caldera, California (2011 to 2014)," *Geophysical Research Letters*, 42, 5250–5257.

Morris, M. D., Mitchell, T. J., and Ylvisaker, D. (1993), "Bayesian design and analysis of computer experiments: use of derivatives in surface prediction," *Technometrics*, 35, 243–255.

Oakley, J. (1999), "Bayesian uncertainty analysis for complex computer codes," Ph.D. thesis, University of Sheffield.

Oakley, J. (2002), "Eliciting Gaussian process priors for complex computer codes," *Journal of the Royal Statistical Society: Series D (The Statistician)*, 51, 81–97.

Oakley, J. and O'Hagan, A. (2002), "Bayesian inference for the uncertainty distribution of computer model outputs," *Biometrika*, 89, 769–784.

Oakley, J. E. and O'Hagan, A. (2004), "Probabilistic sensitivity analysis of complex models: a Bayesian approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66, 751–769.

Park, J. S. (1991), "Tuning complex computer codes to data and optimal designs," .

Patra, A. K., Bauer, A., Nichita, C., Pitman, E. B., Sheridan, M., Bursik, M., Rupp, B., Webber, A., Stinton, A., Namikawa, L., et al. (2005), "Parallel adaptive numerical simulation of dry avalanches over natural terrain," *Journal of Volcanology and Geothermal Research*, 139, 1–21.

Paulo, R. (2005), "Default priors for Gaussian processes," *Annals of statistics*, 33, 556–582.

Paulo, R., García-Donato, G., and Palomo, J. (2012), "Calibration of computer models with multivariate output," *Computational Statistics and Data Analysis*, 56, 3959–3974.

Peng, C.-Y. and Wu, C. J. (2014), "On the choice of nugget in kriging modeling for deterministic computer experiments," *Journal of Computational and Graphical Statistics*, 23, 151–168.

Petris, G., Petrone, S., and Campagnoli, P. (2009), *Dynamic linear models*, Springer.

Picheny, V., Wagner, T., and Ginsbourger, D. (2013), "A benchmark of kriging-based infill criteria for noisy optimization," *Structural and Multidisciplinary Optimization*, 48, 607–626.

Pitman, E. B., Nichita, C. C., Patra, A., Bauer, A., Sheridan, M., and Bursik, M. (2003), "Computing granular avalanches and landslides," *Physics of Fluids (1994-present)*, 15, 3638–3646.

Pujol, G., Iooss, C. C., Michel, F., and Iooss, M. B. (2007), "The sensitivity Package," *R package version*, 1.

Rasmussen, C. E. (2006), *Gaussian processes for machine learning*, MIT Press.

Ren, C., Sun, D., and He, C. (2012), "Objective Bayesian analysis for a spatial model with nugget effects," *Journal of Statistical Planning and Inference*, 142, 1933–1946.

Ren, C., Sun, D., and Sahu, S. K. (2013), "Objective Bayesian analysis of spatial models with separable correlation functions," *Canadian Journal of Statistics*, 41, 488–507.

Rougier, J. (2008), "Efficient emulators for multivariate deterministic functions," *Journal of Computational and Graphical Statistics*, 17, 827–843.

Rougier, J., Guillas, S., Maute, A., and Richmond, A. D. (2009), "Expert knowledge and multivariate emulation: The thermosphere–ionosphere electrodynamics general circulation model (TIE-GCM)," *Technometrics*, 51, 414–424.

Roustant, O., Ginsbourger, D., and Deville, Y. (2012), "DiceKriging, DiceOptim: Two R Packages for the Analysis of Computer Experiments by Kriging-Based Metamodeling and Optimization," *Journal of Statistical Software*, 51, 1–55.

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations," *Journal of the royal statistical society: Series b (statistical methodology)*, 71, 319–392.

Sacks, J., Welch, W. J., Mitchell, T. J., Wynn, H. P., et al. (1989), "Design and analysis of computer experiments," *Statistical science*, 4, 409–423.

Saltelli, A., Chan, K., Scott, E. M., et al. (2000), *Sensitivity analysis*, vol. 1, Wiley New York.

