Asymptotic Multivariate Kriging Using Estimated Parameters with Bayesian Prediction Methods for Non-linear Predictands

Elizabeth C. Shamseldin

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research (Statistics).

Chapel Hill
2008

Approved by

Advisor: Richard L. Smith

Reader: Joseph G. Ibrahim

Reader: Amarjit Budhiraja

Reader: Zhengyuan Zhu

Reader: Chuanshu Ji

ABSTRACT

ELIZABETH C. SHAMSELDIN: Asymptotic Multivariate Kriging Using Estimated Parameters with Bayesian Prediction Methods for Non-linear Predictands
(Under the direction of Richard L. Smith)

The motivation for this dissertation is the need that often arises in spatial settings to perform a data transformation to achieve a stationary process and/or variance stabilization. The transformation may be a non-linear transformation, and the desired predictand may be multivariate in that it is necessary to interpolate predictions at multiple sites. We assume the underlying spatial model is a Gaussian random field with a parametrically specified covariance structure, but that the predictions of interest are for multivariate nonlinear functions of the Gaussian field. This induces new complications in the spatial interpolation known as kriging. For instance, it is no longer possible to derive the predictive distribution function in closed form.

The underlying process of a spatial model is a stochastic process, and spatial prediction techniques are founded on the assumption that the realizations come from a Gaussian process with a known covariance structure, and known, specified parameters. A difficulty that arises with traditional kriging methods is the fact that the standard formula for the mean squared prediction error does not take into account the estimation of the covariance parameters. This generally leads to underestimated prediction errors, even if the model is correct. Smith and Zhu (2004) establish a second-order expansion for predictive distributions in Gaussian processes with estimated covariances. Here, we establish a similar expansion for multivariate kriging with non-linear predictands.

Bayesian methods provide a possible resolution to errors encountered through employing frequentist estimation techniques for obtaining spatial parameters. Bayesian analysis seeks to utilize prior and nonexperimental sources of information. Bayesian methods evalu-

ate procedures for a repeated sampling experiment of parameters drawn from the posterior distribution given a set of observed data. An important property of Bayesian methods is the ability to deal with the uncertainty in a particular model. A Bayesian paradigm enables a more realistic assessment of the variability inherent in estimating parameters of interest. Here we explore a Laplace approximation to Bayesian techniques that provides an alternative to common iterative Bayesian methods, such as Markov Chain Monte Carlo.

The theoretical Bayesian coverage probability bias for prediction intervals is computed and compared with the plug-in method using the restricted maximum likelihood estimates of the covariance parameters. The form of the Bayesian predictive distribution can be expressed as partial derivatives of the restricted log-likelihood. This leads to possible analytical evaluation, and a bootstrap method is explored to obtain predictions for general, non-linear predictands. The main results are asymptotic formulae for a general, non-linear predictand for the expected length of a Bayesian prediction interval, which has possible applications in network design, and for the coverage probability bias, which can lead to the development of a matching prior. The matching prior may be difficult to compute in practice. As an alternative, an asymptotic estimator is developed.

ACKNOWLEDGEMENTS

In loving dedication to my wonderful husband Sam,

and to the world's best parents, Michael and Suzanne Mannshardt.

Words cannot express my love and gratitude for your years of support and encouragement.

Infinite thanks to my dissertation advisor, Richard L. Smith.

Thank you to my dissertation committee.

In addition I would like to thank Amy Grady for her friendship, understanding and laughter. I will forever have Dr. Karen Nylund and Dr. Kirsten Doehler to thank for going through grad school with me, though from a distance. Many thanks to all of the students in the UNC Chapel Hill Statistics Department who have made my journey unforgettable. Ping Bai, Myung Hee Lee, Jeongyoun Ahn, and Jungyeon Yoon will always be my stat girls. Thanks to Francisco Chamu for countless hours of assistance, Brian Lopes and Bret Hanlon for countless hours of ridiculousness, and Rima Izem for countless hours of dancing.

Thank you to the professors of the Mathematics Department at Sonoma State University who encouraged my love of the subject and put on a great barbeque. Specifically Susan Herring, Brian Jerskey, Bill Barnier, and Jean Chan.

Thank you to all teachers, especially those who believe in their students - Brian Monchamp, Don Graham, Nancy Aaberg, Loren Grahn, and Barbara Patterson of Yuba City, CA.

Thank you to the Yuba City Rotary Club as well as Gayle and Mitzi Morrison of Yuba City for their financial support but more importantly their encouragement.

Lastly, thank you to PhDComics.com for preserving my sanity by making me cry hysterically and laugh uproariously - at the same time.

CONTENTS

LIST OF FIGURES

CHAPTER 1

Introduction

Spatial prediction techniques are founded on the assumption that the realizations come from a Gaussian process with a known covariance structure, and known, specified parameters. Interpolation of the spatial process to points across a network is traditionally achieved through *kriging*, a technique for predicting unobserved values of the random field using linear combinations of the observed variables. Universal kriging is employed when the process mean is a linear combination of covariates.

A difficulty that arises with traditional kriging methods is the fact that the standard formula for the mean squared prediction error does not take into account the estimation of the covariance parameters. Typically, the estimation of the spatial covariance model of the underlying parameters is performed first, then the model is developed to obtain predictions. Since the parameter values are treated as known, this generally leads to underestimated prediction errors, even if the model is correct.

Several techniques have been proposed to deal with what is widely assumed to be the underestimation of the prediction standard errors. Zimmerman and Cressie (1992) and Stein (1999) have proposed methods for approximating the mean of the prediction errors. However, these methods are somewhat ad hoc, and it is often assumed that the predictive distribution is normal. Bayesian techniques have been proposed for these problems as well. However, there is no specific proof that Bayesian methods are superior.

Smith and Zhu (2004) establish a second-order expansion for predictive distributions in

Gaussian processes with estimated covariances. These issues are examined using second-order asymptotics to compare the "plug-in" approach to prediction using ordinary and universal kriging, to the intervals constructed using Bayesian methods.

Here, we establish a similar expansion for multivariate kriging with non-linear predictands. The Bayesian coverage probability bias for prediction intervals is computed and compared with the plug-in method using the restricted maximum likelihood estimates of the covariance parameters. The form of the Bayesian predictive distribution can be expressed as partial derivatives of the restricted log-likelihood. This leads to possible analytical evaluation, and a boot-strap method is explored to obtain predictions for general, non-linear predictands.

The main results are explicit formula for a general, non-linear predictand for the expected length of a Bayesian prediction interval, which has possible applications in network design, and for the coverage probability bias. Smith and Zhu also proposed a "matching prior" for which the second-order coverage probability bias goes to zero. Here the existence of a possible "matching prior" for non-linear predictands is also addressed.

CHAPTER 2

Bayesian and Spatial Statistics

2.1 Spatial Statistics

Spatial models are concerned with the underlying covariance structure between a collection of measurements. The underlying process of a spatial model is a stochastic process $Z(s)$, where s is a location in $\mathbb{R}^d$. Usually $d \in \{1, 2, 3\}$. Generally, the process is assumed to be Gaussian with known mean μ and covariance structure Σ which are specified through the parameter vector θ. Realizations from the process can be expressed as:

$$Z(s) = \mu(s) + e(s)$$

where $e(s)$ is a zero mean error process.

The basic model assumes that for a finite-dimensional observation vector Z, we have $Z \sim N(\mu, \Sigma)$, with Σ the covariance matrix for a constant mean μ. For a model assuming a mean as a linear function of covariates, the model takes the form $Z \sim N(X\beta, \Sigma)$, with X a matrix of covariates and β a vector of unknown regression coefficients.

2.1.1 Covariance Structures

Often spatial processes are assumed to be *stationary* and *isotropic*. A process is *stationary* if the joint distribution of the process evaluated at any set of points is not changed if all of the points are shifted by h. Particularly, $E(Z(s)) = E(Z(s+h))$, ie a constant mean. Also for two locations, s_i and s_j, $\sigma(s_i, s_j) = \sigma(s_i + h, s_j + h)$. A process is *isotropic* if it is rotation invariant, ie the properties of the process depend only on the distance between

two points and not on their direction.

If the process is stationary and isotropic, the underlying covariance structure can be modeled using a *variogram*. The variogram can exist under certain conditions where the covariance function cannot. The value of the variogram at distance d is the variance of the difference between two measurements that are distance d apart.

The covariance structure is expressed through a variogram function where

$$\text{var}\{Z(s_1) - Z(s_2)\} = 2\gamma(s_1 - s_2).$$

An example is the general form for the power exponential variogram:

$$\gamma(s_i - s_j) = \begin{cases} 0 & \text{if } s_i = s_j, \\ c_0 + c_1(1 - \exp[-(\frac{d_{ij}}{\phi})^\kappa]) & \text{if } s_i \neq s_j. \end{cases}$$

where d_{ij} is the euclidean distance between sites s_i and s_j and $0 < \kappa \leq 2$.

In the model concerning the mean as a linear combination of covariates, the process is not stationary. However, when constructing prediction using the process of best linear unbiased prediction, stationarity assumptions are not required. Here a covariance structure is introduced. For the power exponential model, the corresponding covariance function can be expressed:

$$cov\{Z(s_i), Z(s_j)\} = \begin{cases} \sigma^2(1 + \nu) & \text{if } s_i = s_j \\ \sigma^2 \exp\left(-\left(\frac{d_{ij}}{\phi}\right)^\kappa\right) & \text{if } s_i \neq s_j \end{cases} \tag{2.1}$$

where $\Sigma(\theta)$ is a vector of standardized covariances determined by the unknown parameter vector $\theta = (\sigma^2, \phi, \kappa)$. The underlying covariance structure is introduced through $V(\theta)$ where $V(\theta)$ is a matrix with diagonal entries $1 + \nu$ and off-diagonal entries $v_{ij} = \exp\left(-(\frac{d_{ij}}{\phi})^\kappa\right)$. ϕ is the range parameter, κ is a smoothness parameter, and ν is the nugget effect. The nugget

effect accounts for the discontinuity at $d_{ij} = 0$, which can be due to measurement error. ν can be estimated as a part of θ, but it is often assumed to be 0. The above equation can be expressed $\Sigma = \sigma^2 V(\theta)$ where σ^2 is an unknown scale parameter.

2.1.2 Kriging

Kriging is a technique for predicting values at unobserved locations within the random field through linear combinations of the observed variables. Kriging refers to the construction of a spatial predictor in terms of known model parameters. *Ordinary kriging* is used when the mean process is an unknown constant. *Universal kriging* is applied when the mean process is a linear combination of covariates.

The universal kriging model can be be written

$$Z(s) = X(s)\beta + e(s),\ E(e(s)) = 0,\ \mathrm{Cov}(e) = V(\theta) \text{ where } V(\theta) = [\sigma(s_i, s_j)] \tag{2.2}$$

If we write the vector of known observations as Y and the value to be predicted as Y_0 where Y_0 is a scalar, then we have

$$\begin{pmatrix} Y \\ Y_0 \end{pmatrix} \sim N\left(\begin{pmatrix} X\beta \\ x_0^T\beta \end{pmatrix}, \begin{pmatrix} V(\theta) & w^T(\theta) \\ w(\theta) & v_0(\theta) \end{pmatrix}\right) \tag{2.3}$$

where X is the $n \times p$ vector of covariates for the observations Y, and x_0 is the $p \times 1$ vector of covariates for the predicted scalar Y_0, β is the vector of regression coefficients, and $\theta = (\sigma^2, \phi, \kappa)$ is the vector of covariance parameters.

Universal kriging aims to find a linear predictor $\hat{Y_0}$:

$$\hat{Y_0} = \lambda^T Y$$

subject to the condition $X^T\lambda = x_0$.

This leads to $E\left\{(Y_0 - \hat{Y_0})\right\} = 0$, so the predictor is unbiased. The best linear unbiased predictor chooses λ to minimize the Mean Squared Prediction Error (MSPE), $E\left\{(Y_0 - \hat{Y_0})^2\right\}$. Using Lagrange multipliers, the optimal λ is:

$$\lambda(\theta) = V^{-1}(\theta)w(\theta) + V^{-1}(\theta)X(X^T V^{-1}(\theta)X)^{-1}(x_0 - X^T V^{-1}(\theta)w(\theta)). \tag{2.4}$$

with corresponding MSPE:

$$\begin{aligned} \sigma_0^2(\theta) &= v_0(\theta) - w(\theta)^T V^{-1}(\theta)w(\theta) \\ &+ (x_0 - X^T V^{-1}(\theta)w(\theta))^T (X^T V^{-1}(\theta)X)^{-1}(x_0 - X^T V^{-1}(\theta)w(\theta)). \end{aligned} \tag{2.5}$$

Thus a definition of the predictive distribution function is

$$\Pr\{Y_0 \leq z \mid Y = y, \theta\} = \psi(z; y, \theta) = \Phi\left(\frac{z - \lambda(\theta)^T y}{\sigma_0(\theta)}\right)$$

Traditional methods estimate the parameters through Least Squares Estimation, Maximum Likelihood methods (MLE), and Restricted Maximum Likelihood (REML) methods. REML estimation is based on the joint density of the vector of contrasts, whose distribution is independent of the population mean (p68, Smith, 2001). The resulting maximum likelihood estimator based on the joint density of contrast is approximately unbiased, as opposed to MLE estimators which can be biased.

To solve for the REML estimates, the log likelihood is written

$$l_n(\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|V(\theta)| - \frac{1}{2}(Y - X\beta)^T V(\theta)^{-1}(Y - X\beta) \tag{2.6}$$

Given θ, we estimate $\hat{\beta}(\theta) = (X^T V^{-1}(\theta)X)^{-1}$ and write $G_r^2(\theta)$ for the corresponding generalized residual sum of squares

$$G_r^2(\theta) = Y^T\{V^{-1}(\theta) - V^{-1}(\theta)X(X^T V^{-1}(\theta)X)^{-1}X^T V^{-1}(\theta)\}Y$$

Then the restricted log likelihood function is given by (Smith, 2001 and Stein, 1999):

$$l_n^*(\theta) = -\frac{n-q}{2}\log(2\pi) + \frac{1}{2}\log|X^TX| - \frac{1}{2}\log|X^TV(\theta)^{-1}X| - \frac{1}{2}\log|V(\theta)| - \frac{1}{2}G_r^2(\theta)$$

Using the REML estimator $\hat{\theta}$, the REML predictive distribution function is given by:

$$\hat{\psi}(w; z, \theta) = \psi(w; z, \hat{\theta})$$

2.2 Bayesian Methods

Frequentist methods imagine repeated sampling from the model, the likelihood model, which defines the probability distribution of the observed data conditional on unknown parameters (Carlin and Louis, 2000, p 4). Bayesian analysis is the approach to statistics that formally seeks to utilize prior and nonexperimental sources of information (Berger, 1980, p 3). The Bayesian approach requires a sampling model and a prior distribution on all unknown parameters in the model. The likelihood and the priors are then used to compute the posterior distribution, ie the conditional distribution of the parameters given the observed data. Bayesian methods evaluate procedures for a repeated sampling experiment of parameters drawn from the posterior distribution given a set of observed data. An important property of Bayesian methods is the ability to deal with the uncertainty in a particular model. A Bayesian paradigm enables a more realistic assessment of the variability inherent in estimating parameters of interest. Specific marginal distributions can focus on specific parameters, and the integration involved ensures that all uncertainties involved influence the spread and shape of the marginal posterior distribution (Carlin and Louis, p 13).

In parametric inference, the likelihood is specified according to its parameters, θ.

$$l(\theta) \propto p(x|\theta)$$

Generally, θ is unknown. We can account for the uncertainty in θ by expressing a prior

distribution for θ, $\pi(\theta)$, which is the density of θ before x is observed. Using Bayes Theorem, we can construct the posterior distribution for θ, $p(\theta|x)$. Bayes Theorem says

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int_\Theta p(x|\theta)\pi(\theta)d\theta}$$

where Θ is the sample space of θ. We can then sample from the posterior in order to obtain estimates for θ.

One of the advantages of Bayesian inference is that it obeys the likelihood principle. The likelihood principle states that if two different sampling designs yield proportional likelihoods for θ, then the inference about θ should be the same for each design. Frequentist inference does not always obey the likelihood principle and different designs can lead to substantially different conclusions based on the same data.

2.2.1 Markov Chain Monte Carlo (MCMC) Methods

Markov Chain Monte Carlo methods provide a way to sample from a posterior distribution using an iterative algorithm. This utilizes the idea that the samples from the posterior distribution will converge to the true distribution.

Gibbs Sampling

Gibbs sampling is a common MCMC method. Gibbs sampling is convenient because it does not require a closed form for the conditional distributions. Thus the conditionals only need to be specified up to a normalizing constant. The Gibbs algorithm assumes that the full conditionals are available for sampling and that under certain conditions (Besag, 1974) the one-dimensional conditional distributions uniquely determine the full joint distribution (Ibrahim, 2000 p 171).

For k random variables $Y = (Y_1, ..., Y_k)$ we write the full conditional distributions as:

$$p(y_i|y_j, i \neq j),\ i = 1, \cdots, k$$

The algorithm starts with a set of arbitrary starting values $\{Y_1^{(0)}, ..., Y_k^{(0)}\}$. $Y_1^{(1)}$ is

sampled from the conditional $[Y_1|Y_2^{(0)}, ..., Y_k^{(0)}]$. Next $Y_2^{(1)}$ is sampled from the conditional $[Y_2|Y_1^{(1)}, Y_3^{(0)}, ..., Y_k^{(0)}]$, etc., through sampling $Y_k^{(1)}$ from $[Y_k|Y_1^{(1)}, ..., Y_{k-1}^{(1)}]$. This completes the first sampling iteration and yields $(Y_1^{(1)}, ..., Y_k^{(1)})$. Then the algorithm is repeated using the first iteration's values as the starting values.

As shown by Geman and Geman (1984), under suitable conditions,

$$(Y_1^{(t)}, ..., Y_k^{(t)}) \xrightarrow{d} [Y_1, ..., Y_k] \text{ as } t \to \infty$$

We are interested in sampling from the joint posterior distribution $[Y_1, ..., Y_k|x]$ where x represents the observed values. The Gibbs sampler requires draws from each of the univariate conditional distributions $p(y_i|y_j, x, i \neq j)$. Thus, using the Rao-Blackwellization technique, Gelfand and Smith (1990), a marginal density estimate of Y_i is:

$$\hat{p}(y_i|x) = \frac{1}{K} \sum_{j=1}^{K} p(y_i|y_{1,j}^{(t)}, ..., y_{i,j}^{(t)}, y_{i+1,j}^{(t)}, ..., y_{k,j}^{(t)}, x)$$

where K is the number of total iterations.

The obvious way to calculate the marginal posterior density of one parameter is to take the MCMC output and use a kernel density estimator. However, it is better to average the sequence of conditional densities, as shown here. This is known as the Rao-Blackwellization technique because of the analogy with the Rao-Blackwell theorem that arises in estimation theory. However this procedure is only available when the conditional densities are available in closed form which will not be the case for all MCMC procedures.

Metropolis-Hastings Algorithm

Another widely used MCMC method is the Metropolis-Hastings algorithm. Here we summarize the technique as outlined in Carlin and Louis (2000.) Assume that the true joint posterior for a parameter U has density $p(u)$ with respect to some measure μ. Choose an auxiliary function $q(v, u)$ as a *candidate* or *proposal* density such that $q(\cdot, u)$ is a pdf with

respect to μ for all u and is symmetric for all u, v. Then generate a Markov chain:

1. For the current state of the Markov chain, draw $v \sim q(;u)$, where $u = U^{(t-1)}$
2. Compute ratio $r = p(v)/p(u)$
3. If $r \geq 1$, set $U^{(t)} = v$;

$$\text{If } r < 1, \text{ set } U^{(t)} = \begin{cases} v, & \text{with probability } r \\ u & \text{with probability } 1 - r \end{cases}$$

Note that the joint posterior density p is only needed up to the proportionality constant, used in computing the acceptance ratio in step 2. In Bayesian appliactions, $p(u) \propto L(u)\pi(u)$ is a typically available form.

The main theorem for the Metropolis-Hastings method is:
For the Metropolis algorithm outlined above, under certain mild conditions, $U^{(t)} \xrightarrow{d} U \sim p$as $t \to \inf$.

CHAPTER 3

Literature Review

3.1 Introduction

With the increasing popularity of Bayesian methods, the field of available literature is rapidly expanding. Several recent papers highlight advances specific to the aim of this paper. Berger, De Oliveira, and Sansó's 2001 paper considers prior definition specific to the spatial setting. They also examine coverage probability bias through simulation. Datta and Ghosh (1995) consider the frequentist coverage probability bias and a general class of priors that satisfies the matching prior criterion. Levine and Casella (2003) expand on Datta and Ghosh to develop an algorithm for establishing matching priors through numerical solutions of partial differential equations.

The field of spatial statistics also encompasses a wide range of methods for dealing with spatial interpolation. Higdon, Swall, and Kern develop a hierarchical model which incorporates the uncertainty involved in model specification. Due to the explicit form developed of the covariance function of the process, the likelihood function for the process can be expressed at any configuration of points. This lends itself to the Bayesian approach developed in the current paper.

A motivation for this paper is the necessity of data transformations, possibly non-linear for multivariate distributions, as well as the univariate transformation explored in Smith, Kolenikov, and Cox (2003.) The exploration of methods for spatially correlated data, including kriging for spatial interpolation, as outlined in Smith, Kolenikov, and Cox is relevant

to the methodology developed here.

Smith and Zhu (2004) consider several properties of predictive inference which while not limited to spatial processes have useful applications in spatial statistics. Here they develop a second-order expansion for predictive distributions in Gaussian processes using estimated covariances where the covariance parameters are obtained using restricted maximum likelihood estimated. Smith and Zhu focus on the estimation of quantiles for the predictive distribution and the application to prediction intervals. They consider both a "plug-in" approach and a Bayesian approach. The Bayesian approach proves superior in the tails of the distribution regardless of the prior implemented. The second-order coverage probability bias is also considered and a frequentist correction is established that has zero second-order coverage probability bias. This is analogous to the existence of a "matching prior" for the Bayesian method. Another key result is an expression for the expected length of a prediction interval. Smith and Zhu provide the original development for the univariate normal predictive distribution of the methods considered in this dissertation for the nonlinear multivariate case.

The design of network criteria is an important consideration in spatial interpolation. Zimmerman considers optimal spatial network design for three design objectives. These include efficient prediction under assumptions of known covariance parameters, estimation of unknown covariance parameters, and a combination of the two where efficient prediction is the objective when the covariance parameters are unknown. The methods developed in this dissertation also consider the case where both covariance parameter estimation and prediction over unobserved sites is the objective.

In order to work in practice with an unknown or perhaps computationally difficult predictive distribution, kernel density estimation can be employed. Silverman (1986) provides a very thorough development of several techniques for kernel density estimation. Here we detail his general overview of density estimation, including various approaches to choosing a density and an appropriate smoothing parameter. We take a more detailed look at the

Epanechnikov kernel which is suitable for the estimation methodology developed in our paper. Park and Marron (1990) more fully compare methods for bandwidth selection, building on the methods introduced in Silverman.

The methodology outlined in our paper relies on a form of a parametric bootstrap. To complete our review, we explore some of the methods and literature pertaining to bootstrapping methods. Efron (2000) provides a basic overview, and compares the accuracy of bootstrapping techniques to older methods relying on Taylor series approximations.

3.2 Berger, Oliveira, and Sansó (2001)

Our paper considers approximations to Bayesian methods, which are dependent on the prior chosen. Berger, Oliveira, and Sansó's 2001 paper considers prior definition specific to the spatial setting. Coverage probability bias is an important result of the approximations developed in our paper, and Berger, Oliveira and Sansó also examine coverage probability bias through simulation.

3.2.1 Summary

Spatial data is often modeled using a Gaussian random field, specified by its mean function and covariance function. The spatial correlation structure is usually specified to be of a given form, such as exponential or Matérn, with a small number of unknown parameters. When considering Bayesian analysis of these spatial models, it is necessary to determine an objective prior distribution for the unknown mean and covariance parameters of the random field.

The aim of this paper is first to show that common choices such as the constant prior and Jeffrey's prior often lead to improper distributions for this model. The reference prior is then developed and shown to yield a proper posterior. Further, the reference prior can be used for computation of posterior probabilities of hypotheses to compare correlation functions.

3.2.2 Introduction

One of the main advantages of Bayesian inference is that the parameter uncertainty is fully accounted for when performing inference and prediction, even in small samples. Here objective Bayesian analysis of spatial data utilizing noninformative or conventional priors for the unknown parameters of the Gaussian field is considered. These priors are often used because of the difficulty interpreting and thus eliminating the correlation parameters. Commonly used noninformative priors can result in improper posterior distributions. The motivation here is to find noninformative priors that lead to proper posterior distributions and have additional desirable properties. One of these properties is the ability to directly compute Bayes factors in order to compare possible spatial covariance functions. The choice of spatial covariance functions can be arbitrary, so it is an easy and powerful method of comparison that is usually computationally intensive for traditional noninformative priors.

The recommendation here is a noninformative "exact" reference prior. This technique also appears to apply to numerous nonspatial models with certain mean and covariance structures, such as standard time series models.

In Berger, Oliveira, and Sansó's paper, the spatial model is presented along with certain spatial correlation functions and explicit results. The general form of the noninformative prior is considered and the posterior impropriety is discussed along with possible solutions. The recommended objective prior is introduced, and the formal development and behavior is presented. The exact reference prior is studied and it is shown that it results in a proper posterior. Applications and generalizations are also discussed.

The Model

Let $\{Z(s), s \in D \subseteq \mathbb{R}^l\}$ be the random field of interest where the data consist of n observations $\mathbf{Z} = (Z(s_1), ..., Z(s_n))$ where $s_1, ..., s_n$ are known sampling locations in D. Here the interest lies in estimation of the mean and covariance functions, and the predic-

tion of an unobserved random vector $\mathbf{Z}_0$. Assume the $Z(\cdot)$ is a Gaussian random field with $E[Z(s)] = \beta^{\mathbf{T}}\mathbf{f}(s)$ where $\beta = (\beta_1, ..., \beta_p)^T$ are unknown regression coefficients and $\mathbf{f(s)} = (f_1(s), ..., f_p(s))^T$ are known location-dependent covariates. The covariance structure is cov{Z(s), Z(u)} $= \sigma^2 K_\theta(||s-u||)$ where $||\cdot||$ denotes the Euclidean distance between locations s and u, $\sigma^2 =$ Var{Z(s)}, and $K_\theta(||s-u||) =$ corr$\{Z(s), Z(u)\}$ is an isotropic correlation function.

The likelihood of the model parameters $(\beta, \boldsymbol{\sigma^2}, \theta)$ based on the observed data z is

$$L(\beta, \sigma^2, \theta; z) = (2\pi\sigma^2)^{-\frac{n}{2}} |\Sigma_\theta^{-\frac{1}{2}}| \times \exp\left\{-\frac{1}{2\sigma^2}(z - X\beta)^T \Sigma_\theta^{-1}(z - X\beta)\right\}$$

where X is the known $n \times p$ matrix defined by $X_{ij} = f_j(s_i)$, assumed to be of full rank, and Σ_θ defined by $\Sigma_{\theta,ij} = K_\theta(||s_i - s_j)||$ is the $n \times n$ covariance matrix.

The results apply to general covariance structures with two properties. The covariance function decreases with distance $d = ||s - u||$ and the limiting values are 1 at $d = 0$ and 0 at $d = \infty$, common associations found in spatial data. A common covariance structure is the Power Exponential model, as given in Equation (2.1) on page 4.

Improper Priors and Possible Solutions

Improper prior densities for $(\beta, \sigma^2, \theta)$ are considered of the form:

$$\pi(\beta, \sigma^2, \theta) \propto \frac{\pi(\theta)}{(\sigma^2)^a},\ a \in \mathbb{R}$$

The most commonly used version of the noninformative prior is with $a = 1$ and $\pi(\theta) = 1$. However, this leads to an improper posterior distribution for $(\beta, \sigma^2, \theta)$. This impropriety also holds for many other common choices of noninformative priors, including the prior with $\pi(\theta) = \frac{1}{\theta}$ and the Laplace prior $\pi(\beta, \sigma^2, \theta) \propto 1$.

There are a variety of possible solutions to consider.

Proper Priors: This is the most obvious way to guarantee a proper posterior. However, in the spatial setting, the correlation parameters can be difficult to interpret and thus elicit.

