Statistical Inferences for Correlated Observations: Prediction and Estimation

by
Xuanyao He

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.

Chapel Hill
2009

Approved by:

Richard L. Smith, Advisor

Zhengyuan Zhu, Advisor

Edward Carlstein, Committee Member

Joseph G. Ibrahim, Committee Member

Yufeng Liu, Committee Member

ABSTRACT

XUANYAO HE: Statistical Inferences for Correlated Observations: Prediction and Estimation.
(Under the direction of Richard L. Smith and Zhengyuan Zhu)

This dissertation has three major parts. Chapter 2 compares Bayesian predictive densities based on different priors and frequentist plug-in type predictive densities when the predicted variables are dependent on the observations. The performance of different inference procedures is measured, by averaging Kullback-Leibler divergence with respect to the true predictive density. The notion of second-order KL dominance is introduced, and an explicit condition is given for a prior to be second-order KL REML-dominant using asymptotic expansions. As an example, it is shown theoretically that for mixed effects models, the Bayesian predictive density with any prior from a particular improper prior family dominates the performance of REML plug-in density, while the Jeffreys prior is not always superior to the REML approach. Simulation studies are included which show good agreement with the asymptotic results for moderate sample size. Chapter 3 considers the asymptotic comparison result for both temporal and spatial AR(1) models, as an important special case of correlated data, using some theoretical results from the previous chapter. We show that all the three candidate priors, the Jeffreys prior, the reference prior and the inverse reference prior, are dominating the performance of the REML estimation for variance parameters, in the sense of the expected KL divergence between the true density and the predictive densities. Simulation results are included, which almost agree with the asymptotic result when sample size is moderately large. Chapter 4 considers estimation and prediction problems in modeling with errors in covariances. Many popular top-

ics and data-driven methods on the shrinkage estimation for covariance matrices are discussed. We also consider a model with semi-parametric covariance matrix, which includes both a parametric and an unstructured part. For this model, we derive a plug-in method for parameter estimation, and also consider how the mean and variance of a kriging predictor are affected if the true matrix V is replaced by an approximation $\hat{V}$. We consider a preliminary estimator for the unstructured error, a linear combination of the sample covariance and the diagonal estimator, and we find that in the case of exponential covariance structure (for both simulation and asymptotic results), this estimator performs better than the sample covariance matrix by comparing the mean square errors of their resulting regression coefficient estimators. In the future, we plan to derive some theoretical proofs and more simulation studies.

Keywords: Mixed effect models; Kullback-Leibler divergence; Jeffreys prior; Predictive density; Prediction fit; Autoregressive Models; Time Series; Spatial Models; Shrinkage Estimation; Covariance Matrix; Asymptotic Approximation.

ACKNOWLEDGEMENTS

I am deeply indebted to my advisors, Professor Richard L. Smith and Professor Zhengyuan Zhu. They led me into the current research field in Statistics. They helped me in learning a lot of knowledge in Statistics, for instance, the Laplace method, the statistical methods in Bayesian data analysis and Spatial Modeling. From them, I improved myself a lot in getting motivated from real problems, in formulating scientific problems, in developing novel methods and related theoretical justifications to solve those problems, etc. Thanks to their stimulating suggestions, encouragement and complete support through all my time and research at Chapel Hill.

Especially, a series of presentations given to a study group organized by Professor Smith and Professor Zhu have increased my, along with other students', knowledge base especially when I started my statistical research. Professor Smith suggested met to study KL divergence as a comparison criterion among different predictive densities. Professor Zhu suggested several asymptotic approaches and put several theoretical considerations which, once I proved, formed the basis of several important theorems and examples in this dissertation. I started to know the field of linear regression, Bayesian Data Analysis and related statistical methods in prediction procedure from two courses provided by Professor Smith, *Applied Statistics I* and *Bayesian Statistics and Generalized Linear Models*. When taking these courses, I related my research into the above fields, and made some contributions as this thesis will show. The course of *Time Series*

and Multivariate Analysis, which is taught by Prof. Smith as well, inspired me to study the asymptotic comparison of REML and Bayesian predictive procedures to the AR model, which is very important in data analysis. We find some interesting priors in the model with temporally correlated error. In developing the asymptotic comparison for different predictive densities with correlated data, Professor Zhu also gave me a lot of help in formulating the theoretical properties, and in organizing the simulation results, etc. One of his courses, *Spatial Statistics* introduced basic concepts about some important spatial models, and kriging. The essential part of the course, inspires me to consider application with respect to models with spatially correlated errors (see 3), and also to investigate the shrinkage estimation for a temporally independent, spatially dependent model (4).

I would like to express my gratitude to my committee members, Professor Edward Carlstein, Professor Yufeng Liu and Professor Joseph Ibrahim, who gave me their full support in completing this thesis. From the course I audited, *Bayesian Statistics* taught by Professor Ibrahim, I learned a lot of background knowledge and basic concepts of Bayesian statistics. Professor Carlstein raised really good questions during my proposal presentation, and I input more necessary background materials to explain the use of REML in our work, for estimation of variance parameters. In addition, Professor Carlstein is a great mentor in terms of teaching. He guided me in becoming a good teacher throughout the undergraduate courses I have taught while at Chapel Hill. Professor Liu always encourages me and gives me many inspiring suggestions for my research. The discussions I had with all of them helped me to better understand the problems I was studying and helped me make a progress.

I have taken courses from almost all the professors in this department. I want to thank them for teaching me and helping me develop my career. Especially, Professor Chuanshu Ji has given

me constant supports in many ways. I also enjoyed my four semesters as a Research Assistant, giving me opportunities to read papers, formulate problems and build my own research career. Everything is charming at Chapel Hill.

For the chapter 2, the author gratefully thanks to the reviewers of *JASA* for their helpful comments and constructive suggestions related to the second-order approximation.

Finally, I would like to give my special thanks to my husband, Dr. Lingsong Zhang, for his strong support and endless love to me, for all the time being.

Contents

List of Figures

Chapter 1

Introduction

Statistical inferences for dependent observations are in general more difficult to obtain than those for independent observations. In this dissertation, we study some inference problems for dependent observations. In particular, Chapter 2 compares Bayesian predictive densities under different priors and frequentist REML estimator-based plug-in estimative densities, based on a criteria related to average of Kullback-Leibler divergences. By asymptotic expansion, we derive explicit conditions for "second-order KL REML-dominance" (see page 24, Section 2.3.3 for the definition). For a simple mixed effect model, we show that the Jeffreys prior is not second-order KL REML-dominant, while an alternative family of improper priors is. This result indicates the improper prior is better than the Jeffreys prior, based on the criterion of averaged KL divergences. Simulation studies are also conducted, which match well to the asymptotic results for moderately large sample sizes. Chapter 3 considers the asymptotic comparison result for both temporal and spatial AR(1) models, as important cases of correlated data, using some theoretical results from Chapter 2. We show that all the three noninformative priors, the Jeffreys prior, the reference prior and the inverse reference prior, dominate the performance of the REML estimation for variance parameters, by the criterion of the expected KL divergence between the true density and the predictive densities. Simulation results are included, which

agree well with the asymptotic result, when sample size is moderately large. Chapter 4 studies the impact of different covariance matrix estimations onto parameter estimation and kriging prediction. In contrast to the structured covariance matrix in Chapter 2, a more complicated covariance matrix is assumed, that is, a semi-parametric one which includes both a structured and an unstructured measurement error part. We derive some preliminary results for the "plug-in" covariance parameter estimation, kriging predictor and prediction error. A brief summary of the three parts is given in Chapter 5.

1.1 Asymptotic Comparison of Predictive Densities for Dependent Observations

The prediction problem is significant in statistics. Suppose $Y = (Y_1, Y_2, \ldots, Y_n)$ is an observation vector from the distribution $f(\mathbf{y}; \theta)$, where θ is the parameter. Assume Z is another variable of interest, whose distribution is also parameterized by θ. Here Y and Z are dependent and we would like to predict Z based on Y. One intuitive way is to derive the predictive density of Z given Y under the Bayesian framework based on some priors. Alternatively, we can consider the plug-in MLE (Maximum likelihood estimator) or REML (Restricted Maximum Likelihood) estimative density as well. There are various ways to determine the goodness-of-fit for the predictors. In Chapter 2, we use Kullback-Leibler divergence as the loss function for the true conditional density and the predictive density. A number of authors have considered comparison of different predictive densities using the same criteria for independent observations, though little has been done for dependent observations. Instead of considering the MLE-based estimative density, we compare the restricted maximum likelihood (REML) estimator-based density (REML is often preferred to MLE, because it takes account of the loss of degrees of

freedom in estimating the mean and also produces unbiased estimating equations for the variance parameters) and Bayesian predictive densities with some objective priors. Since the KL divergence of two densities is usually intractable, we derive a higher order Laplace expansion of the KL divergence, and define the notion of "second-order KL REML-dominant" to compare predictive distributions using the Laplace approximation. A prior on θ is second-order KL REML-dominant if the REML based estimative distribution is no better than the Bayesian predictive distribution under this prior for all θ, in the sense that the leading term of the second order Laplace expansion of their differences for KL divergences is positive. We provide explicit conditions for second-order KL REML-dominance (see page 23 - 24, Section 2.3.3). In addition, for a specific mixed effect model, we show that the Jeffreys prior is not second-order KL REML-dominant, while an alternative family of improper priors is (see page 28 - 29, Section 2.4.2). To our knowledge, this is the first one of such result for the predictive distribution of Z, which is dependent on Y. Simulation studies match well to the theoretical result. This part was one of the winners of the 2008 student paper awards of American Statistical Association (ASA), Section of Bayesian Analysis, and has been submitted for publication.

1.2 Applications with respect to the AR(1) models

Chapter 3 focuses on an intensive exploration with respect to another correlation structure, by the theoretical methods in Chapter 2. Here we consider two important correlated structure: (1) Temporal AR(p) case; (2) Spatial AR(p) case. The p-th order autoregressive (AR(p)) model is one of the most important models in time series analysis. It consists of the data $\{y_t\}$, satisfying $y_t = \sum_{i=1}^{p} a_i y_{t-i} + u_t$, where $\{u_t\}$ is a white noise with mean 0 and variance σ^2. Assuming the stationarity, we focus on both the Bayesian estimation and frequentist estimative method for

the predictive density of the AR(1) model with unknown σ^2 and ρ (which is a_1 when $p = 1$), utilizing the criterion of expected K-L divergence, which we propose in Chapter 2. We also point out that the reference prior is superior to the other two candidate ones with respect to the second-order asymptotic approximation. A related work by Tanaka and Komaki (2005) focused on the Bayesian estimation of the spectral density of the AR(2) model and proposed a superharmonic prior as a noninformative prior. They also considered the more general case, the autoregressive moving average (ARMA) model, focusing on the Bayesian estimation of an unknown spectral density in the ARMA model. They first showed that in the i.i.d. cases, the Bayesian spectral densities based on a superharmonic prior asymptotically dominate those based on the Jeffreys prior, using the asymptotic expansion of the risk difference related to expectation of KL divergence. Actually the stationary Gaussian processes are getting close to the i.i.d. cases as the sample size becomes large, and they obtained the asymptotic expansion of the Bayesian spectral density for the ARMA model, which could be written in the differential-geometrical quantities as in the i.i.d. cases. Finally they obtained the corresponding result in the ARMA model. However, our work directly compares different predictive densities instead of the spectral densities, for the case of AR(1) model.

Meanwhile, models for two-dimensional spatial data where the errors follow a spatial ARMA process have been considered by several authors, e.g. Martin (1990), Cullis and Gleeson (1991), Zimmerman and Harville (1991) and Basu and Reinsel (1994). No past work has considered the prior for this kind of AR(p) spatial model. As a special case, we will explore the AR(1) spatial model for easy presentation, we consider the noninformative priors and the REML estimative density for the model with noise from a spatial multiplicative AR(1) model (see Basu and Reinsel (1993) and Martin (1990)), with fixed σ^2.

General definition of the AR(1) model and the necessary notations for Fisher Information Matrix calculation are briefly reviewed in this chapter. Within this framework, we compare different Bayesian predictive densities and REML plug-in density, based on the expected KL divergence, as proposed in Chapter 2. We provide the asymptotic second-order expression of the differences between their expected KL divergences. In Section 3.3.3 and 3.4.3, we apply this approach to AR(1) models, for both time series and spatial structure, respectively. We consider three candidate priors: the Jeffreys prior, the reference prior and the inverse reference prior. Asymptotically, all the three priors perform quite well as compared to the REML-plug in density. In particular, we prove that the reference prior dominates the other two priors. Berger and Yang (1994) also recommended the reference prior, when comparing it with the Jeffereys prior and the uniform prior (which results in MLE estimator) based on another criterion, the MSE of the resulting estimator. In Section 3.3.4 and 3.4.4, we perform numerical simulation for the AR(1) time series and spatial process, illustrating that the asymptotic results hold, when the sample size is moderately large. Some concluding work can be seen in Section 3.5.

1.3 Parameter Estimation and Prediction with Errors in Covariances

A typical linear mixed effect model is $Y = XB + E$, where Y is the observation vector or matrix, X is the design matrix, B is the regression coefficient matrix and E is the unobserved random errors. In Chapter 2, we compare the Bayesian predictive densities with the REML plug-in estimative density for this model, assuming that the covariance-variance matrix, $V(\theta)$, is an known function of the unknown parameter θ. While in most practical cases, the true covariance matrix is unknown. Therefore estimation of the covariance matrix is necessary,

in order to obtain related statistical inferences. In Chapter 4, we assume that the covariance matrix of given data contains two parts: some known function of the unknown parameter θ, and the measurement error R, which is not parametrically constrained. We consider two problems about this model:

(1) Plug-in estimators of θ, by using sample covariance S or a liner shrinkage estimator $\hat{R}_\nu = \nu S + (1-\nu)S^*, \nu \in [0,1]$, where S^* is the diagonal estimator of R, in the restricted log-likelihood function, where ν is an adjustable tuning parameter that represents shrinkage intensity.

(2) How the kriging performance will be affected, when the true covariance matrix is K_{tt} but we replace it by $\hat{K}_{tt} = K_{tt}(\hat{\theta}, \hat{R})$ (see page 67 in Section 4.1 for definition of K) to derive the likelihood equations.

For problem (1), we suppose θ is a vector and θ^i is the ith element, and we can show that the estimation bias, $\hat{\theta}^i - \theta^i$ depends on the first and the second order moments of the first order derivative of the plug-in restricted log-likelihood function and the first order moment of the second order derivative, where "plug-in" means replacing R by $\hat{R}$ whenever R appears in the definition of these functions or derivatives. Obviously this needs specification for estimation methods of the covariance matrix. As a starting point, we consider a model with an exponential covariance function K and R simultaneously, where R is set to be random, exponential or long range dependence, respectively (see page 78 - 80 in Section 4.3.2). From both theoretical and simulation results, we find that certain linear combination of the sample covariance and the diagonal estimator will result in much smaller mean squared error (MSE) of the plug-in REML estimator $\hat{\theta}$, given the tuning parameter ν^*, which is the optimal value of $\nu \in (0,1)$. We will investigate if there is a unique and explicit way to express this ν^*, in terms of K, R, S and S^*.

For problem (2), we are making extension on the Mean Squared Prediction Error (MSPE)

of empirical kriging predictor. In the future, we plan to get the asymptotic expression of the MSPE, similar to what we have done for problem (1). The next step is to simulate some data and check the effect of different estimators of R on the empirical MSPE.

1.4 Summary

The rest of the dissertation is arranged as following. Chapter 2 compares Bayesian predictive densities with different priors and frequentist REML-based plug-in estimative densities, by averaging the Kullback-Leibler divergences. Chapter 3 applies the theoretical method from Chapter 2, to both temporal and spatial AR(1) models, as important cases of correlated data. We showed that all the three noninformative priors, the Jeffreys prior, the reference prior and the inverse reference prior, dominate the performance of the REML estimation for variance parameters, in the sense of the expected KL divergence between the true density and the predictive densities. The reference prior dominates the other two in terms of expected KL divergence between the true density and the predictive density. Simulation results agree with the asymptotic case when the sample size is not too small. Chapter 4 discusses parameter estimation and kriging prediction problem in spatial modeling with errors in covariances. Further comments and discussions are summarized in Chapter 5.

Chapter 2

Asymptotic Comparison of Predictive Densities for Dependent Observations

This chapter studies Bayesian predictive densities based on different priors and frequentist plug-in type predictive densities when the predicted variables are dependent on the observations. Average Kullback-Leibler divergence to the true predictive density is used to measure the performance of different inference procedures. The notion of second-order KL dominance is introduced, and an explicit condition for a prior to be second-order KL dominant is given using an asymptotic expansion. As an example, we show theoretically that for mixed effects models, the Bayesian predictive density with prior from a particular improper prior family dominates the performance of REML plug-in density, while the Jeffreys prior is not always superior to the REML approach. Simulation studies are included which show good agreement with the asymptotic results for moderate sample size.

2.1 Introduction

Prediction is of great importance in statistics. The general prediction problem can be described as follows. Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be the observation from the distribution $f(\mathbf{y}; \theta)$, where θ

is the parameter, and Z be another random variable with distribution also parameterized by θ. Y and Z may be dependent for time series or spatial data, and we would like to predict Z based on observation Y. In principle, we would like to know the distribution of Z conditional on Y. This is usually characterized by a point predictor and a prediction interval in the frequentist framework, and good prediction means both an accurate point predictor and a narrow prediction interval with correct coverage probability. Alternatively, one can take a decision-theoretic approach and define a loss function between the true conditional density $g(z|y,\theta)$ and the predictive density $\hat{g}(z|y)$. A common measure of discrepancy between two density functions g and $\hat{g}$ is the Kullback-Leibler (KL) divergence:

$$D(g(z|y,\theta),\hat{g}(z|y)) = \int g(z|y,\theta)\log\tfrac{g(z|y,\theta)}{\hat{g}(z|y)}dz.$$

To compare two predictive densities $\tilde{g}_1$ and $\tilde{g}_2$, one can look at the expected difference of KL divergence

$$\int\{D(g,\tilde{g}_1) - D(g,\tilde{g}_2)\}f(y;\theta)dy. \tag{2.1}$$

Numerous authors have considered the comparison of different prediction methods using (2.1) when Z is independent of Y. In terms of this criterion, Aitchison (1975) claimed that in general the Bayesian predictive density based on a vague prior is better than the predictive density $g(Z;\hat{\theta})$, where $\hat{\theta}$ is the maximum likelihood estimate (MLE) (Aitchison refers to it as the estimative density), and proved that Bayesian predictive density is better in the sense of (2.1) for all θ for $g(y;\theta)$ in the gamma and multivariate normal families. These are apparently the only two known cases when it is proven that such claim holds exactly, though it is widely

suspected that the same is true for a much wider class of distributions. Murray (1977) gave a stronger result in his note, using the information measure together with the idea of invariance to derive the vague predictive density as the optimum in a wide class of possible estimates of $p(z|\theta)$. Hartigan (1998) showed that, frequently, in more than one dimension the maximum likelihood estimate plug-in density is asymptotically inadmissible and may be improved upon by using the predictive density corresponding to a least favorable prior. He also provides solutions (if they exist) to certain differential equations as the answer to admissibility questions for the "near ML" estimates. Komaki (1996) considered optimal adjustments of estimative to predictive estimators for exponential families, and confirms Aitchison's conjecture from the viewpoint of asymptotic theory. The general form of the average Kullback-Leibler divergence from the true distribution to a predictive distribution, and the asymptotic expression for Bayesian predictive distributions are obtained under the assumption that Z independent of Y. Adopting Kullback-Leibler divergence as a loss function, Komaki (2006) had also provided some kinds of priors that dominate the Bayesian predictive distributions based on the Jeffreys prior under some differential geometric conditions. George et al. (2006) introduced the Bayesian predictive densities under superharmonic priors that dominate the Bayes predictive density under the uniform prior, for independent p-dimensional multivariate normal vectors. Ren et al. (2006) investigated the estimation and prediction for exponential distributions with unknown rate parameter θ and compared the performances of MLE with Bayesian estimates under several loss functions. They developed second-order asymptotic expressions for the Bayes estimates under these loss functions, one of which is KL divergence. They also assumed that Z is independent of Y.

2.1.1 Background for Restricted log-likelihood estimation Method

Consider a general linear mixed model

$$Y = X\beta + Zb + \epsilon,$$

where X and Z are specified design matrices, β is a vector of fixed effect coefficients, b and ϵ are random, mean zero, and Gaussian if necessary. Usually we can think of b being constant over subjects, ϵ as independent between subjects, possibly correlated within subjects. Let θ denote free parameters in the variance specification.

We observe n r.v.s Y. Once the structure of errors is fully specified, and $\mathrm{Cov}(b, \epsilon) = 0$, we have

$$Y \sim N(X\beta, V(\theta)) \tag{2.2}$$

$$V(\theta) = \mathrm{Var}(\epsilon) + Z\mathrm{Var}(b)Z^T. \tag{2.3}$$

For mixed models, the covariance matrix V is a function of a q-dimensional parameter θ, and is assumed to be positive definite for θ in a neighborhood of the true value. Any estimation or prediction procedure proceeds only if variances and covariances among the observations are known or, more often, after they have been estimated.

One of the recommended methods for estimating variances and covariances is the Restricted Maximum Likelihood (REML) method, which has been used and studied over the past 40 years. In essence, the REML method deals with linear combinations of the observed values whose expectations are zero. These "error contrasts" are free of any fixed effects in the model. In contrast to the maximum likelihood estimator (MLE) of the variance components (which can

be biased downwards in a linear model), REML corrects this problem by using the likelihood of a set of residual contrasts and is generally considered superior. In other words, REML is often preferred to maximum likelihood estimation as a method of estimating covariance parameters in linear models because it takes account of the loss of degrees of freedom in estimating the mean and produces unbiased estimating equations for the variance parameters.

The restricted or residual maximum likelihood (REML) method was proposed by Thompson (1962), as a way of estimating dispersion parameters associated with linear models. Numerous authors have given overviews on REML, e.g.Patterson and Thompson (1971) introduced restricted maximum likelihood estimation (REML) as a method of estimating variance components in the context of unbalanced incomplete block designs. Surveys of REML can be found in articles of Harville (1977), Khuri and Sahai (1985), and Robinson (1987), and in the book by Searle et al. (1992). Alternative and more general derivations of REML were given by Harville (1974), Cooper and Thompson (1977) and Verbyla (1990). In all of these the restricted likelihood is presented as the marginal likelihood of the error contrasts. Or it can be regarded as modified profile likelihood in Barndorff-Nielsenn (1983). Some other areas in which REML has been used include the following: estimating smoothing parameters in penalized estimation Wahba (1990) (see Speed (1991) for related discussions); the estimation of parameters in ARMA processes and other time series in the presence of fixed effects (Cooper and Thompson (1977) and Azzalini (1984)); REML estimation in spatial models (Green (1985) and Gleeson and Cullis (1987)); or the analysis of longitudinal data in Laird and Ware (1982); and REML estimation in empirical Bayes smoothing of the census undercount in Cressie (1992). Therefore, in our work we also consider REML as one candidate method to estimate the variance parameters.

Definition of Log-Likelihood

From equation (2.2) and (2.3), we get the minus twice the log-likelihood is

$$(Y - X\beta)^T V^{-1}(\theta)(Y - X\beta) + \log|V(\theta)|$$

Thus, we can get the MLE of θ (denoted by $\hat{\theta}_{MLE}$) by minimizing the above equation. Furthermore, given $\hat{\theta}$ we can find the MLE of β by generalized least squares.

Definition of Restricted maximum likelihood: (REML)

Suppose that we can find some linear combinations AY whose distribution does not depend on β. In fact we can find up to $n - p$ linearly independent such. One choice is any $n - p$ of the least-squares residuals of the regression of Y on X. In REML we treat AY as the data and use maximum-likelihood estimation of θ (the parameters in V).

The REML estimates do not depend on the choice of A, so this procedure is not as arbitrary as it sounds. Indeed, the REML estimates minimize

$$(Y - X\beta)^T V^{-1}(\theta)(Y - X\beta) + \log|V(\theta)| + \log|X^T V^{-1}(\theta)X|$$

Indeed, the REML fit criterion is the marginal likelihood, integrating β out with a vague prior.

2.1.2 Correlated Data

It is worth noting that in many practical problems the data are correlated. Notable examples include data collected from time series or spatial models, and data that can be modeled by mixed effect models. In particular, a number of authors have considered Bayesian analysis of spatially correlated data. Berger et al. (2001) assumed the spatial correlation structure of the

model is specified with a small number of unknown parameters, and the mean function of the model is a linear function of some unknown covariates. To construct Bayesian analysis of such spatial models, they needed to determine an objective prior distribution for the unknown mean and covariance parameters of the random field. The proposed reference prior for the model was shown to result in a proper posterior. It was also compared with the commonly used Jeffreys prior in terms of the ability to produce confidence sets with good frequentist coverage, indicating that the Jeffreys-rule prior can be quite inadequate. Inspired by this fact and that Jeffreys (1961a) argued for use of the independence Jeffreys prior in problems that involve both linear and covariance parameters, we used the independence Jeffreys prior as one choice of the default prior distributions in this chapter. The model of correlated data can be used in different areas of spatial statistics such as disease mapping (Ferreira and Oliveira (2007)) or image analysis (Besag et al. (1991)). It has also been used for highly structured stochastic models such as spatio-temporal models (Sun et al. (2000)). It can also be used in the CAR model framework, such as in (MacNab (2003)).

In this chapter, we consider the comparison of predictive density using (2.1) when Z and Y are dependent. Instead of considering the MLE-based estimative density, we compare the restricted maximum likelihood (REML) estimator-based estimative density and Bayesian predictive densities with some objective priors. One reason for using the KL divergence is that it has historically been the principal device for developing noninformative priors (Hartigan, 1998). Since (2.1) is in general intractable, we use a higher order Laplace expansion to approximate it, and introduce the notion of "second-order KL REML-dominant" to compare predictive distributions using the Laplace approximation. We focus on comparisons between Bayesian predictive densities and REML-based estimative density, although our result can be used to

compare different class of priors as well. In particular, we call a prior on θ as a second-order KL REML-dominant if the corresponding predictive distribution under this prior is better than the REML based estimative distribution for all θ in the sense that the leading term of the second order Laplace expansion of (2.1) is greater than or equal to zero. We derive explicit conditions for second-order KL REML-dominance, and show for one specific mixed effect model that the Jeffreys prior is not second-order KL REML-dominant, while an alternative family of improper prior is. To our knowledge this is the first of such result for the predictive distribution of Z dependent on Y. Simulation studies are conducted which show good match to the asymptotic results for moderately large sample size.