Särkkä, S. and Hartikainen, J. (2012), "Infinite-dimensional Kalman filtering approach to spatio-temporal Gaussian process regression," in *International Conference on Artificial Intelligence and Statistics*, pp. 993–1001.

Savitsky, T., Vannucci, M., and Sha, N. (2011), "Variable selection for nonparametric Gaussian process priors: Models and computational strategies," *Statistical science: a review journal of the Institute of Mathematical Statistics*, 26, 130–149.

Scarano, M. I., Strazzullo, M., Matarazzo, M. R., and D'Esposito, M. (2005), "DNA methylation 40 years later: Its role in human health and disease," *Journal of cellular physiology*, 204, 21–35.

Schonlau, M. and Welch, W. J. (2006), "Screening the input variables to a computer model via analysis of variance and visualization," in *Screening*, pp. 308–327, Springer.

Severini, T. A. (2000), *Likelihood Methods in Statistics*, Oxford University Press, 1st edn.

Sobol', I. M. (1990), "On sensitivity estimation for nonlinear mathematical models," *Matematicheskoe Modelirovanie*, 2, 112–118.

Sobol', I. M. (2001), "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates," *Mathematics and computers in simulation*, 55, 271–280.

Sobol', I. M., Tarantola, S., Gatelli, D., Kucherenko, S., and Mauntz, W. (2007), "Estimating the approximation error when fixing unessential factors in global sensitivity analysis," *Reliability Engineering & System Safety*, 92, 957–960.

Spiller, E. T., Bayarri, M., Berger, J. O., Calder, E. S., Patra, A. K., Pitman, E. B., and Wolpert, R. L. (2014), "Automating emulator construction for geophysical hazard maps," *SIAM/ASA Journal on Uncertainty Quantification*, 2, 126–152.

Stein, M. (1987), "Large sample properties of simulations using Latin hypercube sampling," *Technometrics*, 29, 143–151.

Stein, M. L. (2012), *Interpolation of spatial data: some theory for kriging*, Springer Science & Business Media.

Sun, Y., Li, B., and Genton, M. G. (2012), "Geostatistics for large datasets," in *Advances and challenges in space-time modelling of natural events*, pp. 55–77, Springer.

Tuo, R. and Wu, C. (2015), "A theoretical framework for calibration in computer models: parametrization, estimation and convergence properties," *arXiv preprint arXiv:1508.07155*.

Tuo, R., Wu, C. J., et al. (2015), "Efficient calibration for imperfect computer models," *The Annals of Statistics*, 43, 2331–2352.

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*, pp. 2655–2675.

Varin, C., Reid, N., and Firth, D. (2011), "An overview of composite likelihood methods," *Statistica Sinica*, 21, 5–42.

Wang, H. and West, M. (2009), "Bayesian analysis of matrix normal graphical models," *Biometrika*, 96, 821–834.

West, M. and Harrison, P. J. (1997), *Bayesian Forecasting & Dynamic Models*, Springer Verlag, 2nd edn.

Whittle, P. (1954), "On stationary processes in the plane," *Biometrika*, pp. 434–449.

Whittle, P. (1963), "Stochastic process in several dimensions," *Bulletin of the International Statistical Institute*, 40, 974–994.

Xiao, M., Breitkopf, P., Coelho, R. F., Knopf-Lenoir, C., Sidorkiewicz, M., and Villon, P. (2010), "Model reduction by CPOD and Kriging," *Structural and multidisciplinary optimization*, 41, 555–574.

Zhang, H. (2004), "Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics," *Journal of the American Statistical Association*, 99, 250–261.

Zhang, W., Spector, T. D., Deloukas, P., Bell, J. T., and Engelhardt, B. E. (2015), "Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements," *Genome biology*, 16, 1–20.

Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T.-Y., Kohlbacher, O., De Jager, P. L., Rosen, E. D., Bennett, D. A., Bernstein, B. E., et al. (2013), "Charting a dynamic DNA methylation landscape of the human genome," *Nature*, 500, 477–481.

Biography

Mengyang Gu was born in 1989 in Nanchang, China. In June 2012, he received the Bachelor degree in the department of mathematics and an honor degree in Chu Kochen Honors College. Mengyang moved to Durham, NC in August 2012 to pursue doctoral studies at the Department of Statistical Science at Duke University.