Truncation of the Parameter Space: To avoid difficulties with improper posteriors, bounds are often placed on the parameter space. This leads to a proper posterior, however often the ensuing inferences are highly dependent on the actual bounds used.

Vague Proper Priors: Using a vague proper prior, such as the inverse gamma distribution (IG(ϵ, ϵ) with ϵ very small positive hyperparameters) does not guarantee a solution to the problem. Often the answer is extremely sensitive to the choice of hyperparameters. For example, the spatial problem that uses IG(ϵ, ϵ) prior distribution for θ results in a posterior for θ that concentrates all its mass near 0.

Transformation of θ This approach transforms θ to a bounded interval through a differentiable, 1-1 transformation g$(\cdot)$ and then placing a noninformative, proper prior on the transformed parameter. However, this approach is essentially equivalent to choosing a subjective proper prior since choosing the transformation g$(\cdot)$ is equivalent to choosing $\pi(\cdot)$.

Jeffreys Prior The most common noninformative prior is the Jeffreys prior. There are two different Jeffreys priors. The first is the *Jeffreys-rule* prior, which is the square root of the determinant of the information matrix. Jeffreys-rule prior uses the information matrix for all of the model parameters. The second, the *independence Jeffreys* prior, assumes that the unknown mean parameters are independent of the unknown covariance parameters and separately applies the Jeffreys-rule prior to each. In the vast majority of cases, the Jeffreys prior leads to a proper posterior. However, in the spatial setting when the random field has an unknown mean level Jeffreys prior fails to yield a proper posterior.

The Reference Prior: The reference prior approach attempts to improve the Jeffreys prior in multiparameter settings. The reference prior decomposes the problem into conditional,

lower dimensional problems for which noninformative priors can be computed. However, in the spatial setting, the reference prior resulted in the same behavior as the independence Jeffreys prior with respect to posterior propriety.

The following algorithm is often used to define a reference prior. The algorithm (1) finds a conditional reference prior for any parameters deemed nuisance parameters, (2) integrates out the nuisance parameters, and (3) finds the marginal reference prior in the integrated model. Traditionally, this has been done through asymptotic approximation, however for the spatial setting as well as a larger class of problems concerning a linear component and particular correlation structures, the approximation fails to give the true reference prior. Therefore, the traditional algorithm doesn't work in this setting for the spatial application.

The correct reference prior can be found be carrying out the algorithm using the exact integrated model. The resulting reference prior, $\pi^R(\beta, \sigma^2, \theta)$ is of the form:

$$\pi^R(\theta) \propto \left\{ \text{tr}[W_\theta^2] - \frac{1}{n-p}(\text{tr}[W_\theta])^2 \right\}^{1/2}$$

where a = 1 and $W_\theta = ((\frac{\partial}{\partial\theta})\Sigma_\theta)\Sigma_\theta^{-1}P_\theta^\Sigma$ and $P_\theta^\Sigma = I - X(X^T\Sigma_\theta^{-1}X)^{-1}X^T\Sigma_\theta^{-1}$. It is shown that this prior always yields a proper posterior for $(\beta, \sigma^2, \theta)$ and is the recommended default prior.

Comparison of the Reference and the Jeffreys-Rule Priors

One way to evaluate default priors is through the study of the resulting frequentist properties. One type of noninformative prior yields credible sets whose frequentist coverage is asymptotically as close to optimal as possible is termed a *matching prior*. Matching priors can be difficult to find in multivariate settings. Experiences have shown there is substantial evidence that reference priors lead to credible sets with satisfactory frequentist coverage.

A small simulation experiment was performed to investigate the frequentist coverage of two-tailed Bayesian credible intervals for the range parameter θ. The reference prior and

the Jeffreys-rule prior are compared. An isotropic Gaussian random field $Z(\cdot)$ is sampled at $n = 25$ lattice locations in $\mathcal{D} = [0,1] \times [0,1]$. Two mean functions are considered, the constant .15 (dimension $p = 1$) and $.15 - .65x - .1y + .9x^2 - xy + 1.2y^2$ (dimension $p = 6$) along with three exponential covariance functions, $C(d) = .12\exp\{-d/\theta\}$, where θ is .2, .5, or 1.0. From each of the 3,000 replications for each of the six possible models, equal-tailed 95% credible intervals are computed for θ based either on the reference or the Jeffreys-rule prior, with the mean functions treated as constants. The resulting approximate frequentist coverage probabilities of the intervals are computed. The results show that for small p, (ie $p = 1$), the empirical coverage probabilities are reasonable for the reference prior in all cases except when $p = 6$ and $\theta = 1.0$. Note here that for $\theta = 1.0$ there is quite strong spatial correlation which effectively reduces the sample size. In contrast, the frequentist results for the Jeffreys-rule prior are highly inadequate when $p = 6$ regardless of the value of θ.

Length and bias of the intervals are also considered. When $p = 1$ and the coverage probabilities are acceptable for either prior, the reference prior generally exhibits shorter intervals. When $p = 6$, the Jeffreys-rule prior exhibits shorter intervals, however the coverage probability for the Jeffreys-rule prior in this case is unacceptable. Examination showed that the intervals were biased to the left, ie the upper interval was less than θ.

This study does not establish decisively that the reference prior generally yields satisfactory frequentist performance. However it does establish strong evidence that the Jeffreys-rule prior can be seriously inadequate in terms of frequentist performance.

Model Selection

Choice of a family of covariance functions for the spatial model is often arbitrary, and even once a family is chosen, choosing the smoothing parameter can be difficult. Bayesian model selection can help with both of these issues. Standard Bayesian model selection often cannot be performed with improper priors and more elaborate techniques must be employed. An exception is when the models being considered have the same invariance structure up to individual model parameters with proper priors. The spatial model fits this when (1) the models compared have mean functions of the same structure, (2) the priors

for the models are of the form $\pi(\beta, \sigma^2, \theta) \propto \frac{\pi(\theta)}{(\sigma^2)^a},\ a \in \mathbb{R}$ and (3) the priors $\pi(\theta)$ are proper. The last two conditions are satisfied by the reference prior when normalized.

When selecting a single model for use, the model with the largest posterior probability would typically be chosen. When prediction is the goal, as in many spatial settings, model averaging can yield considerably better results. Model averaging bases the predictions on the posterior-weighted average of the individual model predictions. The Bayesian approach accommodates model averaging, which is a considerable strength of the approach.

3.2.3 Conclusion

The basic justification for the use of the reference prior is that it yields a proper posterior. An additional feature is the ability to utilize the reference prior to select a spatial correlation function. The simulation study performed also suggests that the reference prior yields credible sets with desirable frequentist coverage in contrast to the Jeffreys-rule prior.

Areas of future interest include various generalizations of the correlation functions to consider additional parameters such as the nugget effect. Another area of interest is allowing the smoothness parameter to be completely unknown. The issues to be considered are whether the resulting posterior is guaranteed to be proper and finding efficient algorithms.

3.3 Datta and Ghosh (1995)

An important results of our paper is the methodology to approximate the coverage probability bias, and the possibility of a matching prior that results in a coverage probability bias of zero. Datta and Ghosh (1995) consider the frequentist coverage probability bias and a general class of priors that satisfies the matching prior criterion.

3.3.1 Summary

Datta and Ghosh compare the reference priors presented by Berger and Bernardo and the reverse reference priors proposed by J. K. Ghosh. These two classes of priors can agree under certain conditions. They are also compared under a criterion that requires the frequentist coverage probability of the posterior region to match a nominal level with a

remainder of $O(n^{-1})$. Finally, a general class of priors that satisfies the matching criterion separately for each parameter is constructed and it is shown how the reference or reverse reference priors fit within this class of priors.

3.3.2 Introduction

Bayesian methods often use noninformative priors. Two of the most widely used have their limitations. A uniform (and possibly improper) prior can be noninvariant under transformation or reparameterization. Jefferys' prior, the positive square root of the determinant of the Fisher information matrix, is invariant under transformation but can encounter difficulties if nuisance parameters are involved.

Bernardo introduced the reference prior approach in 1979 by dividing the parameter vector into parameters of interest and nuisance parameters. This was extended and generalized by Berger and Bernardo (1989, 1992a,b) who split the parameter vector into two or more groups according to their importance and define a general algorithm for the construction of reference priors.

There have been several other proposals in recent years, leading to a wide choice of noninformative priors. Usually the choice is determined by matching the Bayesian solution to the frequentist solution. One of the most widely used class of priors is due to Peers (1965) and are derived by requiring the frequentist coverage probability of the posterior region of a real-valued parametric function to match the nominal level with a remainder of $O(n^{-1})$. A prior satisfying this criterion is deemed a "matching prior." In practice these priors are found through the solution to a partial differential equation.

The reverse reference prior is derived by following the algorithm of Berger and Bernardo, but reversing the parameter of interest and the nuisance parameter. The prior that meets the matching criterion (Welch and Peers, 1963) under orthogonality is the "reverse reference prior." The main focus of this article is to compare the reference priors and the reverse reference priors. Although reverse reference priors are matching priors under orthogonality, this is not necessarily true for reference priors. This is illustrated using a lognormal example. Other proposed priors are also examined to determine if they meet the matching

criteria. Construction of reference priors in random-effects models is examined, including necessary and sufficient conditions for satisfaction of the matching criterion.

Reference Priors, Reverse Reference Priors, and Matching Priors

Let the parameter vector $\theta = (\theta_1, ..., \theta_k)^T$ be group ordered as $\theta = (\theta_{(1)}, ..., \theta_{(m)})$ where $\theta_{(i)}$ has n_i coordinates and $\sum_{i=1}^{m} n_i = k$ and $\theta_{(1)}$ is of more importance than $\theta_{(2)}$, etc. Assume that the Fisher information matrix of θ is

$$I(\theta) = \text{block diagonal } (h_1(\theta), ..., h_m(\theta))$$

where $h_j(\theta)$ is $n_j \times n_j$ and may not be diagonal. Define $\theta_{(j)}^C = (\theta_{(1)}, ..., \theta_{(j-1)}, \theta_{(j+1)}, ..., \theta_{(m)}$ and assume $|h_j(\theta)| = h_{j1}(\theta_{(j)}) h_{j2}(\theta_{(j)}^C)$ for nonnegative h_{j1} and h_{j2}. Then

$$\pi_R(\theta) \equiv \pi_{RR}(\theta) = \prod_{j=1}^{m} h_{j1}^{1/2}(\theta_{(j)})$$

To determine whether the reference and the reverse reference priors satisfy the probability-matching criterion, partition the information matrix as

$$I(\theta_1, \theta_2) = \begin{bmatrix} I_{\theta_1,\theta_1} & I_{\theta_1,\theta_2} \\ I_{\theta_2,\theta_1} & I_{\theta_2,\theta_2} \end{bmatrix}$$

where θ_1 is the real-valued parameter of interest and θ_2 is the nuisance parameter. It follows (Peers, 1965) that a prior π satisfies the matching criterion if and only if it satisfies

$$\partial(\pi I_{\theta_1\theta_1\cdot\theta_2}^{-1/2})/\partial\theta_1 - \sum_i \partial[\pi I_{\theta_1\theta_1\cdot\theta_2}^{-1/2}(I_{\theta_2\theta_2}^{-1} I_{\theta_2\theta_1})_i]/\partial\theta_{i+1} = 0 \tag{3.1}$$

where $I_{\theta_1\theta_1\cdot\theta_2} = I_{\theta_1\theta_1} - I_{\theta_1\theta_2} I_{\theta_2\theta_2}^{-1} I_{\theta_2\theta_1}$, θ_{i+1} is the i^{th} element of θ_2 and $(I_{\theta_2\theta_2}^{-1} I_{\theta_2\theta_1})_i$ denotes the i^{th} element of $I_{\theta_2\theta_2}^{-1} I_{\theta_2\theta_1}$.

3.3.3 Conclusion

In multiparameter problems, reference and reverse reference priors are alternatives to the commonly used Jeffreys prior. Comparing these priors finds sufficient conditions under which they are the same. In comparison with respect to the matching criteria, both require that the parameter of interest be real-valued. Each algorithm for a noninformative prior has its merits and drawbacks, and the statistician should choose the prior depending on the requirements of the given problem.

3.4 Levine and Casella (2003)

In our paper, the possible existence of a matching prior is explored. The solution examined for the possible matching prior requires the solving of partial differential equations. Levine and Casella (2003) expand on Datta and Ghosh to develop an algorithm for establishing matching priors through numerical solutions of partial differential equations.

3.4.1 Summary

Nuisance parameters are handled effectively in Bayesian inference as the posterior distribution can be studied solely in terms of the parameters of interest. There is no general solution for nuisance parameters in the frequentist paradigm. Levine and Casella discuss two approaches to construct a general procedure for frequentist elimination of nuisance parameters through the use of matching priors. Matching priors are developed through solving a partial differential equation. Here Levine and Casella do not present any new theory on the topic, but rather explore numerical algorithms for solving partial differential equations. A numerical/Monte Carlo algorithm for obtaining a matching prior is presented as a solution to the appropriate partial differential equation.

3.4.2 Introduction

Suppose the parameter vector θ is divided into the parameter vector of interest, θ_1, and a vector of nuisance parameters, θ_2. Given an observation X from the sampling distribution $f(x; \theta_1, \theta_2)$ and a prior density $\pi(\theta_1, \theta_2)$ for θ. Bayes rule is used to calculate the posterior

$\pi(\theta_1, \theta_2|x)$ and inferences about θ_1 can be drawn from the marginal distribution $\pi(\theta_1|x)$. The marginal distribution is calculated from:

$$\pi(\theta_1|x) = \int \pi(\theta_1, \theta_2|x)d\theta_2 \propto \int L(\theta_1, \theta_2|x)d\theta_2$$

where $L(\theta_1, \theta_2|x)$ is the likelihood function.

From the frequentist standpoint the removal of the nuisance parameters can be complicated. One approach develops procedures in which the relevant sampling distributions do not depend on the nuisance parameters. For this approach, a general solution is not typically available.

Here the two approaches are combined. The Bayesian marginalization method is used to eliminate the nuisance parameter in a way as to maintain the frequentist inferences. This is done through specifying a prior for which the posterior quantiles have frequentist validity up to $O(n^{-1})$. The prior under which Bayesian inferences have approximate frequentist validity is called a *matching prior*. Matching priors were first introduced by Welsh and Peers (1963) who showed that posterior quantiles have approximate frequentist coverage for priors that satisfy a particular partial differential equation in θ. Except under certain conditions, the solutions to the differential equations can be difficult to obtain analytically and numerical solutions must be considered.

In this paper, methods for constructing matching priors when closed form solutions are not available are described and analytical solutions are presented. The numerical algorithm for obtaining the matching prior is presented and it is shown how to implement Monte Carlo techniques to use these priors for posterior inferences. The techniques are illustrated through examples, including a random effects model, logistic regression, and a beta-binomial model.

Matching Priors

Define $X_1, ..., X_n$ to represent iid random variables with common density $f(x;\theta)$ where $\theta = (\theta_1, \theta_2)$. Consider, for simplicity, the two-parameter model with θ_1 the parameter of interest and θ_2 the nuisance parameter. Let $D_i = \frac{\partial}{\partial \theta_1}$ for $i = 1, 2$ and define:

$$a_{20} = -\mathrm{E}_\theta\{D_1^2 \ln f(X_1;\theta)\} \quad a_{02} = -\mathrm{E}_\theta\{D_2^2 \ln f(X_1;\theta)\}, \quad a_{11} = -\mathrm{E}_\theta\{D_1 D_2 \ln f(X_1;\theta)\}$$
$$B = a_{20}a_{11}^2/a_{02}, \quad g(\theta) = a_{11}/(a_{02}/B^{\frac{1}{2}}), \text{ and } \quad h(\theta) = B^{-\frac{1}{2}}$$

Datta and Ghosh (1995) show that the matching prior satisfies the partial differential equation:

$$D_2\left(\frac{a_{11}}{a_{02}B^{\frac{1}{2}}}\pi(\theta)\right) - D_1\left(\frac{\pi(\theta)}{B^{\frac{1}{2}}}\right) = 0.$$

Note that this is the same results found in Equation (3.1) in Datta and Ghosh (1995), expressed in the former in terms of the log-likelihood.

Thus if $\pi(\theta)$ satisfies the above equation then for $\alpha \in (0, 1)$

$$\alpha = \mathrm{pr}_\pi\{\theta_1 \geq z_\alpha(X)|X\} = \mathrm{pr}_\theta\{\theta_1 \geq z_\alpha(X)|X\} + O(n^{-1})$$

where $z_\alpha(X)$ is the upper α confidence point, $\mathrm{pr}_\pi\{\cdot|X\}$ is the posterior probability of θ_1 under π, and $\mathrm{pr}_\theta\{\cdot\}$ is the probability distribution of X under θ.

The solution to the preceding equation is one way to construct a matching prior. The equation can be solved analytically in simple circumstances including the normal model, random effects models, and exponential regression. A solution to the above equation always exists but closed-form expressions are not readily available in most cases. Here methods are suggested for finding solutions to partial differential equation when analytical solutions are not readily available.

3.4.3 Solutions of the Partial Differential Equation

The partial differential equation can be rewritten as a first-order linear differential equation. Thus,

$$D_2\left(\frac{a_{11}}{a_{02}B^{\frac{1}{2}}}\pi(\theta)\right) - D_1\left(\frac{\pi(\theta)}{B^{\frac{1}{2}}}\right) = 0.$$

becomes

$$g(\theta)D_2\pi(\theta) - h(\theta)D_1\pi(\theta) + \{D_2g(\theta) - D_1h(\theta)\}\pi(\theta) = 0.$$

Partial differential equations of this form are solvable by Colton's algorithm if $g(\theta)$ and $h(\theta)$ are not complex. If $g(\theta)$ and $h(\theta)$ are complex a numerical solution is attainable using the method of characteristics.

In the method of characteristics, the parameters are transformed to continuously differentiable functions, $\xi = \lambda(\theta_1, \theta_2), \eta = \psi(\theta_1, \theta_2)$, with nonzero Jacobian

$$J(\xi, \psi) = \frac{\partial\lambda}{\partial\theta_1}\frac{\partial\psi}{\partial\theta_2} - \frac{\partial\psi}{\partial\theta_2}\frac{\partial\lambda}{\partial\theta_1}$$

$\lambda(\theta_1, \theta_2)$ is the implicit solution to the ordinary differential equation, called the characteristic equation, $\frac{d\theta_2}{d\theta_1} = \frac{g(\theta)}{-h(\theta)}$. If λ is chosen so that $\lambda(\theta_1, \theta_2) = \theta_1$ then $g(\theta)D_2\pi(\theta) - h(\theta)D_1\pi(\theta) + \{D_2g(\theta) - D_1h(\theta)\}\pi(\theta) = 0$ is reduced to:

$$\frac{\partial w(\xi, \eta)}{\partial\xi} + \frac{D_2g(\theta) - D_1h(\theta)}{-h(\theta)}w(\xi, \eta) = 0$$

where the solution is:

$$\pi(\theta_1, \theta_2) = \exp\left\{-\int \frac{D_2g(\xi, \eta) - D_1h(\xi, \eta)}{-h(\xi, \eta)}d\xi\right\}/\kappa(\eta)$$

where $\kappa(\eta)$ is an arbitrary function of η, $g(\xi, \eta) = g(\theta_1, \theta_2(\xi, \eta))$, and $h(\xi, \eta) = h(\theta_1, \theta_2(\xi, \eta))$ where $\theta_2(\xi, \eta)$ is determined by solving the original transformations. However, the transformation may not be easily solvable to yield the inverse transformation $\theta_2(\xi, \eta)$. The

characteristic equation may be solvable only numerically. The transformations between the (ξ, η) and (θ_1, θ_2) spaces may not be trivial. Thus numerical techniques for using the method of characteristics to obtain numerical forms of the matching prior is explored.

The method of characteristics requires the solution of two ordinary differential equations as well as a transformation of variables. Note that the solution to $\frac{d\theta_2}{d\theta_1} = \frac{g(\theta)}{-h(\theta)}$ takes the form $\theta_2 = \theta_2(\theta_1) + c$ for some constant c. Setting c to $\xi(\theta_1, \theta_2)$ the transformation is

$$\xi = \theta_1, \quad \eta = \theta_2 - \theta_2(\theta_1)$$

The algorithm for solving $g(\theta)D_2\pi(\theta) - h(\theta)D_1\pi(\theta) + \{D_2 g(\theta) - D_1 h(\theta)\}\pi(\theta) = 0$ numerically is

Step 1. Solve the characteristic equation numerically to obtain the solution $\theta_2(\theta_1)$.

Step 2. Transform (θ_1, θ_2) to (ξ, η).

Step 3. Solve $\frac{\partial w(\xi,\eta)}{\partial \xi} + \frac{D_2 g(\theta) - D_1 h(\theta)}{-h(\theta)} w(\xi, \eta) = 0$ numerically to obtain the solution $w(\xi, \eta)$.

Step 4. Back-transform (ξ, η) to (θ_1, θ_2) to obtain the solution $\pi(\theta_1, \theta_2)$.

The numerical solution obtained in Step 4 is the matching prior. Appropriate boundary conditions must be imposed to ensure that the resulting solutions to the partial differential equations are distributions.

The matching prior can be used to construct confidence intervals in the presence of nuisance parameters, and provides frequentist validity to Bayesian credible sets. Since the prior that results from the algorithm does not have a closed form expression, a Monte Carlo sampling procedure is used to generate samples from the posterior distribution. Using the fact that the posterior distribution is proportional to the likelihood times the prior,

where a closed form expression for the likelihood is available, a random walk Metropolis-Hastings algorithm can be used to generate the samples from the posterior. A Markov chain $\theta^{(1)}, ..., \theta^{(T)}$ is generated via a Metropolis-Hastings algorithm with the candidate distribution being a random walk. The algorithm generates samples in the transformed space $\lambda = (\xi, \eta)$ and the results are back-transformed to generate the desired sample. Using starting values $\lambda^{(0)} = (\xi^{(0)}, \eta^{(0)})$ the algorithm is

Step 1. Generate a Un(-1,1) random variate U.

Step 2. Set $y_t = \lambda^{(t-1)} + U$.

Step 3. Solve $\frac{\partial w(\xi,\eta)}{\partial \xi} + \frac{D_2 g(\theta) - D_1 h(\theta)}{-h(\theta)} w(\xi, \eta) = 0$ numerically for $w^{(t)}(\xi, y_t(2))$, where $y_t(2)$ is the second component of y_t.

Step 4. Back-transform y_t to $\theta^* = (\theta_1^{(t)}, \theta_2^{(t)})$ to obtain $\pi^{(t)}(\theta^*)$.

Step 5. Take

$$\theta^{(t)} = \begin{cases} \theta^* & \text{with probability } \min\{1, \frac{L(\theta^*|x)\pi^{(t)}(\theta^*)}{L(\theta^{(t-1)}|x)\pi^{(t)}(\theta^{(t-1)})}\}, \\ \theta^{(t-1)} & \text{otherwise.} \end{cases}$$

Step 6. Repeat Steps 1-5 until convergence.

Finally, compute $(1-\alpha)$ posterior percentiles for θ_1 from the generated sample $\{\theta_1^{(t)}\}$.

3.4.4 Conclusion

The problem dealt with in this paper is often solved using a two-stage process. The first stage finds the numerical values of the matching priors and the second stage uses these values to compute the posterior distribution. The numerical approach presented here overcomes the difficulties in the two-stage process by combining them into the unified Metropolis-Hastings algorithm. The routine is of general applicability, and can be extended to the case of three or more parameters, but may require substantial computing time. The matching

prior is sought for its values frequentist properties, but matching priors are also a viable default prior for Bayesian inference.

3.5 Smith, Kolinekov, and Cox (2003)

The exploration of spatial models and methods, including linear and possibly non-linear data transformations for predictands, is a main motivation of the current paper. The EM algorithm developed by Smith, Kolenikov, and Cox is not detailed here as it is a divergence from the methods explored in the current research presented in this paper. However, the exploration of methods for spatially correlated data, including kriging for spatial interpolation, as outlined in Smith, Kolenikov, and Cox is relevant to the methodology developed here.

3.5.1 Summary

Smith, Kolenikov, and Cox propose a method of analyzing spatio-temporal data through decomposition into deterministic nonparametric functions of time and space, linear functions of other covariates, and a random component that is spatially (though not temporally) correlated. To account for missing data, they employ a novel approach through a variant of the expectation-maximization (EM) algorithm. The results are applied to three southeastern U.S. states in the $PM_2.5$ network established by the United States Environmental Protection Agency (USEPA.)

3.5.2 Introduction

Classical theory of spatial statistics is generally primarily concerned with a single realization from a spatially correlated stochastic process. More recently there has been need to extend this to spatio-temporal processes, correlated in both time and space. Many forms of geophysical data fall in between these extremes as space-time processes whose random component is spatially but not temporally correlated. These processes pose various issues, one of which is the treatment of missing data, which is addressed by Smith, Kolenikov, and Cox. The application explored is motivated by the analysis of $PM_2.5$ data, which is

monitored by the USEPA for standards on air pollutants.

Preliminary analysis suggested that the $PM_2.5$ field could be represented as the sum of nonparametric spatial and temporal trends, together with a random component that is spatially (but not temporally) correlated. Smith, Kolenikov, and Cox were able to estimate the weekly average $PM_2.5$ at any point by using geostatistical methods to interpolate the random component. This allowed the estimation of derived quantities such as the long-term average at any site. Questions of interest is the uncertainty of the estimation procedure and how to deal with the somewhat high proportion (28%) of missing data.

Two methods are outlined to address dealing with the missing data. One can calculate an exact likelihood function through computing and inverting the spatial covariance matrix for each week's data for a pure spatial model. This is computationally inefficient, and may not work for spatial models with a temporal component. Alternatively they explore a method that uses the expectation-maximization (EM) algorithm to account for the conditional distribution of the missing observations, which may be used to obtained approximate maximum likelihood estimators.

3.5.3 Building the Model

Preliminary analysis found the need for a variance stabilizing transformation, and ultimately the square root transformation was chosen for the $PM_2.5$ data. Due to the time frame of the data, one year, a full seasonal or long-term trend analysis is not performed. No meteorological effects are incorporated, but two non-parametric approaches are considered. The first is to model each weekly mean as a "week effect" as in standard analysis of variance, and the second is to use a smooth function to represent the weekly trend over the whole year. The smoothing approach used B-splines to approximate an unknown smooth function as the weighted sum of the basis functions:

$$B(x) = \begin{cases} \frac{3|x|^3 - 6x^2 + 4}{6}, & -1 \leq x \leq 1 \\ \frac{(2-|x|)^3}{6}, & 1 < |x| \leq 2 \\ 0, & 2 < |x| \end{cases}$$

$$\hat{f}(t) = \alpha_0 + \Sigma_{k=1}^{K} \alpha_k \delta_k(t),\ t \in [0, T],\ \delta_k(t) = B\left[\frac{K}{T}\left(t - \frac{Tk}{K}\right)\right]$$

Based on this analysis, they provisionally concluded that a common time trend may be applied to all of the stations, but shifted up or down by a constant based on the location and land use of the station.

The spatial trend was also estimated non-parametrically using the bivariate version of splines, *thin-plate splines*. The basis function evaluated at the point (x, y) is

$$\Psi(x, y) = r^2 \log r$$

where $r = \sqrt{x^2 + y^2}$ is the distance from the origin, ie the knot of the spline. The overall spatial trend is represented as:

$$\psi_{x,y} = \beta_0 + \beta_1 x + \beta_2 y + \Sigma_{j=1}^{J} \beta_{j+2} \Psi(x - x^{(j)}, y - y^{(j)})$$

where $(x^{(j)}, y^{(j)})$ denote the coordinates of the j^{th} knot.

One approach, used here, is to take J as the smoothing parameter, which forces $\psi_{x,y}$ to be smoother by restricting the number of knots. For a given J, the 74 monitoring locations are grouped into J clusters, and the cluster centers are used to be the knots of the spline. The additive terms that account for differences in the landscape of the observation sites can be thought of as a component of the spatial trend.

These different models are compared simply by fitting an OLS regression model, ignoring spatial and temporal correlation, and use AIC and BIC for model selection. A few of the

overall conclusions were:

1. The square root transformation was the best in all cases where directly compared.

2. The weekly trend is best modeled by a simple week effect, giving a better fit to the data than any B-splines considered.