The chapter is organized as follows. Section 2.2 provides notation and the model under which we derive our results. Section 2.3 derives the Laplace expansion of the expected difference between the KL divergence of the REML estimative density and the Bayesian predictive density, and gives explicit conditions for a prior to be second-order KL REML-dominant. Section 2.4 applies the results to a specific mixed effect model to show that Jeffreys prior is not second-order KL REML-dominant, while a family of improper priors is. Simulation results are also included. We conclude in Section 2.5 with some discussion and future work.

2.2 Notation and Preliminaries

Let $Y = (Y_1, ..., Y_n)^T$ be an arbitrary $n \times 1$ observation vector. We are interested in predicting some unobserved value Z, which is dependent on Y. In this chapter we only consider univariate Z for simplicity. We assume that Y and Z have joint Gaussian distribution of the form

$$\begin{pmatrix} Y \\ Z \end{pmatrix} \sim N\left[\begin{pmatrix} X\eta \\ x_0\eta \end{pmatrix}, \begin{pmatrix} V(\theta) & w^T(\theta) \\ w(\theta) & v_0(\theta) \end{pmatrix}\right], \tag{2.4}$$

where X and x_0 are assumed to be known matrices of regressors of dimensions $n \times q$ and $1 \times q$ respectively, and η is an unknown $q \times 1$ vector of regression coefficients. Special cases of (2.4) include various linear mixed effect models, and spatial linear models where the errors are assumed to be a realization from a Gaussian random field (GRF). We further assume that the covariance elements $V(\theta), w(\theta)$ and $v(\theta)$ are all known functions of an unknown p -dimensional parameter vector θ. In Bayesian analysis we denote the prior density function by $\pi(\theta)$. To simplify the notation, we write V, w and v without indicating the dependence on θ.

The restricted log likelihood function of Y is given by

$$\ell_n(\theta) = -\tfrac{n-q}{2}(\log 2 + \log \pi) + \tfrac{1}{2}\log|X^TX| - \tfrac{1}{2}\log|V| - \tfrac{1}{2}\log|X^TV^{-1}X| - \tfrac{G^2}{2}, \tag{2.5}$$

where G is the generalized residual sum of squares given by

$$G^2 = G^2(\theta) = Y^T\{V^{-1} - V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}\}Y. \tag{2.6}$$

See for example, Stein (1999) for more discussion on REML and its advantage over regular maximum likelihood (MLE) estimators for covariance parameters.

If θ is known, then the Best Linear Unbiased Predictor (BLUP) of z is given by $\hat{z} = \lambda^T Y$, where

$$\lambda = V^{-1}w^T + V^{-1}X(X^TV^{-1}X)^{-1}(x_0 - X^TV^{-1}w^T), \tag{2.7}$$

and the corresponding prediction error is given by

$$\sigma_0^2 = v_0 - wV^{-1}w^T + (x_0^T - wV^{-1}X)(X^TV^{-1}X)^{-1}(x_0 - X^TV^{-1}w^T). \tag{2.8}$$

Equation (2.7) and (2.8) are also known as the universal kriging formula in geostatistics where Y are observations from a GRF.

If we assume θ is known and η unknown with a uniform prior density, the Bayes rule under least squares loss coincides with the Best Linear Unbiased Predictor (BLUP). In this simple case, the predictive distributions for the frequentist and Bayesian inference are the same, since both reduce to

$$\psi(z; Y, \theta) = \Phi(\tfrac{z-\lambda^T Y}{\sigma_0}), \tag{2.9}$$

where Φ is the standard normal distribution function. This is also the conditional distribution of Z given Y in a Bayesian framework.

In this work, we study a more complicated case, where θ is assumed to be unknown. Furthermore, under the Bayesian framework, we assume that θ has a prior density $\pi(\theta)$ which is differentiable over a region that includes the true value of θ, while the prior of η continues to be assumed uniform (independent of θ). The posterior predictive distribution function of z given Y is given by

$$\widetilde{\psi}(z; Y) = \tfrac{\int e^{l_n(\theta)+Q(\theta)}\psi(z;Y,\theta)d\theta}{\int e^{l_n(\theta)+Q(\theta)}d\theta}, \tag{2.10}$$

with $Q(\theta) = \log \pi(\theta)$.

In contrast to (2.10), we also consider the estimative distribution under the frequentist framework

$$\widehat{\psi}(z; Y) = \psi(z; Y, \widehat{\theta}), \tag{2.11}$$

where $\widehat{\theta}$ is the REML estimator to maximize $\ell_n(\theta)$.

In this chapter (and also the rest of the dissertation) we use superscripts to denote components of θ, e.g. θ^i for the ith component. For scalar functions of θ, such as $\psi(z; Y, \theta)$ (with z and Y fixed for the time being) or $Q(\theta)$, subscripts will indicate derivative with respect to the components of θ, i.e. $Q_i = \frac{\partial Q}{\partial \theta^i}, \psi_{ij} = \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}$, etc. Also we define $U_i = \frac{\partial \ell_n(\theta)}{\partial \theta^i}, U_{ij} = \frac{\partial^2 \ell_n(\theta)}{\partial \theta^i \partial \theta^j}, U_{ijk} = \frac{\partial^3 \ell_n(\theta)}{\partial \theta^i \partial \theta^j \partial \theta^k}$.

All the above quantities are functions of a particular θ, and could be evaluated at the REML estimator $\hat{\theta}$, which we denote with a hat, such as $\hat{U}_i, \hat{\psi}_{ij}$, and so on. By definition, we have $\{\hat{U}_i\} = 0$, and $\{-\hat{U}_{ij}\}$ is the observed information matrix. If the latter matrix is invertible, we denote its inverse matrix with superscripts, i.e., if A is the $p \times p$ matrix with (i, j)-th entry as $\hat{U}_{ij}$, then if A^{-1} exist, $\hat{U}^{ij}$ is its (i, j)-th entry.

We also use the summation convention, where a repeated index appearing as both a subscript and a superscript in the same formula implicitly indicates a summation over that index. For simplicity, it is not explicitly indicated that all the quantities depend on dimension n. The following assumptions are made for the rest of the chapter.

Assumption 1: $\ell_n(\theta)$ and all of its derivatives are of $O_p(n)$. Expectations of them are of $O(n)$.

Assumption 2: Q and ψ and their derivatives are of $O_p(1)$. Their expectations are of $O(1)$.

Assumption 3: $\{\hat{U}_{ij}\}$ is invertible.

Assumption 1 and 2 are made for consistency with regular maximum likelihood theory for i.i.d. observations, and are satisfied by many linear mixed effect models. For spatial linear models, these two assumptions imply that we are working with the framework of "increasing domain asymptotics", given by Mardia and Marshall (1984), instead of the alternative "infill

asymptotics" by Stein (1999).

With the above notations, and also applications of formulae (8.3.50) - (8.3.55) in Chapter 8 of Bleistein and Handelsman (1986), to the numerator and denominator of (7), we get

$$\tilde{\psi} - \hat{\psi} = \tfrac{1}{2}\hat{U}_{ijk}\hat{\psi}_\ell \hat{U}^{ij}\hat{U}^{k\ell} - \tfrac{1}{2}(\hat{\psi}_{ij} + 2\hat{\psi}_i\hat{Q}_j)\hat{U}^{ij} + O_p(n^{-2}). \tag{2.12}$$

We can write $U_i = n^{1/2}Z_i, U_{ij} = n\kappa_{ij} + n^{1/2}Z_{ij}, U_{ijk} = n\kappa_{ijk} + n^{1/2}Z_{ijk}$, where $\kappa_{ij}, \kappa_{ijk}$ are non-random and Z_i, Z_{ij}, Z_{ijk} are random variables with mean 0, and we assume that all these quantities are of $O(1)$ or $O_p(1)$ as $n \to \infty$.

Also, let $\kappa_{i,j} = E(Z_iZ_j) = -\kappa_{ij}, \kappa_{ij,k} = E(Z_{ij}Z_k)$. By standard identity, $\kappa_{i,j}$ is the (i,j)-th entry of the normalized Fisher information matrix, we assume this matrix to be invertible with inverse entries $\kappa^{i,j}$. Explicit formulae exist for calculating these quantities, which can be found in Section 8.2 of Smith and Zhu (2004).

In this notation and by a standard Taylor expansion of ℓ_n, we get the approximation

$$\begin{aligned}\hat{\psi} &= \psi + n^{-1/2}\kappa^{i,j}Z_i\psi_j + n^{-1}(\kappa^{i,j}\kappa^{k,\ell}Z_{ik}Z_j\psi_\ell + \tfrac{1}{2}\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk}Z_rZ_s\psi_t + \tfrac{1}{2}\kappa^{i,j}\kappa^{k,\ell}Z_iZ_k\psi_{j\ell}) \\ &\quad + O_p(n^{-3/2}),\end{aligned} \tag{2.13}$$

and by (2.12) we have

$$\tilde{\psi} = \hat{\psi} + \tfrac{1}{2}n^{-1}\{\kappa_{ijk}\kappa^{i,j}\kappa^{k,\ell}\psi_\ell + (\psi_{ij} + 2\psi_iQ_j)\kappa^{i,j}\} + O_p(n^{-3/2}). \tag{2.14}$$

2.3 Asymptotic Expression of KL divergences

2.3.1 Kullback-Leibler divergence

Suppose we have some observations, which can be modeled by (2.4), and we want to predict Z given Y. In this work we will compare predictive and estimative densities using Kullback-Leibler divergence (KL divergence) as the criterion. Let $\psi(z;Y,\theta)$ be the cumulative distribution function (CDF) of z given Y and θ, then the probability density function (PDF) of z given Y is of the form $\varphi(z;Y,\theta)=\frac{\partial\psi}{\partial z}$, and similarly, for any predictive distribution function $\psi^*(z;Y)$, we let $\varphi^*(z;Y)=\frac{\partial\psi^*}{\partial z}$ be the predictive density function. The KL divergence from $\varphi^*(z;Y)$ to $\varphi(z;Y,\theta)$ (simply written as φ below) is given by

$$D(\varphi(z;Y,\theta),\varphi^*(z;Y))=\int\varphi(z;Y,\theta)\log\tfrac{\varphi(z;Y,\theta)}{\varphi^*(z;Y)}dz \tag{2.15}$$

We say the predictive density function $\varphi^{*(1)}$ is KL dominated by $\varphi^{*(2)}$ if for all $\theta\in R^p$,

$$\begin{aligned}&E_{Y|\theta}[D(\varphi(z;Y,\theta),\varphi^{*(1)})-D(\varphi(z;Y,\theta),\varphi^{*(2)})]\\&=\int[D(\varphi(z;Y,\theta),\varphi^{*(1)})-D(\varphi(z;Y,\theta),\varphi^{*(2)})]\varphi(Y;\theta)dY\geq 0.\end{aligned} \tag{2.16}$$

Equation (2.16) can be used to compare two prediction procedures or two priors under the Bayesian framework. In particular, we can compare the estimative and predictive procedure by taking $\widehat{\varphi}(z;Y)$ as $\varphi^{*(1)}$, and $\widetilde{\varphi}(z;Y)$ as $\varphi^{*(2)}$, respectively. It will be interesting if in the Bayesian framework, there exists some prior such that the Bayesian predictive density function is superior to the REML estimative density function in terms of the KL measure for all θ, and we call such prior "KL REML-dominant prior". It is in general difficult to prove exact KL

REML-dominance. In what follows, we derive the Laplace expansion of (2.16), and use the leading terms to define second-order KL REML-dominance.

2.3.2 Asymptotic approximation to the KL divergences and their difference

Using Taylor expansion of logarithm, we can approximate the Kullback-Leibler divergence from $\tilde{\varphi}$ to φ by

$$\begin{aligned} D(\varphi, \tilde{\varphi}) &= -\int \log(\tfrac{\tilde{\varphi}}{\varphi})\varphi dz = -\int \log(\tfrac{\tilde{\varphi}-\varphi}{\varphi}+1)\varphi dz \\ &= -\int [(\tfrac{\tilde{\varphi}-\varphi}{\varphi}) - \tfrac{1}{2}(\tfrac{\tilde{\varphi}-\varphi}{\varphi})^2 + o(n^{-2}))\varphi dz \\ &= \tfrac{1}{2}\int \tfrac{(\tilde{\varphi}-\varphi)^2}{\varphi} dz + o(n^{-2}). \end{aligned} \tag{2.17}$$

Similarly, we can get $D(\varphi, \hat{\varphi}) = \frac{1}{2}\int \frac{(\hat{\varphi}-\varphi)^2}{\varphi} dz + o(n^{-2})$, and

$$D(\varphi, \hat{\varphi}) - D(\varphi, \tilde{\varphi}) = \tfrac{1}{2}\int \tfrac{(\hat{\varphi}-\varphi)^2 - (\tilde{\varphi}-\varphi)^2}{\varphi} dz + o(n^{-2}). \tag{2.18}$$

The above quantity could be expressed explicitly using (2.13) and (2.14). Let

$$\psi^*(z;Y) = \psi(z;Y,\theta) + n^{-1/2}R(z,Y) + n^{-1}S(z,Y) + n^{-3/2}T(z,Y) + O_p(n^{-2}),$$

where ψ^* denotes the predictive distribution function of Z given Y, and could be either $\widehat{\psi}$ or $\widetilde{\psi}$. We have

$$\varphi^*(z;Y) = \varphi(z;Y,\theta) + n^{-1/2}R'(z,Y) + n^{-1}S'(z,Y) + n^{-3/2}T'(z,y) + O_p(n^{-2}),$$

where φ^* denotes the predictive density function of Z given Y, and could be either $\widehat{\varphi}$ or $\widetilde{\varphi}$. We keep the notation of R since it is identical to both $\widehat{\varphi}$ and $\widetilde{\varphi}$. For $\widehat{\varphi}$, we denote the resulting S and T by S_1 and T_1, respectively. For $\widetilde{\varphi}$, we denote the corresponding S and T by S_2 and T_2.

By (2.13) and (2.14) and their derivatives, we get

$$
\begin{aligned}
R^{'} &= \tfrac{\partial R}{\partial z} = \kappa^{i,j} Z_i \varphi_j, \\
S_1' &= \tfrac{\partial S_1}{\partial z} = \kappa^{i,j}\kappa^{k,l} Z_{ik} Z_j \varphi_l + \tfrac{1}{2}\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk} Z_r Z_s \varphi_t + \tfrac{1}{2}\kappa^{i,j}\kappa^{k,l} Z_i Z_k \varphi_{jl} \text{ for } \hat{\psi}, \\
S_2' &= \tfrac{\partial S_2}{\partial z} = S_1' + \tfrac{1}{2}\kappa^{i,j}\kappa^{k,l}\kappa_{ijk}\varphi_l + (\tfrac{1}{2}\varphi_{ij} + \varphi_i Q_j)\kappa^{i,j} \text{ for } \tilde{\psi}, \\
T_2' &= T_1' + \kappa^{i,j}\kappa^{k,l}(Z_{ijk} + \kappa^{t,r} Z_r \kappa_{ijkt}) + \kappa^{i,j}\kappa^{k,l}\kappa_{ijk}\varphi_{ls}\kappa^{s,r} Z_r \\
&\quad + \kappa_{ijk}\kappa^{k,l}\varphi_l(\kappa^{p,i} Z_{pq}\kappa^{q,j} + \kappa^{p,q} Z_q \kappa^{i,a}\kappa_{abp}\kappa^{b,j}) \\
&\quad + \kappa_{ijk}\kappa^{i,j}\varphi_l(\kappa^{c,k} Z_{cd}\kappa^{d,l} + \kappa^{c,d} Z_d \kappa^{k,e}\kappa_{efc}\kappa^{f,l}) \\
&\quad + \kappa^{i,j}(\kappa^{t,r} Z_r \varphi_{ijt} + 2\kappa^{h,m} Z_m \varphi_i Q_{jh} + 2\kappa^{s,r} Z_r \varphi_{is} Q_j) \\
&\quad + (\varphi_{ij} + 2\varphi_i Q_j)(\kappa^{p,i} Z_{pq}\kappa^{q,j} + \kappa^{p,q} Z_q \kappa^{i,a}\kappa_{abp}\kappa^{b,j}) \text{ for } \tilde{\psi}.
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
\hat{\varphi}(z;Y) &= \varphi(z;Y,\theta) + n^{-1/2} R^{'}(z,Y) + n^{-1} S_1'(z,Y) + n^{-3/2} T_1'(z,Y) + O_p(n^{-2}), \\
\tilde{\varphi}(z;Y) &= \varphi(z;Y,\theta) + n^{-1/2} R^{'}(z,Y) + n^{-1} S_2'(z,Y) + n^{-3/2} T_2'(z,Y) + O_p(n^{-2}), \text{ and}
\end{aligned}
$$

$$
\begin{aligned}
D(\varphi,\widehat{\varphi}) - D(\varphi,\widetilde{\varphi}) &= \tfrac{1}{2}\int \tfrac{n^{-2}(2n^{-3/2}R'(S_1'-S_2')+S_1'^2-S_2'^2)+2n^{-2}R'(T_1'-T_2')}{\varphi} dz + O(n^{-5/2}) \\
&= n^{-3/2}\int \tfrac{R'(S_1'-S_2')}{\varphi} dz + \tfrac{1}{2}n^{-2}\int \tfrac{S_1'^2-S_2'^2}{\varphi} dz + n^{-2}\int \tfrac{R'(T_1'-T_2')}{\varphi} dz + O(n^{-5/2}).
\end{aligned}
\tag{2.19}
$$

From (2.9), we get

$$\varphi_j = \frac{\partial\varphi}{\partial\theta^j} = [-\frac{\sigma_{0j}}{\sigma_0} + (\frac{z-\lambda^T Y}{\sigma_0})(\frac{\lambda_j^T Y}{\sigma_0} + \frac{\sigma_{0j}(z-\lambda^T Y)}{\sigma_0^2})]\varphi, \tag{2.20}$$

and

$$\varphi_{jl} = \frac{\partial^2\varphi}{\partial\theta^j\partial\theta^l} = \frac{\partial\varphi_j}{\partial\theta^l} = \frac{\partial\{[-\frac{\sigma_{0j}}{\sigma_0}+(\frac{z-\lambda^T Y}{\sigma_0})(\frac{\lambda_j^T Y}{\sigma_0}+\frac{\sigma_{0j}(z-\lambda^T Y)}{\sigma_0^2})]\varphi\}}{\partial\theta^l}. \tag{2.21}$$

Applying the above expressions, and the fact that $\frac{z-\lambda^T Y}{\sigma_0} \sim N(0,1)$, we have

$$\int \frac{\varphi_d\varphi_l}{\varphi}dz = \frac{\lambda_d^T Y\lambda_l^T Y}{\sigma_0^2} + 2\frac{\sigma_{0d}\sigma_{0l}}{\sigma_0^2}, \tag{2.22}$$

$$\int \frac{\varphi_{ij}\varphi_d}{\varphi}dz = \frac{\lambda_{ij}^T Y\lambda_d^T Y+2\frac{\sigma_{0d}}{\sigma_0}\lambda_i^T Y\lambda_j^T Y}{\sigma_0^2} + 2\frac{\sigma_{0d}\sigma_{0ij}}{\sigma_0^2} + 2\frac{\sigma_{0i}\sigma_{0j}\sigma_{0d}}{\sigma^3}, \tag{2.23}$$

$$\begin{aligned}\int \frac{\varphi_{ijt}\varphi_d}{\varphi}dz = {} & \frac{\lambda_{ijt}^T Y\lambda_d^T Y+2\sigma_{0d}\sigma_{0ijt}}{\sigma_0^2} + \frac{2\sigma_{0d}}{\sigma_0^3}(\sigma_{0t}\sigma_{0ij} + \sigma_{0j}\sigma_{0it} + \sigma_{0i}\sigma_{0tj} \\ & + \lambda_t^T Y\lambda_{ij}^T Y + \lambda_j^T Y\lambda_{it}^T Y + \lambda_i^T Y\lambda_{jt}^T Y),\end{aligned} \tag{2.24}$$

$$\begin{aligned}\int \frac{\varphi_{ij}\varphi_{bd}}{\varphi}dz = {} & \frac{\lambda_{ij}^T Y\lambda_{bd}^T Y+2\sigma_{0bd}\sigma_{0ij}}{\sigma_0^2} + 26\frac{\sigma_{0i}\sigma_{0j}\sigma_{0b}\sigma_{0d}}{\sigma^4} \\ & + 2\frac{(\sigma_{0b}\sigma_{0d}\sigma_{0ij}+\sigma_{0i}\sigma_{0j}\sigma_{0bd}+\sigma_{0ij}\lambda_b^T Y\lambda_d^T Y+\sigma_{0bd}\lambda_i^T Y\lambda_j^T Y)}{\sigma_0^3} \\ & + 2\frac{(\lambda_i^T Y\lambda_j^T Y\lambda_b^T Y\lambda_d^T Y+\sigma_{0i}\sigma_{0j}\lambda_b^T Y\lambda_d^T Y+\sigma_{0b}\sigma_{0d}\lambda_i^T Y\lambda_j^T Y)}{\sigma_0^4} \\ & + 6\frac{(\sigma_{0i}\lambda_j^T Y+\sigma_{0j}\lambda_i^T Y)(\sigma_{0b}\lambda_d^T Y+\sigma_{0d}\lambda_b^T Y)}{\sigma_0^4}.\end{aligned} \tag{2.25}$$

By substituting the above quantities into the first term on the right hand side of equation (2.19), we get

$$\begin{aligned}n^{-3/2}\int \frac{R'(S_1'-S_2')}{\varphi}dz &= -n^{-3/2}\int \frac{\kappa^{a,b}Z_a\varphi_b[\frac{1}{2}\kappa_{ijk}\kappa^{i,j}\kappa^{k,l}\varphi_l+(\frac{1}{2}\varphi_{ij}+\varphi_i Q_j)\kappa^{i,j}]}{\varphi}dz \\ &= -n^{-3/2}[\frac{1}{2}\kappa^{a,b}Z_a\kappa_{ijk}\kappa^{i,j}\kappa^{k,l}\varphi_l(\frac{\lambda_b^T Y\lambda_l^T Y}{\sigma_0^2} + 2\frac{\sigma_{0b}\sigma_{0l}}{\sigma_0^2})\end{aligned}$$

$$+\tfrac{1}{2}\kappa^{a,b}Z_a\kappa^{i,j}(\tfrac{\lambda_{ij}^TY\lambda_b^TY+2\frac{\sigma_{0b}}{\sigma_0}\lambda_i^TY\lambda_j^TY}{\sigma_0^2}+2\tfrac{\sigma_{0b}\sigma_{0ij}}{\sigma_0^2}+2\tfrac{\sigma_{0i}\sigma_{0j}\sigma_{0b}}{\sigma^3})$$

$$+Q_j\kappa^{a,b}Z_a\kappa^{i,j}(\tfrac{\lambda_b^TY\lambda_i^TY}{\sigma_0^2}+2\tfrac{\sigma_{0b}\sigma_{0i}}{\sigma_0^2})]. \tag{2.26}$$

Similarly we can get the expressions for $\frac{1}{2}n^{-2}\int\frac{({S_1'}^2-{S_2'}^2)}{\varphi}dz$ and $n^{-2}\int\frac{R'(T_1'-T_2')}{\varphi}dz$ as well.

2.3.3 Integration over Y given θ

To compute the leading terms in (2.16), we need to integrate (2.26), $\frac{1}{2}n^{-2}\int\frac{({S_1'}^2-{S_2'}^2)}{\varphi}dz$ and $n^{-2}\int\frac{R'(T_1'-T_2')}{\varphi}dz$, the leading terms of the difference between two KL divergences, over Y conditional on θ. Let $Y_\varepsilon=(X\eta)_\varepsilon+e_\varepsilon$, where the subscript ε is an index between 1 and n. we have $E\{Z_iY_\varepsilon\}=n^{-1/2}E\{U_iY_\epsilon\}$. Note that

$$U_i=\tfrac{1}{2}(\upsilon_{\alpha\beta}\tfrac{\partial\omega^{\alpha\beta}}{\partial\theta^j}-e_\alpha e_\beta\tfrac{\partial\omega^{\alpha\beta}}{\partial\theta^j}),$$

where $\{\omega^{\alpha\beta}\}=W=V^{-1}-V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}$, and also $Y^TWY=e^TWe$, we can get $E\{Z_iY_\varepsilon\}=0$.

Furthermore,

$$E\{Z_eZ_fY_\varepsilon Y_\xi\}=\kappa_{e,f}(X\eta)_\varepsilon(X\eta)_\xi+\kappa_{e,f}\upsilon_{\varepsilon\xi}+\tfrac{1}{n}\{V\tfrac{\partial W}{\partial\theta^e}V\tfrac{\partial W}{\partial\theta^f}V\}_{\varepsilon\xi}+\tfrac{1}{n}\{V\tfrac{\partial W}{\partial\theta^f}V\tfrac{\partial W}{\partial\theta^e}V\}_{\varepsilon\xi},$$

$$E\{Z_eZ_{ij}Y_\varepsilon Y_\xi\}=\kappa_{e,ij}(X\eta)_\varepsilon(X\eta)_\xi+\kappa_{e,ij}\upsilon_{\varepsilon\xi}+\tfrac{1}{n}\{V\tfrac{\partial W}{\partial\theta^e}V\tfrac{\partial^2W}{\partial\theta^i\partial\theta^j}V\}_{\varepsilon\xi}+\tfrac{1}{n}\{V\tfrac{\partial^2W}{\partial\theta^i\partial\theta^j}V\tfrac{\partial W}{\partial\theta^e}V\}_{\varepsilon\xi},$$

$$E\{Z_eZ_{ijk}Y_\varepsilon Y_\xi\}=\kappa_{e,ijk}(X\eta)_\varepsilon(X\eta)_\xi+\kappa_{e,ijk}\upsilon_{\varepsilon\xi}+\tfrac{1}{n}\{V\tfrac{\partial W}{\partial\theta^e}V\tfrac{\partial^3W}{\partial\theta^i\partial\theta^j\partial\theta^k}V\}_{\varepsilon\xi}$$
$$+\tfrac{1}{n}\{V\tfrac{\partial^3W}{\partial\theta^i\partial\theta^j\partial\theta^k}V\tfrac{\partial W}{\partial\theta^e}V\}_{\varepsilon\xi},$$

$$E\{\lambda_j^\xi Y_\xi \lambda_d^\varepsilon Y_\varepsilon\} = \lambda_j^\xi \lambda_d^\varepsilon [(X\eta)_\varepsilon (X\eta)_\xi + \upsilon_{\varepsilon\xi}] = \lambda_j^\xi \lambda_d^\varepsilon \upsilon_{\varepsilon\xi},$$

$$E\{Z_a \lambda_b^\xi Y_\xi \lambda_l^\varepsilon Y_\varepsilon\} = \lambda_b^\xi \lambda_l^\varepsilon E(Z_a e_\varepsilon e_\xi) = -\tfrac{1}{2} n^{-1/2} \lambda_b^\beta \lambda_l^\varepsilon \tfrac{\partial \omega^{\alpha\beta}}{\partial \theta^a} (\upsilon_{\alpha\varepsilon}\upsilon_{\beta\xi} + \upsilon_{\alpha\xi}\upsilon_{\beta\varepsilon}),$$

$$E\{\lambda_i^\epsilon \lambda_j^\xi \lambda_a^\gamma \lambda_b^\delta Y_\epsilon Y_\xi Y_\gamma Y_\delta\} = \lambda_i^\epsilon \lambda_j^\xi \lambda_a^\gamma \lambda_b^\delta (\upsilon_{\varepsilon\xi}\upsilon_{\gamma\delta} + \upsilon_{\epsilon\gamma}\upsilon_{\xi\delta} + \upsilon_{\epsilon\delta}\upsilon_{\epsilon\gamma}).$$

Integration of (2.26) over Y and the other two leading terms can be evaluated using these equations, which leads to the following theorem.

Theorem 2.3.1. *Under Assumption 1-3,*

$$E_{Y|\theta}[D(\varphi, \widehat{\varphi}) - D(\varphi, \tilde{\varphi})] = g_1 - g_2 + o(n^{-2}), \tag{2.27}$$

where

$$\begin{aligned}
g_1 = {} & \tfrac{n^{-2}}{\sigma_0^2}\{Q_j \kappa^{a,b}\kappa^{i,j}\lambda_i^\epsilon \lambda_b^\xi \tfrac{\partial \omega^{\alpha\beta}}{\partial\theta^a}(\upsilon_{\alpha\epsilon}\upsilon_{\beta\xi} + \upsilon_{\alpha\xi}\upsilon_{\beta\epsilon}) \\
& - 3Q_b \kappa^{a,b}\kappa^{i,j}\kappa^{k,l}(\kappa^{ijk} + \kappa^{ik,j})(\lambda_l^\epsilon \lambda_a^\xi \upsilon_{\epsilon\xi} + 2\sigma_{0l}\sigma_{0a}) \\
& - 3Q_b \kappa^{a,b}\kappa^{k,l}(\lambda_{kl}^\epsilon \lambda_a^\xi \upsilon_{\epsilon\xi} + \tfrac{2\sigma_{0a}}{\sigma_0}\lambda_k^\epsilon \lambda_l^\xi \upsilon_{\epsilon\xi} + 2\sigma_{0a}\sigma_{0kl} + 2\tfrac{\sigma_{0a}\sigma_{0k}\sigma_{0l}}{\sigma_0}) \\
& - \tfrac{1}{2}\kappa^{i,j}\kappa^{a,b}(Q_j Q_b + 4Q_{jb})(\lambda_i^\epsilon \lambda_a^\xi \upsilon_{\epsilon\xi} + 2\sigma_{0a}\sigma_{0i})\},
\end{aligned} \tag{2.28}$$

$$\begin{aligned}
g_2 = {} & -\tfrac{n^{-2}}{\sigma_0^2}\{\tfrac{1}{4}(\kappa^{a,b}\kappa^{i,j}\kappa^{k,l}\kappa_{kij}\lambda_b^\xi \lambda_l^\epsilon \tfrac{\partial \omega^{\alpha\beta}}{\partial\theta^a}(\upsilon_{\alpha\epsilon}\upsilon_{\beta\xi} + \upsilon_{\alpha\xi}\upsilon_{\beta\epsilon}) \\
& + \kappa^{a,b}\kappa^{i,j}[\lambda_b^\xi \lambda_{ij}^\epsilon \tfrac{\partial \omega^{\alpha\beta}}{\partial\theta^a}(\upsilon_{\alpha\epsilon}\upsilon_{\beta\xi} + \upsilon_{\alpha\xi}\upsilon_{\beta\epsilon})] \\
& - \tfrac{5}{2}\kappa^{a,b}\kappa^{c,d}\kappa^{i,j}\kappa^{k,l}\kappa_{abc}(\kappa_{ik,j} + \kappa_{ijk})(\lambda_d^\epsilon \lambda_l^\xi \upsilon_{\epsilon\xi} + 2\sigma_{0d}\sigma_{0l}) \\
& - \tfrac{5}{2}\kappa^{a,b}\kappa^{i,j}\kappa^{k,l}(\kappa_{ik,j} + \kappa_{ijk})(\lambda_{ab}^\epsilon \lambda_l^\xi \upsilon_{\epsilon\xi} + 2\tfrac{\sigma_{0l}}{\sigma_0}\lambda_a^\epsilon \lambda_b^\xi \upsilon_{\epsilon\xi} + 2\sigma_{0l}\sigma_{0ab} + 2\tfrac{\sigma_{0l}\sigma_{0a}\sigma_{0b}}{\sigma_0}) \\
& - \kappa^{u,v}\kappa^{i,j}\kappa^{k,l}(\kappa_{u,ijk} + \kappa_{uijk})(\lambda_v^\epsilon \lambda_l^\xi \upsilon_{\epsilon\xi} + 2\sigma_{0v}\sigma_{0l}) - \kappa^{v,t}\kappa^{i,j} \\
& \times [\tfrac{\lambda_{ijt}^\epsilon \lambda_v^\xi \upsilon_{\epsilon\xi} + 2\sigma_{0v}\sigma_{0ijt}}{\sigma_0^2} + \tfrac{2\sigma_{0v}}{\sigma_0^3}(\sigma_{0t}\sigma_{0ij} + \sigma_{0j}\sigma_{0it} + \sigma_{0i}\sigma_{0tj} + \lambda_t^\epsilon \lambda_{ij}^\xi \upsilon_{\epsilon\xi} + \lambda_j^\epsilon \lambda_{it}^\xi \upsilon_{\epsilon\xi} + \lambda_i^\epsilon \lambda_{jt}^\xi \upsilon_{\epsilon\xi})]
\end{aligned}$$

$$
\begin{aligned}
&- \tfrac{3}{8}\kappa^{a,b}\kappa^{i,j}(\lambda^{\epsilon}_{ij}\lambda^{\xi}_{ab}\upsilon_{\epsilon\xi} + 2\sigma_{0ab}\sigma_{0ij} + 26\tfrac{\sigma_{0i}\sigma_{0j}\sigma_{0a}\sigma_{0b}}{\sigma_0^2} + 2\tfrac{\sigma_{0a}\sigma_{0b}\sigma_{0ij}+\sigma_{0i}\sigma_{0j}\sigma_{0ab}+\sigma_{0ij}\lambda^{\epsilon}_{a}\lambda^{\xi}_{b}\upsilon_{\epsilon\xi}+\sigma_{0ab}\lambda^{\epsilon}_{i}\lambda^{\xi}_{j}\upsilon_{\epsilon\xi}}{\sigma_0} \\
&+ 2\tfrac{\lambda^{\epsilon}_{i}\lambda^{\xi}_{j}\lambda^{\gamma}_{a}\lambda^{\delta}_{b}(\upsilon_{\epsilon\xi}\upsilon_{\gamma\delta}+\upsilon_{\epsilon\gamma}\upsilon_{\xi\delta}+\upsilon_{\epsilon\delta}\upsilon_{\xi\gamma})+\sigma_{0a}\sigma_{0b}\lambda^{\epsilon}_{i}\lambda^{\xi}_{j}\upsilon_{\epsilon\xi}+\sigma_{0i}\sigma_{0j}\lambda^{\epsilon}_{a}\lambda^{\xi}_{b}\upsilon_{\epsilon\xi}}{\sigma_0^2} \\
&+ 6\tfrac{\sigma_{0a}\sigma_{0i}\lambda^{\epsilon}_{b}\lambda^{\xi}_{j}\upsilon_{\epsilon\xi}+\sigma_{0a}\sigma_{0j}\lambda^{\epsilon}_{b}\lambda^{\xi}_{i}\upsilon_{\epsilon\xi}+\sigma_{0b}\sigma_{0i}\lambda^{\epsilon}_{a}\lambda^{\xi}_{j}\upsilon_{\epsilon\xi}+\sigma_{0b}\sigma_{0j}\lambda^{\epsilon}_{a}\lambda^{\xi}_{i}\upsilon_{\epsilon\xi}}{\sigma_0^2})\}. \qquad (2.29)
\end{aligned}
$$

Remarks: We say a prior $\pi(\theta)$ is "second-order KL REML-dominant", if under such prior $g_1 \geq g_2$ for all θ, i.e., the leading term of $E_{Y|\theta}[D(\varphi, \widehat{\varphi}) - D(\varphi, \tilde{\varphi})]$ is greater than or equal to zero.

2.4 Example: Mixed Effect Model

In this section we compare the plug-in predictive density and the Bayesian predictive density in terms of their KL divergences to the true conditional density function $\varphi(z; Y, \theta)$, for a simple mixed effect model as an illustration. We consider the Jeffreys prior and a family of improper priors, and show theoretically that the Jeffreys prior is not second-order KL REML-dominant, while there exists α^* such that the improper prior family $\pi(\beta, s_1, s_2) \propto (s_1 s_2)^{-\alpha}$ is second-order KL REML-dominant for $\alpha \in [\alpha^*, 1)$. Simulation studies are conducted which have great agreement with the theoretical results based on asymptotic expansion for moderate sample sizes.

2.4.1 Model and Notation

Consider the simple mixed effect model

$$
y_{i,j} = \beta + \mu_i + \epsilon_{ij}, \qquad i = 1, ..., n, j = 1, ..., m, \qquad (2.30)
$$

where $\mu_i \sim N(0, s_1), \epsilon_{ij} \sim N(0, s_2)$, and $\pi(\beta) \propto$ constant. Presumably μ and ϵ are independent. Without loss of generality, we study the predictive density of μ_1. Using the vector notation in model (2.4), we have $X = \mathbf{1}_{mn}, x_0 = 1, V = Diag\{V_1, V_2, ..., V_n\}$ with $V_i = s_1 \cdot J_m + s_2 \cdot I_m$, $w = (0_{m(n-1)}, s_1 \cdot 1_m)^T$, and $v_0 = s_1$, where $\mathbf{1}_{mn} = (1, ..., 1)^T, I_m$ the identity matrix, $J_m = \mathbf{1}_m \mathbf{1}_m^T, i = 1, ..., n$. By standard computation, we get

$$
\begin{aligned}
|X^T X|^{1/2} &= (mn)^{1/2}, \\
|V|^{-1/2} &= s_2^{-(m-1)n/2}(s_2 + ms_1)^{-n/2}, \\
|X^T V^{-1} X|^{-1/2} &= (s_2 + ms_1)^{1/2}(mn)^{-1/2}, \\
\lambda &= \tfrac{s_2}{mn(s_2+ms_1)} 1_{mn} + \tfrac{s_1}{s_2+ms_1}(0_{m(n-1)}, 1_m)^T, \\
\sigma^2 &= \tfrac{s_2(mns_1+s_2)}{mn(s_2+ms_1)}, \\
W &= \{\omega^{\alpha,\beta}\} = V^{-1} - V^{-1}X(X^T V^{-1} X)^{-1} X^T V^{-1} \\
&= Diag\{V^1, ..., V^n\} - \tfrac{1}{nm(s_2+ms_1)} J_{mn}, \\
V^i &= \tfrac{1}{s_2} I_{m\times m} - \tfrac{s_1}{s_2(s_2+ms_1)} J_m, i = 1, ..., n, \\
K &= \{\kappa^{i,j}\}_{2\times 2} = \begin{pmatrix} \frac{2n(s_2+ms_1)^2}{(n-1)m^2} + \frac{2s_2^2}{m^2(m-1)} & -\frac{2s_2^2}{m(m-1)} \\ -\frac{2s_2^2}{m(m-1)} & \frac{2s_2^2}{m-1} \end{pmatrix}.
\end{aligned}
$$

From the matrix form of K, we can derive that the Jeffreys prior of $\theta = (s_1, s_2)$ in this model is

$$\pi_J(\theta) \propto \tfrac{1}{s_2(ms_1+s_2)},$$

with

$$Q_{1J} = \tfrac{\partial \log \pi}{\partial s_1} = -\tfrac{m}{ms_1+s_2},$$

$$Q_{2J} = -\frac{ms_1+2s_2}{s_2(ms_1+s_2)}.$$

We add the subscript "J" to denote the Jeffreys prior and relevant functions. A family of improper priors is also considered, which we refer to as

$$\pi_I^{\alpha}(\beta, s_1, s_2) \propto (s_1 s_2)^{-\alpha}, \tag{2.31}$$

with $\alpha \in (0, 1)$ to guarantee proper posterior distributions (Hobert and Casella, 1996). Correspondingly, we have

$$Q_{1I} = \frac{\partial \log \pi_I}{\partial s_1} = -\frac{\alpha}{s_1},$$
$$Q_{2I} = -\frac{\alpha}{s_2},$$

where we add "I" as a subscript to relevant functions under the improper priors.

2.4.2 Theoretical results for the mixed effect model

In this section, we show in Theorem 2 and 3 that for the mixed effect model (2.30), there exists α^* such that the family of improper priors π_I with $\alpha \in [\alpha^*, 1)$ are second-order KL REML dominant, while the Jeffreys prior is not. In the remarks we give conditions under which Bayesian predictive density with Jeffreys prior outperforms the REML plug-in estimator, and more explicit results on α^*. All the results are derived under Assumptions 1 - 3. We first consider the Jeffreys prior. Substituting Q_1, Q_2 into equation (2.27) and compare that with (2.28), we have

Theorem 2.4.1. F*or the mixed effect model (2.30), there exists* $r \in (0, \infty)$ *such that* $g_1 \geq g_2$ *iff* $\frac{s_1}{s_2} \geq r$ *as* $n \to \infty$*, i.e., the Jeffreys prior is not second-order KL REML-dominant for this specific model.*

*P*roof: For any fixed n and m, $n^2(g_1 - g_2)$ is the product of a positive number and a function of the ratio $x = \frac{s_1}{s_2}$. We have

$$\lim_{n\to\infty} n^2(g_1 - g_2) = C_J(m, n, s_1, s_2) f_J(x),$$

where $C_J(m, n, s_1, s_2) = \frac{s_2}{s_2 + mns_1}$ is positive for all $m > 1, n > 0, s_1 > 0, s_2 > 0$, and $f_J(x) = x^9 + ax^8 + bx^7 + cx^6 + dx^5 + ex^4 + fx^3 + gx^2 + hx + i$, with

$$\begin{aligned}
a &= \frac{-14704+16809m-30148m^2+43577m^3-26714m^4+5996m^5}{2(m-1)^2m(144+1429m-2716m^2+1200m^3)},\\
b &= \frac{2(-17768+17831m-13642m^2+15744m^3-9334m^4+1985m^5)}{(m-1)^2m^2(144+1429m-2716m^2+1200m^3)},\\
c &= \frac{9(-13702+13337m-6648m^2+5472m^3-3180m^4+635m^5}{2(m-1)^2m^3(144+1429m-2716m^2+1200m^3)},\\
d &= \frac{-41072+41509m-14628m^2+7672m^3-4712m^4+863m^5}{(m-1)^2m^4(144+1429m-2716m^2+1200m^3)},\\
e &= \frac{-20768+18359m+932m^2-5424m^3+2232m^4-515m^5}{2(m-1)^2m^4(144+1429m-2716m^2+1200m^3)},\\
f &= -\frac{4(12+234m-223m^2+93m^3}{m^5(144+1429m-2716m^2+1200m^3)},\\
g &= -\frac{420-518m+355m^2}{2m^6(144+1429m-2716m^2+1200m^3)},\\
h &= -\frac{9}{m^6(144+1429m-2716m^2+1200m^3)},\\
i &= -\frac{9}{2m^6(144+1429m-2716m^2+1200m^3}.
\end{aligned}$$

By the definition, a prior is second-order KL REML-dominant if and only if $f_J(x) \geq 0$ for all $x > 0$. When $n \to \infty, f_J(x)$ is a concave function of x, since $f''_J(x) \leq 0$ for $x > 0$. It is

easy to calculate that $f_J(0) < 0$ and $f_J(1) > 0$, therefore there exists some $r \in (0,1)$ such that $f_J(x) < 0$ for $x \in (0,r)$ and $f_J(x) \geq 0$ for $x \in [r,1)$, which proves the claim. $\square$

Remarks: It can be shown that for any $m, r \in (\frac{1}{2m}, \frac{1}{m})$, as $n \to \infty$, since $f_J(\frac{1}{2m}) < 0$ and $f_J(\frac{1}{m}) > 0$. At $x = \frac{1}{2m}$,

$$\lim_{n\to\infty} n^2(g_1 - g_2) = -\frac{6479184 - 6072235m + 1352974m^2 - 297387m^3 + 182088m^4 + 34992m^5}{583200(m-1)^4 m},$$

which can be shown to be negative for all m, since the denominator is positive and the numerator is negative, for $m \geq 2$. Similarly, we can show $f_J(\frac{1}{m}) > 0$ for all m. Since $f_J(x)$ is a continuous function of x, there exists at least one solution for $f_J(x) = 0$ between $\frac{1}{2m}$ and $\frac{1}{m}$.

Next, we consider the improper prior, which has the property given by the following proposition and theorem. This kind of prior has a very intriguing feature as follows,

Proposition 1: For $m = 2$, under the improper prior family of $\pi_I^\alpha \propto (s_1 s_2)^{-\alpha}$ with $\alpha \in (0,1), g_1 \geq g_2$ always holds as $n \to \infty$, i.e. Bayesian predictive densities under these priors perform better than the REML plug-in density for any $s_1, s_2 \in R^+$. In other words, the improper prior family with $\alpha \in (0,1)$ is second-order KL REML-dominant for $m = 2$.

Theorem 2.4.2. *For any $m \geq 3$, there exists $\alpha^* \in (0, \frac{m-3}{m-2}]$ such that as $n \to \infty$, under the improper prior $\pi_I(\beta, s_1, s_2) \propto (s_1 s_2)^{-\alpha}$ with $\alpha \in [\alpha^*, 1), g_1 \geq g_2$ for all s_1 and s_2, i.e., this improper prior family is second-order KL REML-dominant for $\alpha \in [\alpha^*, 1)$. For $m = 2$, the improper prior is second-order KL REML-dominant for any $\alpha \in (0,1)$.*

Proof: Substituting Q_{1I}, Q_{2I} into equation (2.27), we get

$$\lim_{n\to\infty} n^2(g_1 - g_2) = C_I(m, n, s_1, s_2) f_I(x),$$

where $x = \frac{s_1}{s_2}, C_I(m,n,s_1,s_2) = \frac{s_2^{10}}{4s_1^4(m-1)^6m^9(s_2+ms_1)^6[-48+48m^2+m(9-32\alpha+4\alpha^2)]}, f_I(x) = x^{10} + ax^9 + bx^8 + cx^7 + dx^6 + ex^5 + fx^4 + gx^3 + hx^2 + ix + j$, with

$a = \frac{2[-144+4m^3(-3+2\alpha+\alpha^2)-4m^2(-84+2\alpha+3\alpha^2)+m(-153-72\alpha+16\alpha^2)]}{m[-48+48m^2+m(9-32\alpha+4\alpha^2)]}$,

$b = \{-5184(-3+2m) + (-1+m)^2[-848 + m^2(2818 + 80\alpha - 192\alpha^2) + 4m^3(-28 + 28\alpha + 17\alpha^2) + m(-1669 - 400\alpha + 144\alpha^2)]\}/\{(m-1)^2m^2[-48 + 48m^2 + m(9 - 32\alpha + 4\alpha^2)]\}$,

$c = \{4[2592(7 - 6m + m^2) + (m-1)^2[-376 + m^2(1298 + 212\alpha - 168\alpha^2) + m^3(-33 + 88\alpha + 64\alpha^2) + m(-735 - 292\alpha) + 104\alpha^2]]\}/\{(m-1)^2m^3[-48 + 48m^2 + m(9 - 32\alpha + 4\alpha^2)]\}$,

$d = \{5184(23 - 22m + 5m^2) + (m-1)^2[-1584 + m^2(3830 + 2512\alpha - 1344\alpha^2) + m^3(225 + 672\alpha + 560\alpha^2) + m(-893 - 2640\alpha + 760\alpha^2)]\}/\{(m-1)^2m^4[-48 + 48m^2 + m(9 - 32\alpha + 4\alpha^2)]\}$,

$e = \{2[10368(m-2)^2 + (m-1)^2[-400 + m^3(437 + 448\alpha + 392\alpha^2) - 2m^2(277 - 940\alpha + 420\alpha^2) + m(1767 - 1816\alpha + 432\alpha^2)]]\}/\{(m-1)^2m^5[-48 + 48m^2 + m(9 - 32\alpha + 4\alpha^2)]\}$,

$f = \{[5184(m-2)^2 + (m-1)[-32 + m^3(-5113 + 2288\alpha - 2072\alpha^2) + m(-4901 + 2864\alpha - 592\alpha^2) + m^4(1191 + 896\alpha + 728\alpha^2) + m^2(8919 - 6048\alpha + 1936\alpha^2)]]\}/\{(m-1)^2m^6[-48 + 48m^2 + m(9 - 32\alpha + 4\alpha^2)]\}$,

$g = \frac{8[12+2m(170-75\alpha+14\alpha^2)-2m^2(163-95\alpha+42\alpha^2)+m^3(111+84\alpha+56\alpha^2)]}{m^7[-48+48m^2+m(9-32\alpha+4\alpha^2)]}$,

$h = \frac{4(141-52\alpha+9\alpha^2)-2m(331-184\alpha+96\alpha^2)+m^2(383+352\alpha+176\alpha^2)}{m^7[-48+48m^2+m(9-32\alpha+4\alpha^2)]}$,

$i = \frac{2[-2(9-8\alpha+6\alpha^2)+m(45+56\alpha+20\alpha^2)]}{m^7[-48+48m^2+m(9-32\alpha+4\alpha^2)]}$,

$j = \frac{9+16\alpha+4\alpha^2}{m^7[-48+48m^2+m(9-32\alpha+4\alpha^2)]}$.

When $m \geq 3$, it is a simple but tedious process to show, by checking the extreme points of each of the functions a-j as $\alpha \in [\frac{m-3}{m-2}, 1)$, that $C_I(m,n,s_1,s_2)$ and all the coefficients in $f_I(x)$ are

non-negative, which proves the claim. For $m = 2, f_I(x)$ is always non-negative for any $x > 0$, by checking the extreme point of a -j for $\alpha \in (0, 1)$. $\square$

In Figure 2.1, we plot $n^2(g_1 - g_2)$ as a function of s_1 with $s_2 = 1$, for $n = 10, 20, 50, 100, 1000$ and $m = 2, 5, 10$ respectively. The plots on the left is for the Bayesian predictive densities under the Jeffreys prior, and the plots on the right are for those under the improper prior with $\alpha = 0.9$. These plots numerically show that the asymptotic results in Theorem 2 and 3 are also valid for finite sample size with n as small as 10: The Jeffreys prior are not second order KL REML-dominant while the improper prior with $\alpha = 0.9$ are for all the $m(= 2, 5, 10)$ and $n(= 10, 20, 50, 100, 1000)$ combinations considered. For the Jeffreys prior, the threshold for the Bayesian predictive distribution to be better than the REML based estimative distribution is between $\frac{1}{m}$ and $\frac{1}{2m}$, which is also consistent with the asymptotic results.

In Figure 2.2, we make two 3-dimensional plots from different angles for $n^2(g_1 - g_2)$ with g_1 calculated under the improper prior, when $m = 100$. Both the left and right panel show clearly that, there is some α^* such that **when** $\alpha \in [\alpha^*, 1)$, the difference will always be positive.

2.4.3 Simulation Studies

In this section we compare both the Jeffreys prior and the proposed improper prior with the REML estimative density in terms of $E_{Y|\theta}(D(\varphi, \widehat{\varphi}) - D(\varphi, \tilde{\varphi}))$ using simulation. The following parts gives the implementation details of the numerical procedure for conducting the simulation experiments, and the simulation results are summarized in the last part.

REML estimator for (s_1, s_2)

To evaluate $E_{Y|\theta}(D(\varphi, \widehat{\varphi}))$, we need to plug in the REML estimator from every given data vector and specific $\theta(\theta = (s_1, s_2))$ in each iteration. For the simple random effect model, we can

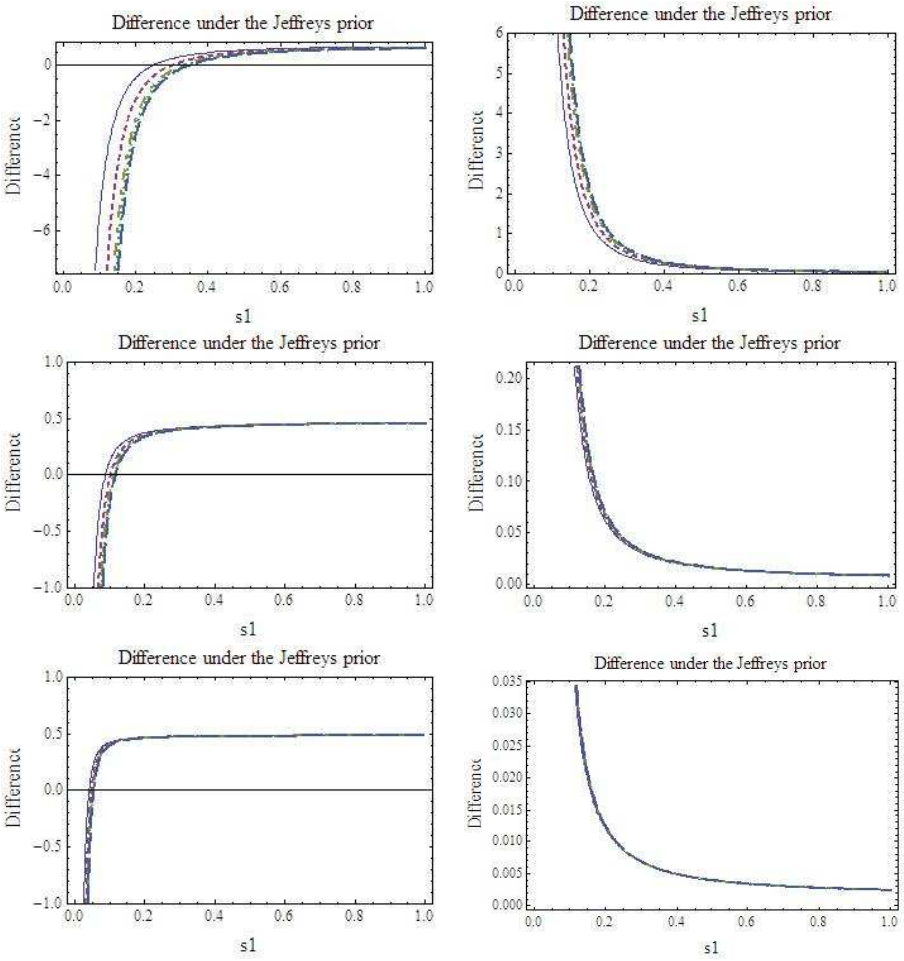

Figure 2.1: Plots of $n^2(g_1 - g_2)$ against s_1 when $s_2 = 1$, for n =10 (thin solid line), 20 (broken line), 50 (broken/dotted line), 100 (dotted line),and 1000 (wider broken line). The plots on the left are for the Bayesian predictive densities under the Jeffreys prior, and those on the right are for those under the improper prior, with $\alpha = 0.9$. The three rows from top to bottom correspond to $m = 2, 5$, and 10 respectively.