It was also concluded that there was no temporal autocorrelation, and a purely spatial analysis was appropriate. Spatial correlation was investigated through the variogram, under assumptions of stationarity and isotropy. A visual inspection of the variograms found significant differences between states and between seasons, there was evidence of a nugget effect but no evidence of a "sill" where the variogram leveled off, and the general characteristics of the variograms of the standardized data were similar.

The EM algorithm developed by Smith, Kolenikov, and Cox is not detailed here as it is a divergence from the methods explored in the current research presented in this paper. However, the consideration of spatial models, including possible data transformations and the investigation of spatial trends through various diagnostics, are a main motivation for this paper.

3.5.4 Conclusion

Smith, Kolenikov, and Cox proposed a model for spatial-temporal data where the field is represented by a sum of three fixed components and a random component, where the random component is spatially but not temporally correlated. A kriging methodology takes into account the fixed as well as random model components, and the estimation procedure emphasizes simultaneous estimation of the fixed and random model components. The EM algorithm developed in their paper produced results comparable to the true MLE.

3.6 Smith and Zhu (2004)

Smith and Zhu establish a second-order expansion for predictive distributions in Gaussian processes. They consider using covariance parameter estimates obtained through REML estimation in a plug-in approach as well as Bayesian methods. The estimation

of quantiles for the predictive distribution and the application to prediction intervals is the main focus. Their development leads to a calculation of the second-order coverage probability bias that lends itself to the possible existence of a matching prior where the coverage probability bias is zero. Also developed is a frequentist correction that leads to a coverage probability bias of zero, which is analogous to the existence of a matching prior. Smith and Zhu provide the original development for the univariate normal predictive distribution of the methods considered in this dissertation for the non-linear multivariate case.

3.6.1 Univariate Normal Case

The coverage probability bias and expected length of a Bayesian prediction interval formulae developed in Smith and Zhu are constructed for the linear predictand. They consider the case of a scalar Y_0, so the $G()$ function of interest is the predictive distribution function, $\psi(z; Z, \theta)$, derived from universal kriging.

Smith and Zhu consider the predictive distribution function $\psi(z; Z, \theta)$. Let ψ^* denote either the plug-in estimator of ψ, which we write as $\hat{\psi}$, or the Bayesian estimator, $\tilde{\psi}$ as defined in Equation (4.3). Assume that ψ^* has an expansion containing a linear term in $\theta - \hat{\theta}$ and a quadratic term in $\theta - \hat{\theta}$:

$$\psi^*(z; Y) = \psi(z; Y, \theta) + n^{-\frac{1}{2}} R(z, Y) + n^{-1} S(z, Y) + o_p(n^{-1}) \tag{3.2}$$

For both the plug-in and the Bayesian method, the components of R and S can be calculated explicitly, using a Taylor expansion for the plug-in approach and a combination of Taylor and Laplace for the Bayesian approach.

3.6.2 Length of the Prediction Interval

Asymptotic arguments show that, for the true and estimated P-quantiles of the predictive distribution z_P and z_P^*, where $\hat{z}_P$ is the plug-in estimate and $\tilde{z}_P^*$ is the Bayesian estimate, the length of the prediction interval is:

$$
\begin{aligned}
z_P^* - z_P &= -n^{-1/2}\frac{R(z_P,Y)}{\psi'(z_P;Y,\theta)} \\
&+ n^{-1}\left[\frac{R(z_P,Y)R^{'}(z_P,Y)}{\psi'^2(z_P;Y,\theta)} - \frac{1}{2}\frac{R^2(z_P,Y)\psi^{''}(z_P;Y,\theta)}{\psi'^3(z_P;Y,\theta)} - \frac{S(z_P,Y)}{\psi'(z_P;Y,\theta)}\right] + o_p(n^{-1})
\end{aligned} \tag{3.3}
$$

where $\psi' = \frac{\partial \psi}{\partial z}$.

3.6.3 Developing the Expansion Terms

Recall the notation from Equation (4.4). Using some further notation, write

$$
\begin{aligned}
U_i &= n^{\frac{1}{2}} Z_i, \\
U_{ij} &= n\kappa_{ij} + n^{\frac{1}{2}} Z_{ij}, \\
U_{ijk} &= n\kappa_{ijk} + n^{\frac{1}{2}} Z_{ijk}
\end{aligned}
$$

where $\kappa_{ij}, \kappa_{ijk}$ are non-random and Z_i, Z_{ij}, Z_{ijk} are random with mean 0.

Also, let $\kappa_{i,j} = \mathrm{E}\{Z_i Z_j\}, \kappa_{ij,k} = \mathrm{E}\{Z_{ij} Z_k\}$. Note by $\kappa_{i,j} = -\kappa_{ij}$ and is the (i,j) entry of the normalized Fisher information matrix which we assume is invertible with inverse entries $\kappa^{i,j}$.

For $\hat{z}_P$,

$$
\begin{aligned}
R &= \kappa^{i,j} Z_i \psi_j \\
S &= \kappa^{i,j}\kappa^{k,l} Z_{ik} Z_j \psi_l + \frac{1}{2}\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk} Z_r Z_s \psi_t + \frac{1}{2}\kappa^{i,j}\kappa^{k,l} Z_i Z_k \psi_{jl}
\end{aligned}
$$

where S for $\hat{z}_P$ is further denoted S_1.

For $\tilde{z}_P$, the corresponding expression is:

$$S_2 = S_1 + \frac{1}{2}\kappa_{ijk}\kappa^{i,j}\kappa^{k,l}\psi_l + (\frac{1}{2}\psi_{ij} + \psi_i Q_j)\kappa^{i,j}$$

where $Q(\theta)$ is the log of the prior, $\pi(\theta)$.

The development of these expansions by Smith and Zhu allows comparison with standard frequentist correction procedures. It also allows for the selection of design criterion based on expected length and coverage probability. As expected, the Bayesian prediction interval provides more accurate coverage but at the cost of a larger prediction interval length.

3.6.4 Coverage Probability Bias

The coverage probability bias is the expected value of $\psi(z_P^*; Y, \theta) - \psi(z_P; Y, \theta)$, where

$$\begin{aligned} \psi(z_P^*; Y, \theta) - \psi(z_P; Y, \theta) &= -n^{-1/2} R(z_P, Y) \\ &+ n^{-1}\left[\frac{R(z_P, Y)R'(z_P, Y)}{\psi'(z_P; Y, \theta)} - S(z_P, Y)\right] + o_p(n^{-1}) \quad (3.4) \end{aligned}$$

3.6.5 Matching Prior

An interesting development is that the coverage probability bias can be reduced to a form (Smith, 2004) that suggests a *matching prior*. It may be possible to chose a prior, π, so that the expectations of the $O(n^{-1/2})$ and $O(n^{-1})$ terms in the second-order coverage probability bias defined in Section (3.6.4) are zero. This an important result because while it may be difficult or impratical to compute the matching prior, it lends itself to assisting in prior selection based on how closely the different forms of standard priors (Jeffreys, reference prior, etc.) come to the matching prior.

Smith and Zhu (2004) also explore an additional estimator as an alternative to solving

possibly complex differential equations in order to find the exact form of the matching prior, similar to the asymptotic frequentist approach to prediction problems of Barndorff-Nielsen and Cox (1996). They consider the estimator $z_P^\dagger$, which is a form of the asymptotic bias expressed in Equation (3.4) and includes a frequentist correction term developed by Harville and Jaske (1992) and Zimmerman and Cressie (1992):

$$z_P^\dagger = \hat{z}_P - n^{-1} \times \frac{\text{asymptotic bias}}{\phi(\Phi^{-1}(P))} \tag{3.5}$$

In order to calculate the coverage probability bias or the expected length of the prediction interval, the calculation of moments of various expressions involving R, S, and their derivatives is needed. By the asymptotic formulae, these can be expressed in terms of the derivatives of ψ and other quantities that are explicit functions of the Gaussian process.

3.6.6 Conclusion

The key results of Smith and Zhu's 2004 paper are expressions for the coverage probability bias and the expected length of a prediction interval for both the plug-in and Bayesian predictors. This is established for a Gaussian process with a mean that is a combination of linear regressors and a parametrically specified covariance. The possible existence of a matching prior is considered, as well as a frequentist correction that allows for a second-order coverage probability bias of zero. This dissertation expands these methods to the analogous non-linear multivariate predictands, such as those motivated by the methods established in Smith, Kolenikov, and Cox (2003).

3.7 Zimmerman, D. L. (2006)

Spatial data is affected by the network configuration of measurement sites. Zimmerman establishes criteria for network design that emphasize the utility of the network for interpolation through kriging of unobserved sites. This is done for the case where spatial covariance parameters are assumed known. This is contrasted with criteria that emphasize estimation of parameters. Examples are outlined that show that the two main design objectives results in quite different optimal designs.

3.7.1 Introduction

Generally, kriging is the ultimate objective of geostatistical analysis. Therefore most developments of network design have emphasized the utility of the designs with regards to prediction, usually under the assumption that the covariance function (or semi-variogram) is known. The design criteria are generally the average kriging variance or the maximum kriging variance over the region of interest. Zimmerman considers three designs. The first two compare the design objectives of prediction assuming known dependence and prediction using efficient estimation of the dependence parameters. These two methods often lead to very different optimal designs. The third is a hybrid design, whose optimality is compared to the other two methods.

3.7.2 Kriging with Known Dependence Parameters

The spatial model framework for this method is identical to the treatment of kriging with known parameters developed in Section (4.2.1). Write $\sigma_K^2(s_0)$ as the kriging variance. Let $\mathcal{S}$ be the set of all possible points where measurements may be taken within the region of interest $\mathcal{D}$. The two most commonly used design criteria are the average kriging variance and the maximum kriging variance, $\max_{s\epsilon\mathcal{S}}\sigma_K^2(s_0)$:

$$\frac{1}{\mathcal{S}}\int_{\mathcal{D}}\sigma_K^2(s_0)ds \text{ or } \frac{1}{\mathcal{S}}\Sigma_{i\epsilon\mathcal{S}}\sigma_K^2(s_i)ds$$

An n-point design is optimal with respect to either criteria if it minimizes the critierion over all possible n-point designs, with each point taken from $\mathcal{S}$.

A design that minimizes $K(\theta)$ for a given θ is called a "locally K-optimal" design, where K represents kriging. Note that if the mean is known, the K-optimal design would minimize the maximum distance between the design points and the prediction sites. This prevents any prediction site from being too far from any design point, which is intuitive.

In the examples outlined in Zimmerman, the designs that minimized the average and the

designs that minimized the maximum kriging variance behaved very similarly. The overall conclusion was that the strength of the spatial correlation had relatively little effect on the K-optimal design. However, the choice of mean had a large effect.

3.7.3 Design for the Estimation of Dependence Parameters

The case is considered where it is desired to design a network that is optimal in estimating θ, the parameter vector pertaining to the covariance function. Miller and Zimmerman (1999) propose maximizing the determinant of the information matrix associated with a nonlinear generalized least squares estimator of θ. Zhu and Stein (2005) consider a similar approach which maximizes the determinant of the information matrix from either the maximum likelihood or residual maximum likelihood estimator of θ. The idea behind these approaches is that under regularity conditions the inverse of the information matrix is the asymptotic covariance matrix of the corresponding estimator of θ. Thus it provides a reasonable approximation to the determinant of the intractable mean squared error matrix for sufficiently large samples.

The overall conclusion from the examples explored in Zimmerman is that for purposes of good estimation of covariance parameters a design should have a greater number of small lags than will occur in a completely random point arrangement or in a K-optimal design. In contrast to the K-optimal design, it appears that the existence of some small lags in the design is not useful for prediction with known covariance parameters, but is very useful in the precise estimation of covariance parameters.

3.7.4 Hybrid Design

Zimmerman also considers a hybrid design, which involves empirical best linear unbiased prediction. The variance of E-BLUP's predicition error, $m_2(s_0, \theta)$ is estimated, and the approximation in Harville and Jeske (1992) and Zimmerman and Cressie (1992) is entailed. The design that minimizes the maximum value of the E-BLUP's asymptotic approximate prediction error variance over all sites in $\mathcal{S}$ is investigated. This design is called an empirical

kriging design, e.g. EK-optimal design.

3.7.5 Conclusion

The comparison of the EK-optimal design to its counterpart K-optimal design shows them very similar on a global scale, in terms of overall spatial coverage and extent. Also, a modest number of very small lags in the design is beneficial for empirical prediction, while of no use for prediction with known parameters, as is seen in the design criteria considered for the design for the estimation of dependence parameters.

3.8 Higdon, Swall, and Kern (1998)

The field of spatial statistics also encompasses a wide range of methods for dealing with spatial interpolation. Higdon, Swall, and Kern develop a hierarchical model which incorporates the uncertainty involved in model specification. Due to the explicit form developed of the covariance function of the process, the likelihood function for the process can be expressed at any configuration of points. This lends itself to the Bayesian approach developed in the current paper.

3.8.1 Summary

Traditional variogram model accounts for spatial dependence. Realistically, assumption of constant (variance) spatial dependence structure is often violated, as is the assumption of stationarity. Higdon, Swall, and Kern present a model that allows the spatial dependence structure to vary as a function of location and also develop a hierarchical model which incorporates the uncertainty involved in the specification. The model is applied in a toxic waste application.

3.8.2 Introduction

The backbone of modeling spatial data is using a Gaussian process. It is common to model spatial dependence through the covariogram function, which models covariance be-

tween two points only as a function of the distance between them. This yields a stationary random field, which if invariant under rotation is also isotropic. Assuming both stationarity and isotropic can be natural if the region of interest is small. Some cases arise where heterogeneous spatial covariance structures need to be considered. Here, an alternative model is proposed which accounts for heterogeneity based on a Moving Average specification of a Gaussian process. A hierarchical modeling structure is used so that uncertainty may be assessed. A large subset of stationary spatial processes may be represented as a moving average of a Gaussian white noise process - ie a white noise process convolved with a kernel. Here the kernel is characterized so that it evolves over spatial location.

3.8.3 Specifying Covariance Structure

If a stationary Gaussian process $z(s)$ has correlogram of the form $\rho(d)$:

$$\rho(d) = \int k(s)k(s-d)ds$$

then the process can be expressed as the convolution of a Gaussian white noise process $x(s)$ with convolution kernel $k(s)$:

$$z(s) = \int k(s-u)x(u)du$$

Higdon, Swall, and Kern focus on the two-dimensional standard normal kernel:

$$k(s) = \frac{1}{2\pi}\exp(-\frac{1}{2}s^T s)$$

This leads to the Gaussian correlation function in two-dimension:

$$\rho(d) = \exp(-d^T d)$$

This representation can be extended to the bivariate normal kernel:

$$k(s;\Sigma) = \frac{1}{2\pi}|\Sigma|^{\frac{-1}{2}}\exp(-\frac{1}{2}s^T\Sigma s)$$

Parameterizing

$$\Sigma(s) = \begin{pmatrix} a^2 & \rho ab \\ \rho ab & b^2 \end{pmatrix}, \ \Sigma(s') = \begin{pmatrix} a'^2 & \rho' a' b' \\ \rho' a' b' & b'^2 \end{pmatrix}$$

This leads to the correlation function

$$\rho(s, s') \propto \frac{1}{q_1} \exp\left\{ -\frac{1}{q_2} (s - s')^T W (s - s') \right\}$$

where

$$W = \begin{pmatrix} b^2 + b'^2 & -(\rho ab + \rho' a' b') \\ -(\rho ab + \rho' a' b') & a^2 + a'^2 \end{pmatrix}$$

$$\begin{aligned} q_1 &= 2\pi aa' bb' sqrt(1 - \rho^2)(1 - \rho'^2) \sqrt{-\frac{(\rho^2 - 1)b^2 + (\rho'^2 - 1)b'^2}{(\rho^2 - 1)(\rho'^2 - 1)b^2 b'^2}} \times \\ & \sqrt{\frac{2\rho\rho' aa' bb' + a^2((\rho^2 - 1)b^2 - b'^2) + a'^2((\rho'^2 - 1)b'^2 - b^2}{a^2 a'^2((\rho^2 - 1)b^2 + (\rho'^2 - 1)b'^2)}} \end{aligned}$$

$$q_2 = -2(\rho\rho' aa' bb' + a^2((\rho^2 - 1)b^2 - b'^2) + a'^2((\rho'^2 - 1)b'^2 - b^2))$$

Since $\Sigma(s)$ depends on location s in a parametrically specified way, then the covariance function can be written down explicitly. This is an explicit example of a nonstationary process that can be written in a fully parametric way, illustrating the point that the methods outlined in the current paper are not restricted to the stationary process.

3.9 Silverman (1986)

In the current paper, we develop the methodology for working with an unknown or perhaps computationally difficult predictive distribution. Kernel density estimation is employed as an appropriate technique for the situation of the multivariate and possibly nonlinear predictand. Silverman (1986) provides a very thorough development of several tech-

niques for kernel density estimation. Here we detail his general overview of density estimation, including various approaches to choosing a density and an appropriate smoothing parameter. We look in detail at the Epanechnikov kernel which is suitable for the estimation methodology developed in our paper.

3.9.1 Density Estimation

If a probability density function $f(z|\theta)$ is unknown or is of a form not easily manipulated, in order to determine $P(a < Z < b)$, a kernel density can be chosen. A kernel density is needed to obtain estimates of the derivatives of the distribution function, with respect to both θ and the prediction point z. Density estimation can be performed utilizing the proper kernel, and the empirical cdf estimated using the kernel density.

For density f,

$$f(z) = \lim_{h \to 0} \frac{1}{2h} P(z - h < z + h)$$

which is approximated by

$$\hat{f}(z) = \frac{1}{2nh}[\text{no.}z_i \text{in}(z - h, z + h)]$$

Let K be the kernel density, with cumulative distribution function K_1 where $K = K_1'$. Therefore, $f(x)$ is approximate by:

$$\hat{f}(z) = \frac{1}{nh} \Sigma_{i=1}^{n} K(\frac{z - x}{h}) \tag{3.6}$$

where h is the bandwidth. Note that the cdf is approximated by $\frac{1}{n}\Sigma_{i=1}^{n} K_1(\frac{z-x}{h})$.

3.9.2 Measures of Discrepancy

When considering estimation at a single point z, a natural measure of the discrepancy between the estimator density $\hat{f}$ and the true density f is the *mean square error*, MSE:

$$MSE_z(\hat{f}) \quad = \quad E\{\hat{f}(z) - f(z)\}^2$$

$$= \quad \{E[\hat{f}(z) - f(z)]\}^2 + var[\hat{f}(z)]$$

The most widely used measure of the *global* accuracy of the estimator density $\hat{f}$ is the *mean integrated square error*, MISE:

$$MISE_z(\hat{f}) = E \int \{\hat{f}(z) - f(z)\}^2 dz$$

Since the integrand is non-negative, the order of the integral and the expectation can be reversed to yield:

$$\begin{aligned} MISE_z(\hat{f}) &= \int E\{\hat{f}(z) - f(z)\}^2 dz \\ &= \int E[\hat{f}(z) - f(z)]^2 dz + \int var[\hat{f}(z)] dz \end{aligned}$$

which expresses MISE as the sum of the integrated squared bias and the integrated variance.

3.9.3 Choosing a Kernel Density

Possible candidates for the kernel density include the univariate normal or the Epanechnikov Kernel, a truncated quadratic of the form (Silverman, 1986):

$$K_e(t) = \begin{cases} \frac{3}{4\sqrt{5}} \left(1 - \frac{1}{5}t^2\right) & -\sqrt{5} \leq t \leq \sqrt{5} \\ 0 & otherwise. \end{cases}$$

Here $t = \frac{(x - x_i)}{h}$, where h is the window width and the x_i's are the values of the independent variable in the data. x is the value of the scalar independent variable for which one seeks an estimate. The Epanechnikov kernel takes the form in Figure (3.1).

Note the the cumulative distribution function, noted here as K_1, of the Epachnenikov kernel is:

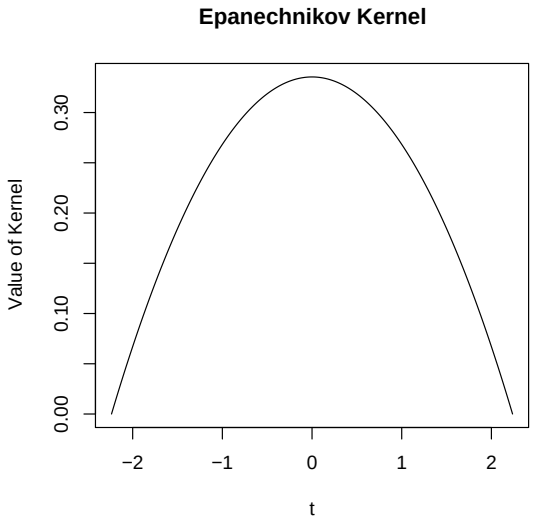

Figure 3.1: Density of the Epanechnikov Kernel

$$K_1(t) = \begin{cases} \frac{3}{4\sqrt{5}}\left(t - \frac{1}{15}t^3\right) + 0.5 & -\sqrt{5} \leq t \leq \sqrt{5} \\ 0 & otherwise. \end{cases}$$

The CDF of the Epanechnikov function takes the form in Figure (3.2).

3.9.4 Choosing a Smoothing Parameter

Choosing h correctly is important. h is determined in part by sample size. For large n, there's a corresponding small h. If a very small value of h is used in an attempt to eliminate bias, the integrated variance will become large. Conversely, choosing a large h will reduce the random variation but possibly at the expense of introducing systematic bias (Silverman, 1986.) The ideal value of the window width h can be chosen by using the h which minimizes the approximate mean integrated squared error. Define h_{opt} (Silverman) where

$$h_{opt} = k_2^{-\frac{2}{5}} \left[\int K(t^2)dt\right]^{\frac{1}{5}} \left[\int f''(x)^2 dx\right]^{-\frac{1}{5}} n^{-\frac{1}{5}} \tag{3.7}$$

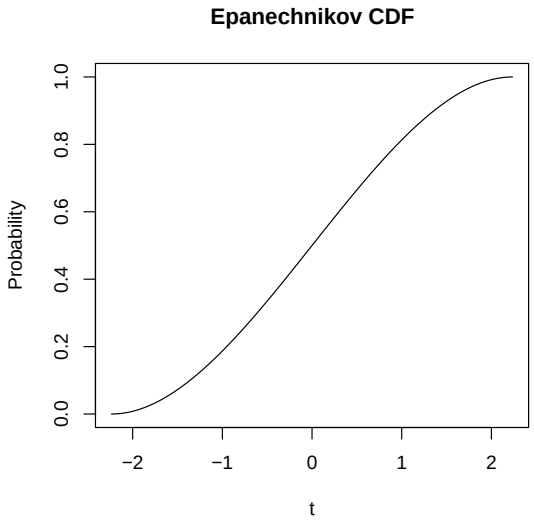

Figure 3.2: Cumulative Distribution of the Epanechnikov Function

where k_2 is the variance of the distribution associated with kernel density K.

Note that the theoretical optimal bandwidth will be a multiple of $\int f''(x)^2 dx$, which is proportional to σ, and is based on sample size. There are several ways to estimate σ. A natural approach is to use a standard family of distributions, such as the normal distribution with variance σ^2, to assign a value to the $\int f''(x)^2 dx$ term. Let ϕ be the standard normal density, then:

$$\begin{aligned} \int f''(x)^2 dx &= \sigma^{-5} \int \phi''(x)^2 dx \\ &= \frac{3}{8} \pi^{-\frac{1}{2}} \sigma^{-5} \\ &\approx 0.212 \sigma^{-5} \end{aligned}$$

If a normal kernel is being used, the smoothing parameter h_{opt} from Equation (3.7) would be:

$$\begin{aligned} h_{opt} &= (4\pi)^{-\frac{1}{10}} \frac{3}{8} \pi^{-\frac{1}{2}} \sigma n^{-\frac{1}{5}} \\ &= \left(\frac{4}{3}\right)^{\frac{1}{5}} \sigma n^{-\frac{1}{5}} \\ &= 1.06 \sigma n^{-\frac{1}{5}} \end{aligned} \tag{3.8}$$

A quick way to select the smoothing parameter is to estimate σ from the data and then substitute the estimate into Equation (3.8).

There are also alternative estimates for h_{opt} using more robust measures of spread that tend to yield better results. For example, Equation (3.8) can be written in terms of the interquartile range, R, of the underlying normal distribution. This leads to:

$$h_{opt} = 0.79 R n^{-\frac{1}{5}}$$

However, basing h_{opt} on the interquartile range can oversmooth for bimodal distributions. Thus, an adaptive estimate of the spread can be used to yield the benefits of both estimates:

$$A = \min(\sigma, \text{IQR}/1.34)$$

For a Gaussian kernel, this leads to the choice

$$h = 0.9 A n^{-1/5}$$

which yields a mean integrated square error within 10% of the optimum for t-distributions, log-normal with skewness up to approximately 1.8, and for the normal mixture with separation up to 3 standard deviations. For samples over size 100, this smoothing parameter will do well for a wide range of densities, is trivial to evaluate, and will usually clearly show skewness or bimodality.

Some use these methods as a basis to choose σ in a non-normal setting. However, σ_h is

not based on a normal kernel, rather it is based on the data density. Theoretical experts in kernel density estimation believe that this method is not optimal. Rather, one should estimate f'' and use $(f'')^2$ to estimate σ. Note, $H(Y_0)$ is possibly a non-linear combination, but is a combination of multivariate normal vectors. Therefore in this context, it is reasonable to assume a normal data density for the selection of h_{opt}. If $f \sim N(\dot{,}\sigma)$, f'' can be determined analytically.

Asymptotic theory uses slightly larger bandwidth. f'' is arbitrarily estimated and the resulting σ is plugged into the formula for h_{opt}.

Least-squares cross-validation

An automatic method for choosing the smoothing parameter is least-squares cross-validation. Given any estimator $\hat{f}$ of density f, the integrated square error can be written:

$$\int (\hat{f} - f)^2 = \int \hat{f}^2 - 2\int \hat{f}f + \int f^2$$

The ideal choice of bandwidth will minimize the integrated square error, which corresponds to minimizing

$$R(\hat{f}) = \int \hat{f}^2 - 2\int \hat{f}f$$

An estimate of $R(\hat{f})$ can be constructed from the data. $\int \hat{f}$ can be determined from the estimate $\hat{f}$. Define $\hat{f}_{-i}$ to be the estimate constructed from all of the data points excluding Z_i:

$$\hat{f}_{-i}(z) = (n-1)^{-1}h^{-1}\Sigma_{j\neq i}K\{h^{-1}(z - Z_j)\}$$

Define

$$M_0(h) = \int \hat{f}^2 - 2n^{-1}\Sigma_i \hat{f}_{-i}(Z_i)$$

The score M_0 depends only on the data. Least-squares cross-validation minimizes the score M_0 over h.

3.10 Park and Marron (1990)

An important consideration in density estimation is the selection of the smoothing parameter. Park, and Marron (1990) more fully compare methods for bandwidth selection, building on the methods introduced in Silverman.

3.10.1 Summary

Kernel density estimation is often used to explore the distribution structure of unknown populations. Practical application of kernel density estimation is very dependent on the proper selection of the smoothing parameter, or bandwidth. There are many methods in use for selecting the appropriate bandwidth. An efficient and objective method of determining an appropriate bandwidth using data would enhance the usefulness of density estimation. In Park and Marron, three data-driven methods are compared through asymptotic rates of convergence and through a simulation study. It is seen that when the underlying density is smooth, the plug-in method is the most efficient. However it is less robust when smoothness is not present.

3.10.2 Overview

Least squares cross-validation is one of the most popular methods of bandwidth selection. Its asymptotic properties provide convergence to the optimum under very weak conditions, however its performance suffers with sample variability. In the plug-in method, parameter estimates are "plugged-in" to an asymptotic representation of the optimal bandwidth. Biased cross-validation is a hybrid of the former two methods which utilizes a score function. The score function is minimized as in cross-validation, but is also data-driven through using the plug-in method. This provides less sample variability than the original least squares cross-validation method.

An understanding of the asymptotic performance of each of the methods is obtained through examining the asymptotic rates of convergence of the bandwidths to the optimum.

The rate of convergence of the density estimator to the density is dependent on the amount of smoothness of the underlying density. The main results shows that under certain assumptions on the underlying density, the plug-in bandwidth dominates the limit. There is some trade-off for this due to the fact that for small amounts of smoothness least squares cross-validation is the most effective, analogous to the trade-off in robustness theory. When smoothness is present, an asymptotic pay-off of reduced variability is present for biased cross-validation and the plug-in rules. However when there is not enough smoothness, it is much less effective. Thus it is found that cross-validation is more robust at the cost of some efficiency whereas the plug-in method is more efficient under stronger assumptions.