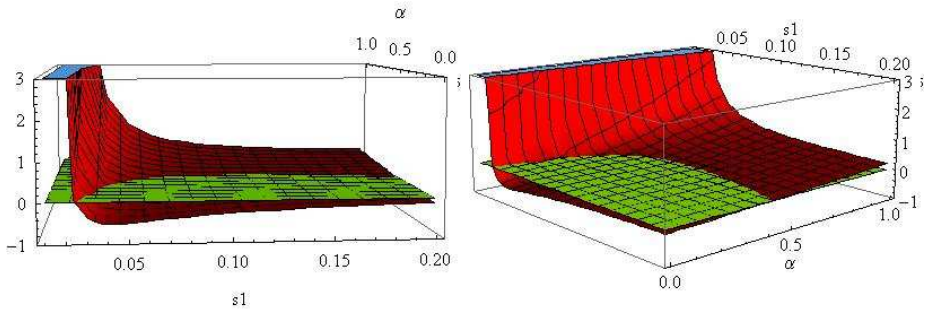

Figure 2.2: Plots of $n^2(g_1 - g_2)$ against s_1 with $s_2 = 1$ when $n \to \infty, m = 100$ under the improper prior. All the plots are made for $s_1 \in (0, 0.2), \alpha \in (0, 1)$, with views from different angles.

explicitly calculate the REML estimator for s_1, s_2. The restricted log-likelihood for $\theta = (s_1, s_2)$ is

$$\ell_n(\theta) = \tfrac{(1-m)n}{2} \log s_2 + \tfrac{1-n}{2} \log(s_2 + ms_1) - \tfrac{G^2}{2}, \tag{2.32}$$

with

$$\begin{aligned} G^2 = G^2(\theta) &= Y^T\{V^{-1} - V^{-1}X(X^TVX)^{-1}X^TV^{-1}\}Y \\ &= \tfrac{1}{s_2}\sum_{i,j} y_{i,j}^2 - \tfrac{1}{mn(s_2+ms_1)}(\sum_{i,j} y_{i,j})^2 - \tfrac{s_1}{s_2(s_2+ms_1)}\sum_i(\sum_j y_{i,j})^2 \quad . \end{aligned} \tag{2.33}$$

For general n and m,

$$\hat{s_1} = \tfrac{n\sum_{i=1}^n(\sum_{j=1}^m y_{i,j})^2 - (\sum_{i=1}^n\sum_{j=1}^m y_{i,j})^2}{m^2n(n-1)} - \tfrac{m\sum_{i=1}^n\sum_{j=1}^m y_{i,j}^2 - \sum_{i=1}^n(\sum_{j=1}^m y_{i,j})^2}{m^2n(m-1)}, \tag{2.34}$$

and

$$\hat{s_2} = \tfrac{m\sum_{i=1}^n\sum_{j=1}^m y_{i,j}^2 - \sum_{i=1}^n(\sum_{j=1}^m y_{i,j})^2}{(m-1)mn}. \tag{2.35}$$

It is easy to check that $\frac{\partial^2\ell_n(\theta)}{\partial s_1^2}, \frac{\partial^2\ell_n(\theta)}{\partial s_2^2} < 0$, and the standard deviation for these above

REML estimators goes asymptotically to 0 as n goes into infinity.

Sampling for (β, s_1, s_2)

To calculate the predictive density function, we need to generate s_1, s_2 from the posterior distributions. For both the Jeffreys and improper priors, the marginal posterior is complicated. However, the Metropolis-Hastings algorithm can be used to generate the posterior distributions as follows:

Step 1. Start with arbitrary s_1^0, s_2^0 from support of the posterior distribution, i.e. $(0, \infty)$.

Step 2. At stage n, generate proposal s_1^*, s_2^* from $q(s_1^*, s_2^* | s_1, s_2)$. The arbitrary proposal distribution is defined as $q(s_1^*, s_2^* | s_1, s_2) = \frac{1}{s_1 s_2} \exp\{-\frac{s_1^*}{s_1} - \frac{s_2^*}{s_2}\}$, the product of two exponentials with means s_1 and s_2.

Step 3. Take $s_1^{n+1} = s_1^*, s_2^{n+1} = s_2^*$ with probability $\alpha = \min\{\frac{q(s_1, s_2 | s_1^*, s_2^*) \pi_J(s_1^*, s_2^*) f(y | s_1^*, s_2^*)}{q(s_1^*, s_2^* | s_1, s_2) \pi_J(s_1, s_2) f(y | s_1, s_2)}, 1\}$. Otherwise, increase n and return to *Step 2*. This random acceptance is done by generating $u \sim$ Uniform (0, 1) and accepting the proposal s_1^*, s_2^* if $u \leq \alpha$.

We burn in 1000 out of 2000 simulations (actually 100-500 is enough) to make sure that there is no influence of the initial values for s_1 and s_2, so only 1000 variates have been used to approximate the posteriors, from which we select one in every ten and make records of 100 pairs of (s_1^*, s_2^*). The convergence is justified by the results of Gelman and Rubin's convergence diagnostics in the "CODA" package (Output analysis and diagnostics for MCMC simulations) of R language. The procedure is as follows:

1. Run two parallel chains, each with 1000 pairs of (s_1, s_2) starting from different initial values.

2. Discard the first 500 draws in each chain.

3. Calculate the "potential scale reduction factor" (see Gelman and Rubin (1992), Brooks and Gelman (1997)) for each parameter (s_1 and s_2) in the chains, together with upper and lower confidence limits. Approximate convergence is diagnosed, since the upper limits are close to 1, indicating both chains have "forgotten" their initial values, and the output from them is indistinguishable.

We also use the Geweke's convergence diagnostic to double-check the convergence: first combine the remaining two chains ($2 \times 500 = 1000$ draws) to produce one chain, and calculate Z-scores for a test of equality of means (see Geweke (1992)) between the first 10% and last 50% (the CODA default values) of the chain for both parameters. The calculated values do not fall in the extreme tails of a standard normal distribution, providing no evidence against convergence.

MC Method for integration of KL divergence

We evaluate $E_{Y|\theta}[D(\varphi, \widehat{\varphi}) - D(\varphi, \tilde{\varphi})]$ for fixed θ by the following algorithm.

Step 1: Generate $Y^{(l)}$ for $l = 1, 2, \ldots, L$ using the model (2.30) for fixed θ.

Step 2: For each $Y^{(l)}$, compute the REML estimator $\hat{\theta} = (\hat{s_1}, \hat{s_2})$ using (2.34) and (2.35), and the corresponding REML predictive density function is given by $\hat{\varphi}(z; Y) = \varphi(z; Y, \hat{\theta})$.

Step 3: Approximate $\tilde{\varphi}(z; Y^{(l)})$ by $\frac{1}{n}\sum_i \varphi(z; Y^{(l)}, \theta^i)$, where θ^i is generated as described in 4.3.2, with Jeffreys and improper priors, respectively.

Step 4: The difference between $D(\varphi, \hat{\varphi})$ and $D(\varphi, \tilde{\varphi})$ is approximated by quadrature integration method.

Step 5: To calculate the expected KL divergence for fixed θ, we approximate it by $\frac{1}{L}\sum_l (D(\varphi, \widehat{\varphi}|Y^{(l)}, \theta) - D(\varphi, \tilde{\varphi}|Y^{(l)}, \theta))$, where the summation is taken over $Y^{(l)}$.

We set L as 100 here, which is also justified by the convergence diagnostic in "CODA"

package of R programming.

Simulation Results

In the simulation studies, we set $s_2 = 1$, $\beta = 0$, $m = 2, 5, 10$, $n = 10, 20, 50, 100$, and $s_1 = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8$, and carried out the computation for both the Jeffreys prior and the improper prior with $\alpha = 0.75$. The results are summarized in Figure 2.3.

The first row of Figure 2.3 describes simulation results for $m = 2$. The left panel shows the results under Jeffreys prior, and the right one under the improper prior with $\alpha = 0.75$. The left panel indicates that under the Jeffreys prior and when s_1 is less than 0.5, the REML plug-in density performs better than the Bayesian predictive density in terms of KL divergence, while the Bayes predictive density performs better than the REML competitor otherwise. The right panel indicates that, under the improper prior the Bayesian predictive density always performs better than the REML estimative density. Both results are consistent with our theoretical findings.

The second row of Figure 2.3 gives simulation results for $m = 5$. The left panel indicates that when $m = 5$, the Bayesian predictive density under Jeffreys prior performs better than REML plug-in estimative density when $\frac{s_1}{s_2}$ is greater than some value around 0.2, and the REML competitor performs better otherwise, which is consistent with the asymptotic results in Figure 2.1. The right panel displays simulation results under the improper prior with $\alpha = 0.75$, which indicates that the Bayesian predictive densities always performs better than the REML estimative density, which are also consistent with the theoretical results.

The third row of Figure 2.3 gives simulation results for $m = 10$. We obtain similar conclusions as for $m = 2$ and 5, except that in the left panel, the change point is around 0.1.

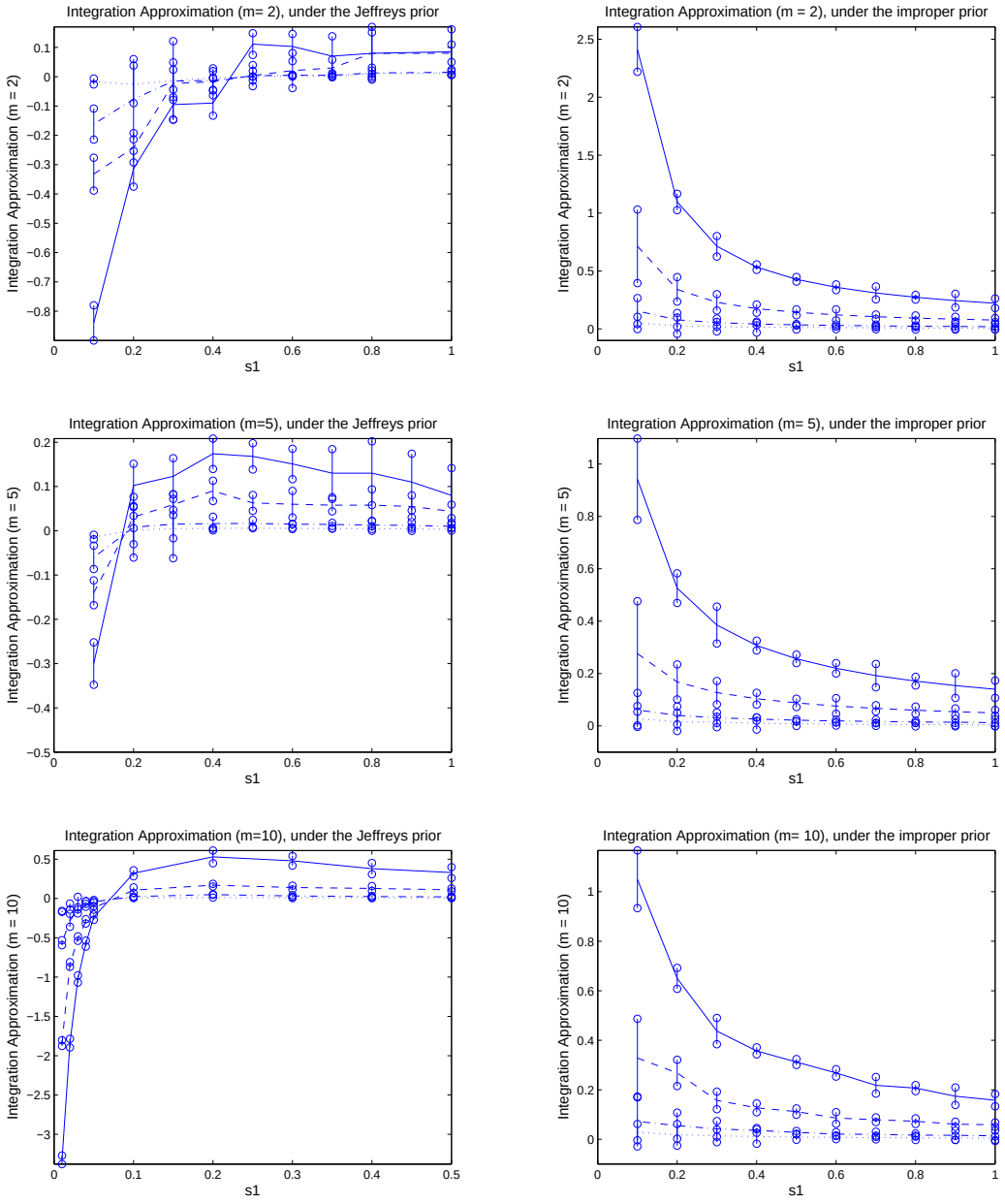

Figure 2.3: Simulation results of expected difference of KL divergence against s_1 with $s_2 = 1$ for n =10 (solid line), 20 (broken line), 50 (broken/dotted line), and 100 (dotted line). The plots on the left are under the Jeffreys prior, and those on the right are under the improper prior, with $\alpha = 0.75$. The three rows from top to bottom correspond to $m = 2, 5$, and 10 respectively.

2.5 Discussion

In this chapter we used the asymptotic expansion of the KL divergence as the main tool to compare different predictive distributions, and derived some explicit results for one-way random effects models. In particular, we find a class of improper priors which leads to predictive distributions that are asymptotically superior to the REML based estimative distributions. Similar results have the potential to hold for more general mixed effects models, including the spatial linear models commonly used in geostatistics, nevertheless the proof will be explored elsewhere. The asymptotic expressions we derived for KL divergence is quite general, and can be used for other purpose, such as spatial sampling design in the context of spatial linear model. Vidoni (1995) introduced a simple form to express the predictive densities by approximating the sampling distribution of the maximum likelihood estimator with the p^*-formula and then using Laplace approximation to integrate out the parameter, for exponential families and for location models. We could also consider the similar idea when computing the KL divergence by integrating out the parameter. In particular, we can introduce a design criteria that takes into account of the Kullback-Leibler divergence between the true density and the REML plug-in density or the Bayesian predictive density, with respect to the point or block predictor. To achieve the optimal design, it is reasonable to consider employing the asymptotic approximation to the KL divergence to the second order, which we obtain as equation (2.25). This might give some explicit form for the integration of Kullback-Leibler divergences, and possibly reduces the computation workload.

Garcia-Donato and Sun (2007) discussed objective priors for hypothesis testing for one-way random effects models, and derived the divergence based (DB) prior and the intrinsic prior. Their work is related with ours, while their emphasis is on the use of these priors to develop

consistent objective Bayesian factors, which is different from our purpose. It is interesting to check whether their priors are also second order KL REML dominate priors, and whether some of their priors can dominant other priors in the sense of second order KL divergence.

Chapter 3

Applications to regression models with temporally or spatially correlated errors

3.1 Introduction

We continue to make intensive exploration on another correlation structure: AR(p) process, instead of the linear mixed effect model, by using theoretical methods in Chapter 2. The p-th order autoregressive (AR(p)) model is widely known in time series analysis. It consists of the data $\{y_t\}$, satisfying $y_t = -\sum_{i=1}^{p} a_i y_{t-i} + u_t, t = 1, \cdots, T$, where $\{u_t\}$ is a white noise with mean 0 and variance σ^2. The point estimation of the AR parameters $a_1, a_2, \cdots, a_p$ is well understood and has been vigorously investigated for a long time, e.g. page $238 - 241$ in Chapter 8 of Brockwell and Davis (1991). However, the AR(1) model, the simplest case of AR(p) processes, is surprisingly challenging to objective Bayesians, even in comparatively simple case of known σ^2, as was discussed in Phillips (1991). Phillips and some other discussants highlighted the issues and controversies in developing a noninformative prior for AR(p) models. Berger and Yang (1994) investigated the AR(1) model from a Bayesian point of view. Their approach was based on the reference prior method and the stationarity was not assumed.

They compared different noninformative priors based on the mean squared error or coverage probability. They considered three candidate priors when making frequentist comparison: the uniform prior (which results in MLE estimator of ρ), the Jeffreys prior, and the symmetrized reference prior, which is the reference prior for $|\rho| < 1$. According to their simulation results for mean squared error, the symmetrized reference prior seemed generally superior with exception of the explosive case ($|\rho| > 1$). On the other aspect, the coverage for the symmetrized reference prior was generally more attractive as well. Therefore they highly recommended symmetrized reference prior as the "default" prior for the AR(1) model.

Tanaka and Komaki (2005) looked at the Bayesian estimation of the spectral density of the AR(2) model and proposed a superharmonic prior as a noninformative prior. They also considered a more general case, the autoregressive moving average (ARMA) model, focusing on the Bayesian estimation of an unknown spectral density in the ARMA model. They first showed that in the i.i.d. cases, the Bayesian spectral densities based on a superharmonic prior asymptotically dominate those based on the Jeffreys prior, using the asymptotic expansion of the risk difference. Then they obtained the asymptotic expansion of the Bayesian spectral density for the ARMA model, which could be written in the differential-geometrical quantities as in the i.i.d. cases. Finally they obtained a similar result in the ARMA model.

In addition, models for two-dimensional spatial data where the errors follow a spatial ARMA process have been noticeably considered by several authors, e.g. Martin (1990), Zimmerman and Harville (1991), Cullis and Gleeson (1991) and Basu and Reinsel (1994). The analysis of such spatial processes is of interest in many different fields and they have been studied in such disciplines as geography, geology, biology and agriculture. Many of the developments have been summarized in the books by Bartlett (1975), Ripley (1981) and Cliff and Ord (1981).

In particular, Basu and Reinsel (1993) considered the spatial processes defined on a regular rectangular grid in two dimensions with sites labeled (i, j), with an associated random variable Y_{ij} defined at each site. Examples of such phenomena include data collected on a regular grid of size $m \times n$ from satellites and from agricultural field trials. They concentrated on the first-order unilateral models of interest, including a special case of the multiplicative (or linear-by-linear) first-order spatial models, which have proved to be of practical use in modeling of two-dimensional spatial lattice data. No previous work has considered the prior for this kind of linear-by-linear spatial model before.

Our work assumes the stationarity, and focuses on both the Bayesian estimation and frequentist estimative method for the predictive density of the AR(1) model with unknown σ^2, by utilizing the criterion of expected K-L divergence, as proposed in Chapter 2. We point out that the reference prior is superior to the Jeffreys prior and the reference inverse prior, with respect to the second-order asymptotic approximation (see Sections 3.3.3 - 3.3.4). We also consider the noninformative priors and REML estimative density for the model with noise from a spatial multiplicative AR(1) model (see Basu and Reinsel (1993) and Martin (1990)), with fixed σ^2 for simplicity (see Sections 3.4.3 - 3.4.4).

General definition of the AR(1) model and the necessary notations for Fisher Information Matrix calculation are briefly reviewed in Section 3.3.1 - 3.3.2 (temporal case) and 3.4.1 - 3.4.2 (spatial case), respectively. Our argument is mainly based on the expected K-L divergence, which is proposed in Chapter 2, with respect to different Bayesian predictive densities or REML plug-in density within the framework of this type of model. In Chapter 2 we proposed using the averaged K-L divergence when comparing Bayesian predictive densities with REML plug-in one and provided the second-order expression of it. In Section 3.3.3 and 3.4.3, we apply this

approach to the AR(1) model, for time series and spatial structure, respectively. We consider three candidate priors: the Jeffreys prior, the reference prior and the inverse reference prior. In an asymptotic sense, all the three priors perform quite well when compared to the REML-plug in density. In particular, we prove that the reference prior dominates the other two priors. In Section 3.3.4 and 3.4.4, we perform the numerical simulation for the AR(1) time series and spatial process, illustrating that the asymptotic results hold,when the sample size is moderately large. In Section 3.5, we make some concluding remarks for our work.

3.2 Review of Noninformative Priors

3.2.1 Background

The use of noninformative priors has an extensive tradition in statistics, starting with Bayes (1763) and Laplace (1812) who used the "uniform" prior

$$\pi_U(\theta) = 1. \tag{3.1}$$

In developing the Bayesian methodology, use of π_U was generally very successful, although there were concerns about its lack of invariance to transformation (because one cannot, for instance, be simultaneously "uniform" in θ and $\eta = \log(\theta)$). Also, a number of counterexamples to its use have been encountered, see for example, Mitchell (1967), Monette et al. (1984) and Ye and Berger (1991).

Jeffreys (1961b) sought to overcome the lack of invariance of π_U through the development of the now-famous Jeffreys prior

$$\pi_J(\theta) = \sqrt{\det(I(\theta))}, \tag{3.2}$$

where $I(\theta)$ is the Fisher information matrix with (i, j) entry

$$I(\theta) = -E_\theta[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(\,Y|\theta)], \tag{3.3}$$

where f is the likelihood function of Y given θ, and E_θ stands for expectation over X, given θ. This prior is invariant to reparameterization of the problem, and this method of deriving a noninformative prior seems to correct a number of the counterexamples to use of $\pi_U(\theta) = 1$, especially those arising from nonintegrability of the posterior distribution, $\pi_U(\theta|\text{data})$. For instance, for the AR(1) model, with $\theta = (\rho, \sigma^2)$ where we assume $|\rho| < 1$ (the stationary case),

$$\pi_J(\theta) = \sqrt{I(\theta)} \propto \sigma^{-2}(1-\rho^2)^{-1/2}, \tag{3.4}$$

(see Jeffreys (1961b), Zellner (1971), and Box and Jenkins (1976), for the $|\rho| < 1$ case and Phillips (1991) for the general case). At $|\rho| = 1$, $\pi_J(\theta)$ can be defined by continuity (see Phillips (1991)). In our work we only consider the stationary case.

3.2.2 The Reference Prior Approach

Bernardo (1979) initiated an information-based approach to the development of noninformative priors, called the reference prior approach. A review and discussion of the current status of the approach can be found in Berger and Bernardo (1992).

The motivation for developing this approach was the acknowledged problems of the Jeffreys prior in higher dimensions. Even Jeffreys would often alter $\pi_J(\theta)$ in multiparameter problems to remove perceived inadequacies. The reference prior approach sought to overcome these difficulties by breaking up multiparameter problems into a series of conditional one-parameter

problems, for which reasonable noninformative priors could be determined. The approach has been proven to be remarkably successful in overcoming the inadequacies of Jeffreys prior in multiparametr problems.

Unfortunately, the motivation for the reference prior method was primarily based on the i.i.d. asymptotics. An attempt to generalize this to the dependent-data AR(1) model met with only partial success: the reference prior exists for the stationary case ($|\rho| < 1$) but not for the explosive case ($|\rho| > 1$). This problem could be solved by a symmetrized version of the stationary case reference, which was ultimately recommended in Berger and Yang (1994).

Indeed, the reference prior algorithm consists of four components (see Berger and Yang (1994) for more details):

(i) information maximization,

(ii) maximizing asymptotic missing information,

(iii) finding limits of reference priors on compact sets, and

(iv) dealing with multiparameter problems by conditional decompositions (This is the original motivation for the algorithm).

Following the above steps, we can find the reference prior for the stationary AR(1) model. In particular when making applications to the AR(1) model, the multiparameter issue can be avoided entirely by considering only the case $\sigma^2 = 1$, then

$$\pi_R(\theta) = \frac{1}{\sqrt{1-\rho^2}}.$$

The only change in the reference prior analysis that would result from having σ^2 unknown would be to introduce a multiplicative factor of $1/\sigma$ in the prior.

We consider the above reference prior suggested by Berger and Yang (1994), which had also

previously been given in Zellner (1977) as an approximate Jeffreys prior. Interestingly, Zellner (1977) also suggested the inverse of this prior,

$$\pi_{RI} = \tfrac{1}{\sigma}(1-\rho^2)^{1/2},$$

based on his "Maximal Data Information Prior" (MDIP) approach, which we also consider for comparison.

3.3 One-Dimensional AR(1) Case

3.3.1 The Temporal AR(1) Model

We consider the linear regression model:

$$Y_t = X_t^T\beta + \epsilon_t, \quad t = 1, ..., n, \tag{3.5}$$

where Y_t represents observation in time t, $X_t = (x_{1t}, ..., x_{rt})^T$ is an r-dimensional vector of explanatory variable associated with time t, and $\beta = (\beta_1, \ldots, \beta_r)^T$ is the vector of unknown regression parameters.

Suppose that the errors $\{\epsilon_t\}$ are stationary and follow a time series AR(1) process, i.e. the errors are generated by the following scheme:

$$\epsilon_1 = u_1/\sqrt{1-\rho^2},$$

$$\epsilon_t = \rho\epsilon_{t-1} + u_t, \quad t = 2, ..., n,$$

where $|\rho| < 1$, assumed to be an unknown autoregressive coefficient; and $\{u_t\}$ is a white noise

process with zero mean and variance σ^2. In addition, we assume the errors have a Gaussian distribution. For the AR(1) model, we know that $\gamma(0) = \text{Var}(\epsilon_t) = \sigma^2/(1-\rho^2)$. Given a sample of n observations, let $Y = (Y_1, Y_2, ..., Y_n)^T$ and $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^T$ be the $n \times 1$ data and error vectors, respectively. Define the $n \times r$ matrix $X = (X_1 X_2 ... X_n)^T$, and assume X is of full rank, i.e. $rank(X) = r$. Then the regression model (1) may be expressed in matrix form as

$$Y = X\beta + \epsilon,$$

with $\text{Cov}(\epsilon) = \sigma^2 V^*$ that $|V^*| = (1-\rho^2)^{-1}$. It can be easily shown that under given assumptions the explicit form of the variance-covariance matrix $V = \text{Cov}(\epsilon)$ is given by

$$V = \sigma^2 V^* = \frac{\sigma^2}{1-\rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix}$$

Note that the matrix V is non-linear in its parameters, especially in the autoregressive coefficient ρ. Under the constraints $|\rho| < 1$ and $\sigma^2 > 0$, matrix V always remains positive definite. As a function of the parameters ρ and $\sigma^2, V(\rho, \sigma^2)$ belongs to the class $\mathcal{C}^2$, i.e. to the class of twice differentiable functions, satisfying the existence condition for $U_i, U_{ij}...$ in Chapter 2. The restricted log likelihood function of Y is

$$\ell_n(\theta) = -\tfrac{n-r}{2}(\log 2 + \log \pi) + \tfrac{1}{2}\log|X^T X| - \tfrac{1}{2}\log|V| - \tfrac{1}{2}\log|X^T V^{-1} X| - \tfrac{G^2}{2}, \qquad (3.6)$$

where G is the generalized residual sum of squares, given by

$$G^2 = G^2(\theta) = Y^T\{V^{-1} - V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}\}Y. \tag{3.7}$$

3.3.2 Fisher Information Matrix for the AR(1) model

Now we are interested in predicting the density function of the unobserved process Y_{n+1}, which is dependent on Y. We can show that Y and Y_{n+1} have joint Gaussian distribution of the form

$$\begin{pmatrix} Y \\ Y_{n+1} \end{pmatrix} \sim N\left[\begin{pmatrix} X\beta \\ x_0\beta \end{pmatrix}, \begin{pmatrix} V(\theta) & w^T(\theta) \\ w(\theta) & v_0(\theta) \end{pmatrix}\right], \tag{3.8}$$

where $\theta = (\rho, \sigma^2)$.

The covariances $V(\theta), w(\theta)$ and $v(\theta)$ are all known functions of an unknown 2 -dimensional parameter vector θ. To simplify the notation, we write V, w and v without indicating the dependence on θ.

To make comparison of Bayesian predictive density and frequentist plug-in type density, we use the same notations as in Chapter 2 (page 22). Superscripts are used to denote components of θ, e.g. θ^i for the ith component. For scalar functions of θ, for example the distribution function $\psi(Y_{n+1}; Y, \theta)$ (with Y_{n+1} and Y fixed for the time being) or $Q(\theta)$ (the logarithm of prior distributions), subscripts will indicate derivative with respect to the components of θ, i.e. $Q_i = \frac{\partial Q}{\partial \theta^i}, \psi_{ij} = \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}$, etc. Also we define $U_i = \frac{\partial \ell_n(\theta)}{\partial \theta^i}, U_{ij} = \frac{\partial^2 \ell_n(\theta)}{\partial \theta^i \partial \theta^j}, U_{ijk} = \frac{\partial^3 \ell_n(\theta)}{\partial \theta^i \partial \theta^j \partial \theta^k}$.

All the above quantities are functions of a particular θ, and could be evaluated at the REML estimator $\hat{\theta}$, which we denote with a hat, such as $\hat{U}_i, \hat{\psi}_{ij}$, and so on. By definition, we have $\{\hat{U}_i\} = 0$, and $\{-\hat{U}_{ij}\}$ is the observed information matrix. If the latter matrix is invertible, we

denote its inverse matrix with superscripts, i.e., if A is the $p \times p$ matrix with (i, j) entry as $\hat{U}_{ij}$, then if A^{-1} exist, $\hat{U}^{ij}$ is its (i, j)th entry. We also use the summation convention, where a repeated index appearing as both a subscript and a superscript in the same formula implicitly indicates a summation over that index.

For simplicity, it is not explicitly indicated that all the quantities depend on dimension n. It is straightforward to show that for the AR(1) model (both temporal and spatial model), all the assumption conditions 1-3 in Chapter 2 are satisfied.

We can write $U_i = n^{1/2} Z_i, U_{ij} = n\kappa_{ij} + n^{1/2} Z_{ij}, U_{ijk} = n\kappa_{ijk} + n^{1/2} Z_{ijk}$, where $\kappa_{ij}, \kappa_{ijk}$ are non-random and Z_i, Z_{ij}, Z_{ijk} are random variables with mean 0, and we assume that all these quantities are of $O(1)$ or $O_p(1)$ as $n \to \infty$.

Also, let $\kappa_{i,j} = E(Z_i Z_j) = -\kappa_{ij}, \kappa_{ij,k} = E(Z_{ij} Z_k)$. By standard identity, $\kappa_{i,j}$ is the (i, j) entry of the normalized Fisher information matrix, we assume this matrix to be invertible with inverse entries $\kappa^{i,j}$. Explicit formulae exist for calculating these quantities, which can be found in Section 8.2 of Smith and Zhu (2004).

If θ is known, then the Best Linear Unbiased Predictor (BLUP) of Y_{n+1} is given by $\hat{y}_{n+1} = \lambda^T y$, where

$$\lambda = V^{-1} w^T + V^{-1} X (X^T V^{-1} X)^{-1} (x_0 - X^T V^{-1} w^T), \tag{3.9}$$

with $w = \frac{\sigma^2}{1-\rho^2}(\rho^n, \rho^{n-1}, \cdots, \rho)$. And the corresponding prediction error variance is given by

$$\sigma_0^2 = v_0 - w^{V^{-1}} w^T + (x_0^T - w V^{-1} X)(X^T V^{-1} X)^{-1} (x_0 - X^T V^{-1} w^T). \tag{3.10}$$

$v_0 = \frac{\sigma^2}{1-\rho^2}$ here.

For simplicity, we can assume $X = \mathbf{1}_n$, and β is an unknown constant. By standard computation, we get

$$
\begin{aligned}
|X^T X|^{1/2} &= n^{1/2}, \\
|V|^{-1/2} &= \tfrac{(1-\rho^2)^{1/2}}{\sigma^n}, \\
|X^T V^{-1} X|^{-1/2} &= \sigma\{(1-\rho)[2+(n-2)(1-\rho)]\}^{-1/2}, \\
\lambda &= \tfrac{1}{2+(n-2)(1-\rho)}(1-\rho, (1-\rho)^2, \cdots, (1-\rho)^2, 1+\rho[1+(n-2)(1-\rho)])^T, \\
\sigma_0^2 &= \tfrac{\sigma^2[2+(n-1)(1-\rho)]}{2+(n-2)(1-\rho)}, \\
W = \{\omega^{\alpha,\beta}\} &= V^{-1} - V^{-1}X(X^T V^{-1} X)^{-1} X^T V^{-1} \\
&= \tfrac{1}{\sigma^2} \begin{pmatrix} 1 & -\rho & 0 & \cdots & 0 & \\ -\rho & 1+\rho^2 & -\rho & 0\cdots & 0 & \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & -\rho & 1+\rho^2 & -\rho \\ 0 & \vdots & \vdots & \vdots & -\rho & 1 \end{pmatrix} \\
&\quad - \tfrac{1-\rho}{\sigma^2[2+(n-2)(1-\rho)]} \begin{pmatrix} 1 & 1-\rho & \cdots & 1-\rho & 1 \\ 1-\rho & (1-\rho)^2 & \cdots & (1-\rho)^2 & 1-\rho \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1-\rho & (1-\rho)^2 & \cdots & (1-\rho)^2 & 1-\rho \\ 1 & 1-\rho & \cdots & 1-\rho & 1 \end{pmatrix}.
\end{aligned}
$$

Furthermore, we can get the normalized Fisher information matrix

$$I(\theta) = \{\kappa^{i,j}\}_{2\times 2} = \begin{pmatrix} \frac{1}{2\sigma^4} & 0 \\ 0 & \frac{1}{1-\rho^2} \end{pmatrix}.$$

The general expression for the Fisher Information matric on the AR(p) model has been provided by Tanaka and Komaki (2005), by using another coordinate system (see Amari (1987)).

3.3.3 Comparison of Noninformative Priors and Estimative Method

We consider the following candidate noninformative priors:

(i) Jeffreys prior: based on the Fisher information matrix $I(\theta)$, we can derive the Jeffreys prior in this model is

$$\pi_J(\theta) \propto |I(\theta)|^{1/2} \propto \delta^{-1}(1-\rho^2)^{-1/2},$$

where $\delta = \sigma^2$.

(ii) Reference prior: Inspired by the symmetrized reference prior in Berger and Yang (1994), we consider a noninformative prior as follows

$$\pi_R(\theta) = \frac{1}{\pi\sigma\sqrt{1-\rho^2}}.$$

This prior is proper with respect to ρ.

(iii) Inverse reference prior: $\pi_{RI}(\theta) \propto \sigma^{-1}(1-\rho^2)^{1/2}$, using Maximal data information prior densities (MDIPs) approach suggested by Zellner (1977), which is similar to the Jeffreys prior in the sense that it is also invariant to linear transformations of the data and parameters.

Theorem 3.3.1. *Within the AR(1) model, we have the following results for the above given*

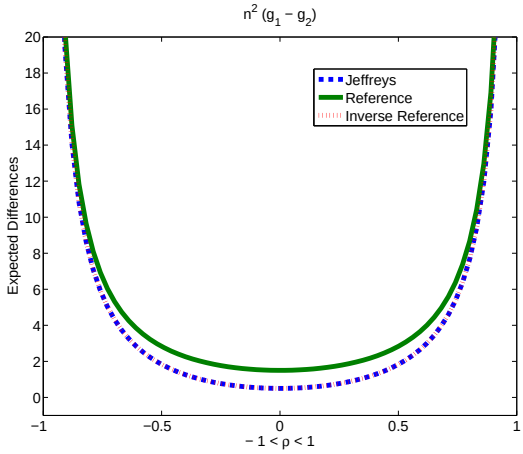

Figure 3.1: Plots of $n^2(g_1 - g_2)$ against $\rho \in (-1, 1)$ as $n \to \infty$.
Dashed: Jeffreys prior. Solid: Reference Prior. Dotted: Inverse Reference prior

priors:

(1) The Jeffreys prior is second-order KL REML-dominant.

(2) The reference prior is second-order KL REML-dominant.

(3) The inverse reference prior is also second-order KL REML-dominant.

(4) The reference prior is second-order KL dominant to the Jeffreys prior, and better than the Inverse prior in most stationary cases.

Proof: Let $a_{\cdot} = \lim_{n\to\infty} n^2(g_1 - g_2)$, we have

$$a_J = \tfrac{1+7\rho^2}{2(1-\rho^2)} > 0,$$

$$a_R = (3 + 5\rho^2)/(2 - 2\rho^2) > 0,$$

$$a_{IR} = (500 + 3501\rho^2)/[1000(1 - \rho^2)] > 0,$$

$$\lim_{n\to\infty} n^2(E_{Y|\theta}[D(\varphi, \tilde{\varphi}_J) - D(\varphi, \tilde{\varphi}_R)]) = a_R - a_J = 1 > 0,$$

$$\lim_{n\to\infty} n^2(E_{Y|\theta}[D(\varphi, \tilde{\varphi}_{IR}) - D(\varphi, \tilde{\varphi}_R)]) = a_R - a_{IR} = \tfrac{1000-1001\rho^2}{1000(1-\rho^2)} > 0 \text{ for } \rho \in (-0.9995, 0.9995).$$

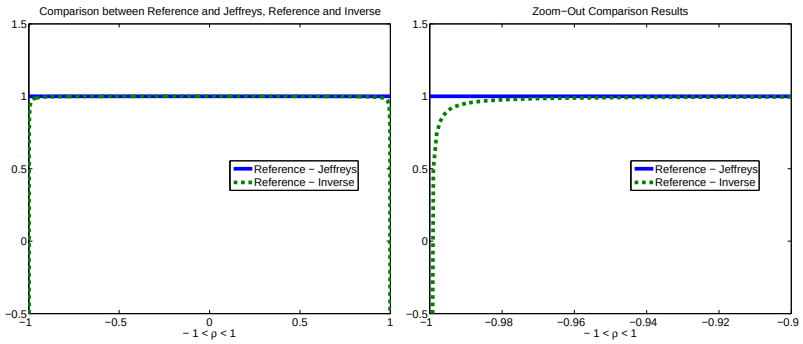

Figure 3.2: Left: Comparison between reference prior and Jeffreys prior (Solid), and reference prior and inverse prior (Dashed). Right: Zoom Out Image

Because all the (i) - (iii) candidate priors dominate the REML estimative density, we compare the two priors by the difference between the corresponding expected KL divergence differences (to the second order) in (4). In particular, comparisons are constructed between reference prior and the Jeffreys prior, the reference prior and its inverse prior, respectively. Note that, positive values of equations, e.g. equation (3.12), indicate the Bayesian predictive density based on the reference prior is closer to the true conditional density than the other one is. Clearly, we can see within this model, the reference prior dominates the other two candidate priors (Jeffreys prior and the inverse reference prior) in the sense of averaged KL divergence. Berger and Yang (1994) also concluded a similar result: $\pi_R(\theta)$ is superior to $\pi_J(\theta)$ with respect to the mean squared error criterion.

In the left panel of Figure 3.2, we plot the asymptotic differences related to the Jeffreys prior and Reference prior, and the Inverse Reference prior and Reference prior, respectively, for $\rho \in (-1, 1)$ and $\sigma^2 \in (0, 1)$. Zoom-Out version is also involved in the right panel.

3.3.4 Simulation results

Our simulation work goes as follows:

REML estimator for (ρ, σ^2)

To evaluate $E_{Y|\theta}(D(\varphi, \widehat{\varphi}))$, we need to plug in the REML estimator from every given data vector and specific $\theta = (\rho, \sigma^2)$ in each iteration. For the temporal AR(1) model, the restricted log-likelihood for $\theta = (\rho, \sigma^2)$ can be written as

$$\ell_n(\rho, \sigma^2) = -\tfrac{n-r}{2} \log(\sigma^2) + \tfrac{1}{2} \log(1 - \rho^2) - \tfrac{1}{2\sigma^2} S(\rho, \tilde{\beta}) - \tfrac{1}{2} \log |X^T V^{-1} X|, \tag{3.11}$$

where the sum of squares function

$S(\rho, \beta) = (Y - X\beta)^T V^{-1} (Y - X\beta)$, and $\tilde{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y$ is the GLS estimator of β. The restricted likelihood equation for σ^2 is

$$\tfrac{\partial \ell_n}{\partial \sigma^2} = -(n - r)/(2\sigma^2) + S(\rho, \tilde{\beta})/(2\sigma^4) = 0, \tag{3.12}$$

which leads to $\hat{\sigma}^2 = S(\rho, \tilde{\beta})/(n - r)$ as the REML estimator of σ^2.

From (3.14), the restricted likelihood equation for ρ, is given by

$$\tfrac{\partial \ell_n}{\partial \rho} = -\tfrac{\rho}{1-\rho^2} - \tfrac{1}{2\sigma^2} \partial S(\rho, \tilde{\beta})/\partial \rho - \tfrac{1}{2} \partial \log |X^T V^{-1} X|/\partial \rho = 0, \tag{3.13}$$

which, with (10) together can be solved simultaneously after substituting for $\hat{\sigma}^2$ and $\hat{\rho}^2$.

We check that $\frac{\partial^2 \ell_n(\theta)}{\partial \rho}, \frac{\partial^2 \ell_n(\theta)}{\partial \sigma^2} < 0$, and the standard deviation for these above REML estimators goes asymptotically to 0 as $n \to \infty$.

Sampling for (ρ, σ^2)

To calculate the predictive density function, we need to generate $\rho^*, \sigma^{2(*)}$ from the posterior distributions. For all the Jeffreys , reference and reference inverse priors, the marginal posterior

is complicated. However, Metropolis-Hastings algorithm can be used to generate the posterior distribution for θ, which is analog to the work in Chapter 2. We burn in 500 out of 1000 simulations to make sure there is no influence of the initial values for ρ and σ^2. Only 500 variates have been used to approximate the posteriors, from which we select one in every ten and make records of 50 pairs of (ρ, σ^2). In addition, the convergence is justified by the result of Gelman and Rubin's convergence diagnostic in the "CODA" package of R language.

Simulation

Here we compare the above three priors and also the REML plug-in density by utilizing the following simulation for the time series AR(1) model from (3.8).

First of all, since the REML estimators of θ are invariant with respect to the true value of β, without loss of generality, we take the regression coefficients as $\beta = 0$ in generating the simulated response data $Y = \epsilon$.

Fix $\sigma^2 = 1$. Set $n = 5, 50, 100, 200, 300, 400, 500$ and $\rho = -0.9, -0.8, ..., 0, 0.1, ..., 0.9$. For every fixed pair of (n, ρ), we evaluate $E_{Y|\theta}[D(\varphi, \widehat{\varphi}) - D(\varphi, \tilde{\varphi})]$ for fixed θ by the following steps.

Step 1: Generate 100 groups of observation $Y = (Y_1, ..., Y_n)$ under the AR(1) process.

Step 2: For each specific observation set, Y, compute the REML estimator $\hat{\theta} = (\hat{\rho}, \hat{\sigma}^2)$ from equations (3.12) and (3.13), and the corresponding REML predictive density function is given by $\hat{\varphi}(Y_{n+1}; Y) = \varphi(Y_{n+1}; Y, \hat{\theta})$.

Step 3: Approximate $\tilde{\varphi}(Y_{n+1}; Y)$ by $\frac{1}{M} \sum_i \varphi(Y_{n+1}; Y, \theta^i)$, where θ^i is generated as described in previous subsection, with respect to Jeffreys, reference and reference inverse priors, respectively.

Step 4: The difference between $D(\varphi, \hat{\varphi})$ and $D(\varphi, \tilde{\varphi})$ is approximated by quadrature integration method.

Step 5: To calculate the expected KL divergence for fixed θ, we approximate it by $\frac{1}{L}\sum_l (D(\varphi, \widehat{\varphi}|Y^{(l)}, \theta) - D(\varphi, \tilde{\varphi}|Y^{(l)}, \theta))$, where the summation is taken over $Y^{(l)}$.$L = 100$ is also justified by the convergence diagnostic in "CODA" package in R.

The numerical results (with the confidence interval for each approximation) are plotted in Figure 3.3. These results illustrate the validity of our method in a practical application.

In summary, we simulate some temporal AR(1) data, approximating the expected K-L divergence by MCMC method. The simulation results can be seen in Figure 3.3, note that for the Jeffreys prior and the inverse reference prior, when n is larger than 500, we can achieve positive comparison result. For the reference prior, the result is positive when $n \geq 400$.

3.4 Two-Dimensional AR(1) Case

3.4.1 Spatial AR(1) model

Besides the temporal AR(1) model, we also consider a spatial process defined on a regular rectangular grid of size $m \times m$ in two dimensions with sites labeled (i, j). Let Y_{ij} be the response variable at location (i, j), and suppose the Y_{ij} follow a linear regression model of the form

$$Y_{ij} = X_{ij}^T \beta + \epsilon_{ij}, \tag{3.14}$$

with $i, j = 1, ..., n$. $X_{ij} = (x_{ij1}, ..., x_{ijr})^T$ is an r-dimensional vector of explanatory variables associated with the location (i, j), and $\beta = (\beta_1, ..., \beta_r)^T$ is the unknown regression parameter vector. The errors $\{\epsilon_{ij}\}$ are assumed to be stationary and follow a spatial multiplicative first-

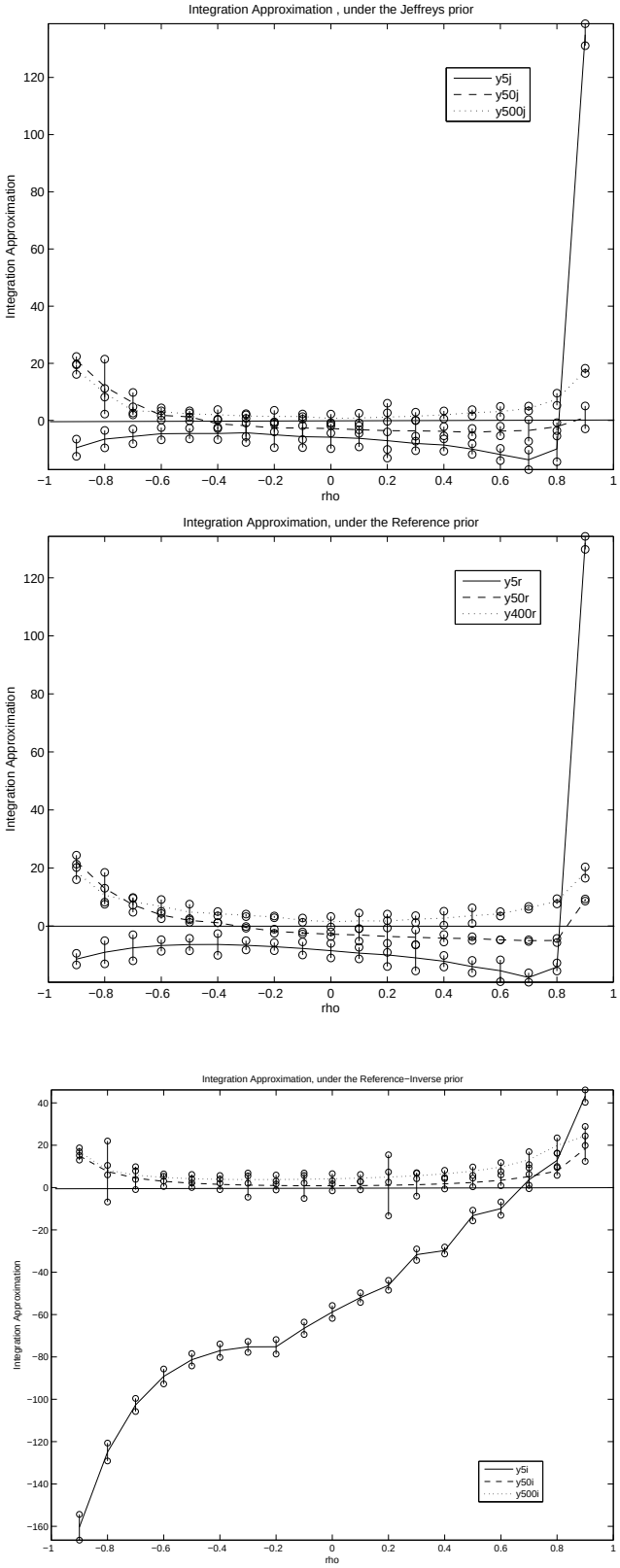

Figure 3.3: Simulation results for $n^2 E_{Y|\theta}[D(\varphi;\hat{\varphi}) - D(\varphi;\tilde{\varphi})]$, where $\tilde{\varphi}$ is constructed under the Jeffreys prior, the reference prior and the inverse reference prior, respectively

order autoregressive model, considered by Martin (1990) and Basu and Reinsel (1993):

$$\epsilon_{ij} = \rho_1 \epsilon_{i-1,j} + \rho_2 \epsilon_{i,j-1} - \rho_1 \rho_2 \epsilon_{i-1,j-1} + u_{ij}, \tag{3.15}$$

where $|\rho_k| < 1, k = 1, 2$, and the $\{u_{ij}\}$ are independent random variables with mean zero and variance σ^2. We also assume $\{u_{ij}\}$ are from Gaussian distribution. The bivariate covariance of the process $\{\epsilon_{ij}\}$ at spatial lag (s, t) is given by

$$\mathrm{Cov}(\epsilon_{ij}, \epsilon_{i-s,j-t}) = \gamma \rho_1^{|s|} \rho_2^{|t|}, s, t \in \mathbf{Z},$$

where $\gamma = Var(\epsilon_{ij}) = \sigma^2/\Delta$, with $\Delta = (1 - \rho_1^2)(1 - \rho_2^2)$. Given a sample of $n = m^2$ observations, let $Y = (Y_{11}, Y_{21}, \ldots, Y_{m1}, \ldots, Y_{1m}, \ldots, Y_{mm})^T$ and $\epsilon = (\epsilon_{11}, \epsilon_{21}, \ldots, \epsilon_{m1}, \ldots, \epsilon_{1m}, \ldots, \epsilon_{mm})^T$ be the $n \times 1$ data and error vectors, respectively. Define the $n \times r$ matrix $X = [X_{11} X_{21} \ldots X_{m1} \ldots X_{1m} \ldots X_{mm}]^T$, and assume X is of full rank r. Then the regression model can be expressed to the matrix form as

$$Y = X\beta + \epsilon,$$

with $\mathrm{Cov}(\epsilon) = \sigma^2 V^*$. Actually, $V^* = V_2 \bigotimes V_1$, where V_k is an $m \times m$ matrix with (i, j)th element $\rho_k^{|i-j|}/(1 - \rho_k^2), k = 1, 2$, and $\bigotimes$ is notation of the Kronecker product. $V_k^{-1} = P_k P_k^T$ has a special patterned form. In particular, $|V^*| = |V_1|^m |V_2|^m = (1 - \rho_1^2)^{-m}(1 - \rho_2^2)^{-m}$. We can get the restricted log-likelihood function of Y as

$$\ell_N(\theta) = -\tfrac{n-r}{2}(\log 2 + \log \pi) + \tfrac{1}{2} \log |X^T X| - \tfrac{1}{2} \log |V| - \tfrac{1}{2} \log |X^T V^{-1} X| - \tfrac{G^2}{2}, \tag{3.16}$$

where G is the generalized residual sum of squares, given by

$$G^2 = G^2(\theta) = Y^T\{V^{-1} - V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}\}Y. \tag{3.17}$$

3.4.2 Fisher Information Matrix for the AR(1) Model

Now we are interested in predicting the density function of the unobserved process $Y_{n+1,1}$, which is dependent on Y. We can show that Y and $Y_{n+1,1}$ have joint Gaussian distribution of the form

$$\begin{pmatrix} Y \\ \\ Y_{n+1,1} \end{pmatrix} \sim N\left[\begin{pmatrix} X\beta \\ \\ x_0\beta \end{pmatrix}, \begin{pmatrix} V(\theta) & w^T(\theta) \\ \\ w(\theta) & v_0(\theta) \end{pmatrix}\right], \tag{3.18}$$

where $\theta = (\rho_1, \rho_2, \sigma^2)$.

The covariances $V(\theta), w(\theta)$ and $v(\theta)$ are all known functions of an unknown 3 -dimensional parameter vector θ. To simplify the notation, we write V, w and v without indicating the dependence on θ.

To make comparison of Bayesian predictive density and frequentist plug-in type density, we use the same notations as in Section 3.2 and Chapter 2 (page 22). It is easy to show that all the assumption conditions in Chapter 2 are validated.