3.10.3 Details

The mean integrated squared error, MISE(h) = $\mathrm{E}\int(\hat{f}_h - f)^2$ is a common method of assessing the performance of a density estimator. h_{MISE} is the minimizer of MISE(h). $\hat{h}_{\mathrm{CV}}$, the least squares cross-validation bandwidth, is the minimizer of the cross validation function

$$CV(h) = R(\hat{f}_h) - 2n^{-1}\Sigma_{j=1}^{n}\hat{f}_{j,h}(X_j) \tag{3.9}$$

where for $g(x)$, $R(g) = \int g(x)^2 dx$ and $\hat{f}_{j,h}$ denotes the leave-one-out kernel estimator constructed with X_j deleted. For definiteness, $\hat{h}_{\mathrm{CV}}$ is taken to be the largest local minimizer over the range.

The biased cross-validated bandwidth utilizes the fact that when f has a Hölder-continuous, square-integrable second-derivative, the asymptotic representation of MISE(h) is

$$\mathrm{AMISE}(h) = n^{-1}h^{-1}\int K^2 + h^4\sigma_K^4 R(f'')/4$$

The biased cross-validated bandwidth $\hat{h}_{\mathrm{BCV}}$ is the minimizer over the range of the estimate of AMISE(h) obtained by replacing R(f") by R($\hat{f}_h$") - $n^{-1}h^{-5}R(K'')$.

The plug-in estimator for the bandwidth, $\hat{h}_{PI}$, is the root (when it exists, the largest root is there is more than one) over the range of

$$h = \{R(K)/\sigma_K^4 \hat{R}_{f''}(a_{\hat{\lambda}}(h))\}^{1/5} n^{-1/5}$$

where $\hat{\lambda}$ is an $n^{1/2}$ consistent estimate of λ, the scale of f and

$$a_\lambda(h) = C_3(K) C_4(g_1) \lambda^{3/13} h^{10/13}$$

with g_1 taken to be the normal density and

$$C_3(K) = \{18R(K^{(4)}\sigma_K^8/\sigma_{K*K}^4 R(K)^2\}^{1/13}$$

and

$$C_4 = \{R(f)R(f'')^2/R(f^{(3)})^2\}^{1/13}$$

3.10.4 Conclusion

The results of the simulation study in Park and Marron showed superior performance of the plug-in estimator $\hat{h}_{\mathrm{PI}}$. This was also demonstrated theoretically in the derivations found in the paper. The cross-validation bandwidth estimator, $\hat{h}_{\mathrm{CV}}$ yielded unreasonable results that were undersmoothed, possibly due to small-scale clustering in the data. The biased cross-validation estimator, $\hat{h}_{\mathrm{BCV}}$, performed much better, giving reasonable bandwidths. However it was too variable to allow for construction of the density estimates as shown in the paper's Figure 2. Overall, the plug-in estimator $\hat{h}_{\mathrm{PI}}$ performed much better than expected, and yielded big dividends in the real data situations explored in Park and Marron's examples.

3.11 Efron (2000)

The methodology outlined in our paper relies on a form of a parametric bootstrap in order to obtain estimates of the prediction density. To complete our review, we explore some of the methods and literature pertaining to bootstrapping methods. Efron (2000)

provides a basic overview, and compares the accuracy of bootstrapping techniques to older methods relying on Taylor series approximations.

3.11.1 Summary

Efron looks at the changing pace of mathematics in the face of ever increasing computing power. The capabilities of modern statistical computing, which are infinite relative to past methods, can be verified through a bootstrap example. Much effort has been spent to justify the theoretical basis for the bootstrap. Here, Efron outlines the basic principles, summarizing an application of nonparametric maximum likelihood estimation.

1. Assume the data have been obtained by random sampling from some unknown probability distribution F.
2. The goal is the estimation of the parameter of interest θ with some statistic $\hat{\theta}$.
3. We wish to know σ_F, the standard error of $\hat{\theta}$ when sampling from F.
4. σ_F is approximated by $\sigma_{\hat{F}}$

The Monte Carlo routine for estimating $\hat{\theta}^*$ provides a way of evaluating $\sigma_{\hat{F}}$ without going through the computationally more complex Taylor series approximations used in other methods. Some of the advantages of this approach when compared to the Taylor series approach include the ease of use, higher accuracy, and generality.

The bootstrap was first developed as "an explanation for the success of an older methodology, the jackknife." The bootstrap and the Quenouille-Tukey jackknife undertake the same underlying tasks: routine calculation of biases and standard errors. The more recent development has been the automatic calculation of bootstrap confidence intervals. Of concern is how to obtain second-order accuracy with the bootstrap methods. Hall (1988) verified the second-order accuracy of the bias-corrected and accelerated bootstrap confidence interval developed by Efron (1987). One limitation of the bootstrap is less accurate coverage in small-sample non-parametric situations. The discussion and process of connecting the bootstrap with the fundamental ideas of statistical theory continues.

Efron compares the frequentist bootstrap method to the Bayesian Markov chain Monte Carlo method. He connects the use of uninformative priors to the theoretical basis for the bootstrap. Efron concludes with an example using the bootstrap to obtain a bias-corrected estimate of R^2 for a regression model that required quite a bit of data mining to obtain the original R^2.

CHAPTER 4

Prediction of Gaussian Fields with Unknown Covariance Structure

4.1 Introduction

Many spatial prediction techniques are founded on the assumption that the realizations come from a Gaussian process with a known covariance structure and known, specified parameters. Interpolation of $Z(s)$ to points of the measuring network is traditionally achieved through the linear prediction technique known as kriging. Traditional methods estimate the parameters through Least Squares Estimation, Maximum Likelihood methods (MLE), and Restricted Maximum Likelihood (REML) methods. However, kriging using a covariance structure with covariance parameter estimates obtained through these techniques ignores possible error introduced through estimating the covariance parameters. Due to this, interval estimates constructed using mean squared prediction errors (MSPE) are expected to be too short.

We propose Bayesian methods as a possible resolution. However, it is unknown in general whether a Bayesian Prediction Interval (PI) or a frequentist Prediction Interval comes closer to nominal coverage probability. A simulation experiment using an Gaussian process with an underlying power exponential covariance structure is employed to test three approaches. The first approach employs traditional kriging methods using Restricted Maximum Likelihood (REML) estimators for the covariance parameters. The second approach implements Bayesian methods via Markov Chain Monte Carlo (MCMC) methods. A third approach, which could provide an alternative implementation, using an analytical Laplace

approximation method is outlined. Prediction Intervals are generated and compared for each approach. All three approaches are run 100 times and the empirical coverage probabilities are computed.

4.2 Estimation

This sections outlines the methods behind each of the three estimation techniques.

4.2.1 Traditional Kriging Using REML Estimates

To consider traditional kriging methods, parameter estimates $\theta = (\phi, \sigma^2, \kappa)$, for the power exponential spatial covariance structure, were obtained using Restricted Maximum Likelihood estimation. REML estimation is based on the joint density of the vector of contrasts, whose distribution is independent of the population mean. The resulting maximum likelihood estimator based on the joint density of contrast is approximately unbiased. The log-likelihood is written as in Equation (2.6), and the parameter estimates are obtained as outlined following the expression of Equation (2.6). Point predictions are obtained via the kriging techniques outlined in Section (2.1.2) using the REML parameter estimates.

4.2.2 Bayesian Prediction

The main advantage of using Bayesian prediction methods is the ability to account for parameter uncertainty, even with small sample sizes as we have here with $n = 16$.

For notational convenience we define $Y^* = \begin{pmatrix} Y \\ Y_0 \end{pmatrix}$ where Y is an n-dimensional vector of observations, and Y_0 is an unobserved scalar prediction. X and x_0 are dimensions $n\ timesq$ and $q \times 1$ matrices of known regressors, with β an unknown $q \times 1$ vector of regression coefficients. $V(\theta), w(\theta)$, and $v(\theta)$ are covariance elements that are known functions of an unknown p-dimensional parameter vector θ.

$$\begin{pmatrix} Y \\ Y_0 \end{pmatrix} \sim N\left(\begin{pmatrix} X\beta \\ x_0^T\beta \end{pmatrix}, \begin{pmatrix} V(\theta) & w^T(\theta) \\ w(\theta) & v_0(\theta) \end{pmatrix}\right)$$

becomes

$$Y^* \sim N(X^*\beta, V^*)$$

We assume a uniform prior on β, and a prior density on θ of the form $e^{Q(\theta)}$ for some function Q. Recall $\theta = (\sigma^2, \phi, \kappa)$ is the vector of covariance parameters. The Bayesian predictive density of Y_0 given Y is then given by

$$p(Y_0|Y) = \frac{\int f(Y^*|\beta,\theta)e^{Q(\theta)}d\beta d\theta}{\int f(Y|\beta,\theta)e^{Q(\theta)}d\beta d\theta}$$

Through analytic integration with respect to β and routine manipulation it can be shown (Smith and Zhu, 2004) that

$$p(Y_0|Y) = \frac{\int \psi(Y_0;Y,\theta)e^{Q(\theta)}e^{l_n(\theta)}d\theta}{\int e^{Q(\theta)}e^{l_n(\theta)}d\theta}$$

where $l_n(\theta)$ is the restricted log likelihood and the function of θ to be predicted is $\psi(Y_0;Y,\theta) = (2\pi\sigma_0^2)^{-1/2}\exp\left\{-\frac{1}{2}\left(\frac{Y_0-\lambda^T Y}{\sigma_0}\right)^2\right\}$ with

$$\lambda(\theta) = V^{-1}(\theta)w(\theta) + V^{-1}(\theta)X(X^TV^{-1}(\theta)X)^{-1}(x_0 - X^TV^{-1}(\theta)w(\theta))$$

and corresponding MSPE:

$$\sigma_0^2 = v_0 - w^TV^{-1}w + (x_0 - X^TV^{-1}w)^T(X^TV^{-1}X)^{-1}(x_0 - X^TV^{-1}w).$$

where $V(\theta), v_0(\theta)$ and $w(\theta)$ are written as V, v_0, and w for ease of notation.

4.2.3 Laplace's Method

Bayesian posterior approximation techniques are first order approximations. Laplace's method offers an expansion technique which uses Taylor series to obtain posterior estimates (Ibrahim, 2000). Laplace's integral formula offers a second order approximation for a multi-dimensional integral that includes a leading term and an expansion term. Consider the formula:

$$I = \int f(\theta) e^{-ng(\theta)} d\theta$$

where $g(\theta) = \frac{-ln(\theta)}{n}$ with $g(\theta)$ is minimized uniquely at $\hat{\theta}$ and g and f are at least four times continuously differentiable on a neighborhood of $\hat{\theta}$. Here we write f for $f(\hat{\theta})$, f_i for $\frac{\partial f}{\partial \theta^i}|_{\theta=\hat{\theta}}$, f_{ij} for $\frac{\partial^2 f}{\partial \theta^i \partial \theta^j}|_{\theta=\hat{\theta}}$, etc. In the case of g we assume $g_i = 0$ for all i and that the matrix with entries $\{g_{ij}\}$, denoted by G, is positive definite and has an inverse G^{-1} with entries $\{g^{ij}\}$, using summation notation. The general form and regularity conditions can be found in Bleistein and Handelsman's book (1986), which was the source of the formula below

$$\begin{aligned} I &= \left(\frac{2\pi}{n}\right)^{p/2} e^{-ng} |G|^{-1/2} \cdot \Big\{ f - \frac{1}{8n} f g_{ijk\ell} g^{ij} g^{k\ell} + \frac{1}{8n} f g_{ijk} g_{\ell mq} g^{ij} g^{k\ell} g^{mq} \\ &+ \frac{1}{12n} f g_{ijk} g_{\ell mq} g^{i\ell} g^{jm} g^{kq} + \frac{1}{2n} f_{ij} g^{ij} - \frac{1}{2n} g_{ijk} f_\ell g^{ij} g^{k\ell} + O(n^{-2}) \Big\}. \end{aligned} \tag{4.1}$$

This is the main result in Chapter 8 of Bleistein and Handelsman (1986), specifically, equations (8.3.50)–(8.3.55).

The application we are interested in here is the form found in the Bayesian predictive posterior:

$$\tilde{\psi} = \frac{\int \psi(Y_0; Y, \theta) e^{Q(\theta)} e^{-ng(\theta)} d\theta}{\int e^{Q(\theta)} e^{-ng(\theta)} d\theta} = \frac{I_1}{I_2} \tag{4.2}$$

ψ is the function of θ whose posterior expectation we wish to evaluate, $e^{Q(\theta)}$ is the prior, and $e^{-ng(\theta)}$ is the likelihood expressed in terms of n.

Applying (4.1) separately to I_1 and I_2 and rearranging terms, Smith (1999) derived the formula

$$\tilde{\psi} = \hat{\psi} + \hat{\mathcal{D}} + O_p(n^{-2}) \tag{4.3}$$

where

$$\mathcal{D} = \frac{1}{2}U_{ijk}\psi_l U^{ij}U^{kl} - \frac{1}{2}(\psi_{ij} + 2\psi_i Q_j)U^{ij} \tag{4.4}$$

and $\hat{\mathcal{D}}$ indicates the evaluation of $\mathcal{D}$ at $\hat{\theta}$. Notationally, superscripts denote components of vectors, such as θ^i denoting the ith component of θ. Subscripts will indicate differentiation with respect to the components of θ, ie $Q_i = \frac{\partial Q}{\partial \theta^i}$. Also, let $U_i = \frac{\partial l_n(\theta)}{\partial \theta^i}$, $U_{ij} = \frac{\partial^2 l_n(\theta)}{\partial \theta^i \partial \theta^j}$, etc, and U^{ij} is the (i,j) entry of the inverse of the matrix whose (i,j) entry is U_{ij}.

This result was derived earlier by Lindley (1980) using different reasoning.

This direct approximation for a Bayes estimator is an alternative to MCMC methods. Consider the distribution function of a random variable Z, $\psi(z_\alpha;\theta)$, where z_α is the α-quantile of the distribution. This leads to construction of prediction intervals. For the Bayesian prediction intervals, we define $\tilde{z}_\alpha$, an approximation to the Bayesian quantile with the same order of accuracy as Lindley's formula for the Bayesian distribution function. $\tilde{z}_\alpha$ is based on a Taylor expansion of the inverse distribution function, such that $\tilde{\psi}(\tilde{z}_\alpha) = \alpha$ and

$$\tilde{z}_\alpha = \hat{z}_\alpha - \frac{\hat{D}(\hat{z}_\alpha)}{\hat{\psi}'(\hat{z}_\alpha)} + O_p(n^{-3/2}) \tag{4.5}$$

where $\hat{\psi}'(\hat{z}_\alpha) = \frac{\partial \psi}{\partial z}$.

4.3 Simulation

A random plane of 16 location values was simulated on a plane $(0,2)X(0,2)$. This was done in R using the function *Psim* to simulate a Binomial spatial point process.
Using the 16 location values denoted 1-G in Figure (4.1), corresponding observation values Y were simulated:

$$Y(s) = X^T(s)\beta + S(s)$$

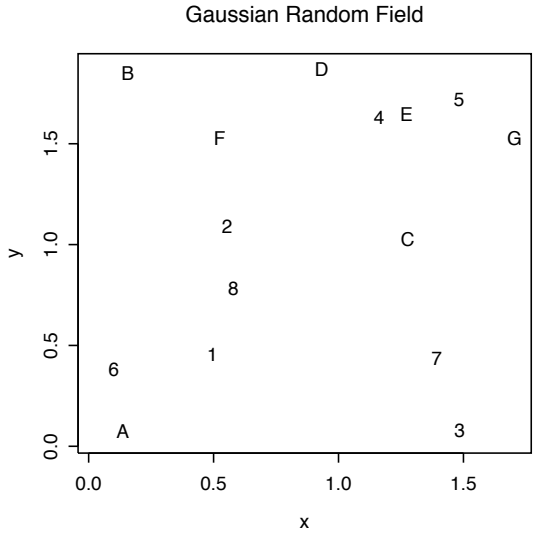

Figure 4.1: Gaussian Random Field of Prediction Locations

$X(s)$ is column vector with entries $1, s_1, s_2$ where s_1 and s_2 are the coordinates at site s. $\beta = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ and $S(s)$ is a stationary Gaussian process with mean 0 and variance $\sigma^2 = 1$ (partial sill) and a correlation function parametrized by $\phi = 1$ (range parameter) through a powered exponential spatial covariance structure:

$$cov\{Y(s_1), Y(s_2)\} = \sigma^2(\exp(-\frac{d_{ij}}{\phi})^\kappa)$$

Values $\kappa = 0.5, \kappa = 1.0$, and $\kappa = 1.5$ were initially considered for the scale parameter. No nugget effect was incorporated. As detailed in Section (4.3.4), the κ parameter proved to be unstable in the Laplace approximation technique, so for the sake of comparison, κ was held fixed at 1.0. Note that fixing $\kappa = 1.0$ in the powered exponential model is equivalent to using an exponential model:

$$cov\{Y(s_1), Y(s_2)\} = \sigma^2(\exp(-\frac{d_{ij}}{\phi}))$$

For each of the estimation techniques, the model was re-estimated 16 times to predict

Y at each of the 16 locations using the other 15 sites. $\hat{Y}_{(i)} = X_{(i)}\beta + e_{(i)}$ represents the prediction at site i using the covariate matrix and parameter estimates constructed without the i^{th} observation.

4.3.1 Prediction with Traditional Kriging

Traditional universal kriging methods were used to find prediction estimates for each of the 16 locations. Kriging predictions were obtained at each site using the REML covariance parameter estimates $\hat{\sigma}$ and $\hat{\phi}$ obtained through the joint density of the 15 remaining sites. Using the corresponding MSPEs, 95% prediction intervals were constructed for each site. The simulated Y value was then compared to the prediction interval to determine whether or not the prediction interval contained the true value. The simulation of Y values and REML estimation leading to kriging predictions was run 100 times. The resulting empirical coverage probabilities for the theoretical 95% prediction intervals were computed.

4.3.2 Kriging with True Parameter Values

To investigate the error introduced into the model through estimating the parameters in the covariance structure, the kriging predictions using the true parameters $\phi = 1$ and $\sigma = 1$ used in the original simulation were also found. κ here is treated as fixed at 1.0.

The resulting empirical coverage probabilities were computed and compared to the empirical coverage probabilities for kriging using the REML estimates.

4.3.3 Prediction using Bayesian Methods

To consider Bayesian methods, Gibbs sampling was used to find prediction estimates for each of the 16 locations using WinBUGS software. For each value of κ, three chains of length 1000 with a burn-in of 250 were run, with the REML estimates used as initial values for the chains. β was assigned a flat prior, and flat priors were assigned to the remaining parameters as follows:

$$\sigma \sim unif(0.0005, 5)$$

$$\phi \sim unif(0.0005, 12)$$

Improper priors, such as $\pi(\tau) \sim 1/\tau$, and noninformative priors, such as $\pi(\phi) = 1$, can lead to improper posterior distributions. It can also lead to non-convergence of the MCMC simulations, with parameter values going to infinity (Berger, De Oliveira, and Sansó, 2001). One of the main advantages of Bayesian inference is that the parameter uncertainty is fully accounted for when performing inference and prediction, even in small samples. Here Bayesian analysis of spatial data utilizing noninformative conventional priors for the unknown parameters of the Gaussian field is considered, and ultimately yields satisfactory results. A possible area of future research is to consider the performance of the Bayesian techniques explored here under various different priors.

Values for the predicted $\hat{Y}$ were sampled from the above posterior distribution under these conditions. The expression for the Bayesian predictive density,

$$p(Y_0|Y) = \frac{\int \psi(Y_0; Y, \theta) e^{Q(\theta)} e^{ln(\theta)} d\theta}{\int e^{Q(\theta)} e^{l_n(\theta)} d\theta}$$

can be approximated through MCMC samples, expressed as:

$$\mathrm{E}_{\theta|Y} \psi(Y_0; Y, \theta) \approx \frac{1}{N} \sum_{j=1}^{N} \psi(Y_0|\theta^{(j)}) \tag{4.6}$$

where the right-hand side of Equation (4.6) is an estimate of the left-hand side based on the MCMC sample and $\psi(Y_0; Y, \theta)$ is the predictive distribution function.

95% prediction intervals were constructed for each site. The simulated Y value was then compared to the prediction interval to determine whether or not the prediction interval contained the true value. Gibbs sampling was run for all 100 simulated measurements. The resulting empirical coverage probabilities for the theoretical 95% prediction intervals were

computed.

4.3.4 Prediction through Laplace Approximation

A Laplace approximation was performed by applying the correction term defined in Equation (4.4) to the $\hat{z}_P$ estimates obtained through REML estimation. Due to erratic behavior in the estimation of the parameter κ in the powered exponential model, a simpler model, defined as the exponential model where $\kappa = 1$, was used. Under the exponential model, the scale parameter σ and the range parameter ϕ are estimated using REML estimation while κ is treated as fixed and known.

4.4 Laplace, Bayesian and Plug-In Methods Prediction Empirical Coverage Results

For the exponential covariance structure, where κ is fixed at 1.0, the empirical prediction interval coverage for the kriging prediction using the true parameters is near the theoretical 95% coverage as expected. The Bayesian methods also produced empirical coverage results around 95%, with a range of 86% to 99%, and an Average Empirical Coverage (AEC) of 94.8%. The kriging prediction empirical coverage probabilities obtained using the REML estimates for the covariance parameters σ and ϕ are much lower than 95%, with values ranging from 64% to 89% and an Average Empirical Coverage of 81.4%. This discrepancy can be attributed to the error introduced into the model through the estimates of the covariance parameters. It can also be due to the fact that REML theory is based on asymptotics and becomes unreliable for small samples.

The Laplace approximation technique shows a definite improvement over the empirical coverage probability obtained through kriging with the REML parameter estimates of σ and ϕ, with κ fixed at 1.0 for the exponential covariance model. The Laplace approximation technique increased the empirical coverage probability to an Average Empirical Coverage of 92.9% with a range of 87% to 98% coverage.

An issue to note with the Laplace approximation technique is the sometimes erratic

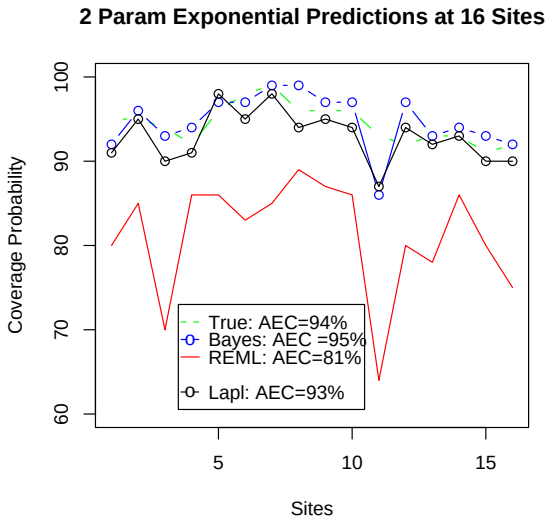

Figure 4.2: Comparison of Laplace, Bayesian, and Plug-In Methods Empirical Coverage Probabilities

behavior of the estimation. In some cases, the correction produced extremely large corrections to the REML estimates of the percentiles, often reversing the direction of the lower and upper bounds of the prediction intervals. This can perhaps be attributed to the REML estimates hitting the bounds set in the optimization algorithm, and also to the fact that the theory is based on asymptotics and may be unreliable for small samples. In the cases where the Laplace approximation produced unreliable estimates, specifically the approximation led to lower bounds for the prediction interval that were larger than the upper bounds, the REML percentile values were used for the empirical coverage probabilities. An area for future study is the cause of and adjustment for the sometimes erratic behavior of the Laplace approximation. If a suitable correction can be found, it is reasonable to assume that the empirical coverage probabilities may improve even beyond the improvements over the REML coverage probabilities seen here.

It is also to be noted that some improvement can be seen within the REML estimation when calculating empirical probabilities after taking into account whether or not the REML estimates of the parameters hit or exceeded the imposed bounds. Bounds on the covariance

parameters were imposed for the sake of the simulation to avoid numerical overflow in the approximation calculations and also to allow for a comparison between the interpolation methods. Slightly higher overall empirical coverage probabilities is seen, ranging from 64% to 94% across the 16 sites, with an Average Empirical Coverage of 85.4%. A topic for future investigation is the robustness of the Bayesian methods and the Laplace approximation technique to different boundary conditions.

4.4.1 Future Considerations for the Laplace Approximation Technique

Another area where an improvement over the Laplace approximation may benefit is in the 3-parameter structure, the powered exponential model. If the cause of the sometimes erratic results can be found, there is evidence that the Laplace technique can greatly improve on the kriging results obtained with REML estimates. This can be seen in the simulation comparing the Bayesian method to the REML results in the 3-parameter model with $\kappa = 0.5, 1.0$, and 1.5, and is a problem for future investigation. This section compares the results of kriging with REML estimates for all three parameters σ, ϕ, and κ with Bayesian methods for the 3-parameter powered exponential model. Under the 3 parameter model, priors were assigned as follows in the Bayesian method:

$$\tau \sim gamma(0.001, 0.001)$$

$$\phi \sim unif(0.0005, 12)$$

$$\kappa \sim unif(0.0005, 2)$$

where τ is the inversion parameter $\frac{1}{\sigma^2}$.

For $\kappa = 0.5$, both the true parameter values and the Bayesian methods produce empirical coverage probabilities around 95%, as seen in the 2 parameter case where κ is fixed at 1.0. The kriging prediction empirical coverage probabilities obtained using the REML estimates

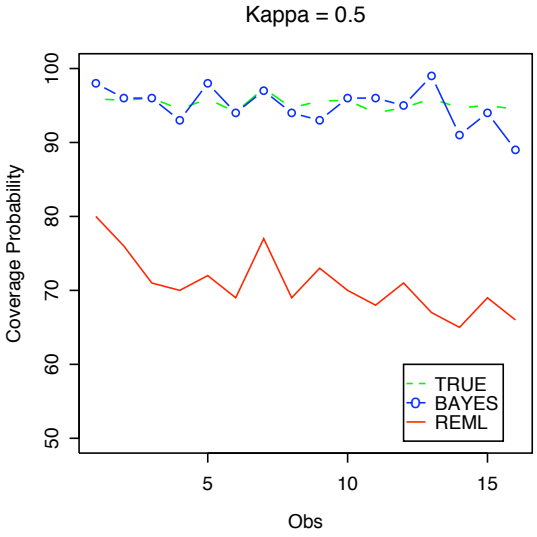

Figure 4.3: Comparison of Bayesian and Plug-In Methods Empirical Coverage Probabilities for $\kappa = 0.5$

for the covariance parameters are much lower, with values ranging from 65% to 80%. The empirical coverage probabilities can be seen in Figure (4.3).

For $\kappa = 1.0$, the results for kriging using the true parameter values and the Bayesian methods are around 95% as seen in the case of $\kappa = 0.5$. As before, the kriging prediction empirical coverage probabilities obtained using the REML estimates for the covariance parameters are much lower than 95%, with values ranging from 77% to 89%. Once again, this discrepancy can be attributed to the error introduced into the model through the estimates of the covariance parameters and the unrealiability of REML when dealing with small sample sizes. Both the true parameter values and the Bayesian methods produce empirical coverage probabilities around 95%, as seen in the 2 parameter case where κ is fixed at 1.0. The empirical coverage probabilities can be seen in Figure (4.4).

For $\kappa = 1.5$, the results for kriging using the true parameter values and the Bayesian methods are around 95% as seen in the case of $\kappa = 0.5$ and $\kappa = 1.5$. The empirical coverage probabilities can be seen in Figure (4.5).

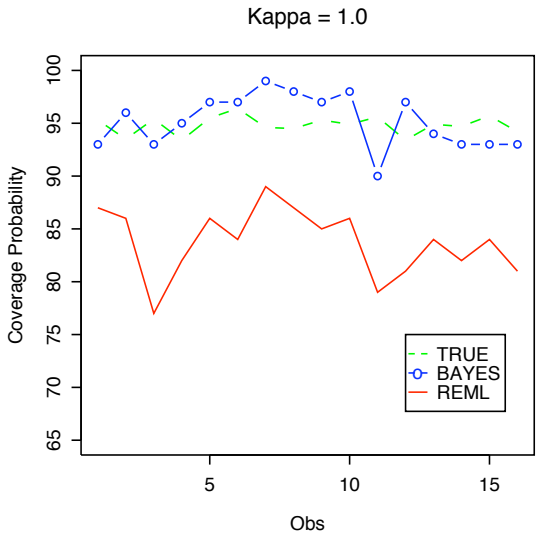

Figure 4.4: Comparison of Bayesian and Plug-In Methods Empirical Coverage Probabilities for $\kappa = 1.0$

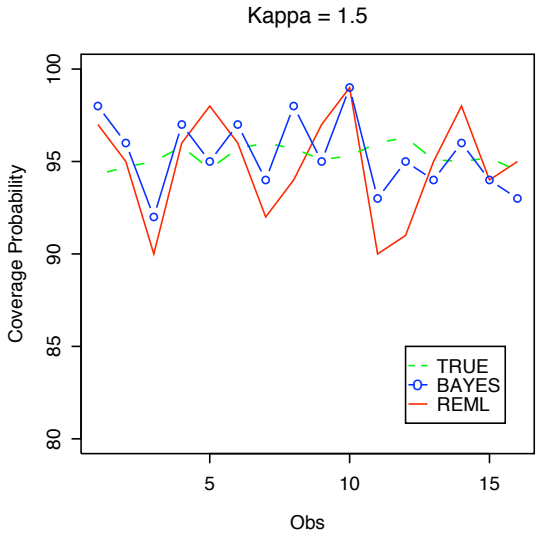

Figure 4.5: Comparison of Bayesian and Plug-In Methods Empirical Coverage Probabilities for $\kappa = 1.5$

4.5 Conclusion

The empirical coverage probabilities for the intervals produced using the known true parameters are close to 95% as expected. However, for $\kappa = 0.5$ and $\kappa = 1.0$, the empirical coverage probabilities show that the intervals produced using the parameters obtained through REML estimation are exhibiting undercoverage.

In contrast, the empirical coverage probabilities for the Bayesian estimates show that the prediction intervals obtained are exhibiting coverage probabilities close to the theoretical coverage probability. The empirical coverage probabilities obtained by applying the Laplace approximation technique also exhibit results close to the theoretical coverage probabilities.

Although this is a short preliminary study, the main point is to show that prediction coverage probabilities can be substantially underestimated when using estimated parameters by the REML method, but appear to be fully satisfactory when using the Bayesian method and the Laplace approximation. These results for the Bayesian methods are consistent with earlier studies on kriging with estimated parameters, such as Zimmerman and Cressie (1992) or Stein (1999). The idea of using Bayesian methods to obtain superior covarage probability has been suggested previously (e.g. by Handcock and Stein (1993) and Berger, et al. 2001). Berger et al and Stein both suggested that there is small coverage probability bias in the Bayesian method, and their corresponding simulations supported this. To our knowledge, this has never been systematically verified by theoretical arguments, except for in special cases, as in those outlined in Stein (1985). Questions for future study include whether the use of reference priors, as advocated by Berger *et al.* (2001), would lead to even better agreement between the nominal and actual coverage probabilities of Bayesian prediction intervals.

CHAPTER 5

Non-linear Predictand and Multivariate Kriging

5.1 Motivation

The motivation behind consideration of the non-linear predictand in the spatial setting is found in considering a necessary preliminary transformation. Ordinary Least Squares Regression, the simplest way of performing model comparison, ignores the possible effects of spatial and/or temporal correlation. In kriging, it is common to apply an initial transformation, such as the square root or a logarithmic transformation, to achieve better goodness of fit to a Gaussian process. The prediction intervals for a single observation are easily calculated. First a prediction interval is calculated for a normally distributed variable and then the transformation is inverted. However for multiple predictions or a prediction that depends on multiple locations, it is not as straightforward. Consider $Z = \sum_{j=1}^{m} h(Y_{0,j})$ where $Y_{0,j}$ is the jth component of the vector Y_0 of predictions and $h()$ is nonlinear. There is no direct application of Gaussian-process theory to find an exact prediction interval for Z in the case of a nonlinear predictand.

An example arose recently in Smith, Kolenikov, and Cox (2003) where a variance-stabilizing transformation, the square root transformation, was desired. In this case, the nonlinear h function was a square and the quantity being predicted was an annual mean of $PM_{2.5}$, equivalent to a sum of daily values, at a particular location. A Gaussian process model was fitted to the square root of $PM_{2.5}$ itself and it is not possible to represent the annual mean $PM_{2.5}$ as a linear function of a Gaussian process. The issues raised are quite typical in applications of spatial statistics to atmospheric pollution data.

This expands on Smith and Zhu's examination of the coverage probability bias in the univariate predictand case.

The starting point for Bayesian analysis for scalar Y_0:

$$P(Y_0 < z|Y) = \frac{\int\int P(z|Y,\beta,\theta) f(\theta|Y,\beta)\pi(\beta,\theta)d\beta d\theta}{\int\int f(\theta|Y,\beta)\pi(\beta,\theta)d\beta d\theta}$$

Through analytic integration with respect to β and routine manipulation it can be shown that

$$P(Y_0 < z|Y) = \frac{\int G(z|Y,\theta)e^{l_n^*(\theta|Y)}e^{Q(\theta)}d\theta}{\int e^{l_n*(\theta|Y)}e^{Q(\theta)}d\theta}$$

where $l_n^*(\theta)$ is the restricted log likelihood, $Q(\theta)$ =log($\pi(\theta)$), and define $G(z|Y,\theta)$ as the predictive distribution function of θ based universal kriging to be predicted.

5.2 Non-linear Predictand

In the multivariate case, the predictand can be written as

$$\mathbf{H(Y_0)} = \sum_{j=1}^{m} h(Y_{0,j}) \tag{5.1}$$

where $h(\cdot)$ is a linear kriging function, such as $h(y) = y^2$ as in Smith, Kolenikov and Cox, and $H(\cdot)$ is a more general transformation function, not necessarily of additive linear functions.

Of interest is the cumulative predictive distribution function:

$$G(z;Y,\theta) = P\{H(Y_o) \leq z|Y,\theta\} \tag{5.2}$$

where $H = \sum_{j=1}^{m} h(Y_{0,j})$ or $H = H(Y_0)$ is the predictand of interest, such as the data transformation used.

5.3 Multivariate Universal Kriging

The universal kriging model can be be written as in Equation (2.2) on page 5.

For multivariate kriging where Y_0 is a vector of m predictions

$$Y_0 = \begin{pmatrix} Y_{0,1} \\ . \\ . \\ Y_{0,m} \end{pmatrix} \text{ and } \text{Cov}\left\{ \begin{pmatrix} Y_{0,1} \\ . \\ . \\ Y_{0,m} \end{pmatrix} \right\} = \begin{pmatrix} \sigma_{11} & ... & \sigma_{1m} \\ . & ... & . \\ \sigma_{1m} & ... & \sigma_{mm} \end{pmatrix}$$

Let V_0 denote the covariance matrix of Y_0 above.

The joint distribution of Y and Y_0 is

$$\begin{pmatrix} Y \\ Y_0 \end{pmatrix} \sim N\left[\begin{pmatrix} X\beta \\ x_0\beta \end{pmatrix}, \begin{pmatrix} V & \tau \\ \tau^T & V_0 \end{pmatrix} \right]$$

where x_0 is a matrix of predictand locations, and possibly other regressors, and τ is a matrix of covariance elements.

The predictive distribution of $Y_0|Y$ is $N[\Lambda^T Y, V_0]$ where Λ is a matrix of universal kriging weights such that $\text{E}(Y_j|Y) = \Lambda_j^T Y$ and $\text{E}(Y_0|Y) = \Lambda^T Y$:

$$\lambda(\theta) = V^{-1}(\theta)w(\theta) + V^{-1}(\theta)X(X^T V^{-1}(\theta)X)^{-1}(x_0 - X^T V^{-1}(\theta)w(\theta))$$

as derived in Equation (2.4).

The prediction error covariance matrix is then derived from:

$$\begin{aligned} \text{E}\left[(z_i - \lambda_i^T Y)(z_j - \lambda_j^T Y)\right] &= \sigma_{ij} - \lambda_i^T \tau_j - \lambda_j^T \tau_i + \lambda_i^T V \lambda_j \\ &= \sigma_{ij} - \tau_j^T V^{-1} \tau_i \end{aligned}$$

$$+ \quad (x_i - X^T V^{-1}\tau_i)^T (X^T V^{-1} X)^{-1} (x_j - X^T V^{-1}\tau_j)$$

and can be written

$$\begin{aligned} \mathrm{E}\left[(Y_0 - \Lambda^T Y)(Y_0 - \lambda^T Y)^T\right] &= V_0 - \Lambda^T \tau - \tau^T \Lambda + \Lambda^T V \Lambda \\ &= V_0 - \tau^T V^{-1} \tau \\ &+ (x_0 - X^T V^{-1}\tau)^T (X^T V^{-1} X)^{-1} (x_0 - X^T V^{-1}\tau) \end{aligned}$$

$\mathrm{E}\big[(Y_0 - \Lambda^T Y)(Y_0 - \lambda^T Y)^T\big]$ is subsequently referred to as $\Sigma(\theta)$.

5.3.1 A Key Identity

A key identity, detailed in this section, is that the non-linear case, defining $G(z|Y,\theta) = P(H(Y_0) \leq z|Y,\theta)$ as the universal kriging function to be predicted for non-linear predictand $H(Y_0)$ and Y_0 a vector of predictions, takes a form similar to the form in Smith and Zhu (2004) for the univariate case:

$$P\left(H(Y_0) < z|Y\right) = \frac{\int G(z|Y,\theta) e^{l_n^*(\theta|Y)} \pi(\theta) d\theta}{\int e^{l_n^*(\theta|Y)} \pi(\theta) d\theta} \tag{5.3}$$

In this section we prove that the restricted likelihood formula for the multivariate case works out analogously to the univariate predictand case.

A simplification of Harville's formula is used to make the same simplification in multivariate case as done in Smith and Zhu (2004) for the univariate case. Assuming the bivariate structure defined in (2.3) holds and $f_n(Y;\beta,\theta)$ defines the density of Y, Harville derives

$$\int f_n(Y;\beta,\theta) d\beta = |X^T X|^{-1/2} e^{ln(\theta)} \tag{5.4}$$

which is extended to the joint density of Y and Y_0, $f_{n+1}(Y, Y_0;\beta,\theta)$

$$\int f_{n+1}(Y, Y_0; \beta, \theta) d\beta = |X^T X|^{-1/2} e^{ln(\theta)} \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{ -\frac{1}{2} \left(\frac{Y_0 - \lambda^T Y}{\sigma_0} \right)^2 \right\} \tag{5.5}$$

where λ and σ_0^2 are, respectively, the vector of kriging weights and the MSPE as defined in Equations (2.4) and (2.5).

5.3.2 Univariate Kriging Prediction

Let $f_n(Y; \beta, \theta), f_{n+1}(Y, Y_0; \beta, \theta)$ denote the density of Y and the joint density of Y and Y_0, respectively. Using the identity

$$(y - X\beta)^T H^{-1} (y - X\beta) = (y - X\hat{\beta})^T H^{-1} (\beta - X\hat{\beta}) + (y - \hat{\beta})^T X^T H^{-1} X (\beta - \hat{\beta}),$$

Harville shows

$$f_n(Y; \beta, \theta) = |X^T X|^{-\frac{1}{2}} \mathrm{e}^{l_n(\theta)} (2\pi)^{-\frac{q}{2}} |X^T V^{-1} X|^{\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\hat{\beta} - \beta)^T X^T V^{-1} X (\hat{\beta} - \beta) \right\}$$

where $e^{l_n(\theta)} = (2\pi)^{-\frac{n-q}{2}} |X^T X|^{-\frac{1}{2}} V(\theta)|^{-\frac{1}{2}} |X^T V(\theta)^{-1} X|^{-\frac{1}{2}} \exp\left(-\frac{G^2(\theta)}{2} \right)$, the restricted log-likelihood (Smith, 2001 and Stein, 1999). This leads to

$$f_n(Y; \beta, \theta) = (2\pi)^{-\frac{q}{2}} |X^T X|^{-\frac{1}{2}} |V|^{-\frac{1}{2}} |X^T V^{-1} X|^{\frac{1}{2}} \exp\left\{ -\frac{1}{2} (Y - X\hat{\beta})^T X^T V^{-1} X (Y - X\hat{\beta}) \right\}$$

Lemma 1

$$\int f_n(Y; \beta, \theta) d\beta = |X^T X|^{-\frac{1}{2}} \mathrm{e}^{\ln(\theta)} \tag{5.6}$$

where $\ln(\theta)$ is as in $\ln^*(\theta)$ previously defined:

$$l_n^*(\theta) = -\frac{n-q}{2} \log(2\pi) + \frac{1}{2} \log |X^T X| - \frac{1}{2} \log |X^T V(\theta)^{-1} X| - \frac{1}{2} \log |V(\theta)| - \frac{1}{2} G^2(\theta)$$

and due to an expansion of Harville (1974) in Smith and Zhu (2004) for univariate prediction.

Lemma 2

$$\int f_{n+1}(Y, Y_0; \beta, \theta) d\beta = |X^T X|^{-\frac{1}{2}} e^{l_n(\theta)} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{ -\frac{1}{2} \left(\frac{Y_0 - \lambda^T Y}{\sigma_0} \right)^2 \right\} \tag{5.7}$$

Thus the predictive density of univariate Y_0 given Y and θ as given in Smith and Zhu is

$$\psi(Y_0; Y, \theta) = \frac{\int f_{n+1}(Y, Y_0; \beta, \theta) d\beta}{\int f_n(Y; \beta, \theta) d\beta} = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{ -\frac{1}{2} \left(\frac{Y_0 - \lambda^T Y}{\sigma_0} \right)^2 \right\} \tag{5.8}$$

Thus, as used in Smith and Zhu (2004), in the univariate case the Bayesian predictor agrees with the universal kriging predictor.

Multivariate Kriging Prediction

Lemma 3

The multivariate prediction density of $Y_0 | Y, \theta$ is

$$\frac{\int f_{n+m}(Y, Y_0; \beta, \theta) d\beta}{\int f_n(Y; \beta, \theta) d\beta} = (2\pi)^{-\frac{m-q}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (Y_0 - \Lambda^T Y)^T \Sigma^{-1} (Y_0 - \Lambda^T Y) \right\} \tag{5.9}$$

where

$$\Sigma = V_0 - \tau^T V^{-1} \tau + (x_0 - X^T V^{-1} \tau)^T (X^T V^{-1} X)^{-1} (x_0 - X^T V^{-1} \tau). \tag{5.10}$$

Thus in the multivariate case the Bayesian predictor also agrees with the universal kriging predictor.

Proof of Lemma 3

For multivariate prediction at m sites, let $T = Y_0 - \Lambda^T Y$. Since $Y_0|Y \sim N[x_)^T\beta + \tau^T V_0^{-1}(Y - X\beta), V_0 - \tau^T V^{-1}\tau]$,

$$T|Y \sim N[x_0^T\beta + \tau^T V^{-1}(Y - X\beta) - \Lambda^T Y, V_0 - \tau^T V^{-1}\tau]$$

where $\mathrm{E}(T|Y)$ can be rewritten as $-(X_0^T - \tau^T V^{-1}X)(\hat{\beta} - \beta)$.

$$\begin{aligned} f_{n+m}(Y, Y_0; \beta, \theta) &= f_n(Y; \beta, \theta) \times f_m(Y_0|Y; \beta, \theta) \\ &= |X^T X|^{-\frac{1}{2}} \exp^{ln}(2\pi)^{-\frac{q}{2}} |X^T V^{-1} X|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\hat{\beta} - \beta)^T X^T V^{-1} X(\beta \hat{-} \beta)\right\} \\ &\times \quad (2\pi)^{-\frac{m}{2}} |V_0 - \tau^T V^{-1}\tau|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} T_{V\beta}\right\} \end{aligned}$$

where $T_{V\beta} = (T + (X_0^T - \tau^T V^{-1}X)(\hat{\beta} - \beta))^T |V_0 - \tau^T V^{-1}\tau|^{-1}(T + (X_0^T - \tau^T V^{-1}X)(\hat{\beta} - \beta))$.

The exponent term

$$\exp\left\{-\frac{1}{2}(\hat{\beta} - \beta)^T X^T V^{-1} X(\beta \hat{-} \beta)\right\} \times \exp\left\{-\frac{1}{2} T_{V\beta}\right\}$$

can be expressed as

$$\begin{aligned} &(\hat{\beta} - \beta)^T \left[X^T V^{-1} X + (X_0 - X^T V^{-1}\tau)(V_0 - \tau^T V^{-1}\tau)^{-1}(X_0 - X^T V^{-1}\tau)^T\right](\hat{\beta} - \beta) \\ &+2(\hat{\beta} - \beta)^T (X_0 - X^T V^{-1}\tau)(V_0 - \tau^T V^{-1}\tau)^{-1} T + T^T (V_0 - \tau^T V^{-1}\tau)^{-1} T \end{aligned}$$

Writing

$$\begin{aligned} A &= X^T V^{-1} X + (X_0 - X^T V^{-1}\tau)(V_0 - \tau^T V^{-1}\tau)^{-1}(X_0 - X^T V^{-1}\tau)^T, \\ B &= (X_0 - X^T V^{-1}\tau)(V_0 - \tau^T V^{-1}\tau)^{-1} T, \end{aligned}$$

$$C \;=\; T^T(V_0 - \tau^T V^{-1}\tau)^{-1}T$$

and using the identity

$$Z^T AZ + 2Z^T B + C = (Z + A^{-1}B)^T A(Z + A^{-1}B) - B^T A^{-1}B + C$$

for completing the square, this becomes

$$\begin{aligned}
& (\hat{\beta} - \beta + A^{-1}B)^T \\
\times \quad & \left[X^T V^{-1}X + (X_0 - X^T V^{-1}\tau)^T (V_0 - \tau^T V^{-1}\tau)^{-1}(X_0 - X^T V^{-1}\tau)\right](\hat{\beta} - \beta + A^{-1}B) \\
- \quad & \left[(X_0 - X^T V^{-1}\tau)(V_0 - \tau^T V^{-1}\tau)^{-1}T\right]^T \\
\times \quad & \left[X^T V^{-1}X + (X_0 - X^T V^{-1}\tau)^T (V_0 - \tau^T V^{-1}\tau)^{-1}(X_0 - X^T V^{-1}\tau)\right]^{-1} \\
\times \quad & (X_0 - X^T V^{-1}\tau)(V_0 - \tau^T V^{-1}\tau)^{-1}T \\
+ \quad & T^T(V_0 - \tau^T V^{-1}\tau)^{-1}T
\end{aligned}$$

Note that the first terms are a normal kernal in β with mean term $\hat{\beta} + A^{-1}B$. Integrating the exponential term with respect to β leaves

$$\begin{aligned}
& |X^T X|^{-\frac{1}{2}} e^{l_n(\theta)} (2\pi)^{-\frac{m-q}{2}} \\
* \quad & |V_0 - \tau^T V^{-1}\tau|^{-\frac{1}{2}} |X^T V^{-1}X|^{\frac{1}{2}} |X^T V^{-1}X| \\
+ \quad & (X_0 - X^T V^{-1}\tau)^T (V_0 - \tau^T V^{-1}\tau)^{-1}(X_0 - X^T V^{-1}\tau)|^{-\frac{1}{2}} \\
* \quad & \exp\left\{-\frac{1}{2} * - \left[(X_0 - X^T V^{-1}\tau)(V_0 - \tau^T V^{-1}\tau)^{-1}T\right]^T\right. \\
* \quad & \left[X^T V^{-1}X + (X_0 - X^T V^{-1}\tau)^T (V_0 - \tau^T V^{-1}\tau)^{-1}(X_0 - X^T V^{-1}\tau)\right]^{-1} \\
* \quad & \left.(X_0 - X^T V^{-1}\tau)(V_0 - \tau^T V^{-1}\tau)^{-1}T + T^T(V_0 - \tau^T V^{-1}\tau)^{-1}T\right\}
\end{aligned}$$

This reduces algebraically to

$$|X^TX|^{-\frac{1}{2}}e^{l_n(\theta)}(2\pi)^{-\frac{m-q}{2}}|V_0-\tau^TV^{-1}\tau|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}T^T(V_0-\tau^TV^{-1}\tau)^{-1}T\right\}$$

So the multivariate prediction distribution of $Y_0|Y$ is

$$\frac{\int f_{n+m}(Y,Y_0;\beta,\theta)d\beta}{\int f_n(Y;\beta,\theta)d\beta}=(2\pi)^{-\frac{m}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(Y_0-\Lambda^TY)^T\Sigma^{-1}(Y_0-\Lambda^TY)\right\} \qquad (5.11)$$

where $\Sigma=V_0-\tau^TV^{-1}\tau$.

This proves that the multivariate predictive distribution in Equation (5.11) is analogous to the univariate development in Equation (5.8).

5.4 Estimation Technique for a Multivariate Non-Linear Predictand

The objective is to evaluate $P(H(Y_0)\leq z|y;\theta)$ and its partial derivatives with respect to both θ and z. For the case of the multivariate, possibly non-linear predictand, the exact form of G may not be easily manipulated. (However, it must be twice-differentiable with respect to z and θ.) In addition, the issue of dependence between Y and expressions involving Y_0 is not easily resolved through the identity methods used in the case of the univariate normal predictand.

Once the corresponding theoretical expressions have been developed in terms of G, its derivatives, and $\frac{\partial}{\partial\theta}lnf(\theta)$, numerical methods can be used to evaluate each of the terms. In order to solve the system of equations for the coverage probability bias and other desired quantities, a bootstrap method is employed.

The Laplace approach considered here is a second order asymptotic approximation to the integral of the prediction distribution. It involves differentiation with respect to the components of θ. Recall the expression, developed in Equation (4.2), for the univariate

Bayesian predictive posterior:

$$\tilde{\psi} = \frac{\int \psi(Y_0; Y, \theta) e^{Q(\theta)} e^{l_n(\theta)} d\theta}{\int e^{Q(\theta)} e^{-ng(\theta)} d\theta} \tag{5.12}$$

where ψ is the function of θ whose posterior expectation we wish to evaluate, $e^{Q(\theta)}$ is the prior, and $e^{l_n(\theta)}$ is the likelihood. Smith (1999) expressed this as

$$\tilde{\psi} = \hat{\psi} + \hat{\mathcal{D}} + O_p(n^{-2})$$

where

$$\mathcal{D} = \frac{1}{2} U_{ijk} \psi_l U^{ij} U^{kl} - \frac{1}{2} (\psi_{ij} + 2\psi_i Q_j) U^{ij}$$

Subscripts indicate differentiation with respect to the components of θ, and $U_i = \frac{\partial l_n(\theta)}{\partial \theta^i}$, $U_{ij} = \frac{\partial^2 l_n(\theta)}{\partial \theta^i \partial \theta^j}$, etc, and U^{ij} is the (i, j) entry of the inverse of the matrix whose (i, j) entry is U_{ij}.

In the univariate linear development established in Smith and Zhu ψ and its derivatives with respect to θ (ψ_i, ψ_{ij} etc) and with respect to z (ψ‘) can be written down explicitly using standard identities. This is not necessarily true for the non-linear multivariate case. Thus the need for a method to determine the derivatives of the predictive distribution function is readily apparent.

5.4.1 Bootstrap Method

In order to develop methodology to construct the predictive distribution and its derivatives, a bootstrap method is considered. The issue of dependence can also be addressed by running a type of double bootstrap, which takes the form of an inner and outer loop. The outer loop takes B samples from the multivariate vector $Y|\theta$ where θ has been obtained through REML. The inner bootstrap loop samples from the predictive distribution of $Y_0|Y$.

5.4.2 Bootstrap Details

The following outlines the steps of the bootstrap procedure more in detail.

First, a parametric bootstrap is run sampling from the parametric distribution of the data, Y, which depends on θ. Prior to the bootstrap, $\hat{\theta}$ has been estimated by some means, in this case by restricted maximum likelihood estimation (REML.) B parametric samples are obtained from

$$Y \sim N(X\beta, V(\theta))$$

Second, a bootstrap is run of B_1 samples from $Y_0|Y, \theta$ from the universal kriging distribution, which is a multivariate normal distribution:

$$p(Y_0|Y) = \frac{\int G(z|Y,\theta) e^{l_n^*(Y|\theta)} \pi(\theta) d\theta}{\int e^{l_n^*(Y|\theta)} \pi(\theta) d\theta} \tag{5.13}$$

where $l_n^*(Y|\theta)$ is the restricted log likelihood and $G(z|Y,\theta)$ is a known density of θ to be predicted. For each Y a sample of B_1 Y_0's is obtained. $H(Y_0)$ is a non-linear function of a normal random variable, and the empirical cdf is computed for each z.

Using the Plug-In method, θ is estimated by REML. If $H(Y_0)$ is a quadratic form, the distribution could in principle be calculated analytically using results for quadratic forms of the multivariate normal distribution. But for general, non-linear $H(Y_0)$, a bootstrap approximation is needed. A bootstrap method can be performed to obtain predictions.

5.4.3 Monte Carlo Simulation of the Distribution Function and Its Derivatives

Treating derivatives of $G(z|Y,\theta) = P\{H(Y_0) \leq z|Y,\theta)\}$ as known functions of θ for the analytical solution and in practice using a numerical bootstrap method as functions of $\hat{\theta}$, can be explored. Let G_B denote estimation of G through a bootstrap method.

1. For $b = 1, \ldots, B$ replications generate $Y_0^{(b)} \sim N[\Lambda^T Y, V_0]$.

2. Calculate $G_B(z|Y,\theta) = \frac{1}{B}\sum_b I\{H(Y_0^{(b)}) \leq z\}$.

In practice, if $G(z|Y,\theta)$ is unknown or is of a form not easily manipulated, a method may need to be employed that allows for considering the predictive density function as well as the empirical predictive distribution derived above.

5.4.4 Employing Kernel Density Estimation

If G is unknown or is of a form that is not easily manipulated, in order to determine $P[H(Y_0) \leq z]$ for the non-linear predictand $H(Y_0)$, a kernel density can be chosen. The kernel density needs to be differentiable with respect to z and with respect to the components of θ. A kernel density is needed to obtain estimates of the derivatives of the distribution function, with respect to both θ and the prediction point z. The inner loop of the bootstrap can be run using the proper kernel, and the empirical cdf estimated using the kernel density.

Here we consider the form of kernel density estimation outlined in Section (3.9). For the ease of consistent notation, the sample size is expressed as B, where here we are considering a bootstrap sample for approximating the predictive density. Let K be the kernel density, with cumulative distribution function K_1 where $K = K_1'$. Therefore, the density of z, $f(z)$, is approximated by:

$$\hat{f}(z) = \frac{1}{Bh}\Sigma_{b=1}^{B} K(\frac{z - H(Y_0^b)}{h}) \tag{5.14}$$

where h is the bandwidth.

Using the Epanechnikov kernel described in Chapter 2 we have:

$$K(t) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}t^2\right) & -\sqrt{5} \leq t \leq \sqrt{5} \\ 0 & otherwise. \end{cases}$$

where $t = \frac{z - H(Y_0^b)}{h}$.

The Epanechnikov density is used to estimate the predictive density, here denoted in the multivariate non-linear case by $G_B^{'}(z|Y,\theta)$.

Using the Epanechnikov cumulative distribution as described in Chapter 2 to approximate the predictive distribution $G(z|Y,\theta)$, note that the cdf is approximated by $\frac{1}{B}\Sigma_{b=1}^{B}K_1(\frac{z-x}{h})$.

$$K_1(t) = \begin{cases} \frac{3}{4\sqrt{5}}\left(t - \frac{1}{15}t^3\right) + 0.5 & -\sqrt{5} \leq t \leq \sqrt{5} \\ 0 & otherwise. \end{cases}$$

where $t = \frac{z-H(Y_o^b)}{h}$ and the smoothing parameter h is estimated arbitrarily or using the methods described in Section (3.9.4) for choosing the optimal bandwidth.

We use the cumulative distribution of the kernel density estimate to approximate $G(z|Y,\theta)$, denoted G_B for the bootstrap estimate.

$$\begin{aligned} G_B(z|Y,\theta) &= \frac{1}{B}\Sigma I\{H(Y_o^b) \leq z\} \\ &\approx \frac{1}{B}\Sigma_{b=1}^{B}K_1(\frac{z - H(Y_o^b)}{h}) \end{aligned} \tag{5.15}$$

5.4.5 Laplace Approach

Developing Laplace is a mechanical calculation. The theory behind the Laplace estimation was developed in Chapter 3. The Laplace approximation of an integral relies on a Taylor expansion about its maximum. This uses partial derivatives. Therefore there is need for an expression for the partial derivatives. The application we are interested in here is the form found in the Bayesian predictive posterior expressed in Equation (4.2).