The following derivations are quite similar to the previous part in this Chapter. If θ is known, then the Best Linear Unbiased Predictor (BLUP) of $Y_{m+1,1}$ is given by $\hat{Y}_{m+1,1} = \lambda^T Y$, where

$$\lambda = V^{-1}w^T + V^{-1}X(X^TV^{-1}X)^{-1}(x_0 - X^TV^{-1}w^t), \tag{3.19}$$

with $w = \frac{\sigma^2}{(1-\rho_1^2)(1-\rho_2^2)}(\rho_1^m, \rho_1^m\rho_2, \ldots, \rho_1^m\rho_2^{m-1}, \rho_1^{m-1}, \ldots, \rho_1^{m-1}\rho_2^{m-1}, \ldots, \rho_1, \ldots, \rho_1\rho_2^{m-1})^T$. And

the corresponding prediction error variance is given by

$$\sigma_0^2 = v_0 - wV^{-1}w^T + (x_0^T - wV^{-1}X)(X^TV^{-1}X)^{-1}(x_0 - X^TV^{-1}w^T), \tag{3.20}$$

with $v_0 = \gamma = Var(\epsilon_{ij}) = \sigma^2/\Delta$, with $\Delta = (1-\rho_1^2)(1-\rho_2^2)$.

For simplicity, we can assume $X = \mathbf{1}_n$, and β is an unknown constant. By standard computation, we get

$$\begin{aligned}
|X^TX|^{1/2} &= n^{1/2} = n, \\
|V|^{-1/2} &= \tfrac{(1-\rho_1^2)^{m/2}(1-\rho_2^2)^{m/2}}{\sigma^n}, \\
|X^TV^{-1}X|^{-1/2} &= \sigma\{(1-\rho_1)(1-\rho_2)[2+(n-2)(1-\rho_1)(1-\rho_2)]\}^{-m/2}, \\
\lambda &= \tfrac{1}{[2+(n-2)(1-\rho_1)(1-\rho_2)]}((1-\rho_1)(1-\rho_2), (1-\rho_1)^2(1-\rho_2)^2, \\
&\quad \ldots, (1-\rho_1)^2(1-\rho_2)^2, 1+(\rho_1+\rho_2)[1+(n-2)(1-\rho_1)(1-\rho_2)])^T, \\
\sigma_0^2 &= \tfrac{\sigma^2[2+(n-1)(1-\rho_1)(1-\rho_2)]}{[2+(n-2)(1-\rho_1)(1-\rho_2)]},
\end{aligned}$$

For the spatial AR(1) model, we can get the normalized Fisher information matrix

$$I(\theta) = \{\kappa^{i,j}\}_{3\times 3} = \begin{pmatrix} \frac{1}{2\sigma^4} & 0 & 0 \\ 0 & \frac{1}{1-\rho_1^2} & 0 \\ 0 & 0 & \frac{1}{1-\rho_2^2} \end{pmatrix}.$$

3.4.3 Comparison of Noninformative Priors and Estimative Method

As a candidate noninformative prior, we have encountered the following:

(i) By similar process as above for the temporal AR(1) model, we can get the Jeffreys prior for

this spatial process as

$$\pi_{SJ}(\theta) \propto |I(\theta)|^{1/2} \propto \delta^{-1}(1-\rho_1^2)^{-1/2}(1-\rho_2^2)^{-1/2},$$

where $\delta = \sigma^2$.

(ii) The reference prior

$$\pi_{SR}(\theta) \propto \sigma^{-1}(1-\rho_1^2)^{-1/2}(1-\rho_2^2)^{-1/2}.$$

(iii) The inverse of the reference prior

$$\pi_{SIR}(\theta) \propto \sigma^{-1}(1-\rho_1^2)^{1/2}(1-\rho_2^2)^{1/2}.$$

Theorem 3.4.1. *Within the spatial AR(1) model, we have*

(1) The Jeffreys prior is second-order REML-KL dominant.

(2) The reference prior is second-order KL REML-dominant.

(3) The inverse reference is second-order KL REML-dominant.

(4) The reference prior is second-order KL dominant to the Jeffreys prior, and better than the Inverse prior in most stationary cases.

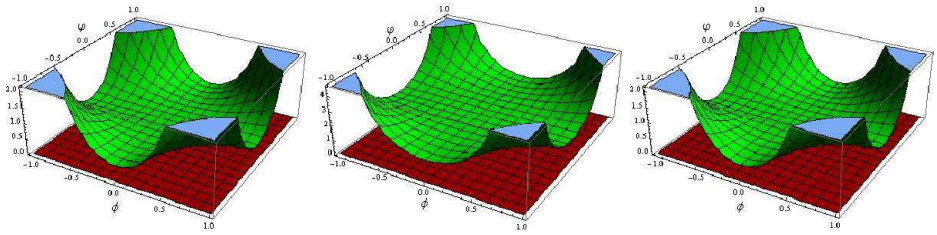

Figure 3.4: Plots of $n^2(g_1 - g_2)$ against $\rho_1 \in (-1, 1)$, $\rho_2 \in (-1, 1)$ as $n \to \infty$. Left: Jeffreys prior. Mid: Reference Prior. Right: Inverse Reference prior

Proof: Let $a_{\cdot} = \lim_{n\to\infty} n^2(g_1 - g_2)$, we have

$$a_{SJ} = \frac{1+7(\rho_1^2+\rho_2^2-\rho_1^2\rho_2^2)}{2(1-\rho_1^2)(1-\rho_2^2)} > 0,$$

$$a_{SR} = \frac{3+5(\rho_1^2+\rho_2^2-\rho_1^2\rho_2^2)}{2(1-\rho_1^2)(1-\rho_2^2)} > 0,$$

$$a_{SIR} = \frac{500+3501(\rho_1^2+\rho_2^2-\rho_1^2\rho_2^2)}{1000(1-\rho_1^2)(1-\rho_2^2)} > 0,$$

$$\lim_{n\to\infty} n^2(E_{Y|\theta}[D(\varphi, \tilde{\varphi}_{SJ}) - D(\varphi, \tilde{\varphi}_{SR})]) = a_{SR} - a_{SJ} = 1 > 0,$$

$$\lim_{n\to\infty} n^2(E_{Y|\theta}[D(\varphi, \tilde{\varphi}_{SIR}) - D(\varphi, \tilde{\varphi}_{SR})]) = a_{SR} - a_{SIR} = \frac{1000-1001(\rho_1^2+\rho_2^2-\rho_1^2\rho_2^2)}{1000(1-\rho^2)(1-\rho_2^2)} > 0, \text{ for most}$$

$$\rho_k \in (0, 1), k = 1, 2.$$

When σ^2 is fixed (to be 1), we get the comparison results for finite sample size as follows:

3.4.4 Simulation Results

We simulate some spatial AR(1) data with fixed $\sigma^2 = 1$, approximating the expected value by MCMC method. The simulation steps are analog to that in the previous section (set $\rho_1, \rho_2 = -0.9, -0.8, \ldots, 0, 0.1, \ldots, 0.9$), except now we are simulating 100 groups of $n = m \times m$ spatial AR(1) observation for $m = 5, 10, 11, 12, \ldots, 30$ respectively (we do simulation for each fixed n), instead of one-dimensional $AR(1)$ process. The simulation results can be seen in Figure 3.5, note that for the Jeffreys prior and the inverse reference prior, when n is larger than 20×20,

we can achieve positive comparison result. For the reference prior, the simulaiton numerical value is positive when $n \geq 12 \times 12$.

3.5 Conclusions

We investigate how to predict the conditional density within the AR(p) process in the Bayesian framework. As a starting point, we consider the case of AR(1) model. We introduce some noninformative priors, all of which are proved to be second-order KL REML-dominant. We also compare among the three candidate priors themselves, illustrating that in the asymptotic sense the reference prior is superior to the other two ones, in both the temporal and spatial AR(1) models. We simulate data for both the time series and spatial AR(1) models, of which the results agree with the asymptotic study when the sample size is moderately large. We also notice that, when the sample size is very small, the simulation numerical values differ substantially from the asymptotic second-order approximation, possibly due to the biases of higher-order with respect to averaged KL divergence. In particular, the smaller the sample size is, the greater the influence of higher-order biases are.

Our attention has only been paid to the AR(1) model now. For the general AR(p) model, especially when the order is moderately large, there is another way to estimate the model: utilizing the Bayesian estimation of the spectral density of the model, e.g. Tanaka and Komaki (2005) showed that in i.i.d. case the Bayesian estimation of spectral densities based on a superharmonic prior (if exists) asymptotically dominate thosed based on the Jeffreys prior, using the asymptotic expansion of the risk difference. Tanaka and Komaki (2008) focused on the AR(2) process and proposed an explicit form of such a superharmonic prior. On the other hand, the moving average (MA) model is also one of the most important models in data analysis.

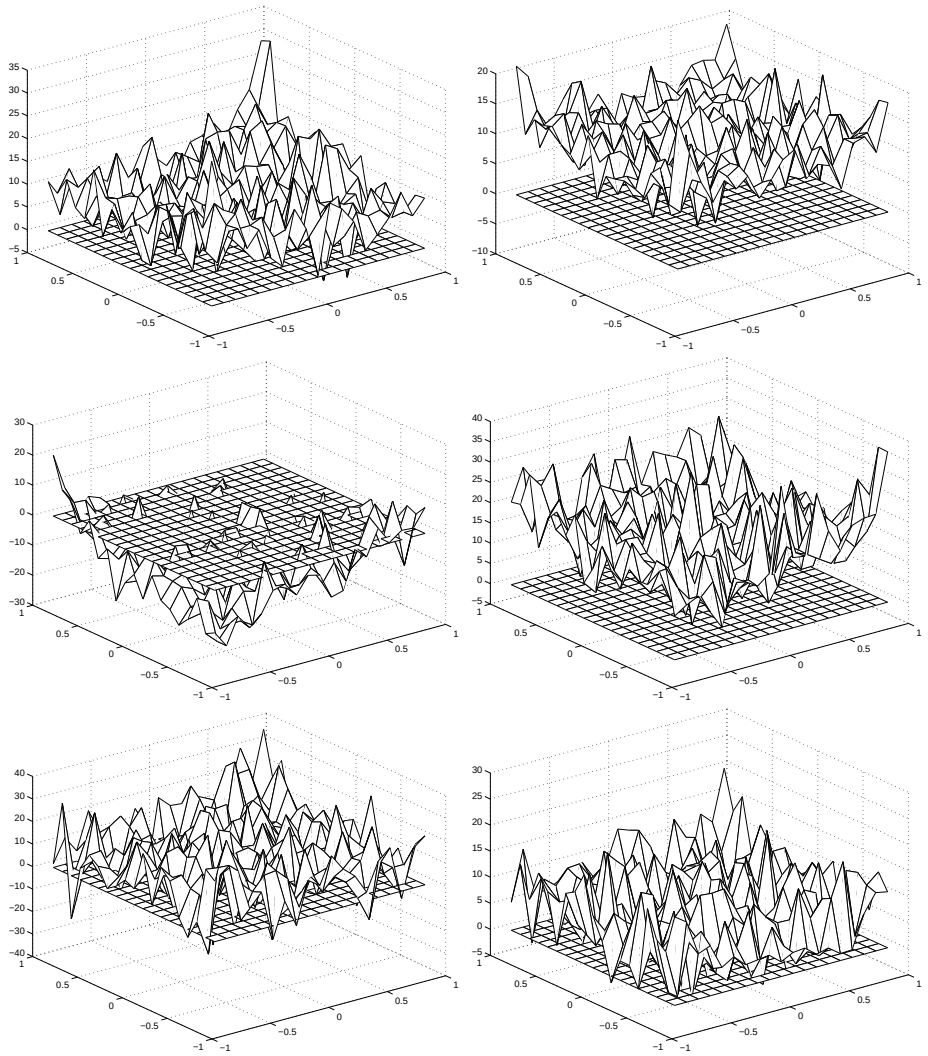

Figure 3.5: Simulation results for $N^2 E_{Y|\theta}[D(\varphi;\hat{\varphi}) - D(\varphi;\tilde{\varphi})]$, where $\tilde{\varphi}$ is constructed under the Jeffreys prior (Top Row), the reference prior (Middle Row) and the inverse reference prior (Bottom Row), respectively. Left: $n = 5 \times 5$. Right: $n = 20 \times 20, 12 \times 12, 20 \times 20$, respectively.

The MA models are completely different from the AR models as a stochastic process and in the information geometrical viewpoint they are known to have different structures. We can also consider the ARMA model for the most general situation. In addition, it seems that the result of the special spatial AR(1) model relies heavily on the fact that the covariance function is just a Kronecker product of two AR(1) time series and therefore, a general spatial process would still be quite a bit harder to characterize. We will explore this problem in the future.

Chapter 4

Estimation and Prediction with Errors in Covariances

Many statistical data analysis involve covariance matrices or their estimations, e.g. kriging in spatial statistics, time series analysis, multivariate data analysis techniques, etc. In this chapter, we want to explore parameter estimation and kriging prediction problems with measurement errors. Some related work on covariance estimation is reviewed, and a motivational example is given. Furthermore, we introduce a framework for this type of problems. Some theoretical results on the effect of regularized covariance estimation on parameter estimation and preliminary results of its effect on kriging prediction are provided.

4.1 Introduction

Estimation of covariance matrices in small samples has been studied by many authors. Standard estimators, like the unstructured sample covariance matrix, can be very unstable when the sample size is relatively small as compared to the data dimension. In that case, the smallest estimated eigenvalues tend to be too small. Numerous papers have explored better alternative estimators for covariance matrices, in both the frequentist and Bayesian frameworks (see for

example, James and Stein (1961), Haff (1977), Chen (1979), Haff (1980), Haff (1991) and Daniels and Kass (1999)). Many of these estimators gave substantial risk reductions compared to the sample covariance estimator in small sample sizes. A common underlying property of many of these estimators is that, they are shrinkage estimators in the sense of James-Stein (see James and Stein (1961) and Stein (1956)). In particular, the Bayesian approach often yields estimators which "shrink" towards a structure associated with a pre-specified prior. One of the first papers to exploit this idea is Chen (1979) who showed that if the prior used on the inverse covariance matrix is the standard conjugate, i.e. a Wishart distribution, then for an appropriate choice of the shape (or shrinkage) and scale hyperparameters, the posterior mean for the covariance matrix is a linear combination of the sample covariance matrix and the prior mean, as Rajaratnam et al. (2008) showed that the eigenvalues of such estimators are also shrinkage estimators of the eigenvalues of covariance matrix. Daniels and Kass (1999) shrunk the matrix toward a diagonal structure and obtained estimates (and posterior distributions) using combinations of importance sampling and Markov chain Monte Carlo (MCMC). For a similar approach in the context of a covariance function in time series data analysis, see Daniels and Cressie (1999). Daniels and Kass (2001) extended their previous work, by considering two general shrinkage approaches to estimate the covariance matrix and regression coefficients with correlated (or longitudinal) data. One method was based on shrinking the eigenvalues of the unstructured ML or REML estimator. The second involved shrinking an unstructured estimator toward a structured estimator. For both cases, the amount of shrinkage was data-driven. Both estimators were consistent and gave consistent and asymptotically efficient estimates for regression coefficients. Finally, they proposed a combination of both shrinkage approaches, i.e., shrinking the eigenvalues and then shrinking toward structure.

There are several other ways to shrink and estimate the covariance matrices. One approach is to apply the Cholesky decomposition as in Daniels and Pourahmadi (2002), or a modified Choleskey factorization (see Wu and Pourahmadi (2003) and Huang et al. (2007)), which was referred as regularized MLE. Another way is to use the penalized likelihood method, which shrinks the Choleskey factor by adding an L_p penalty to the negative log-likelihood function. Relatively few work has been done on shrinkage estimation for spatial problems. Zhu and Liu (2007) described a penalized likelihood method for estimating the spatial covariance structure. The effect of shrinkage covariance estimation on kriging prediction has not been studied before and will be the focus of our current work.

4.1.1 A motivating example

Our current work is motivated by the problem of estimating the regional trend of sulfur-dioxide. The problem faced by the Environmental Protection Agency (EPA) and other environmental organizations is how to characterize the trend in certain pollutants. For example, during the 1980's and 1990's there was a concerted effort to solve the acid rain problem by reducing levels of sulfur in the atmosphere, the EPA were interested in measuring trends of gaseous sulfur dioxide (SO2) during this period. However, the trend differs from one place to another, and there is an additional complication in that the measured SO2 at any particular time depends on a number of factors unrelated to long-term trends: there is a clear seasonal pattern and it is also affected by temperature, humidity, wind speed and direction, as well as the long-term trend. To take account of this, Holland et al. (2000) fitted a two-stage model, where in the first stage they fitted a generalized additive model (GAM), in which the seasonal and annual factors as well as various meteorological variables were modeled nonparametrically; while the long-term

trend was estimated separately for each site, with the estimated standard error. In the second stage of the model, a spatial process was assumed for the true underlying trend, in which the spatial parameters were estimated followed by kriging to obtain an optimal reconstruction of the true spatial trend surface.

In particular, in the second stage, $Z(s)$ was modeled as a spatial process at any location s, with

$$E\{Z(s)\} = \sum_{j=1}^{q} \beta^j X_j(s), \quad \text{Cov}\{Z(s), Z(u)\} = \alpha M_\gamma(\|s-u\|), \tag{4.1}$$

where β is a q-vector of unknown regression parameters, X_j's are known functions of location s, $\alpha = \text{Var}\{Z(s)\}$, $\quad M_\gamma(\cdot,\cdot)$ is a correlation function in R^2 parameterized by γ, for instance, γ could be the range parameter for the exponential, spherical, or the Gaussian correlation function, or the range and smoothness parameters for the Matern correlation function. $\|s-u\|$ denotes the Euclidean distance between sites s and u. Assume the covariance function parameter $\theta = (\alpha, \gamma)^T$ in this example. They assumed that the true value of Z is unobserved and unknown, but estimated by $Y(s_i)$, along with the standard error $\tilde{\sigma}_i$ for each site s_i. The variables were related by the equation

$$Y(s_i) = Z(s_i) + e_i, \tag{4.2}$$

where $e_i \sim N(0, \tilde{\sigma_i}^2)$ were interpreted as measurement errors, independent of the random field $\{Z(\cdot)\}$. Denote $\text{Var}(e) = R$, in which e_i's could be either independent or correlated. In their paper, they considered two estimators of R,

(i) where R was assumed to be diagonal with entries $\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_n^2$ (simple but not so realistic), and

(ii) where R was simply estimated by the sample covariance of the regression errors.

Suppose $Z = \{Z_i\} = \{Z(s_i)\}, i = 1, ..., n$, where s_i is the site for the spatial data. To make inference about the value of Z at unmonitored site s_0, they applied an extension of kriging analysis to estimate θ and make empirical prediction of $Z(s_0)$, in which the covariance matrix of measurement error, R, was replaced by an estimator $\hat{R}$ from the first stage.

Here is how they estimated the covariance function parameter θ: since equations (4.1) and (4.2) provided a hierarchical model for the "data", $Y = (Y(s_1), \ldots, Y(s_n))^T$, they directly considered the log-likelihood of the model parameters, which is given by

$$\ell(\beta, \alpha, \gamma; Y) = -\tfrac{n}{2}\log(2\pi) - \tfrac{1}{2}\log(|D|) - \tfrac{1}{2}(Y - X\beta)^T D^{-1}(Y - X\beta), \tag{4.3}$$

where X is the known $n \times q$ matrix of regressors, $D = \mathrm{Var}(Y) = K(\theta) + R$, with $K(\theta)_{ij} = \alpha M_\gamma(\|s_i - s_j\|)$ and $R = \mathrm{Var}(e)$. The approximate maximum likelihood estimators for model parameters β, α and γ can be obtained from equation (4.3), by replacing R by $\hat{R}$.

Next, given values of $Y(s_i), i = 1, \ldots, n$, the kriging predictor of $Z(s_0)$ at unmonitored location s_0 is

$$\hat{Z}(s_0) = x_0^T\hat{\beta} + \tau_0^T D^{-1}(Y - X\hat{\beta}) \tag{4.4}$$

where x_0 is a known vector, $\hat{\beta} = (X^T D^{-1} X)^{-1} X^T D^{-1} Y$, and $\tau_0 = (\mathrm{Cov}\{Z(s_0), Y(s_1)\}, \ldots,$ $\mathrm{Cov}\{Z(s_0), Y(s_n)\})^T$, with $\mathrm{Cov}\{Y(s), Z(u)\} = \alpha M_\gamma(||s - u||)$.

The mean-squared prediction error (MSPE) of $\hat{Z}(s_0)$ was given by

$$\sigma^2(s_0) = \alpha - \tau_0^T D^{-1}\tau_0 + (x_0 - X^T D^{-1}\tau_0)^T (X^T D^{-1} X)^{-1}(x_0 - X^T D^{-1}\tau_0). \tag{4.5}$$

Clearly, both $\hat{Z}$ and $\sigma^2(s_0)$ depend on R and θ, and they were replaced by the corresponding

estimators to obtain the empirical BLUP and MSPE in equations (4.4) and (4.5).

In summary, the kriging predictor was obtained, based on estimate of $D = K(\theta) + R$, where $K(\cdot)$ is a known function of the parameter vector θ and R is the covariance matrix of some measurement error that is not parametrically constrained. In order to estimate θ, an estimator of R has to be supplied, then kriging prediction can be done by replacing D with the estimator $\hat{D} = K(\hat{\theta}) + \hat{R}$. In their work, they used the diagonal estimator and sample covariance matrix as two alternative ways to estimate R. A natural question is: can we find a better estimator of R, which will reduce the uncertainty in covariance parameter estimation and kriging prediction?

4.1.2 Our Work

We provide a framework to study the effect of regularized covariance estimators on parameter estimation and kriging prediction. Inspired by the work of Ledoit and Wolf (2004), we consider the following estimator of the measurement error covariance-variance matrix R,

$$\hat{R}_\nu = \nu S + (1 - \nu) S^*, \quad \nu \in [0, 1],$$

where S is the sample covariance matrix and S^* is the diagonal estimator. The sample covariance and diagonal estimator are two special cases of $\hat{R}_\nu$, with $\nu = 1$ and 0, respectively.

Theoretical results are provided for the parameter estimation for a mixed effect model, and certain preliminary results are provided for kriging prediction. Denoting the covariance parameter (vector) by θ and assuming θ^i is the ith element, we show that the estimation bias, $\hat{\theta}^i - \theta^i$, depends on the first and the second order moments of the first order derivative of the plug-in restricted log-likelihood function and the first order moment of the second order derivative, where "plug-in" means replacing R by $\hat{R}_\nu$ whenever R appears in the definition of

these functions or derivatives. As a starting point, we consider a model with an exponential covariance function K, and measurement error matrix R , which is set to be random, exponential or long range dependent, respectively (see page 78 - 80 in Section 4.3.2). We find that certain linear combination of the sample covariance and the diagonal estimator will result in much smaller mean squared error (MSE) of the plug-in REML estimator $\hat{\theta}$. Simulation results also indicate that there always exists certain $\nu^* \in (0, 1)$ such that the approximate MSE of resulting $\hat{\theta}$ based on $\hat{R}_{\nu^*}$ is the substantially smaller than that based on the sample covariance. In the future, we will investigate if there is a unique and explicit way to express this ν^*, in terms of K, R, S and S^*.

In Section 4.2, we introduce the framework of a general model which captures the essential features of the problem and allows for more explicit calculations. In Section 4.3, we derive the asymptotic approximation for the parameter estimation bias and the MSE of both $\hat{\theta}_S$ and $\hat{\theta}_{\hat{R}}$ with respect to specific K and R and $\hat{R}_\nu$. In Section 4.4, we give some preliminary results about how to perform empirical kriging, and estimate the Mean Squared Prediction Error (MSPE) with estimator $\hat{R}$.

4.2 A general Model

Motivated by Holland et al. (2000), we study a linear mixed effect model, which is similar to Holland et al. (2000) and allows more explicit theoretical calculation. Let

$$Y = XB + E, \tag{4.6}$$

where $Y = \{Y_{ti}\}$ is an $N \times n$ observation matrix at n spatial locations and N times, with $Y_{ti} = Y(t, s_i), \quad t \in \{1, \ldots, N\}$, and $i \in \{1, \ldots, n\}$. $X = (X_1, ..., X_p)$ is an $N \times p$ design matrix with p known covariates. We could assume each entry X_{ij} as a function of location s, however for simplicity, we just consider the spatially independent case. B is a $p \times n$ matrix of unknown regression parameters, and $E = \{\epsilon_{ti}\}$ is an $N \times n$ matrix of unobserved random errors. In the remaining parts of this chapter, we assume $p = 2$ for simplicity. The methodology developed is quite general and can be applied to models with p arbitrary variables.

We can write $X = (X^1, \ldots, X^N)^T$, assume $X^t = (1, t)^T$. Let $\beta^j = (\beta_{1j}, \ldots, \beta_{nj})^T, j = 1 \text{ or } 2, \quad \beta_i = (\beta_{i1}, \beta_{i2})^T, \quad i = 1, \ldots, n$. Thus $\beta_{ji} = \mu_j + \alpha_{ji}, B = (\beta_1, \ldots, \beta_n) = (\beta^1, \beta^2)^T$. Assume $\beta_i = (\mu_1, \mu_2)^T + (\alpha_{i1}, \alpha_{i2})^T$ are spatially dependent, with $\alpha^j = (\alpha_{1j} \ldots \alpha_{nj})^T \sim N(0, K^j(\theta)), j = 1 \text{ or } 2$. For convenience, we assume $\alpha^1 \perp \alpha^2$, but they are not necessarily orthogonal in some other cases. The covariance for β_i is given by parametric function $C^j(\cdot, \theta)$, i.e. $K^j_{lm}(\theta) = C^j(s_l - s_m, \theta)$. Let $\epsilon_i = (\epsilon_{1i}, \ldots, \epsilon_{Ni})^T, \epsilon^t = (\epsilon_{t1}, \ldots, \epsilon_{tn})^T$, thus $E = (\epsilon_1, \ldots, \epsilon_n) = (\epsilon^1, \ldots, \epsilon^N)^T$. Furthermore, we assume $\epsilon^t \overset{i.i.d.}{\sim} N(0, R)$ with R an unstructured covariance matrix, i.e. $\{\epsilon_{ti}\}$ are spatially dependent and temporally independent. Let $Y_i = (Y_{1i}, \ldots, Y_{Ni})^T = X\beta_i + \epsilon_i$ and $Y^t = (Y_{t1}, \ldots, Y_{tn})^T$, therefore $Y = \{Y_{ti}\} = (Y_1, \ldots, Y_n) = (Y^1, \ldots, Y^N)^T$, where $Y_{ti} = \beta_{1i} + t\beta_{2i} + \epsilon_{ti}$. Thus at time t and t', the covariance-variance matrix between Y^t and $Y^{t'}$ is $K_{tt'}(\theta) = K^1(\theta) + tt'K^2(\theta) + I_{(t=t')}R$. We are interested in estimating θ based on the restricted log likelihood function, and also making the spatial prediction for either α^j or $Z(t, s_0) = (1, t)\beta_0$, as in the motivation example, with R a nuisance parameter. We want to investigate the effect of R estimators on covariance parameter estimation and kriging prediction.

For estimating the covariance function parameter θ, it is natural to consider the method

of restricted maximum log likelihood. Under the assumption of the random regression coefficients and the unstructured errors, the restricted log likelihood for the data Y in terms of the parameters B, θ and R is given by

$$\ell(\theta; B, R) = -\tfrac{nN-2}{2}(\log 2+\log \pi)+\tfrac{1}{2}\log|(X^*)^T X^*|-\tfrac{1}{2}\log|V|-\tfrac{1}{2}\log|(X^*)^T V^{-1} X^*|-\tfrac{1}{2}(Y^*)^T W Y^*, \tag{4.7}$$

where $Y^* = ((Y^1)^T, \ldots, (Y^N)^T))^T, X^* = X \otimes 1_n$.

The (i,j)th element of V is given by

$$\mathrm{Cov}(Y_{ti}, Y_{t'i'}) = \begin{cases} \sum_{j=1}^{2} x_{tj}(s_i)x_{t'j}(s_{i'})K^j_{ii'}(\theta) = K^1_{ii'}(\theta) + tt'K^2_{ii'}(\theta), & \text{if } t \neq t' \\ \sum_{j=1}^{2} x_{tj}(s_i)x_{tj}(s_{i'})K^j_{ii'}(\theta) + R_{ii'} = K^1_{ii'}(\theta) + t^2K^2_{ii'}(\theta) + R_{ii'}, & \text{if } t = t'. \end{cases},$$

and we can get $W = V^{-1} - V^{-1}X^*((X^*)^T V^{-1} X^*)^{-1}(X^*)^T V^{-1}$. We could also write

$$V = \begin{pmatrix} K_{11}(\theta) & \ldots & K_{1N}(\theta) \\ \ldots & K_{tt'}(\theta) & \ldots \\ K_{N1}(\theta) & \ldots & K_{NN}(\theta) \end{pmatrix}.$$

To simplify the analysis with respect to the above log-likelihood function, we estimate B and R by $\hat{B} = (X^T X)^{-1} X^T Y$ and $\hat{R} = N^{-1} Y^T P Y$, with $P = I_{N\times N} - X(X^T X)^{-1} X^T$. This coincides with the MLE of B and R when B are fixed coefficients instead of random variables.

Let $\beta = ((\beta^1)^T, (\beta^2)^T)^T$, we can get $\hat{\beta} = ((\hat{\beta}^1)^T, (\hat{\beta}^2)^T)^T \sim N(\beta \otimes 1_n, V_B)$, where

$$V_B = \begin{pmatrix} K^1(\theta) & 0 \\ 0 & K^2(\theta) \end{pmatrix} + (X^T X)^{-1} \otimes R.$$

This is the key equation. Since we can compute $\hat{\beta}$ (also written as $\hat{B}$) without knowing R, then we use the distributional equation for $\hat{\beta}$ to estimate the parameters of V_B knowing R. Then we can explore different estimators for R in terms of how they affect the estimation of θ and then the subsequent kriging, for instance, the kriging prediction for α^j. Here β_i and $\hat{\beta}_i$ correspond to $Z(s)$ and $Y(s)$ in model (4.1), $(X^TX)^{-1} \otimes R$ corresponds to R in the motivational example.

For given R, the restricted log likelihood of θ as a function of $\hat{B}$ in terms of θ is given by

$$\ell(\theta; \hat{B}, R) = -\tfrac{2n-2}{2}(\log 2 + \log \pi) + \tfrac{1}{2}\log|X_B^T X_B| - \tfrac{1}{2}\log|V_B| - \tfrac{1}{2}\log|X_B^T V_B^{-1} X_B| - \tfrac{1}{2}\hat{\beta}^T W_B \hat{\beta}, \tag{4.8}$$

where $X_B = \begin{pmatrix} 1_n & 0 \\ 0 & 1_n \end{pmatrix}, W_B = \{w^{\alpha\beta}\} = V_B^{-1} - V_B^{-1} X_B (X_B^T V_B^{-1} X_B)^{-1} X_B^T V_B^{-1}$.

The REML estimator of θ based on above equation is much easier to compute than that from equation (4.7), since the covariance matrix V_B in (4.8) is of much lower dimension when $N > 2$. Just like the idea from the motivation example, we need to know R or its estimator for intensive calculations. In Section 4.3 - 4.4, we will study the effects of different estimators of R onto the Mean Squared Error (MSE) of resulting $\hat{\theta}$.

Another problem we study is to predict α^0 or $Z(t, s_0) = (1, t)\beta_0$, which is similar to the regional trend prediction in previous example. Prediction problem is of great interest in spatial statistics, a commonly used method is the kriging predictor. From model (4.5), it is well known that the predictor of so-defined $Z(t, s_0)$ at location s_0 and time t, using the universal kriging method (or BLUP, the best linear unbiased predictor) is given by

$$\hat{Z}(t, s_0) = x_0^T \widehat{\beta_0} + \tau_0^T (K_{tt})^{-1} (Y^t - X_0 \widehat{\beta_0}) \tag{4.9}$$

where $x_0 = (1, t)^T$, $X_0 = (x_0, \ldots, x_0)^T$ is an $n \times 2$ design matrix, $\widehat{\beta_0}$ is the kriging predictor of β_0 based on V_B and $\hat{\beta}$, and $\tau_0 = (\text{Cov}\{Z(t, s_0), Y_{t1}\}, \ldots, \text{Cov}\{Z(t, s_0), Y_{tn}\})^T$. The mean-squared prediction error (MSPE) of $\hat{Z}(t, s_0)$ is given by

$$\sigma^2(t, s_0) = c - \tau_0^T K_{tt}^{-1} \tau_0 + (x_0 - X_0^T K_{tt}^{-1} \tau_0)^T (X_0^T K_{tt}^{-1} X_0)^{-1} (x_0 - X_0^T K_{tt}^{-1} \tau_0), \tag{4.10}$$

where $c = c(t, \theta) = \text{Var}\{Z(t, s_0)\} = C^1(0, \theta) + t^2 C^2(0, \theta)$.

The right-hand sides of Equations (4.9) and (4.10) both depend on K_{tt}, i.e., on the covariance parameter θ and unstructured covariance matrix R. We can derive the empirical BLUP and MSPE by replacing θ and R with their estimates. Here are the steps:

- Estimate R by $\hat{R}$.
- Plug $\hat{R}$ into the restricted log-likelihood function in equation (4.3). Let $\hat{\ell}(\theta) = \ell(\theta; \hat{B}, \hat{R})$. The approximate REML estimators for covariance function parameter θ can be derived by maximizing $\hat{\ell}(\theta)$ with respect to θ, which we denote by $\hat{\theta}_{\hat{R}}$.
- Apply $\hat{\theta}_{\hat{R}}$, $\hat{B}$ and $\hat{R}$ to equation (4.4) and (4.5) to get empirical kriging predictor and the corresponding MSPE.

We will investigate the effect of different $\hat{R}$'s onto the estimation of the covariance function parameter θ and on kriging prediction, including the point prediction and corresponding MSPE. We define the meaning of a "good" $\hat{R}$ as follows:

Consider the following estimator of the measurement error covariance-variance matrix R,

$$\hat{R}_\nu = \nu S + (1 - \nu) S^*, \quad \nu \in [0, 1],$$

where S is the sample covariance matrix and S^* is the diagonal estimator. The sample covariance and diagonal estimator are two special cases of $\hat{R}_\nu$, with $\nu = 1$ and 0, respectively.

This linear combination estimator is inspired by Ledoit and Wolf (2004), while they focused on shrinking toward the identity matrix instead of the diagonal matrix here and considered the optimal $\hat{R}$ based on smallest MSE of itself. In this chapter, we are interested in good estimator of R and also investigating its effect on the estimation of parameter θ and on kriging prediction. For example, both of the two above estimators' performances can be compared by computing MSE of $\hat{\theta}_{\hat{R}}$. If certain shrinkage estimator results in smaller MSE of $\hat{\theta}$, we consider it as a more efficient estimator of R. Furthermore, we can use the expression of the MSE to select the optimal tuning parameter ν. Similar things could be done for predicting α^j in kriging calculation.

4.3 Results for Covariance Parameter Estimation

4.3.1 Theoretical Results

To estimate the covariance parameter θ, our analytical procedure is as follows: consider model (4.6) at time t. Using equation (4.8) for the restricted log-likelihood in terms of V_B and the definition about W, e, U_i, U_{ij} in Chapter 2 (page 17 - 18 in Section 2.2), we can rewrite equation (4.8) as

$$\ell_n = -\tfrac{2n-2}{2}(\log 2 + \log \pi) + \tfrac{1}{2}\log|X_B^T X_B| - \tfrac{1}{2}\log|V_B| - \tfrac{1}{2}\log|X_B^T V_B^{-1} X_B| - \tfrac{1}{2} e_\alpha e_\beta w^{\alpha\beta}, \tag{4.11}$$

where $W_B = \{w^{\alpha\beta}\} = V_B^{-1} - V_B^{-1}X_B(X_B^TV_B^{-1}X_B)^{-1}X_B^TV_B^{-1}$, and $\hat{\beta}^TW_B\hat{\beta} = e^TW_Be$ with $e = \{e_\alpha\} = \hat{\beta} - X_B\beta$. Also we have defined $U_i = \frac{\partial \ell_n(\theta)}{\partial\theta^i}, U_{ij} = \frac{\partial^2\ell_n(\theta)}{\partial\theta^i\partial\theta^j}$.

From expression (4.11), we get

$$U_i = \tfrac{1}{2}v_{\alpha\beta}\tfrac{\partial w^{\alpha\beta}}{\partial\theta^i} - \tfrac{1}{2}e_\alpha e_\beta\tfrac{\partial w^{\alpha\beta}}{\partial\theta^i}, \tag{4.12}$$

$$U_{ij} = \tfrac{1}{2}\tfrac{\partial v_{\alpha\beta}}{\partial\theta^j}\tfrac{\partial w^{\alpha\beta}}{\partial\theta^i} + \tfrac{1}{2}v_{\alpha\beta}\tfrac{\partial^2 w^{\alpha\beta}}{\partial\theta^i\partial\theta^j} - \tfrac{1}{2}e_\alpha e_\beta\tfrac{\partial^2 w^{\alpha\beta}}{\partial\theta^i\partial\theta^j}. \tag{4.13}$$

Let $\hat{R}$ be an estimator of R such that $A = \frac{1}{\epsilon_n}(X_B^TX_B)^{-1}\otimes(\hat{R} - R) = O_p(1)$, where $\epsilon_n \to 0$ as $n \to \infty$. One example of such $\hat{R}$ is the MLE of R, for which $\epsilon_n = \frac{1}{\sqrt{n}}$. We write

$$\hat{V}_B = \begin{pmatrix} K^1(\theta) & 0 \\ 0 & K^2(\theta) \end{pmatrix} + (X_B^TX_B)^{-1}\otimes\hat{R} = V_B + \epsilon_nA.$$

From the definition of W_B in terms of V_B, we have

$$\tfrac{dW_B}{d\epsilon_n} = -W_B\tfrac{dV_B}{d\epsilon_n}W_B, \tag{4.14}$$

$$\tfrac{d^2W_B}{d\epsilon_n^2} = 2W_B\tfrac{dV_B}{d\epsilon_n}W_B\tfrac{dV_B}{d\epsilon_n}W_B - W_B\tfrac{d^2V_B}{d\epsilon_n^2}W_B. \tag{4.15}$$

Then if we replace W_B by $\hat{W}_B$ in terms of V_B, we get the corresponding $\hat{W}_B = \hat{V}_B^{-1} - \hat{V}_B^{-1}X_B(X_B^T\hat{V}_B^{-1}X_B^{-1})X_B^T\hat{V}_B^{-1}$, and

$$\hat{W}_B = W_B - \epsilon_nW_BAW_B + \epsilon_n^2W_BAW_BAW_B + o_p(n^{-1}). \tag{4.16}$$

Let $\hat{U}_i$ and $\widehat{U}_{ij}$ be U_i and U_{ij} with V_B or W_B appearing in the definitions replaced by $\hat{V}_B$ or $\hat{W}_B$, respectively, where $\hat{U}^{ij}$ is the (i,j)th element of the inverse matrix of $\{\hat{U}_{ij}\}$. Define the

estimator $\hat{\theta}$ as the one such that $\{U_i(\theta)\} = \mathbf{0}_p$ and $\hat{\theta}$ the one such that $\{\hat{U}_i(\theta)\} = \mathbf{0}_p$, then we get

$$0 = \hat{U}_i(\hat{\theta}) = \{\hat{U}_i(\theta)\} + (\hat{\theta} - \theta)\{\hat{U}_{ij}(\theta)\} + O_p(n^{-1}),$$

therefore the bias of the estimator $\hat{\theta}$ is

$$\hat{\theta}^i - \theta^i = -\hat{U}^{ij}\hat{U}_j + O_p(n^{-2}). \tag{4.17}$$

When $n \to \infty$, we denote

$$n^{-1}U_{ij} \xrightarrow{p} B_{ij},$$

$$n^{-1}\hat{U}_{ij} \xrightarrow{p} \hat{B}_{ij},$$

$$n^{-1/2}U_i \xrightarrow{d} N[0, C],$$

$$n^{-1/2}\hat{U}_i \xrightarrow{d} N[\hat{b}, \hat{C}],$$

where $\hat{B}_{ij} - B_{ij} = O(n^{-1/2}), \hat{C} - C = O(n^{-1/2}), \hat{b} = O(n^{-1/2})$. Thus as $n \to \infty$, we get

Theorem 4.3.1. *Under regularity conditions for increasing domain asymptotics, we have*

$$\begin{aligned}\sqrt{n}(\hat{\theta} - \theta) &= -\sqrt{n}\{\hat{U}^{ij}\}\{\hat{U}_j\} + O_p(n^{-3/2}) \\ &\approx -\sqrt{n}\{\hat{U}_{ij}\}^{-1}\{\hat{U}_j\} \xrightarrow{d} N[-\hat{B}^{-1}\hat{b}, \hat{B}^{-1}\hat{C}\hat{B}^{-1}].\end{aligned} \tag{4.18}$$

Remark: Using the expression of the above theorem, we can calculate the Mean Squared Error (MSE) of the estimation of covariance parameter vector θ, depending on both K and R.

In order to maintain the expression of everything in the last term of equation (4.18), we need to develop asymptotic approximation to $\hat{b}, \hat{B}$ and $\hat{C}$ as $n \to \infty$, which are the approximations to the first and second order moments of U_i, and the first order moments of U_{ij}. We get

$$
\begin{aligned}
E\{\hat{U}_i\} &= \tfrac{1}{2} E\{\hat{v}_{\alpha\beta} - v_{\alpha\beta}\} E\{\tfrac{\partial \hat{W}_B^{\alpha\beta}}{\partial \theta^i}\} \\
&\approx \tfrac{\epsilon_n}{2} A^*_{\alpha\beta} \{\tfrac{\partial W_B^{\alpha\beta}}{\partial \theta^i} - \epsilon_n \tfrac{\partial}{\partial \theta^i}(W_B^{\beta\gamma} A^*_{\gamma\delta} W_B^{\delta\alpha})\} \\
&= \tfrac{\epsilon_n}{2} A^*_{\alpha\beta} \tfrac{\partial W_B^{\alpha\beta}}{\partial \theta^i} - \tfrac{\epsilon_n^2}{2} A^*_{\alpha\beta} (\tfrac{\partial W_B^{\beta\gamma}}{\partial \theta^i} A^*_{\gamma\delta} W_B^{\delta\alpha} + W_B^{\beta\gamma} A^*_{\gamma\delta} \tfrac{\partial W_B^{\delta\alpha}}{\partial \theta^i}), \\
E\{\hat{U}_{ij}\} &= \tfrac{1}{2} E\{\tfrac{\partial \hat{v}_{\alpha\beta}}{\partial \theta^j} \tfrac{\partial \hat{W}_B^{\alpha\beta}}{\partial \theta^i}\} + \tfrac{1}{2} \epsilon_n E\{(\hat{v}_{\alpha\beta} - e_\alpha e_\beta) \tfrac{\partial^2 \hat{W}_B^{\alpha\beta}}{\partial \theta^i \partial \theta^j}\} \\
&= \tfrac{1}{2} E\{\tfrac{\partial \hat{v}_{\alpha\beta}}{\partial \theta^j} \tfrac{\partial \hat{W}_B^{\alpha\beta}}{\partial \theta^i}\} + \tfrac{\epsilon_n}{2} E\{\tfrac{\partial v_{\alpha\beta}}{\partial \epsilon_n} \tfrac{\partial^2 \hat{W}_B^{\alpha\beta}}{\partial \theta^i \partial \theta^j}\} \\
&\approx \tfrac{1}{2} \{\tfrac{\partial \hat{v}_{\alpha\beta}}{\partial \theta^j} [\tfrac{\partial W_B^{\alpha\beta}}{\partial \theta^i} - \epsilon_n \tfrac{\partial}{\partial \theta^i}(W_B^{\beta\gamma} A^*_{\gamma\delta} W_B^{\delta\alpha}) + \epsilon_n^2 \tfrac{\partial}{\partial \theta^i}(W_B^{\beta s} A^*_{st} W_B^{t\delta} A^*_{\delta\gamma} W_B^{\gamma\alpha})] \\
&+ \epsilon_n^2 A^*_{\alpha\beta} [\tfrac{\partial^2 W_B^{\alpha\beta}}{\partial \theta^i \partial \theta^j} - \epsilon_n \tfrac{\partial^2}{\partial \theta^i \partial \theta^j}(W_B^{\beta\gamma} A^*_{\gamma\delta} W_B^{\delta\alpha})]\} \\
&= \tfrac{1}{2} \{\tfrac{\partial \hat{v}_{\alpha\beta}}{\partial \theta^j} [\tfrac{\partial W_B^{\alpha\beta}}{\partial \theta^i} - \epsilon_n (\tfrac{\partial W_B^{\beta\gamma}}{\partial \theta^i} A^*_{\gamma\delta} W^{\delta\alpha} + W_B^{\beta\gamma} A^*_{\gamma\delta} \tfrac{\partial W_B^{\delta\alpha}}{\partial \theta^i}) \\
&+ \epsilon_n^2 (\tfrac{\partial W_B^{\beta s}}{\partial \theta^i} A^*_{st} W_B^{t\delta} A^*_{\delta\gamma} W_B^{\gamma\alpha} + W_B^{\beta s} A^*_{st} \tfrac{\partial W_B^{t\delta}}{\partial \theta^i} A^*_{\delta\gamma} W_B^{\gamma\alpha} + W_B^{\beta s} A^*_{st} W_B^{t\delta} A^*_{\delta\gamma} \tfrac{\partial W_B^{\gamma\alpha}}{\partial \theta^i})] \\
&+ \epsilon_n A^*_{\alpha\beta} \tfrac{\partial^2 W_B^{\alpha\beta}}{\partial \theta^i \partial \theta^j}\}, \qquad (4.19) \\
\mathrm{Var}\{\hat{U}_i\} &= E\{\hat{U}_i^2\} - E\{\hat{U}_i\}^2 \\
&= \tfrac{1}{4} E[(\hat{v}_{\alpha\beta} - e_\alpha e_\beta)(\hat{v}_{\gamma\delta} - e_\gamma e_\delta)] E\{\tfrac{\partial \hat{W}_B^{\alpha\beta}}{\partial \theta^i}\} E\{\tfrac{\partial \hat{W}_B^{\gamma\delta}}{\partial \theta^i}\} - E\{\hat{U}_i\}^2 \\
&\approx \tfrac{1}{4} [\epsilon_n^2 A^*_{\alpha\beta} A^*_{\gamma\delta} + v_{\alpha\gamma} v_{\beta\delta} + v_{\alpha\delta} v_{\beta\gamma}] [\tfrac{\partial W_B^{\alpha\beta}}{\partial \theta^i} - \epsilon_n \tfrac{\partial}{\partial \theta^i}(W_B^{\beta s} A^*_{st} W_B^{t\alpha}) + \epsilon_n^2 \tfrac{\partial}{\partial \theta^i}(W A^* W A^* W)^{\alpha\beta}] \\
&[\tfrac{\partial W_B^{\gamma\delta}}{\partial \theta^i} - \epsilon_n \tfrac{\partial}{\partial \theta^i}(W_B^{\gamma s} A^*_{st} W_B^{t\delta}) + \epsilon_n^2 \tfrac{\partial}{\partial \theta^i}(W A^* W A^* W)^{\gamma\delta}] - E\{\hat{U}_i\}^2 \\
&\approx \tfrac{1}{4} [v_{\alpha\gamma} v_{\beta\delta} + v_{\alpha\delta} v_{\beta\gamma}] [\tfrac{\partial W_B^{\alpha\beta}}{\partial \theta^i} \tfrac{\partial W_B^{\gamma\delta}}{\partial \theta^i} - \epsilon_n (\tfrac{\partial W_B^{\alpha\beta}}{\partial \theta^i} \tfrac{\partial}{\partial \theta^i}(W_B^{\gamma s} A^*_{st} W_B^{t\delta}) + \tfrac{\partial W_B^{\gamma\delta}}{\partial \theta^i} \tfrac{\partial}{\partial \theta^i}(W_B^{\beta s} A^*_{st} W_B^{t\alpha})) \\
&+ \epsilon_n^2 (\tfrac{\partial}{\partial \theta^i}(W_B^{\beta s} A^*_{st} W_B^{t\alpha}) \tfrac{\partial}{\partial \theta^i}(W_B^{\gamma s} A^*_{st} W_B^{t\delta}) + \tfrac{\partial W_B^{\alpha\beta}}{\partial \theta^i} \tfrac{\partial}{\partial \theta^i}(W A^* W A^* W)^{\gamma\delta} \\
&+ \tfrac{\partial W_B^{\gamma\delta}}{\partial \theta^i} \tfrac{\partial}{\partial \theta^i}(W A^* W A^* W)^{\alpha\beta})], \qquad (4.20)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Cov}\{\hat{U}_i, \hat{U}_j\} &= E\{\hat{U}_i\hat{U}_j\} - E\{\hat{U}_i\}E\{\hat{U}_j\} \\
&= \tfrac{1}{4}E[(\hat{v}_{\alpha\beta} - e_\alpha e_\beta)(\hat{v}_{\gamma\delta} - e_\gamma e_\delta)]E\{\tfrac{\partial \hat{W}_B^{\alpha\beta}}{\partial\theta^i}\}E\{\tfrac{\partial \hat{W}_B^{\gamma\delta}}{\partial\theta^j}\} - E\{\hat{U}_i\}E\{\hat{U}_j\} \\
&\approx \tfrac{1}{4}[\epsilon_n^2 A^*_{\alpha\beta}A^*_{\gamma\delta} + v_{\alpha\gamma}v_{\beta\delta} + v_{\alpha\delta}v_{\beta\gamma}][\tfrac{\partial W_B^{\alpha\beta}}{\partial\theta^i} - \epsilon_n\tfrac{\partial}{\partial\theta^i}(W_B^{\beta s}A^*_{st}W_B^{t\alpha}) \\
&+ \epsilon_n^2\tfrac{\partial}{\partial\theta^i}(WA^*WA^*W)^{\alpha\beta}][\tfrac{\partial W_B^{\gamma\delta}}{\partial\theta^j} - \epsilon_n\tfrac{\partial}{\partial\theta^j}(W_B^{\gamma s}A^*_{st}W_B^{t\delta}) + \epsilon_n^2\tfrac{\partial}{\partial\theta^j}(WA^*WA^*W)^{\gamma\delta}] \\
&- E\{\hat{U}_i\}E\{\hat{U}_j\} \\
&\approx \tfrac{1}{4}[v_{\alpha\gamma}v_{\beta\delta} + v_{\alpha\delta}v_{\beta\gamma}][\tfrac{\partial W_B^{\alpha\beta}}{\partial\theta^i}\tfrac{\partial W_B^{\gamma\delta}}{\partial\theta^j} - \epsilon_n(\tfrac{\partial W_B^{\alpha\beta}}{\partial\theta^j}\tfrac{\partial}{\partial\theta^i}(W_B^{\gamma s}A^*_{st}W_B^{t\delta}) \\
&+ \tfrac{\partial W_B^{\gamma\delta}}{\partial\theta^i}\tfrac{\partial}{\partial\theta^j}(W_B^{\beta s}A^*_{st}W_B^{t\alpha})) + \epsilon_n^2(\tfrac{\partial}{\partial\theta^i}(W_B^{\beta s}A^*_{st}W_B^{t\alpha})\tfrac{\partial}{\partial\theta^j}(W_B^{\gamma s}A^*_{st}W_B^{t\delta}) \\
&+ \tfrac{\partial W_B^{\alpha\beta}}{\partial\theta^i}\tfrac{\partial}{\partial\theta^j}(WA^*WA^*W)^{\gamma\delta} + \tfrac{\partial W_B^{\gamma\delta}}{\partial\theta^j}\tfrac{\partial}{\partial\theta^i}(WA^*WA^*W)^{\alpha\beta})], \qquad (4.21)
\end{aligned}
$$

where $A^*_{\alpha\beta} = E\{A_{\alpha\beta}\}$. Furthermore, we need the following expressions of derivatives to plug into the above equations:

$$
\begin{aligned}
\tfrac{\partial W_B}{\partial\theta^i} &= -W_B\tfrac{\partial V_B}{\partial\theta^i}W_B, \\
\tfrac{\partial^2 W_B}{\partial\theta^i\partial\theta^j} &= W_B\tfrac{\partial V_B}{\partial\theta^i}W_B\tfrac{\partial V_B}{\partial\theta^j}W_B + W_B\tfrac{\partial V_B}{\partial\theta^j}W_B\tfrac{\partial V_B}{\partial\theta^i}W_B - W_B\tfrac{\partial^2 V_B}{\partial\theta^i\partial\theta^j}W_B.
\end{aligned}
$$

Using Theorem 4.3.1, and the fact that

$$
\begin{aligned}
\mathrm{MSE}(\hat{\theta}) = \mathrm{bias}^2 + \mathrm{Var}(\hat{\theta}) &= E^2(\hat{\theta} - \theta) + \mathrm{Var}(\hat{\theta}) \\
&= \tfrac{1}{n}[(\hat{B}^{-1}\hat{b})^2 + \hat{B}^{-1}\hat{C}\hat{B}^{-1}] + o(n^{-2}),
\end{aligned}
$$

we can calculate equations (4.19) - (4.22) (for specific K, R and $\hat{R}$), and substitute them into the above equation to approximate the MSE of $\hat{\theta}_{\hat{R}}$, where $\hat{R}$ indicates the estimator is derived

by plugging $\hat{R}$ into the restricted log-likelihood function (4.8) or (4.11). $\hat{R}$ could be any possible estimator of R, e.g. the sample covariance S, or the linear shrinkage estimator $\hat{R}_\nu$. In addition, we can use our calculation results by Theorem 4.3.1 to choose the optimal tuning parameter for $\hat{R}_\nu$. Simulation studies are also necessary for finite sample sizes.