The direct approximation for a Bayes estimator is an alternative to MCMC methods:

$$\tilde{\psi} = \hat{\psi} + \hat{\mathcal{D}} + O_p(n^{-2})$$

where $\mathcal{D}$ is as expressed in Equation (4.4):

$$\mathcal{D} = \frac{1}{2}U_{ijk}\psi_l U^{ij}U^{kl} - \frac{1}{2}(\psi_{ij} + 2\psi_i Q_j)U^{ij} \tag{5.16}$$

U_i, U_{ij}, U_{ijk}, and U^{ij} etc can be directly calculated from the restricted log-likelihood. However, ψ_i, ψ_{ij}, etc need additional methodology for calculation.

Consider the distribution function of a random variable Y, $\psi(z_\alpha;\theta)$, where z_α is the α-quantile of the distribution. This leads to construction of prediction intervals. For the Bayesian prediction intervals, we define $\tilde{z}_\alpha$ as in Equation(4.5).

Note that $\tilde{\psi} = \frac{\int \psi(\theta)e^{ng(\theta)}d\theta}{\int e^{ng(\theta)}d\theta}$ involves $\hat{\theta}, \psi(\hat{\theta}) = \hat{\psi}^*, \psi_i(\hat{\theta}), \psi_{ij}(\hat{\theta})$, etc. Once you derive $\psi(\hat{\theta})$, $\psi_i(\hat{\theta})$, in the linear case $\psi_{ij}(\hat{\theta})$, etc can be derived with minimal work.

These expressions hold for any function $\psi(\theta)$ whose expectation needs to be evaluated, including the distribution function of a scalar Y_0 and the distribution function of a nonlinear function $H(Y_0)$. The key is the evaluation of of $\psi(\theta)$ and its partial derivatives of up to order 2. The partial derivatives can be expressed as expectations with respect to the predictive distribution function.

5.4.6 Calculation of Partial Derivatives

In order to evaluate the Laplace approximation, partial derivatives up to the second order with respect to θ are necessary. Here we consider expressing the partial derivatives as expectations with respect to the predictive distribution function in order to evaluate the non-linear multivariate case, where the predictive distribution function may be of a form that is unknown or not easily manipulated. The derivative of the predictive distribution function with respect to z will also be needed. A method for approximating the derivative with respect to z using kernel density estimation is outlined in Section (6.1.4).

Note: $\frac{\partial(\ln f(\theta))}{\partial\theta^i} = \frac{\partial f(\theta)}{\partial\theta^i f(\theta)} \implies \frac{\partial(\ln f(\theta))}{\partial\theta^i} f(\theta) = \frac{\partial f(\theta)}{\partial\theta^i}$.

Thus for the derivative of the expected value of $P[H(Y_0) \leq z|Y,\theta]$ with respect to the restricted likelihood $f(Y_0|Y,\theta)$ we have

$$\begin{aligned}\frac{\partial}{\partial\theta_i}\int I\{H(Y_0)\leq z\}f(Y_0|Y,\theta)dY_0 &= \int I\{H(Y_0)\leq z\}\left[\frac{\partial}{\partial\theta^i}\ln f(Y_0|Y,\theta)\right]f(Y_0|Y,\theta)dY_0 \\ &= \mathrm{E}_{f(Y_0|Y,\theta)}\left[I\{H(Y_0)\leq z\}\frac{\partial}{\partial\theta^i}\ln f(Y_0|Y,\theta)\right]\end{aligned}$$

where $\frac{\partial}{\partial\theta^i}\ln f(Y_0|Y,\theta)$ can be analytically evaluated.

For the second derivative of the expected value of $P[H(Y_0) \leq z|Y,\theta]$, note that $\frac{\partial f(\theta)}{\partial\theta^i} = \frac{\partial(\ln f(\theta))}{\partial\theta^i}f(\theta)$.

Thus

$$\begin{aligned}\frac{\partial}{\partial\theta^j}\left[\frac{\partial f(\theta)}{\partial\theta^i}\right] &= \frac{\partial}{\partial\theta^j}\left[\frac{\partial(\ln f(\theta))}{\partial\theta^i}f(\theta)\right] \\ &= \frac{\partial}{\partial\theta^j}\left[\frac{\partial(\ln f(\theta))}{\partial\theta^i}\right]f(\theta)+\frac{\partial}{\partial\theta^j}f(\theta)\frac{\partial(\ln f(\theta))}{\partial\theta^i} \\ &= \frac{\partial^2}{\partial\theta^i\partial\theta^j}(\ln f(\theta))f(\theta)+\frac{\partial(\ln f(\theta))}{\partial\theta^i}\frac{\partial(\ln f(\theta))}{\partial\theta^j}f(\theta)\end{aligned}$$

Therefore

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\int I\{H(Y_0)\leq z\}f(Y_0|Y,\theta)dY_0 =$$

$$\int I\{H(Y_0)\leq z\}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln f(Y_0|Y,\theta)+\frac{\partial}{\partial\theta^i}\ln f(Y_0|Y,\theta)\frac{\partial}{\partial\theta^j}\ln f(Y_0|Y,\theta)\right]f(Y_0|Y,\theta)dY_0 =$$

$$\mathrm{E}_{f(Y_0|Y,\theta)}\left[I\{H(Y_0)\leq z\}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln f(Y_0|Y,\theta)+\frac{\partial}{\partial\theta^i}\ln f(Y_0|Y,\theta)\frac{\partial}{\partial\theta^j}\ln f(Y_0|Y\theta)\right)\right]$$

The first and second derivatives of $\ln f(Y_0|Y,\theta)$ and $\psi(\hat{\theta})$ can be analytically evaluated in the case of a linear predictand. In the multivariate non-linear setting, methodology is developed to construct analogous results to the univariate case established in Smith and Zhu. In practice, a bootstrap method can be performed to obtain the prediction estimates in conjunction with the bootstrap method that may be needed for density estimation.

5.5 Research Questions

The fundamental question is whether the non-linear multivariate case works out analogously to the linear case developed in Smith and Zhu. Specifically we consider the coverage probability bias for the Bayesian prediction method and the plug-in approach using estimated covariance parameters. Also considered is how to compare different priors from the same viewpoint, specifically whether there is evidence that suggests the existence of a matching prior. The Laplace expansion expresses the Bayesian predictors in a form where this question can be answered analytically, using asymptotic arguments.

We use the universal kriging identity in the linear case:

$$P(Y_0 < z|Y) = \frac{\int G(z|Y,\theta)e^{l_n^*(Y|\theta)}\pi(\theta)d\theta}{\int e^{l_n^*(Y|\theta)}\pi(\theta)d\theta}$$

where $e^{l_n^*(Y|\theta)}$ is the restricted log likelihood and $G(z|Y,\theta)$ is a known universal kriging function of θ to be predicted. We can evaluate $G()$ and as many derivatives as needed using a bootstrap method. An appropriate method has been detailed in the preceding sections.

In order to calculate the coverage probability bias or the expected length of the prediction interval, the calculation of moments of various expressions involving R, S, and their derivatives is needed. By the asymptotic formulae, these can be expressed in terms of the derivatives of ψ and other quantities that are explicit functions of the Gaussian process.

It is important to note that the nonlinear case is in principle no different from the linear

case, just G as the analog of ψ is harder to manipulate. That is the principal reason for saying that eventually the nonlinear case should be no different from the linear case.

We consider the linear case developed in Section (3.6.1) in order to develop the multivariate analogs to those defined in Smith and Zhu. The specific research questions addressed are:

1. To find the coverage probability bias for a known G function, what are the multivariate prediction forms of E[R], E[$\frac{RR^{'}}{\psi^{'}}$], E[S], and E[$-\frac{1}{2}\frac{R^2\psi^{''}}{\psi^{'3}}$] in terms of G, its derivatives, and restricted likelihood. For a non-linear predictand, to what extent do the Bayesian procedure properties compare with the frequentist perspective?

2. How do the analogous multivariate R,S, $\frac{RR^{'}}{\psi^{'}}$, and $-\frac{1}{2}\frac{R^2\psi^{''}}{\psi^{'3}}$ terms work out and how do we evaluate them? The investigation of the bootstrap method, treating derivatives of $G(z|Y,\theta) = P\{H(Y_0 \leq z|Y,\theta)\}$ as known functions of θ for the analytical solution and in practice using a numerical bootstrap method as functions of $\hat{\theta}$, can be explored.

3. What is the multivariate non-linear analog of the univariate prediction formula such that the coverage probability bias = $o_p(n^{-1})$?

4. Can coverage probability bias be expressed as a function of the prior? Ie, is there evidence of the existence of a matching prior for the non-linear predictand?

5. Is there a multivariate non-linear analog to the alternative estimator $z_P^{\dagger}$?

CHAPTER 6

Coverage Probability Bias

Of interest is the cumulative predictive distribution function:

$$G(z; Y, \theta) = P\{H(Y_o) \leq z | Y, \theta\} \tag{6.1}$$

where $H = \sum_{j=1}^{m} h(Y_{0,j})$ or $H = H(Y_0)$ is the predictand of interest, such as the data transformation used.

Also of interest is the P-quantile z_P where $z_P = z_P(Y)$ such that $G(z_P; Y, \theta) = P$ for some given quantile P between 0 and 1. The plug-in estimator, $\hat{z}_P$ can be obtained by simply plugging in the estimate $\hat{\theta}$ obtained through restricted maximum likelihood estimation. The Bayes estimator, $\tilde{z}_P$ is determined by the P-quantile of the Bayesian predictive distribution function. z_P^* will denote either $\hat{z}_P$ or $\tilde{z}_P$.

Let $G^*(z; Y)$ denote an estimator of the true distribution $G(z; Y, \theta)$, specifically $\hat{G}$ or $\tilde{G}$. Assume G^* has an expansion:

$$G^*(z; Y) = G(z; Y, \theta) + n^{-\frac{1}{2}} R(z, Y) + n^{-1} S(z, Y) + o_p(n^{-1})$$

where R and S can be expressed using components of the Fisher information matrix and derivatives of the distribution function as outlined in Section (3.6.3).

This expression of the predictive distribution function involves $\hat{\theta}, G(\hat{\theta}) = \hat{G}, G_i(\hat{\theta}), G_{ij}(\hat{\theta}),$

etc. Once you derive $G(\hat{\theta})$, which can be expressed $\hat{G}$, you can obtain $G_i(\hat{\theta}), G_{ij}(\hat{\theta})$, etc through the methods developed in Chapter Four. Theoretically, the quantities of interest involve expectations with respect to the predictive distribution of the observed Y's. In practice, the expectations can be evaluated as the average over many bootstrapped samples.

G can be estimated empirically with G_B^*:

$$G_B(z|Y,\theta) = \frac{1}{B}\Sigma_b I\{H(Y_o^b) \leq z\} = E[I\{H(Y_o^b) \leq z\}] \tag{6.2}$$

In addition, several derivatives of G are needed for the coverage probability bias as well as the expected length of a bayesian prediction interval. These are outlined in Sections (6.1.2) and (6.1.1).

6.1 Derivatives for the Coverage Probability Bias

The coverage probability bias is the expected value of:

$$\begin{aligned} G(z_P^*;Y,\theta) - G(z_P;Y,\theta) &= -n^{-\frac{1}{2}}R(z_P,Y) \\ &+ n^{-1}\left[\frac{R(z_P,Y)R^{'}(z_P,Y)}{G^{*'}(z_P;Y,\theta)} - S(z_P,Y)\right] + o_p(n^{-1}) \end{aligned} \tag{6.3}$$

The coverage probability bias represents the difference between the $P\{H(Y_0) \leq z_P^*|Y,\theta)\}$ and the target probability P, where z_P^* is the plug-in estimate $\hat{z_P}$ or the bayesian estimate $\tilde{z_P}$ of the P-quantiles of the target distribution.

6.1.1 Derivatives with respect to θ

Recall from Section (3.6.3) that for $\hat{z_P}$ and considering G:

$$R_G = \kappa^{i,j} Z_i G_j \tag{6.4}$$

$$S_G = \kappa^{i,j}\kappa^{k,l}Z_{ik}Z_jG_l + \frac{1}{2}\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk}Z_rZ_sG_t + \frac{1}{2}\kappa^{i,j}\kappa^{k,l}Z_iZ_kG_{jl} \tag{6.5}$$

We will further denote $S_G(\hat{z_P})$ as S_{1G}.

For $\tilde{z_P}$ we have:

$$S_G = S_{1G} + \frac{1}{2}\kappa_{ijk}\kappa^{i,j}\kappa^{k,l}G_l + \left(\frac{1}{2}G_{ij} + G_iQ_j\right)\kappa^{i,j} \tag{6.6}$$

We will further denote $S_G(\tilde{z_P})$ as S_{2G}.

Looking at Equations (6.4) - (6.6), first and second derivatives with respect to θ, G_i and G_{ij}, are required.

Using the derivations outlined in Section (5.4.6) we find the first derivatives with respect to θ:

$$\begin{aligned} G_i = \frac{\partial G_i}{\partial \theta_i} &\approx \frac{\partial}{\partial \theta_i}\frac{1}{B}\Sigma_b I\{H(Y_o^b) \leq z\} \\ &= \frac{\partial}{\partial \theta_i}\int I\{H(Y_0) \leq z\}f(Y_0|Y,\theta)dY_0 \\ &= \mathrm{E}_{f(Y_0|Y,\theta)}\left[I\{H(Y_0) \leq z\}\frac{\partial}{\partial \theta^i}\ln f(Y_0^b|Y,\theta)\right] \end{aligned} \tag{6.7}$$

The second derivatives with respect to θ can be expressed:

$$\begin{aligned} G_{ij} = \frac{\partial^2 G_i}{\partial \theta_i \partial \theta_j} &\approx \frac{\partial^2}{\partial \theta_i \partial \theta_j}\frac{1}{B}\Sigma_b I\{H(Y_o^b) \leq z\} \\ &= \frac{\partial^2}{\partial \theta_i \partial \theta_j}\int I\{H(Y_0) \leq z\}f(Y_0|Y,\theta)dY_0 \\ &= \mathrm{E}_{f(Y_0|Y,\theta)}\left[A_{G_{ij}}\right] \end{aligned} \tag{6.8}$$

where

$$A_{G_{ij}} = I\{H(Y_0) \leq z\} \left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln f(Y_0^b|Y,\theta) + \frac{\partial}{\partial\theta^i} \ln f(Y_0^b|Y,\theta) \frac{\partial}{\partial\theta^j} \ln f(Y_0^b|Y\theta) \right)$$

6.1.2 Derivatives of G with respect to z

Looking at it piece-by-piece, the additional terms require first and second derivatives with respect to z: G', $(G')^2$, $(G')^3$, and G''. Additionally, the derivative of the predictive distribution function with respect to both z and θ, $G_j^{'}$, is needed.

Here we have a generic density function G. For practical applications and the remaining derivations, assume G is known and thus $G^{'}$, $(G^{'})^2$, $(G^{'})^3$ and $G^{''}$ can be found easily. An alternative method using kernel density estimation is outlined in Section (5.4.1).

6.1.3 Components of the G Expansion

Expanding further on the components of R_G and S_G, we consider the expectation with respect to the predicitve distribution function of Y:

$$\begin{aligned} E[R] &= E\left[\kappa^{i,j} Z_i G_j\right] \\ &= \kappa^{i,j} E[Z_i G_j] \\ &= n^{-\frac{1}{2}} \kappa^{i,j} E[U_i G_j] \end{aligned}$$

$$\begin{aligned} E[S_{1G}] &= E\left[\kappa^{i,j}\kappa^{k,l} Z_{ik} Z_j G_l + \frac{1}{2}\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk} Z_r Z_s G_t + \frac{1}{2}\kappa^{i,j}\kappa^{k,l} Z_i Z_k G_{jl}\right] \\ &= \kappa^{i,j}\kappa^{k,l} E[Z_{ik} Z_j G_l] + \frac{1}{2}\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk} E[Z_r Z_s G_t] + \frac{1}{2}\kappa^{i,j}\kappa^{k,l} E[Z_i Z_k G_{jl}] \end{aligned}$$

Further algebraic manipulation of the identities $U_i = n^{\frac{1}{2}} Z_i$ and $U_{ij} = n\kappa_{ij} + n^{\frac{1}{2}} Z_{ij}$ yields

$$Z_{ik}Z_j = (\kappa_{ik} - n^{-1}U_{ik})U_j$$

This leads to:

$$\begin{aligned} E\left[S_{1G}\right] &= \kappa^{i,j}\kappa^{k,l}\kappa_{ik}E\left[U_jG_l\right] - n^{-1}\kappa^{i,j}\kappa^{k,l}E\left[U_{ik}U_jG_l\right] \\ &+ \frac{1}{2}\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk}n^{-1}E\left[U_rU_sG_t\right] + \frac{1}{2}n^{-1}\kappa^{i,j}\kappa^{k,l}E\left[U_iU_kG_{jl}\right] \end{aligned}$$

If $z_P^* = \tilde{z}_P$ we have

$$\begin{aligned} E\left[S_{2G}\right] &= E\left[S_{1G}\right] + E\left[\frac{1}{2}\kappa_{ijk}\kappa^{i,j}\kappa^{k,l}G_l + \left(\frac{1}{2}G_{ij} + G_iQ_j\right)\kappa^{i,j}\right] \\ &= E\left[S_{1G^*}\right] + \frac{1}{2}\kappa_{ijk}\kappa^{i,j}\kappa^{k,l}E\left[G_l\right] + \frac{1}{2}\kappa^{i,j}E\left[G_{ij}\right] + \kappa^{i,j}E\left[G_iQ_j\right] \end{aligned}$$

where $Q(\theta) = \log \pi(\theta)$ from the Bayesian framework.

The final term needed for the coverage probability bias from Equation (6.3) is the expectation of $\frac{RR'}{G^{*'}}$ with respect to the predicitve distribution function of Y:

$$\begin{aligned} E\left[\frac{RR'}{G'}\right] &= E\left[\frac{\kappa^{i,j}Z_iG_j\kappa^{k,l}Z_kG_l'}{G'}\right] \\ &= \kappa^{i,j}\kappa^{k,l}E\left[\frac{Z_iG_jZ_kG_l'}{G'}\right] \\ &= n^{-1}\kappa^{i,j}\kappa^{k,l}E\left[\frac{U_iG_jU_kG_l'}{G'}\right] \end{aligned}$$

6.1.4 Kernel Density Estimation for Derivatives of G

In order to solve for the coverage probability bias, we need to find the expected value of several derivatives of G with respect to the prediction z. In order to evaluate the derivative terms, we consider kernel density estimation as outlined in Section (3.9).

Recall that K is the kernel density, with cumulative distribution function K_1 where $K = K_1'$. Note that the cdf is approximated empirically by

$$\frac{1}{B}\Sigma_{b=1}^{B} K_1(\frac{z - H(Y_0^b)}{h}) \tag{6.9}$$

where B is the number of bootstrap samples used in the density estimation procedure.

We use the kernel density estimate to approximate $G(z|Y,\theta)$.

$$G(z|Y,\theta) = \frac{1}{B}\Sigma I\{H(Y_o^b) \leq z\} \approx \frac{1}{B}\Sigma_{b=1}^{B} K_1(\frac{z - H(Y_o^b)}{h}) \tag{6.10}$$

where b represents the specific iteration out of B iterations in the the bootstrap estimation. The bootstrap estimator can be expressed G_B.

To solve for the first derivative of G, we consider

$$G_B^{'} = K_1' = \hat{f}(z) = \frac{1}{Bh}\Sigma_{b=1}^{B} K(\frac{z - H(Y_o^b)}{h}) \tag{6.11}$$

Kernel Density Estimation for the Derivatives wrt θ

Using this development of the kernel density estimate, we can now substitute it into the derivatives of G with respect to θ.

First derivatives with respect to θ:

$$G_i(z|Y,\theta) \approx \frac{1}{B}\Sigma_{b=1}^{B} K_1(\frac{z - H(Y_o^b)}{h})\frac{\partial}{\partial\theta^i}\ln f(Y_0^b|Y,\theta) \tag{6.12}$$

Second derivatives with respect to θ:

$$\begin{aligned} G_{ij}(z|Y,\theta) &\approx \frac{1}{B}\Sigma_{b=1}^{B} K_1(\frac{z - H(Y_o^b)}{h}) \\ &\times \left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln f(Y_0^b|Y,\theta) + \frac{\partial}{\partial\theta^i}\ln f(Y_0^b|Y,\theta)\frac{\partial}{\partial\theta^j}\ln f(Y_0^b|Y\theta)\right) \end{aligned} \tag{6.13}$$

Derivatives with respect to z and θ

The derivative with respect to both z and θ is needed:

$$G'_{Bi}(z|Y,\theta) \approx \frac{1}{Bh}\Sigma_{b=1}^{B}K(\frac{z-H(Y_o^b)}{h})\frac{\partial}{\partial\theta^i}\ln f(Y_0^b|Y,\theta) \tag{6.14}$$

6.1.5 Kernel Density Estimation within the Expansion Terms

Using the preceding development of the kernel density estimate within the expressions for the derivatives of G* with respect to θ we can develop the expansion terms.

$$E\left[R\right] = n^{-\frac{1}{2}}\kappa^{i,j}E\left[U_iG_j\right] \tag{6.15}$$

with which the substitution of the first derivative of G* with respect to θ:

$$G_{Bj}(z|Y,\theta) \approx \frac{1}{B}\Sigma_{b=1}^{B}K_1(\frac{z-H(Y_o^b)}{h})\frac{\partial}{\partial\theta^j}\ln f(Y_0^b|Y,\theta) \tag{6.16}$$

leads to:

$$E\left[R\right] = n^{-\frac{1}{2}}\kappa^{i,j}E\left[U_i\frac{1}{B}\Sigma_{b=1}^{B}K_1(\frac{z-H(Y_o^b)}{h})\frac{\partial}{\partial\theta^j}\ln f(Y_0^b|Y,\theta)\right] \tag{6.17}$$

For the E[S_{1G}] (when $z_P^* = \hat{z}_P$):

$$\begin{aligned} E\left[S_1\right] &= n^{-1}\kappa^{i,j}\kappa^{k,l}\kappa_{ik}E\left[U_jG_l\right] - \kappa^{i,j}\kappa^{k,l}E\left[U_{ik}U_jG_l\right] \\ &+ \frac{1}{2}n^{-1}\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk}E\left[U_rU_sG_t\right] + \frac{1}{2}n^{-1}\kappa^{i,j}\kappa^{k,l}E\left[U_iU_kG_{jl}\right] \end{aligned} \tag{6.18}$$

with the substitution of Equation (6.16) and the second derivative of G with respect to θ:

$$\begin{aligned} G_{Bij}(z|Y,\theta) &\approx \frac{1}{B}\Sigma_{b=1}^{B}K_1(\frac{z-H(Y_o^b)}{h}) \\ &\times \left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln f(Y_0^b|Y,\theta) + \frac{\partial}{\partial\theta^i}\ln f(Y_0^b|Y,\theta)\frac{\partial}{\partial\theta^j}\ln f(Y_0^b|Y\theta)\right) \end{aligned} \tag{6.19}$$

leads to:

$$\begin{aligned} E\left[S_{1G}\right] &= \kappa^{i,j}\kappa^{k,l}\kappa_{ik}E\left[U_j\right] \\ &- n^{-1}\kappa^{i,j}\kappa^{k,l}E\left[U_{ik}U_j\frac{1}{B}\Sigma_{b=1}^{B}K_1(\frac{z-H(Y_o^b)}{h})\frac{\partial}{\partial\theta^l}\ln f(Y_0^b|Y,\theta)\right] \\ &+ \frac{1}{2}n^{-1}\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk}E\left[U_rU_s\frac{1}{B}\Sigma_{b=1}^{B}K_1(\frac{z-H(Y_o^b)}{h})\frac{\partial}{\partial\theta^s}\ln f(Y_0^b|Y,\theta)\right] \\ &+ \frac{1}{2}n^{-1}\kappa^{i,j}\kappa^{k,l}E\left[U_iU_k\frac{1}{B}\Sigma_{b=1}^{B}K_1(\frac{z-H(Y_o^b)}{h})\times A_{S_{1ln}}\right] \end{aligned}$$

where

$$A_{S_{1ln}} = \left(\frac{\partial^2}{\partial\theta_j\partial\theta_l}\ln f(Y_0^b|Y,\theta) + \frac{\partial}{\partial\theta^j}\ln f(Y_0^b|Y,\theta)\frac{\partial}{\partial\theta^l}\ln f(Y_0^b|Y\theta)\right)$$

when $z_P^* = \tilde{z}_P$, we need $\mathrm{E}[S_{2G}] = E[S_{1G}] + \frac{1}{2}\kappa_{ijk}\kappa^{i,j}\kappa^{k,l}E[G_l] + \frac{1}{2}\kappa^{i,j}E[G_{ij}] + \kappa^{i,j}E[G_iQ_i]$.

With the substitution of Equations (6.16) and (6.19) this leads to:

$$\begin{aligned} E\left[S_{2G}\right] &= E\left[S_{1G}\right] \\ &+ \frac{1}{2}\kappa_{ijk}\kappa^{i,j}\kappa^{k,l}E\left[\frac{1}{B}\Sigma_{b=1}^{B}K_1(\frac{z-H(Y_o^b)}{h})\frac{\partial}{\partial\theta^l}\ln f(Y_0^b|Y,\theta)\right] \\ &+ \frac{1}{2}\kappa^{i,j}E\left[\frac{1}{B}\Sigma_{b=1}^{B}K_1(\frac{z-H(Y_o^b)}{h})A_{S2G^*}\right] \\ &+ \kappa^{i,j}E\left[\frac{1}{B}\Sigma_{b=1}^{B}K_1(\frac{z-H(Y_o^b)}{h})\frac{\partial}{\partial\theta^i}\ln f(Y_0^b|Y,\theta)\right]Q_j \end{aligned} \tag{6.20}$$

where $Q(\theta) = \log(\pi(\theta))$ from the Bayesian framework and

$$A_{S2G} = \left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln f(Y_0^b|Y,\theta) + \frac{\partial}{\partial\theta^i}\ln f(Y_0^b|Y,\theta)\frac{\partial}{\partial\theta^j}\ln f(Y_0^b|Y\theta)\right)$$

The final term needed for the coverage probability bias is E$[\frac{RR'}{G'}]$:

$$E\left[\frac{RR'}{G'}\right] = n^{-1}\kappa^{i,j}\kappa^{k,l}E\left[\frac{U_i G_j U_k G_l'}{G'}\right] \tag{6.21}$$

Substituting the first derivative of G with respect to a component of θ:

$$G_{Bi}(z|Y,\theta) \approx \frac{1}{B}\Sigma_{b=1}^{B} K_1(\frac{z - H(Y_o^b)}{h})\frac{\partial}{\partial\theta^i}\ln f(Y_0^b|Y,\theta)$$

and substituting the derivative of G with respect to z:

$$G_B' \approx \frac{1}{Bh}\Sigma_{b=1}^{B} K(\frac{z - H(Y_o^b)}{h})$$

and the derivative of G with respect to both z and θ is:

$$G'_{Bi}(z|Y,\theta) \approx \frac{1}{Bh}\Sigma_{b=1}^{B} K(\frac{z - H(Y_o^b)}{h})\frac{\partial}{\partial\theta^i}\ln f(Y_0^b|Y,\theta)$$

we have:

$$E\left[\frac{RR'}{G'}\right] = n^{-1}\kappa^{i,j}\kappa^{k,l}E\left[\frac{A_{RR'G}}{B_{RR'G}}\right] \tag{6.22}$$

where

$$\begin{aligned} A_{RR'G} &= U_i\frac{1}{B}\Sigma_{b=1}^{B} K_1(\frac{z - H(Y_o^b)}{h})\frac{\partial}{\partial\theta^j}\ln f(Y_0^b|Y,\theta)U_k \times \\ &\times \Sigma_{b=1}^{B}\frac{1}{Bh}K(\frac{z - H(Y_o^b)}{h})\frac{\partial}{\partial\theta^l}\ln f(Y_0^b|Y,\theta) \end{aligned} \tag{6.23}$$

and

$$B_{RR'G} = \frac{1}{Bh}\Sigma_{b=1}^{B} K\left(\frac{z - H(Y_o^b)}{h}\right) \tag{6.24}$$

6.1.6 Kernel Density Estimation for the Coverage Probability Bias

Ultimately, we are interested in the Bayesian coverage probability bias, $\mathrm{E}[G(z_P^*) - G(z_P)]$, where $z_p^* = \tilde{z}_P$, and where

$$G(z_P^*; Y, \theta) - G(z_P; Y, \theta) = -n^{-\frac{1}{2}} R(z_P, Y) + n^{-1} \left[\frac{R(z_P, Y) R^{'}(z_P, Y)}{G^{*'}(z_P; Y, \theta)} - S(z_P, Y) \right] + o_p(n^{-1})$$

Putting together Equations (6.17), (6.22), and (6.20) we have, for $z_p^* = \tilde{z}_P$:

$$\begin{aligned} E\left[G(z_P^*) - G(z_P)\right] &= n^{-\frac{1}{2}} \left(n^{-\frac{1}{2}} \kappa^{i,j} E\left[U_i \frac{1}{B} \Sigma_{b=1}^{B} K_1(\frac{z - H(Y_o^b)}{h}) \frac{\partial}{\partial \theta^i} \ln f(Y_0^b | Y, \theta) \right] \right) \\ &+ n^{-1} \left(n^{-1} \kappa^{i,j} \kappa^{k,l} E\left[\frac{A_{RR'G}}{B_{RR'G}} \right] - E\left[S_{2G}\right] \right) \end{aligned} \tag{6.25}$$

where $A_{RR'G}$ and $B_{RR'G}$ are as expressed in Equations (6.23) and (6.24) and $\mathrm{E}[S_{2G}]$ is in Equation (6.20).

6.2 Matching Prior Development

The form of Equation (6.20), with $Q_j = \frac{\partial}{\partial \theta^j} \log \pi(\theta)$, suggests the possibility of a matching prior. If π can be chosen such that $E[G(z_P^*) - G(z_P)] = 0$, then the second-order coverage probability bias of the Bayesian predictor z_P^* is 0. While in practice the derivations may be tedious and impractical, the concept of a matching prior is important. Different priors can be compared and chosen according to how well they approach a coverage probability bias of zero.

6.2.1 Asymptotic Frequentist Correction Alternative to Matching Prior

It is not necessary to find the exact form of the matching prior. Finding an appropriate estimator, which we shall denote as $z_P^{\dagger}$, can correspond to the matching prior. This is

seen in Smith and Zhu (2004) and expressed in Equation (3.5). We construct the artificial predictor $z_P^\dagger$, which is equivalent to the Bayesian predictor, as an alternative to solving Equation (6.25) by obtaining the matching prior. For percentile P, define $z_P^\dagger$ by expressing:

$$\begin{aligned} & n\left[E[G(z_P^\dagger)] - E[G(z_P)\right] \\ \text{as} \quad & n\left[E[G(z_P^\dagger)] - E[G^*(\hat{z}_P)] + E[G^*(\hat{z}_P)] - E[G(z_P)]\right] \\ = \quad & n\left[(z_P^\dagger - \hat{z}_P)E[G^{'}(z_P)] + \text{Equation}(6.25)\right] \end{aligned}$$

which leads to an alternative estimator, corresponding to the matching prior, constructed to be:

$$z_P^\dagger = \hat{z}_P - n^{-1}\left[\frac{\text{Equation (6.25)}}{G'(G^{-1}(P))}\right] \tag{6.26}$$

This is a function of the asymptotic bias as seen in Equation (3.5), and is an analog to the univariate normal case.

CHAPTER 7

Expected Length of a Bayesian Prediction Interval

The prediction interval is $(z^*_{P_1}, z^*_{P_2})$, with the Bayesian prediction interval found when $z^*_P = \tilde{z}_P$. A Bayesian prediciton interval minimizes the coverage probability bias as expressed in Equation (6.25). The expected length of the Bayesian prediction interval has applications in network design.

As explained in Smith and Zhu (2004) the expected length of a Bayesian prediction interval may be used as the basis for network design when choosing locations for a network designed to predict a quantity $H(Y_0)$. $Y_1, ..., Y_n$ may be measurements at n monitors at which it is needed to predict Y_0 as accurately as possible. Because precise measurement is not possible a prediction interval for Y_0 can be obtained based on spatial interpolation from $Y_1, ..., Y_n$. A Bayesian prediction interval can be constructed to minimize the coverage probability bias. A particular prior may turn out to produce better estimates, but it may be possible to construct an experimental design to produce a shorter prediction interval. If the first term, $\Sigma \times [G^{-1}(P_2) - G^{-1}(P_1)]$, is of interest the interval can be chosen to minimize Σ, the predictive standard error of Y_0, based on a known θ. The remaining terms account for the prediction error due to the estimation of θ. The benefit of using the expected length of a Bayesian prediction interval is that it accounts for both the predictive and estimation error.

We consider the comparison of the plug-in estimate, $\hat{z}_P$, to the Bayesian estimate, $\tilde{z}_P$. A Taylor expansion leads to the approximation:

$$\tilde{z}_P - \hat{z}_P = -\frac{\tilde{G}(\hat{z}_P) - \hat{G}(\hat{z}_P)}{\hat{G}'(\hat{z}_P)} + O_p(n^{-2}) \tag{7.1}$$

This is analogous to Equation (4.5), and the numerator of the left side of Equation (7.1) can be evaluated using the analogous expression to Equation (4.4) where

$$D = U_{ijk} G_l U^{ij} U^{kl} - \frac{1}{2}(G_{ij} + 2G_i Q_j) U^{ij}$$

The formula for the predictive distribution in Equation (8.1) or its inverse Equation (7.1) provide an alternative to MCMC methods for Bayesian computation.

The expected value of Equation (7.1) can be used in obtaining the expected length of the prediction interval under predictive distribution G.

7.1 Expected Length

Once again as in Chapter Five we consider the expectations with respect to the predictive distribution of Y. In practice, the expectation can be evaluated as the average over bootstrapped samples. For the true and estimated P-quantiles of the predictive distribution z_P and z_P^*, we have:

$$\begin{aligned} E[z_P^* - z_P] &= -n^{-1/2} E\left[\frac{R(z_P, Y)}{G'(z_P; Y, \theta)}\right] \\ &+ n^{-1} E\left[\frac{R(z_P, Y) R'(z_P, Y)}{G'^2(z_P; Y, \theta)}\right] - \frac{1}{2} E\left[\frac{R^2(z_P, Y) G''(z_P; Y, \theta)}{G'^3(z_P; Y, \theta)}\right] \\ &- E\left[\frac{S(z_P, Y)}{G'(z_P; Y, \theta)}\right] + o_p(n^{-1}) \end{aligned} \tag{7.2}$$

where $G' = \frac{\partial G}{\partial z}$.

Several derivatives of G with respect to both z and θ are needed. Developing the deriva-

tives of G is outlined in Sections (5.4.1) and (6.1.5).

In a similar expansion to the terms outlined in Section (6.1.5), Equation(7.2) requires $\mathrm{E}[\frac{R}{G'}]$, $\mathrm{E}[\frac{R^2G''}{G'^3}]$, $\mathrm{E}[\frac{RR'}{G'^2}]$, and $\mathrm{E}[\frac{S}{G'}]$. $G(z_P; Y\theta), R(z_P, Y)$, and $S(z_P, Y)$ are expressed as G, R, and S for simplicity.

7.1.1 Kernel Density Estimation within the Expansion Terms

Recall the first derivative of G with respect to z can be expressed using kernel density estimation as:

$$G'_B = \hat{f}(z) = K'_1 = \frac{1}{Bh}\Sigma_{b=1}^{B} K(\frac{z - H(Y_o^b)}{h})$$

The second derivative of G_B with respect to z can be expressed as:

$$G''_B = \frac{1}{Bh^2}\Sigma_{b=1}^{B} K' \left(\frac{z + h - H(Y_o^b)}{h}\right) \tag{7.3}$$

where K' is the derivative of the chosen kernel density K with respect to z.

Upon the substitution of the G derivative terms with respect to z, and also including the expansions for R, R', and S outlined in Equations (6.4) - (6.6), the terms in Equation (7.2) become

$$\begin{aligned} E\left[\frac{R}{G'}\right] &= E\left[\frac{\kappa^{i,j} Z_i G_j}{\frac{1}{Bh}\Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h})}\right] \\ &= \kappa^{i,j} Bh \times E\left[\frac{Z_i G_j}{\Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h})}\right] \end{aligned}$$

$$\begin{aligned} E\left[\frac{RR'}{G'^2}\right] &= E\left[\frac{\kappa^{i,j} Z_i G_j \kappa^{i,j} Z_k (G_l)'}{(\frac{1}{Bh}\Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h}))^2}\right] \\ &= (\kappa^{i,j})^2 B^2 h^2 \times E\left[\frac{Z_i G_j Z_k (G_l)'}{(\Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h}))^2}\right] \end{aligned}$$

$$
\begin{aligned}
E\left[\frac{R^2G''}{G'^3}\right] &= E\left[\frac{(\kappa^{i,j}Z_iG_j)^2\left(\frac{1}{Bh^2}\Sigma_{b=1}^{B}K'\left(\frac{z-H(Y_o^b)}{h}\right)\right)}{(\frac{1}{Bh}\Sigma_{b=1}^{B}K(\frac{z-H(Y_o^b)}{h}))^3}\right] \\
&= (\kappa^{i,j})^2B^2h\times E\left[\frac{Z_i^2G_j^{*2}\Sigma_{b=1}^{B}K'\left(\frac{z-H(Y_o^b)}{h}\right)}{(\Sigma_{b=1}^{B}K(\frac{z-H(Y_o^b)}{h}))^3}\right]
\end{aligned}
$$

$$
\begin{aligned}
E\left[\frac{S}{G'}\right] &= E\left[\frac{\kappa^{i,j}\kappa^{k,l}Z_{ik}Z_jG_{il}+\frac{1}{2}\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk}Z_rZ_sG_t+\frac{1}{2}\kappa^{i,j}\kappa^{k,l}Z_iZ_kG_{jl}}{\frac{1}{Bh}\Sigma_{b=1}^{B}K(\frac{z-H(Y_o^b)}{h})}\right] \\
&= \kappa^{i,j}\kappa^{k,l}Bh\times E\left[\frac{Z_{ik}Z_jG_l}{\Sigma_{b=1}^{B}K(\frac{z-H(Y_o^b)}{h})}\right] \\
&+ \frac{\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk}Bh}{2}\times E\left[\frac{Z_rZ_sG_t}{\Sigma_{b=1}^{B}K(\frac{z-H(Y_o^b)}{h})}\right] \\
&+ \frac{\kappa^{i,j}\kappa^{k,l}Bh}{2}\times E\left[\frac{Z_iZ_kG_{jl}}{\Sigma_{b=1}^{B}K(\frac{z-H(Y_o^b)}{h})}\right]
\end{aligned}
$$

If $z_P^*=\tilde{z}_P$, $S_{G^*}=S_{G^*2}=S_{G^*1}+\frac{1}{2}\kappa_{ijk}\kappa^{i,j}\kappa^{k,l}G_l+\left(\frac{1}{2}G_{ij}+G_iQ_j\right)\kappa^{i,j}$, so this last term also includes:

$$
+\frac{\kappa_{ijk}\kappa^{i,j}\kappa^{k,l}Bh}{2}\times E\left[\frac{G_l}{\Sigma_{b=1}^{B}K(\frac{z-H(Y_o^b)}{h})}\right]+\kappa^{i,j}Bh\times E\left[\frac{\frac{1}{2}G_{ij}+G_iQ_i}{\Sigma_{b=1}^{B}K(\frac{z-H(Y_o^b)}{h})}\right] \tag{7.4}
$$

For the purpose of solving practical applications numerically, it is more useful to express these asymptotic expansions in terms of U_i and U_{ij} rather than Z_i and Z_{ij}, where $U_i=n^{\frac{1}{2}}$ and $U_{ij}=n\kappa_{ij}+n^{\frac{1}{2}}Z_{ij}$.

$$
E\left[\frac{R}{G'}\right]=\frac{\kappa^{i,j}Bh}{n^{1/2}}\times E\left[\frac{U_iG_j}{\Sigma_{b=1}^{B}K(\frac{z-H(Y_o^b)}{h})}\right]
$$

$$
E\left[\frac{RR'}{G'^2}\right]=\frac{\kappa^{i,j}\kappa^{k,l}B^2h^2}{n}\times E\left[\frac{U_iG_jU_k(G_l)'}{(\Sigma_{b=1}^{B}K(\frac{z-H(Y_o^b)}{h}))^2}\right]
$$

$$E\left[\frac{R^2G''}{G'^3}\right] = \frac{(\kappa^{i,j})^2B^2h}{n} \times E\left[\frac{U_i^2G_j^{*2}\Sigma_{b=1}^{B}K'\left(\frac{z-H(Y_o^b)}{h}\right)}{(\Sigma_{b=1}^{B}K(\frac{z-H(Y_o^b)}{h}))^3}\right]$$

$$\begin{aligned} E\left[\frac{S}{G'}\right] &= \frac{\kappa^{i,j}\kappa^{k,l}Bh}{n^{1/2}} \times E\left[\frac{(n^{-1/2}U_{ik} - n^{1/2}\kappa_{ij})U_jG_l}{\Sigma_{b=1}^{B}K(\frac{z-H(Y_o^b)}{h})}\right] \\ &+ \frac{\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk}Bh}{2n} \times E\left[\frac{U_rU_sG_t}{\Sigma_{b=1}^{B}K(\frac{z-H(Y_o^b)}{h})}\right] \\ &+ \frac{\kappa^{i,j}\kappa^{k,l}Bh}{2n} \times E\left[\frac{U_iU_kG_{jl}}{\Sigma_{b=1}^{B}K(\frac{z-H(Y_o^b)}{h})}\right] \end{aligned}$$

The first term in the E$[\frac{S}{G'}]$ expression becomes

$$\frac{\kappa^{i,j}\kappa^{k,l}Bh}{n} \times E\left[\frac{U_{ik}}{\Sigma_{b=1}^{B}K(\frac{z-H(Y_o^b)}{h})}\right] - \kappa^{i,j}\kappa^{k,l}\kappa_{ij}Bh \times E\left[\frac{U_jG_l}{\Sigma_{b=1}^{B}K(\frac{z-H(Y_o^b)}{h})}\right]$$

If $z_P^* = \tilde{z}_P$, $S_G = S_{G^*2} = S_{G1} + \frac{1}{2}\kappa_{ijk}\kappa^{i,j}\kappa^{k,l}G_l + \left(\frac{1}{2}G_{ij} + G_iQ_j\right)\kappa^{i,j}$, the last term also includes Equation (7.6), which requires no substitution of U_i terms.

7.1.2 Derivatives within Expansion Terms

To further develop the expansion to reach a form that can be solved analytically, we substitute the G^* derivative terms with respect to both z and θ as developed in Equations (6.7) - (6.8) and Equation (6.14):

$$E\left[\frac{R}{G'}\right] = \frac{\kappa^{i,j}Bh}{n^{1/2}} \times E\left[\frac{U_i\frac{1}{B}\Sigma_{b=1}^{B}K_1(\frac{z-H(Y_o^b)}{h})\frac{\partial}{\partial\theta^j}\ln f(Y_0^b|Y,\theta)}{\Sigma_{b=1}^{B}K(\frac{z-H(Y_o^b)}{h})}\right]$$

$$E\left[\frac{RR'}{G'^2}\right] = \frac{\kappa^{i,j}\kappa^{k,l}Bh}{n} \times E\left[\frac{U_i\frac{1}{B}\Sigma_{b=1}^{B}K_1(\frac{z-H(Y_o^b)}{h})\frac{\partial}{\partial\theta^j}\ln f(Y_0^b|Y,\theta)U_k \times A_{Z_{Pln}}}{(\Sigma_{b=1}^{B}K(\frac{z-H(Y_o^b)}{h}))^2}\right]$$

$$E\left[\frac{R^2G''}{G'^3}\right] = \frac{(\kappa^{i,j})^2B^2h}{n} \times E\left[\frac{U_i^2\Sigma_{b=1}^{B}K'\left(\frac{z-H(Y_o^b)}{h}\right) \times (A_{Z_{Pln}})^2}{(\Sigma_{b=1}^{B}K(\frac{z-H(Y_o^b)}{h}))^3}\right]$$

where

$$A_{Z_{Pln}} = \Sigma_{b=1}^{B} K(\frac{z - H(Y_o^b)}{h}) \frac{\partial}{\partial \theta^l} \ln f(Y_0^b | Y, \theta)$$

For the last term, $\mathrm{E}[\frac{S}{G'}]$, we have:

$$\begin{aligned}
E\left[\frac{S}{G'}\right] &= \frac{\kappa^{i,j}\kappa^{k,l}Bh}{n} \times E\left[\frac{U_{ik}}{\Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h})}\right] \\
&- \kappa^{i,j}\kappa^{k,l}\kappa_{ij}h \times E\left[\frac{U_j \Sigma_{b=1}^{B} K_1(\frac{z-H(Y_o^b)}{h}) \frac{\partial}{\partial \theta^l} \ln f(Y_0^b|Y,\theta)}{\Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h})}\right] \\
&+ \frac{\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk}h}{2n} \times E\left[\frac{U_r U_s \Sigma_{b=1}^{B} K_1(\frac{z-H(Y_o^b)}{h}) \frac{\partial}{\partial \theta^t} \ln f(Y_0^b|Y,\theta)}{\Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h})}\right] \\
&+ \frac{\kappa^{i,j}\kappa^{k,l}Bh}{2n} \times E\left[\frac{U_i U_k G_{jl}}{\Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h})}\right]
\end{aligned}$$

where

$$G_{jl} = \frac{1}{B}\Sigma_{b=1}^{B} K_1(\frac{z - H(Y_o^b)}{h}) \left(\frac{\partial^2}{\partial \theta_j \partial \theta_l} \ln f(Y_0^b|Y,\theta) + \frac{\partial}{\partial \theta^j} \ln f(Y_0^b|Y,\theta) \frac{\partial}{\partial \theta^l} \ln f(Y_0^b|Y\theta)\right)$$

In the Bayesian case where $S = S_{2G}$, $\mathrm{E}[\frac{S}{G'}]$ also includes the term:

$$\begin{aligned}
&+ \frac{\kappa_{ijk}\kappa^{i,j}\kappa^{k,l}h}{2} \times E\left[\frac{\Sigma_{b=1}^{B} K_1(\frac{z-H(Y_o^b)}{h}) \frac{\partial}{\partial \theta^l} \ln f(Y_0^b|Y,\theta)}{\Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h})}\right] \\
&+ \kappa^{i,j}h \times E\left[\frac{\frac{1}{2}\Sigma_{b=1}^{B} K_1(\frac{z-H(Y_o^b)}{h}) G_{ij}}{\Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h})}\right] \\
&+ \kappa^{i,j}h \times E\left[\frac{\Sigma_{b=1}^{B} K_1(\frac{z-H(Y_o^b)}{h}) \frac{\partial}{\partial \theta^i} \ln f(Y_0^b|Y,\theta) Q_i}{\Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h})}\right]
\end{aligned}$$

where G_{ij} is as expressed above for G_{jl} with respect to i and j.

Thus we have:

$$
\begin{aligned}
E[z_P^* - z_P] &= -\frac{\kappa^{i,j}Bh}{n} \times E\left[\frac{U_i \frac{1}{B}\Sigma_{b=1}^{B} K_1(\frac{z-H(Y_o^b)}{h})\frac{\partial}{\partial\theta^j}\ln f(Y_0^b|Y,\theta)}{\Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h})}\right] \\
&+ \frac{\kappa^{i,j}\kappa^{k,l}Bh}{n^2} \times E\left[\frac{U_i \frac{1}{B}\Sigma_{b=1}^{B} K_1(\frac{z-H(Y_o^b)}{h})\frac{\partial}{\partial\theta^j}\ln f(Y_0^b|Y,\theta)U_k \times A_{Z_{Pln}}}{(\Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h}))^2}\right] \\
&- \frac{(\kappa^{i,j})^2 B^2 h}{2n} \times E\left[\frac{U_i^2 \Sigma_{b=1}^{B} K'\left(\frac{z-H(Y_o^b)}{h}\right) \times (A_{Z_{Pln}})^2}{(\Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h}))^3}\right] \\
&- \frac{\kappa^{i,j}\kappa^{k,l}Bh}{n} \times E\left[\frac{U_{ik}}{\Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h})}\right] \\
&+ \kappa^{i,j}\kappa^{k,l}\kappa_{ij}h \times E\left[\frac{U_j \Sigma_{b=1}^{B} K_1(\frac{z-H(Y_o^b)}{h})\frac{\partial}{\partial\theta^l}\ln f(Y_0^b|Y,\theta)}{\Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h})}\right] \\
&- \frac{\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk}h}{2n} \times E\left[\frac{U_r U_s \Sigma_{b=1}^{B} K_1(\frac{z-H(Y_o^b)}{h})\frac{\partial}{\partial\theta^t}\ln f(Y_0^b|Y,\theta)}{\Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h})}\right] \\
&- \frac{\kappa^{i,j}\kappa^{k,l}Bh}{2n} \times E\left[\frac{U_i U_k G_{jl}}{\Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h})}\right] \qquad (7.5)
\end{aligned}
$$

where

$$A_{Z_{Pln}} = \Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h})\frac{\partial}{\partial\theta^l}\ln f(Y_0^b|Y,\theta)$$

and

$$G_{jl} = \frac{1}{B}\Sigma_{b=1}^{B} K_1(\frac{z-H(Y_o^b)}{h})\left(\frac{\partial^2}{\partial\theta_j\partial\theta_l}\ln f(Y_0^b|Y,\theta) + \frac{\partial}{\partial\theta^j}\ln f(Y_0^b|Y,\theta)\frac{\partial}{\partial\theta^l}\ln f(Y_0^b|Y\theta)\right)$$

If $S = S_2$ this also includes:

$$
\begin{aligned}
&+ \frac{\kappa_{ijk}\kappa^{i,j}\kappa^{k,l}h}{2} \times E\left[\frac{\Sigma_{b=1}^{B} K_1(\frac{z-H(Y_o^b)}{h})\frac{\partial}{\partial\theta^l}\ln f(Y_0^b|Y,\theta)}{\Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h})}\right] \\
&+ \kappa^{i,j}h \times E\left[\frac{\frac{1}{2}\Sigma_{b=1}^{B} K_1(\frac{z-H(Y_o^b)}{h})G_{ij}}{\Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h})}\right]
\end{aligned}
$$

$$+ \quad \kappa^{i,j} h \times E \left[\frac{\Sigma_{b=1}^{B} K_1(\frac{z-H(Y_o^b)}{h}) \frac{\partial}{\partial \theta^i} \ln f(Y_0^b|Y,\theta) Q_i}{\Sigma_{b=1}^{B} K(\frac{z-H(Y_o^b)}{h})} \right]$$

where G_{ij} is as expressed above for G_{jl} with respect to i and j.

The Bayesian prediction interval provides more accurate coverage probability than the plug-in estimator's prediction interval, but at the cost of a longer interval.

7.2 Expected Length of a Prediction Interval

Ultimately, we are interested in the the expected length of a Bayesian Prediction Interval, $\mathrm{E}[z_p^* - z_p]$, where $z_P^* = \tilde{z}_P$. The prediction interval is $(z_{P_1}^*, z_{P_2}^*)$. The expected length is:

$$\begin{aligned} E[z_{P_2}^* - z_{P_1}^*] &= E[z_{P_2} - z_{P_1}] + E[z_{P_2}^* - z_{P_2}] - E[z_{P_1}^* - z_{P_1}] \\ &= \Sigma[G^{-1}(P_2) - G^{-1}(P_1)] + E[z_{P_2}^* - z_{P_2}] - E[z_{P_1}^* - z_{P_1}] \end{aligned} \tag{7.6}$$

where $\Sigma = V_0 - \tau^T V^{-1} \tau + (x_0 - X^T V^{-1} \tau)^T (X^T V^{-1} X)^{-1} (x_0 - X^T V^{-1} \tau)$ is the covariance matrix as expressed in Equation (5.10).

When $z_P^* = \tilde{z_P}$, we can find the expected length of the Bayesian prediction interval:

$$\begin{aligned} E[\tilde{z}_{P_2} - \tilde{z}_{P_1}] &= \Sigma \times [G^{-1}(P_2) - G^{-1}(P_1)] + E[\tilde{z}_{P_2} - z_{P_2}] - E[\tilde{z}_{P_1} - z_{P_1}] \\ &= \Sigma \times [G^{-1}(P_2) - G^{-1}(P_1)] + E[\tilde{z}_{P_2} - \tilde{z}_{P_1}] - E[z_{P_2} - z_{P_1}] \\ &= \Sigma \times [G^{-1}(P_2) - G^{-1}(P_1)] + \mathrm{ELPI}(z_{P_2}) - \mathrm{ELPI}(z_{P_1}) \end{aligned} \tag{7.7}$$

where ELPI = Equation(7.5) and includes S_{2G} as expressed in Equation (7.6).

7.2.1 Applications of the ELPI

As explained in Smith and Zhu, 2004, Equation (7.7) may be used as the basis for network design when choosing locations for a network designed to predict a quantity $H(Y_0)$. A Bayesian prediction interval can be constructed to minimize the coverage probability bias, and it may be possible to construct an experimental design to produce a shorter prediction interval. The benefit of using Equation (7.7) is that it accounts for both the predictive and estimation error.

CHAPTER 8

Simulation: Comparing the Laplace Approximation and Plug-In Method

In order to compare the Laplace approximation technique to the standard Plug-In approach using REML estimates, a simulation was constructed. The original motivation for this development is when a data transformation is needed to achieve a stationary process or variance stabilization. This data transformation may lead to a non-linear predictand and may be multivariate due to the need to interpolate a single predictand over multiple sites. Here we look specifically at the sum of the squares of predictions over multiple sites.

The simulation is run over N iterations, where here $N = 100$ is used. The simulation can be thought of as a double loop. The outer loop generates predictions using the REML parameter estimates in universal kriging. The inner loop uses kernel density estimation to obtain an empirical predictive distribution across the prediction sites and calculate an estimate using the Laplace approximation technique.

n_1 sites of (x, y) coordinates are generated using the uniform random function "runif" in R. Over these coordinates a random field $Y(S)$ is generated of the form

$$Y(s) = X^T(s)\beta + S(s)$$

where $X(s)$ is column vector with entries s_1, s_2 where s_1 and s_2 are the coordinates at site s. $\beta = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and $S(s)$ is a stationary Gaussian process with mean 0 and variance $\sigma^2 = 1$ (partial sill). The correlation function is parametrized by an exponential covariance

structure

$$cov\{Y(s_1), Y(s_2)\} = \sigma^2(\exp(-\frac{d_{ij}}{\phi}))$$

where $\sigma = 1.0$ and the range parameter $\phi = 0.2$. This is done using the "grf" function from the geoR package in R.

These n_1 Y values (here $n_1 = 30$) are treated as the observed values across the original n_1 sites. An additional $n_2 = 5$ sites are generated, and the corresponding simulated field values, denoted as Y_0, are treated as the true values at these sites. The objective of the simulation is to interpolate a non-linear, multivariate prediction $H(Y_0)$ across the n_2 sites. The function $H(Y_0)$ here is the sum of squares across the n_2 sites.

$$H(Y_0) = \Sigma_{i=1}^{n_2} Y_0^2$$

Of interest is the coverage probability bias, which can be directly computed using Equation (6.25). Also of interest is the empirical coverage of prediction intervals computed as $(1-\alpha)\%$ prediction intervals. Here we are interested in a standard 95% prediction interval and thus the quantiles z_P^* where $P = 0.025$ and 0.975 are needed.

The n_1 observed sites are used to obtain parameter estimates for both the one parameter case, where the range parameter ϕ from the exponential covariance structure is estimated, and the two parameter case where both the range parameter ϕ and the shape parameter σ are estimated. The one parameter case is an artificial case that is considered simply to illustrate the comparison between the plug-in and Laplace approximation methods. Parameter estimates for ϕ and σ are obtained using restricted maximum likelihood estimation (REML). (σ is assumed known and set to $\sigma = 1.0$ in the one parameter case.) The REML estimates are then "plugged-in" to the universal kriging methods to produce quantile estimates for the sum of squares across the n_2 sites for $P = 0.025$ and 0.975. The empirical prediction interval can then be tested against the true Y_0 values to determine an empirical coverage probability.

8.1 Laplace Approximation Development

In order to investigate the Laplace approximation technique, an estimation of the predictive distribution is needed. The bootstrap method described in Section (5.4.1) is employed. Recall that the inner bootstrap loop samples from the predictive distribution of $Y_0|Y$. B samples are generated from $Y_0|Y$, which is a multivariate normal distribution:

$$p(Y_0|Y) = \frac{\int G(z|Y,\theta) e^{l_n^*(Y|\theta)} \pi(\theta) d\theta}{\int e^{l_n^*(Y|\theta)} \pi(\theta) d\theta} \tag{8.1}$$

where $e^{l_n^*(Y|\theta)}$ is the restricted log likelihood. For each Y a sample of B Y_0's is obtained and the empirical cdf is computed for each sum of squares predictand across n_2 z_P^* values. For the simulation considered here, $B = 100$.

In summary

1. For $b = 1, \ldots, B$ replications generate $Y_0^{(b)} \sim N[\Lambda^T Y, V_0]$.
2. Calculate $G_B^*(z|Y,\theta) = \frac{1}{B} \sum_b I\{H(Y_0^{(b)}) \leq z\}$.

The true distribution function of $H(Y_0)$ is not easily manipulated and the forms of the first and second order derivatives with respect to both θ and z are not readily obtained. Thus in order to determine $P[H(Y_0) \leq z]$ and the respective derivatives needed for the approximation, a kernel density can be chosen. The kernel density needs to be differentiable with respect to z and with respect to the components of θ. The empirical cdf as can be estimated using the kernel density.

Here we consider the Epachenikov kernel outlined in Section (3.9).

$$K(t) = \begin{cases} \frac{3}{4\sqrt{5}} \left(1 - \frac{1}{5} t^2\right) & -\sqrt{5} \leq t \leq \sqrt{5} \\ 0 & otherwise. \end{cases}$$

where $t = \frac{z - H(Y_o^b)}{h}$.

The Epachnenikov density is used to estimate the predictive density, denoted by $G_B^{*'}(z|Y,\theta)$.

$$\begin{aligned} G_B^*(z|Y,\theta) &= \frac{1}{B}\Sigma I\{H(Y_o^b) \leq z\} \\ &\approx \frac{1}{B}\Sigma_{b=1}^{B} K_1(\frac{z - H(Y_o^b)}{h}) \end{aligned} \tag{8.2}$$

The cumulative distribution of the Epachenikov kernel is needed to approximate the predictive distribution $G_B^*(z|Y,\theta)$:

$$G_B^*(z|Y,\theta) = \begin{cases} \frac{1}{B}\Sigma_{i=1}^{B} \frac{3}{4\sqrt{5}}\left(t - \frac{1}{15}t^3\right) + 0.5 & -\sqrt{5} \leq t \leq \sqrt{5} \\ 0 & otherwise. \end{cases}$$

where $t = \frac{z - H(Y_o^b)}{h}$.

The smoothing parameter h is estimated within each outer iteration of the kriging loop, across all B iterations of the inner loop. h is selected by comparing $t = \frac{z - H(Y_o^b)}{h}$ to the support of the Epachenikov kernel $-\sqrt{(}5) \leq t \leq \sqrt{(}5)$. The simulated data is truncated at the 10^{th} and 90^{th} quantile in order to ensure that unusually small or large values of the difference $z - H(Y_o^b)$ do not produce extreme values of the bandwidth h.

The theory behind the Laplace estimation was developed in Chapter 3. In order to evaluate the Laplace approximation, partial derivatives up to order 2 are necessary. Here we express the partial derivatives as expectations with respect to the predictive distribution function, which is estimated by the kernel distribution function. For practicality, the simulated values are used in place of the theoretical expected values.

$$\frac{\partial}{\partial\theta_i}\int I\{H(Y_0) \leq z\} f(Y_0|Y,\theta) dY_0 = \mathrm{E}_{f(Y_0|Y,\theta)}\left[I\{H(Y_0) \leq z\}\frac{\partial}{\partial\theta^i}\ln f(Y_0|Y,\theta)\right]$$

where $f(Y_0|Y,\theta)$ is the restricted likelihood and $\frac{\partial}{\partial\theta^i}\ln f(Y_0|Y,\theta)$ can be analytically evaluated.

The second derivatives can be expressed:

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\int I\{H(Y_0)\leq z\}f(Y_0|Y,\theta)dY_0 =$$

$$\mathrm{E}_{f(Y_0|Y,\theta)}\left[I\{H(Y_0)\leq z\}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln f(Y_0|Y,\theta)+\frac{\partial}{\partial\theta^i}\ln f(Y_0|Y,\theta)\frac{\partial}{\partial\theta^j}\ln f(Y_0|Y\theta)\right)\right]$$

When calculating the derivative terms, $I\{H(Y_0)\leq z\}\frac{\partial}{\partial\theta^i}\ln f(Y_0|Y,\theta)$ and $I\{H(Y_0)\leq z\}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln f(Y_0|Y,\theta)+\frac{\partial}{\partial\theta^i}\ln f(Y_0|Y,\theta)\frac{\partial}{\partial\theta^j}\ln f(Y_0|Y\theta)\right)$ are empirically estimated and averaged over the B iterations. This is done using a numerical approximation for the derivatives of the restricted log-likelihood within the inner bootstrap step.

The coverage probability bias is estimated as

$$\begin{aligned} G^*(z_P^*;Y,\theta)-G^*(z_P;Y,\theta) &= -n^{-\frac{1}{2}}R(z_P,Y) \\ &+ n^{-1}\left[\frac{R(z_P,Y)R^{'}(z_P,Y)}{G^{*'}(z_P;Y,\theta)}-S(z_P,Y)\right]+o_p(n^{-1}) \end{aligned}$$

and is calculated using the expansion developed in Equation (6.25).

Also of interest is the estimator $z^\dagger$ developed as an alternative to the matching prior in Section (6.2.1):

$$z_P^\dagger = \hat{z}_P - n^{-1}\left[\frac{\text{Equation (6.25)}}{G^{*'}(G^{*-1}(P))}\right]$$

Here $z_P^\dagger$ is estimated for $P=0.025$ and 0.975 in order to form a 95% prediction interval for the sum of squares across the n_2 sites. A flat prior for the covariance parameters θ is

assumed, where here $\theta = \{\phi\}$ in the one parameter case and $\theta = \{\phi, \sigma\}$ in the two parameter case. The empirical coverage probability across the $N = 100$ simulations is computed and compared to the empirical coverage probability of the 95% prediction intervals generated using the un-adjusted $\hat{z}_P$ predictions using universal kriging with the REML estimates.

8.2 Results

This section details the results of the bootstrap simulation.

8.2.1 Empirical Coverage Probabilities: 1 Parameter Case

The 2.5% and 97.5% quantiles were computed $N = 100$ times for the plug-in estimate of the sum of squares across the $n_2 = 5$ prediction sites, using the restricted likelihood estimate for the covariance parameter ϕ. Within each iteration of the computation, the bootstrap procedure was run $B = 100$ times in order to obtain an empirical estimate of the predictive density and to calculate its respective derivatives based on the methods outlined in Chapter Four. Equation (6.26) was then used to calculate the Laplace estimators $z_P^{\dagger}$ of the quantiles in order to obtain Bayesian prediction intervals using the Laplace approximation method. The "true" simulated values were then compared to the plug-in and Laplace intervals in order to obtain an empirical coverage probability for each method.

Non-Linear Plug-In vs Laplace Prediction Intervals

The empirical 95% prediction intervals for the plug-in method result in severe under-coverage. Across $N = 100$ simulated prediction intervals for the sum of squares over $n_2 = 5$ sites, an Average Empirical Coverage (AEC) of 75.8% was found.

Using the Laplace approximation technique resulted in an improvement in the empirical coverage probabilities. The average empirical coverage for the Laplace approximation prediction intervals was 78.7%.

To further compare the empirical coverages from the plug-in and Laplace approximation methods, the empirical probabilities from the Laplace technique are plotted against the

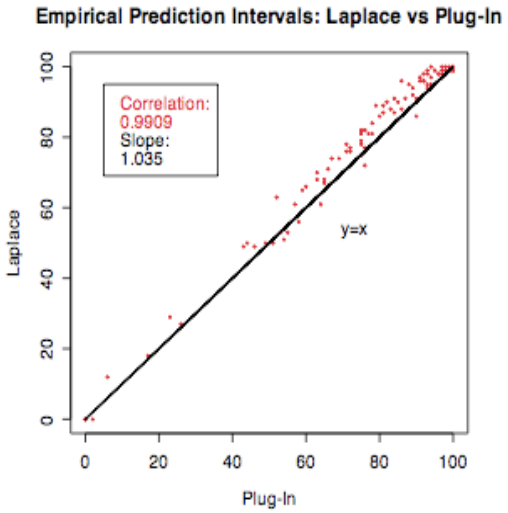

Figure 8.1: Empirical Coverage Probabilities for the Laplace versus Plug-In Methods for Non-linear Multivariate Predictand: 1 Parameter

plug-in coverages in Figure (8.1). As expected there is a strong positive correlation. A line corresponding to $y = x$ shows that the majority of the Laplace coverages are larger than the plug-in coverages.

8.2.2 Coverage Probability Bias Estimates: 1 Parameter Case

The coverage probability bias is also considered for both the plug-in prediction interval bounds and the bounds obtained using the Laplace approximation. The coverage probability bias for the plug-in method is seen to be about the same as the coverage probability bias from the Laplace approximation technique. This is shown in the plots of the empirical coverage probabilities of the plug-in predictions versus the Laplace approximations empirical coverage probabilities for both the lower and upper bounds in Figures (8.2) - (8.3). As expected there is an extremely strong correlation between the plug-in and Laplace coverage probability biases and there is little distinction between the specific values across the two estimation methods.

The plots for each of the lower and upper bound of the coverage probability biases versus the empirical coverage probabilities for both the plug-in predictions and the Laplace

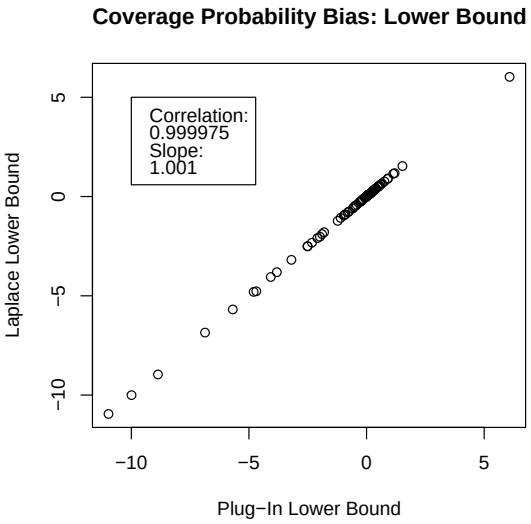

Figure 8.2: Coverage Probability Bias for the Laplace versus Plug-In Methods for 1 Parameter

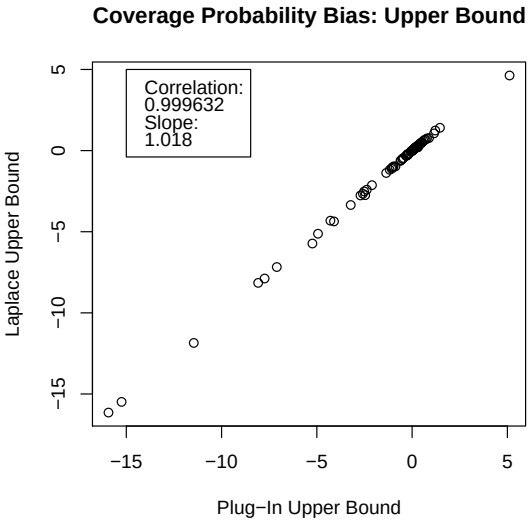

Figure 8.3: Coverage Probability Bias of the Upper Bound for the Laplace versus Plug-In Methods for 1 Parameter

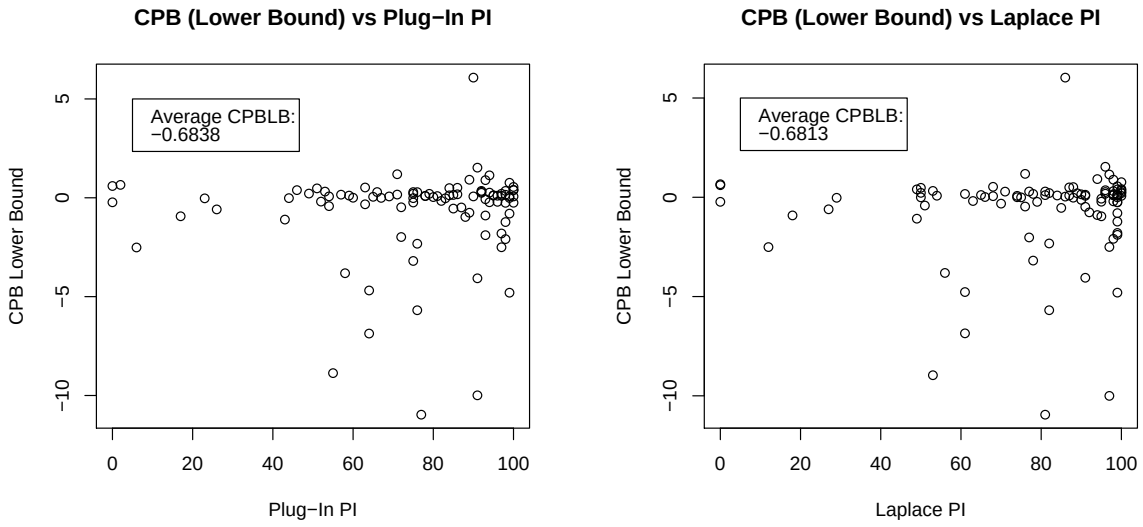

Figure 8.4: Coverage Probability Bias Comparison - Lower Bound: 1 Parameter. Left - Plug-In Method; Right - Laplace Method

predictions are shown in Figures (8.4) - (8.5). They both show an increase in magnitude and the dispersion of the coverage probability bias as the empirical coverage probability increases. The general trend that is seen is the increase in the magnitude and dispersion of the coverage probability bias as the empirical coverage probability increases for both the plug-in and Laplace methods.

In a few instances the coverage probability bias was extremely large in magnitude. These cases were removed for the purposes of calculating the average coverage probability bias and assessing an overall trend.

8.2.3 Empirical Coverage Probabilities: 2 Parameter Case

As in the one parameter case, the 2.5% and 97.5% quantiles were computed $N = 100$ times for the plug-in estimate of the sum of squares across the $n_2 = 5$ prediction sites, using the restricted likelihood estimate for the covariance parameters ϕ and σ. Within each iteration of the computation, the bootstrap procedure was run $B = 100$ times in order to obtain an empirical estimate of the predictive density and to calculate its respective derivatives based on the methods outlined in Chapter Four. Equation (6.26) was then used to calculate

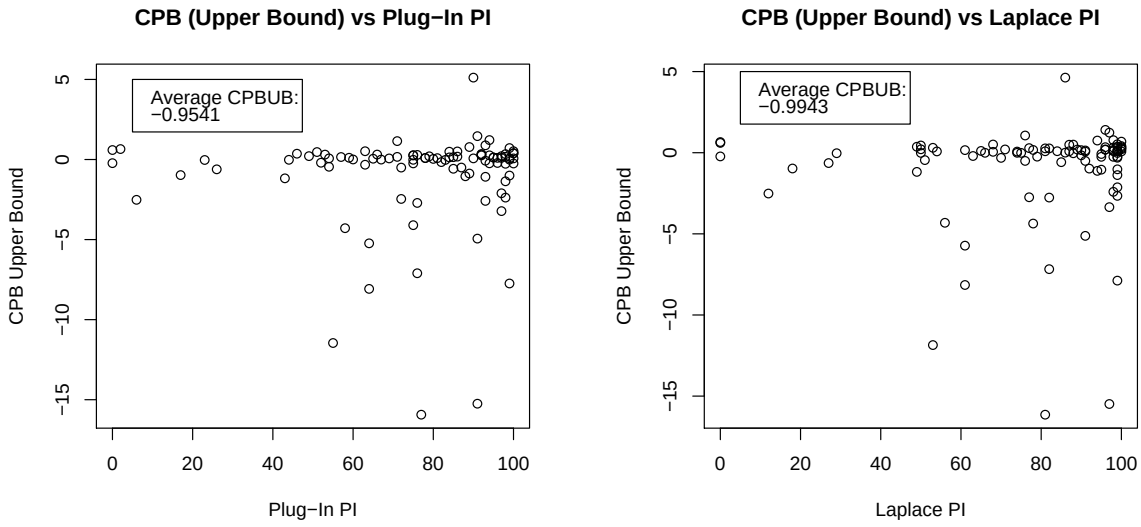

Figure 8.5: Coverage Probability Bias Comparison - Upper Bound: 1 Parameter. Left - Plug-In Method; Right - Laplace Method

the Laplace estimators $z_P^\dagger$ of the quantiles in order to obtain Bayesian prediction intervals using the Laplace approximation method. The "true" simulated values were then compared to the plug-in and Laplace intervals in order to obtain an empirical coverage probability for each method.

Non-Linear Plug-In vs Laplace Prediction Intervals

The empirical 95% prediction intervals for the plug-in method result in undercoverage. Across $N = 100$ simulated prediction intervals for the sum of squares over $n_2 = 5$ sites, an Average Empirical Coverage (AEC) of 91.91% was found.

Using the Laplace approximation technique resulted in a slight improvement in the empirical coverage probabilities. The average empirical coverage for the Laplace approximation prediction intervals was 92.22%.

To further compare the empirical coverages from the plug-in and Laplace approximation methods, the empirical probabilities from the Laplace technique are plotted against the

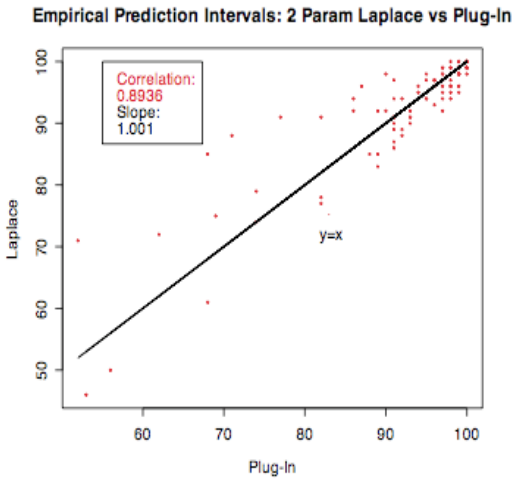

Figure 8.6: Empirical Coverage Probabilities: Laplace versus Plug-In Methods for 2 Parameters

plug-in coverages in Figure (8.6). As expected there is a strong positive correlation. A line corresponding to $y = x$ shows that the the empirical coverages of the Laplace approximation are larger than the plug-in coverages in about half of the simulated intervals.

8.2.4 Coverage Probability Bias Estimates: 2 Parameter

The coverage probability bias is also considered for both the plug-in prediction interval bounds and the bounds obtained using the Laplace approximation. As in the one parameter case, the coverage probability bias for the plug-in method is seen to be about the same as the coverage probability bias from the Laplace approximation technique. This is shown in Figures (8.7) - (8.8) of the coverage probabilities of the plug-in predictions vs the Laplace predictions for both the lower and upper bounds below. As expected there is an extremely strong correlation between the plug-in and Laplace coverage probability biases and there is little distinction between the specific values across the two estimation methods.

The plots for each of the lower and upper bound of the coverage probability biases for the Laplace predictions are shown in Figure (8.9). They both show an increase in magnitude and the dispersion of the coverage probability bias as the empirical coverage probability increases. Once again, the general trend that is seen is the increase in the magnitude and

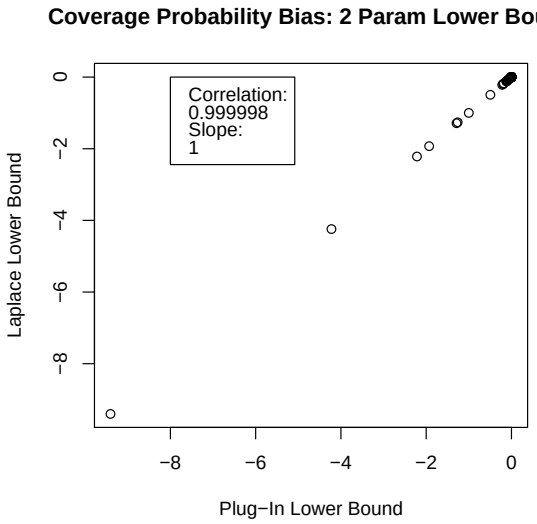

Figure 8.7: Coverage Probability Bias Comparison - Lower Bound: Laplace versus Plug-In Methods for 2 Parameters

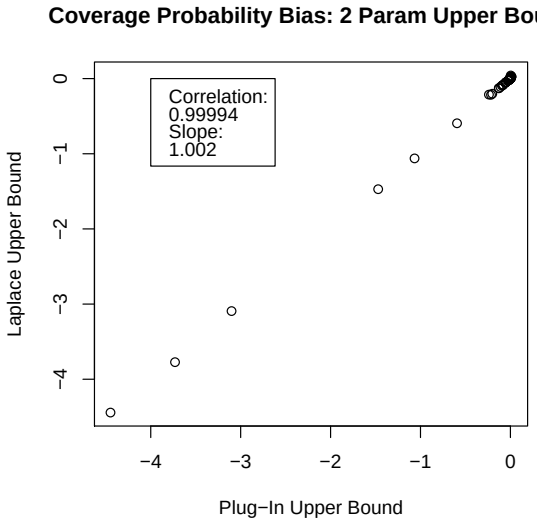

Figure 8.8: Coverage Probability Bias Comparison - Upper Bound: Laplace versus Plug-In Methods for 2 Parameters

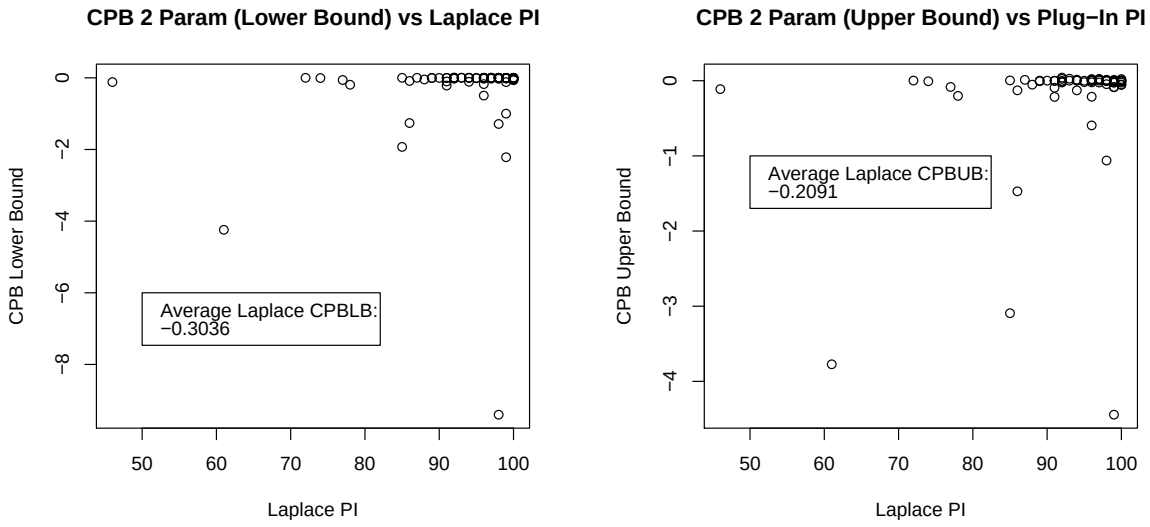

Figure 8.9: Coverage Probability Bias: Laplace Method for 2 Parameters. Left - Lower Bound; Right - Upper Bound

dispersion of the coverage probability bias as the empirical coverage probability increases for both the plug-in and Laplace methods.

In a few instances the coverage probability bias was extremely large in magnitude. These cases were removed for the purposes of calculating the average coverage probability bias and assessing an overall trend.

8.2.5 Conclusions Drawn from Simulation

This simulation shows promise for the Laplace approximation technique as an improvement over standard methods which essentially treat estimated covariance parameters as known. This simulation considered the restricted likelihood estimation of the range parameter ϕ and variance parameter σ^2. Of interest for future study is the incorporation of the restricted likelihood estimation of the exponential parameter κ through the incorporation of a powered exponential covariance structure.

Here the simulation considered flat priors for the covariance parameters θ. The investigation of various priors remains a possibility for future research. Of particular interest is

Jeffrey's prior as detailed in Berger, De Oliveira, and Sansó (2001). Of specific relevance to this paper is the possible construction of a matching prior, which would theoretically reduced the coverage probability bias to zero.

REFERENCES

Berger, J.O. (1980), *Statistical Decision Theory and Bayesian Analysis*, 2nd Ed. Springer Series in Statistics, New York.

Berger, J.O., De Oliveira, V., and Sansó, B. (2001), Objective Bayesian Analysis of Spatially Correlated Data. *Journel of the American Statistical Association.* **96** 1361 - 1374.

Carlin, Bradley P. and Louis, Thomas A. (2000), *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman and Hall/CRC Text in Statistical Science Series, New York.

Christensen, Ronald (1991), *Linear Models for Multivariate, Time Series, and Spatial Data*, Springer Series in Statistics, Springer-Verlag, New York.

Datta, Gauri Sankar and Ghosh, Malay (1995), Some Remarks on Noninformative Priors. *JASA*, **90**, **Nó 432**.

Handcock, M.S. and Stein, M. (1993), A Bayesian analysis of kriging. *Technometrics*, **35**, pp 403–410.

Harville, D.A. (1974), Bayesian nference for variance components using only error contrasts. *Biometrika* **61**, 383-385.

Higdon, D., Swall, J. and Kern, J. (1999), Non-stationary spatial modeling. In *Bayesian Statistics 6*, eds. J.M. Bernardo *et al.*, Oxford University Press, pp. 761–768.

Ibrahim, J.G. (2000), Course Notes for Bayesian Statistics. Harvard School of Public Health. pp 137-177.

Levine, Richard A. and Casella, George (2003), Implementing matching priors for frequentist inference. *Biometrika*, **90** pp 127-137.

Lindley, D.V. (1980), Approximate Bayesian methods. In *Bayesian Statistics*, edited by J.M. Bernardo, M.H. Degroot, D.V. Lindley and A.F.M. Smith. Valencia University Press, pp. 223–245.

Park, B. U., and Marron, J. S. (1990), Comparison of Data-Driven Bandwidth Selectors. *JASA* **85**, **No. 409**, pp 66-72.

Ribeiro, JR., P.J. and Diggle, P.J. (2001), geoR: A package for geostatistical analysis. *R-NEWS* **Vol 1, No 2.** ISSN 1609-3631

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, Monographs on Statistics and Applied Probability, Chapman and Hall, New York.

Smith, R.L., Kolenikov, S. and Cox, L.H. (2003), Spatio-temporal modeling of $PM_{2.5}$ data with missing values. *J. Geophys. Res.*, **108(D24)**, 9004, doi:10.1029/2002JD002914, 2003.

Smith, R.L. and Zhu, Z. (2004), Asymptotic theory for kriging with estimated parameters and its application to network design. Preliminary version, online at http://www.stat.unc.edu/postscript/rs/supp5.pdf.

Stein, Charles M. (1985), On the Coverage Probability of Confidence Sets Based on a Prior Distribution. *Sequential Methods in Statistics, Banach Center Publications*, **Vol 16**, Polish Scientific Publishers, Warsaw. pp 485-514.

Spiegelhalter, D.J., Thomas A., and Best, N.G. (1999), WinBUGS Version 1.2 User Manual. *MRC Biostatistics Unit*

Smith, R.L., (2001), Environmental Statistics, version 5.0. Department of Statistics, UNC-CH.

Smith, R.L., (1999), Bayesian and frequentist approaches to parametric predictive inference (with discussion). In *Bayesian Statistics 6*, edited by J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Oxford University Press, pp. 589-619.

Stein, M.L. (1999), *Interpolation of Spatial Data: Some Theory of Kriging.* Springer Verlag, New York.

Zimmerman, D.L. (2006), Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics.* **17** , 635-652.

Zimmerman, D.L. and Cressie, N. (1992), Mean Squared Prediction Error in the Spatial Linear Model with Estimated Covariance Parameters. *Annals of Institute of Statistical Mathematics.* **44** , 27-43.