4.3.2 Simulation Results

As a starting point, we consider a model with a specific covariance function K, where both K^1 and K^2 are exponential covariance functions with the same parameters, with the (i,j)'th element of K^l set to be $K^l(i,j) = \phi \cdot \exp(-\frac{|i-j|}{\rho}), \quad l = 1, 2$. Assume the parameters ϕ and ρ are independent of each other. We set $\phi = .5, 1, 1.5, 2,\ \rho = 0.25, 0.5, 0.75, 1$ and $n = 5, N = 30$, respectively. The data are simulated and analyzed as follows:

Step 1: Simulate the data with covariance structure K and random error R together, where R could be a random positive-definite $n \times n$ matrix, an exponential covariance (where the (i,j)'th element of R is set to be $R(i,j) = \phi_R \cdot \exp(-|i-j|/\rho_R), \phi_R = 1, \rho_R = 1$) or a long range dependence structure (where the (i,j)'th element of R is constructed as $R(i,j) = \frac{\sigma_R^2}{2}[(|i-j|+1)^{2H} - 2|i-j|^{2H} + (|i-j|-1)^{2H}]$, with $\sigma_R^2 = 1, H = 3/4$), respectively. In particular, all of these three types of R and the exponential covariance structure could be constructed directly in matlab code.

Step 2: For each pair of possible K and R with specific parameter values ($4 \times 4 \times 3 = 48$ in total), we simulate the data for 100 runs. For each $i = 1, \ldots, 100$, we consider the following estimators of R:

$$\hat{R}_\nu^i = \nu S_i + (1-\nu) S_i^*,$$

where $\nu = 0, 0.05, 0.1, \ldots, 1$ (corresponding to the sample covariance). S_i^* is the diagonal

estimator of R.

This family of estimators is inspired by the work in Ledoit and Wolf (2004), while their motivation is to shrink S toward a re-scaled identity matrix instead of the diagonal matrix here, and to find the optimal estimator $\hat{R}$ based on the smallest mean squared error of the covariance matrix.

On one hand, applying Theorem 4.3.1, we can calculate the theoretical MSE for $\hat{\theta}_{\hat{R}}$.On the other hand, we can calculate the MSE for $\hat{\theta}_{\hat{R}}$ in the simulation as well, since we can get $\hat{R}_{\nu}^{i}, i = 1, \ldots, 100$ and then we can get the corresponding $\hat{\theta}_{\hat{R}}$, so the simulation result for its MSE should be $\frac{1}{100}\sum_{i=1}^{100}(\hat{\theta}_{\hat{R}} - \theta)^2$, here $\theta = \phi$ or ρ.

Step 3: Since the structure of K is fixed to be exponential here, for each one of the three types of R (random, exponential and long range dependence, respectively), we make $4 \times 4 = 16$ plots for the MSE of $\hat{\theta}_{\hat{R}}$ versus. $\nu = 0, 0.05, \ldots, 1.\theta = \phi$ or ρ. Figure 4.1 - Figure 4.2 is for random R, Figure 4.3 - 4.4 for R with the exponential covariance, Figure 4.5 - 4.6 is for the long range dependence case of R.

From Figure 4.1 - 4.6, we can see clearly that :

- There exists some $\nu^* \in [0, 1]$ such that $\hat{R}_{\nu^*}$ will result in the smallest MSE for $\hat{\theta}$ ($\theta = \rho$ or ϕ), among all the linear combinations of S and S^*. In addition, in most cases the MSE improvements from the sample covariance are very substantial, say, more than 50%.

- In most cases, the diagonal estimator (when $\nu = 0$) is better than the sample covariance, resulting in smaller MSE of $\hat{\theta}$.

- It seems that the simulation result for the case of finite sample size agrees well with calculation by Theorem 4.3.1, which approximates the MSE of $\hat{\theta}$. We could use this

theorem to choose the optimal tuning parameter for $\hat{R}_\nu$, corresponding to the smallest mean squared error of $\hat{\theta}_{\hat{R}}$.

- Interestingly, the optimal ν which result in smallest MSE of both ϕ and ρ seem close to each other, for specific K and R.

The next work we should do is to investigate if there is a unique and explicit way to express this ν^*, in terms of K, R and $\hat{R}$.

4.4 Preliminary Results for kriging performance

To make kriging prediction by using $\hat{R}_\nu$ and $\hat{\theta}$ in place of R and θ respectively, we can proceed as follows.

Suppose we want to predict the scalar variable $Z(s_0, t) = x_0\beta_0$ from model (4.6), with variance c and $\text{Cov}(Z, Y^t)^T = \tau_0$. The standard kriging predictor for $Z(s_0, t)$ is $\hat{Z} = \lambda^T Y^t$ with mean squared prediction error σ_0^2, where

$$\lambda = K_{tt}^{-1}\tau_0 + K_{tt}^{-1}X_0(X_0^T K_{tt}^{-1} X_0)^{-1}(x_0 - X_0^T K_{tt}^{-1}\tau_0), \tag{4.22}$$

$$\sigma_0^2 = c - \tau_0^T K_{tt}^{-1}\tau_0 + (x_0^T - \tau_0^T K_{tt}^{-1}X_0)(X_0^T K_{tt}^{-1}X_0)^{-1}(x_0 - X_0^T K_{tt}^{-1}\tau_0). \tag{4.23}$$

When R is substituted by $\hat{R}$, it is natural to replace λ by

$$\hat{\lambda} = \hat{K}_{tt}^{-1}\tau_0 + \hat{K}_{tt}^{-1}X_0(X_0^T \hat{K}_{tt}^{-1} X_0)^{-1}(x_0 - X_0^T \hat{K}_{tt}^{-1}\tau_0), \tag{4.24}$$

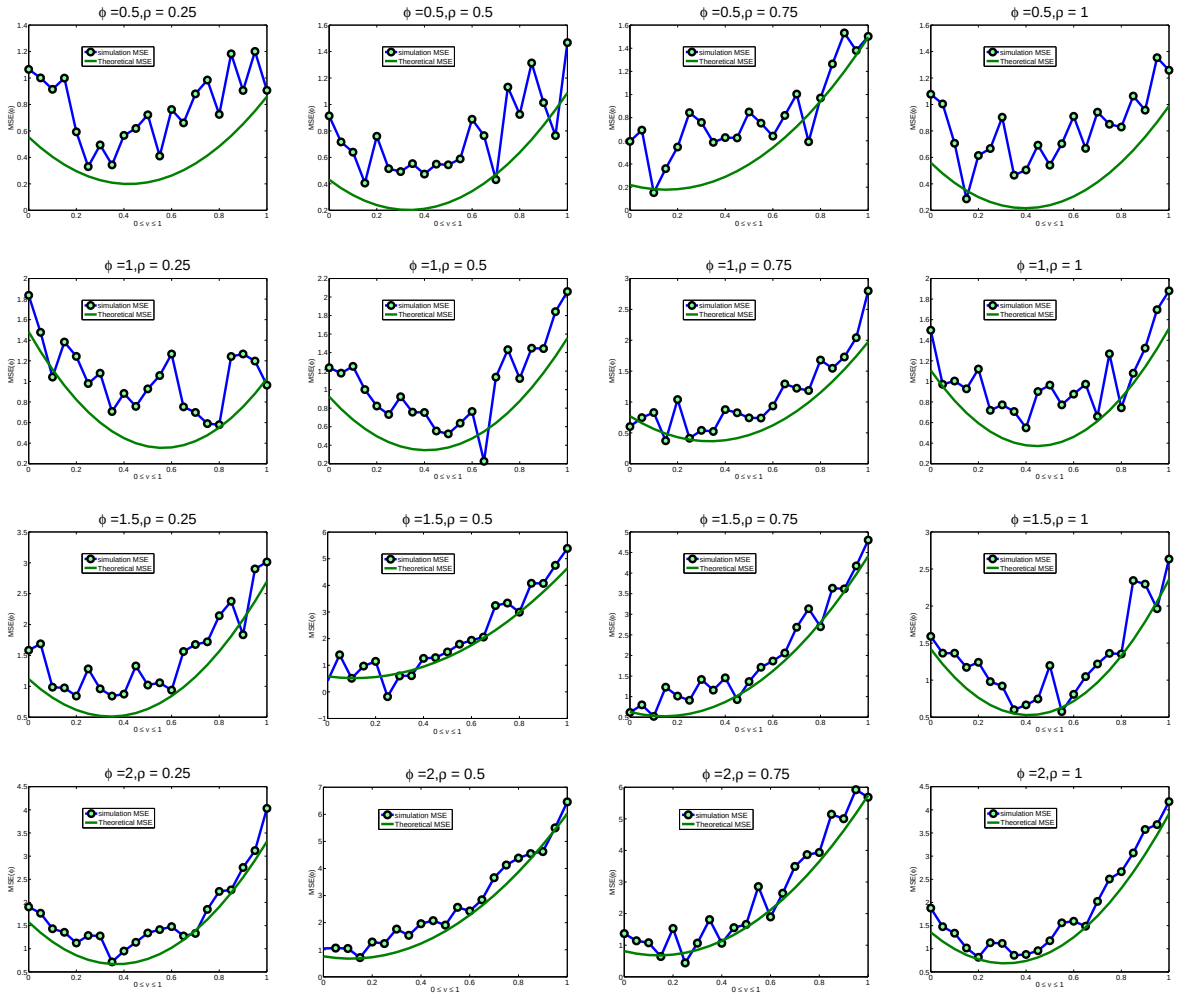

Figure 4.1: When R is random. Covariance function $K^l = \phi \exp(-d/\rho), l = 1, 2$. Plots of $MSE(\hat{\phi}_{\hat{R}})$ vs. ν, where $\hat{R}_\nu = \nu S + (1-\nu)S^*, \nu = 0, 0.05, 0.1, \ldots, 1$. Solid line: theoretical result by Theorem 4.3.1, Circled line: simulation results (100 iterations).

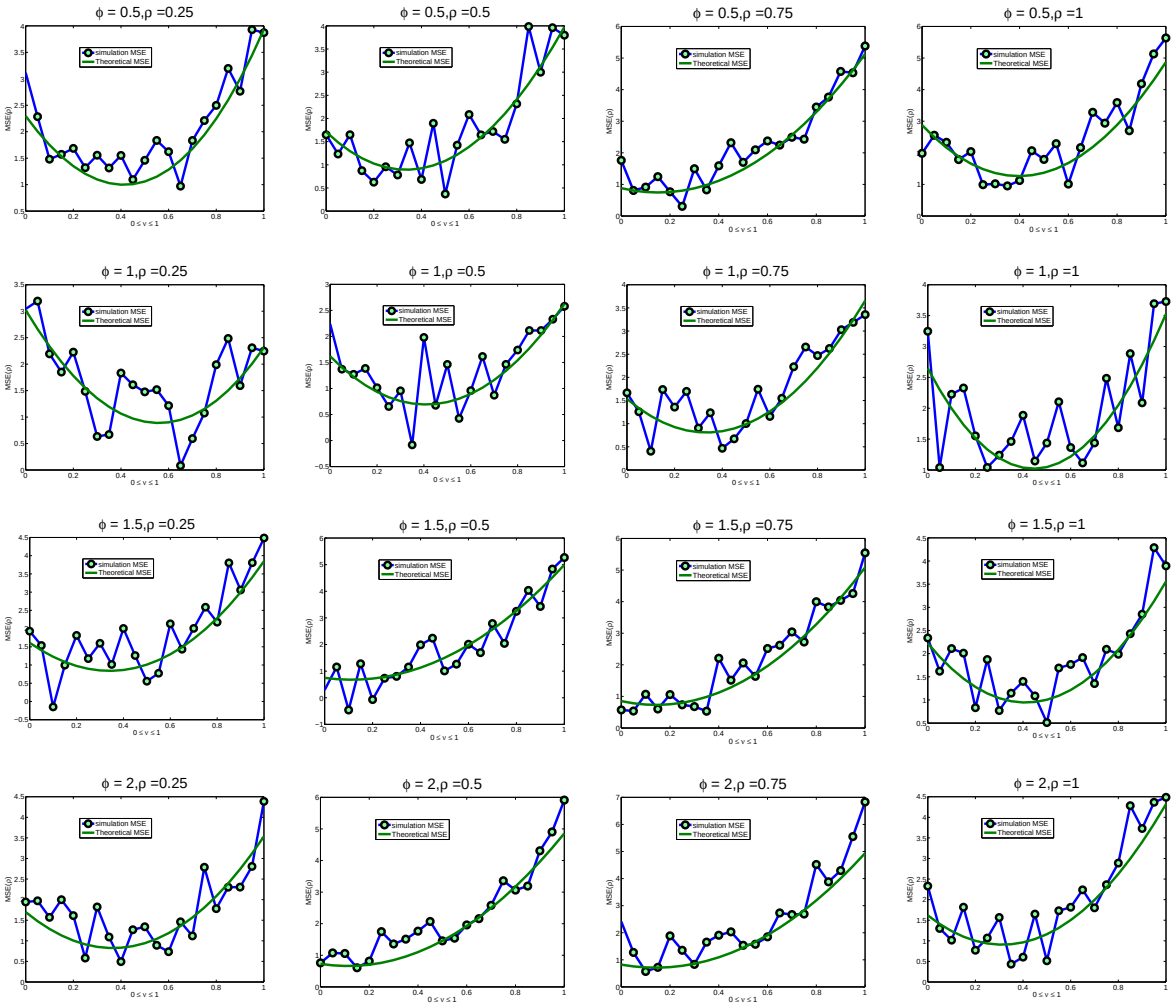

Figure 4.2: When R is random. Covariance function $K^l = \phi \exp(-d/\rho), l = 1, 2$. Plots of $MSE(\hat{\rho}_{\hat{R}})$ vs. ν, where $\hat{R}_\nu = \nu S + (1-\nu)S^*, \nu = 0, 0.05, 0.1, \ldots, 1$. Solid line: theoretical result by Theorem 4.3.1, Circled line: simulation results (100 iterations).

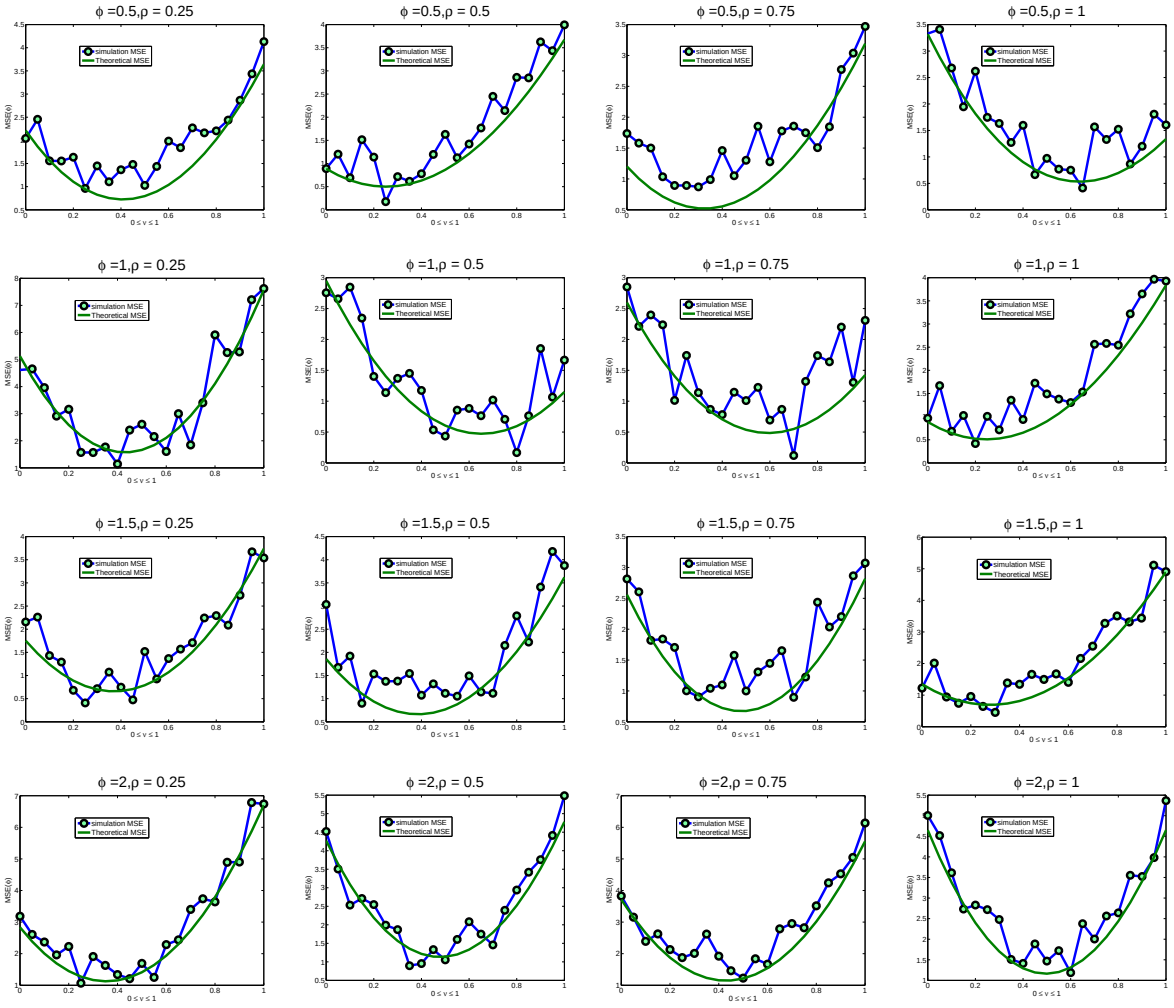

Figure 4.3: When R is exponential. Covariance function $K^l = \phi \exp(-d/\rho), l = 1, 2$. Plots of $MSE(\hat{\phi}_{\hat{R}})$ vs. ν, where $\hat{R}_\nu = \nu S + (1-\nu)S^*, \nu = 0, 0.05, 0.1, \ldots, 1$. Solid line: theoretical result by Theorem 4.3.1, Circled line: simulation results (100 iterations).

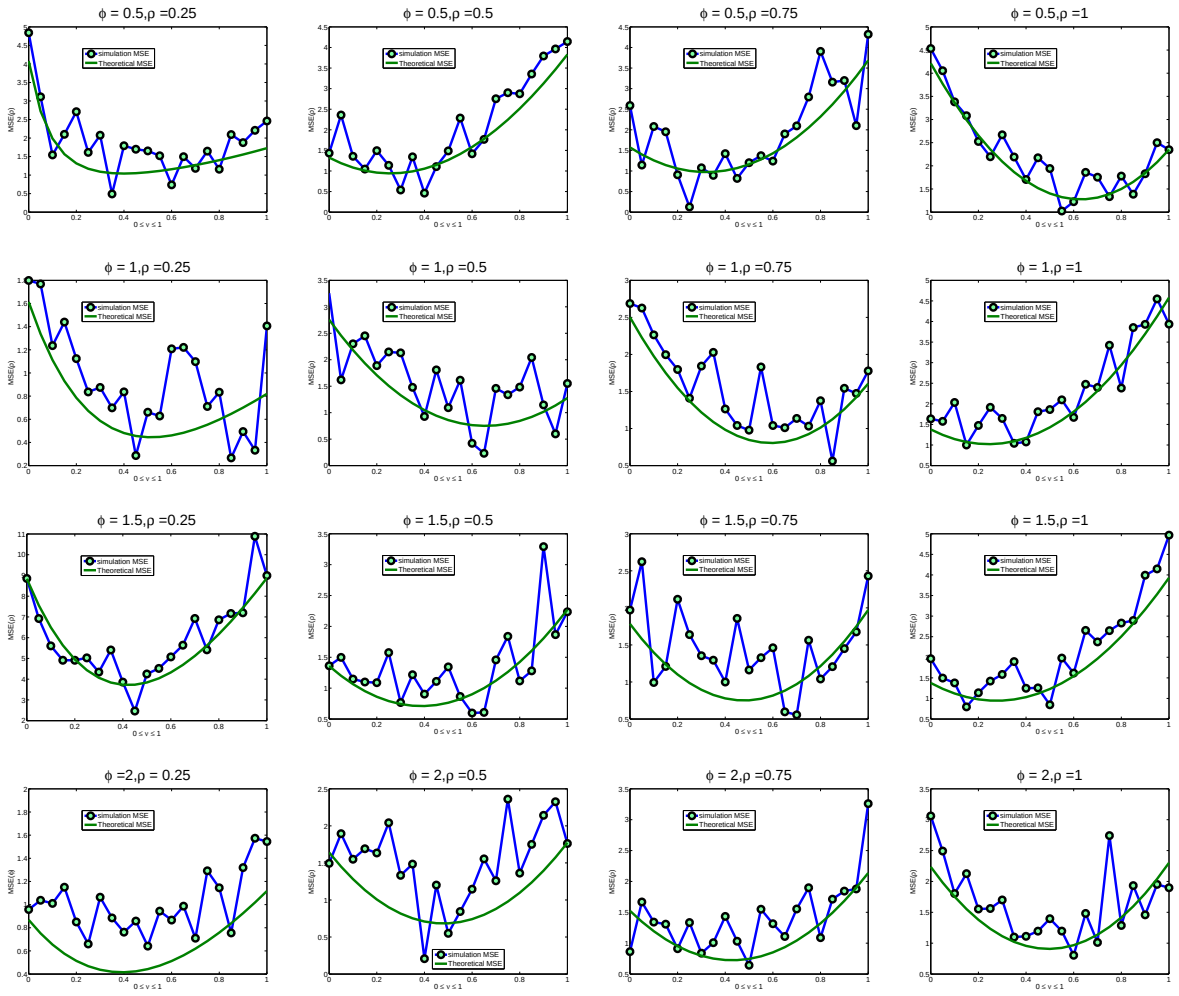

Figure 4.4: When R is exponential. Covariance function $K^l = \phi \exp(-d/\rho), l = 1, 2$. Plots of $MSE(\hat{\rho}_{\hat{R}})$ vs. ν, where $\hat{R}_\nu = \nu S + (1-\nu)S^*, \nu = 0, 0.05, 0.1, \ldots, 1$. Solid line: theoretical result by Theorem 4.3.1, Circled line: simulation results (100 iterations).

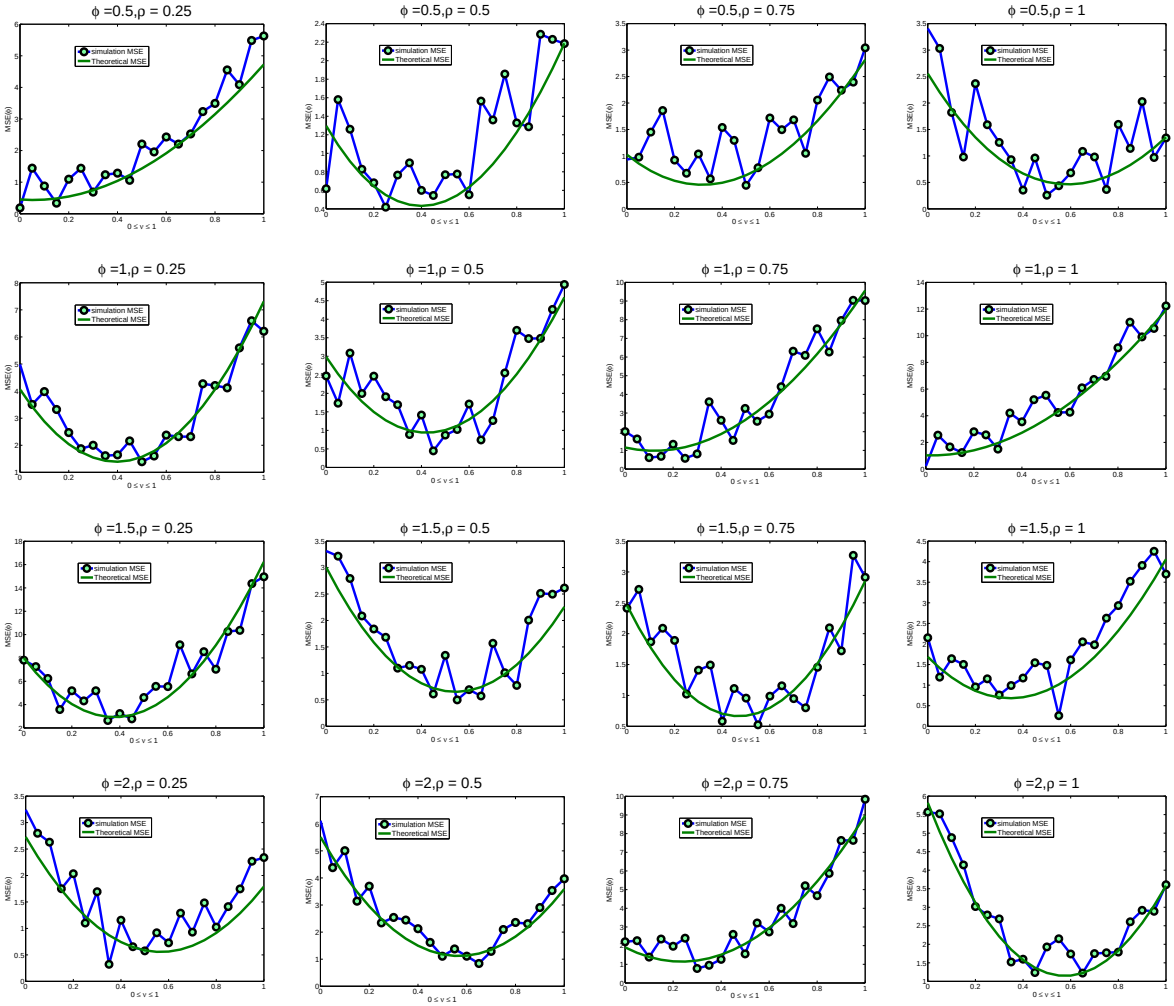

Figure 4.5: When R is long range dependence. Covariance function $K^l = \phi \exp(-d/\rho), l = 1, 2$. Plots of $MSE(\hat{\phi}_{\hat{R}})$ vs. ν, where $\hat{R}_\nu = \nu S + (1-\nu)S^*, \nu = 0, 0.05, 0.1, \ldots, 1$. Solid line: theoretical result by Theorem 4.3.1, Circled line: simulation results (100 iterations).

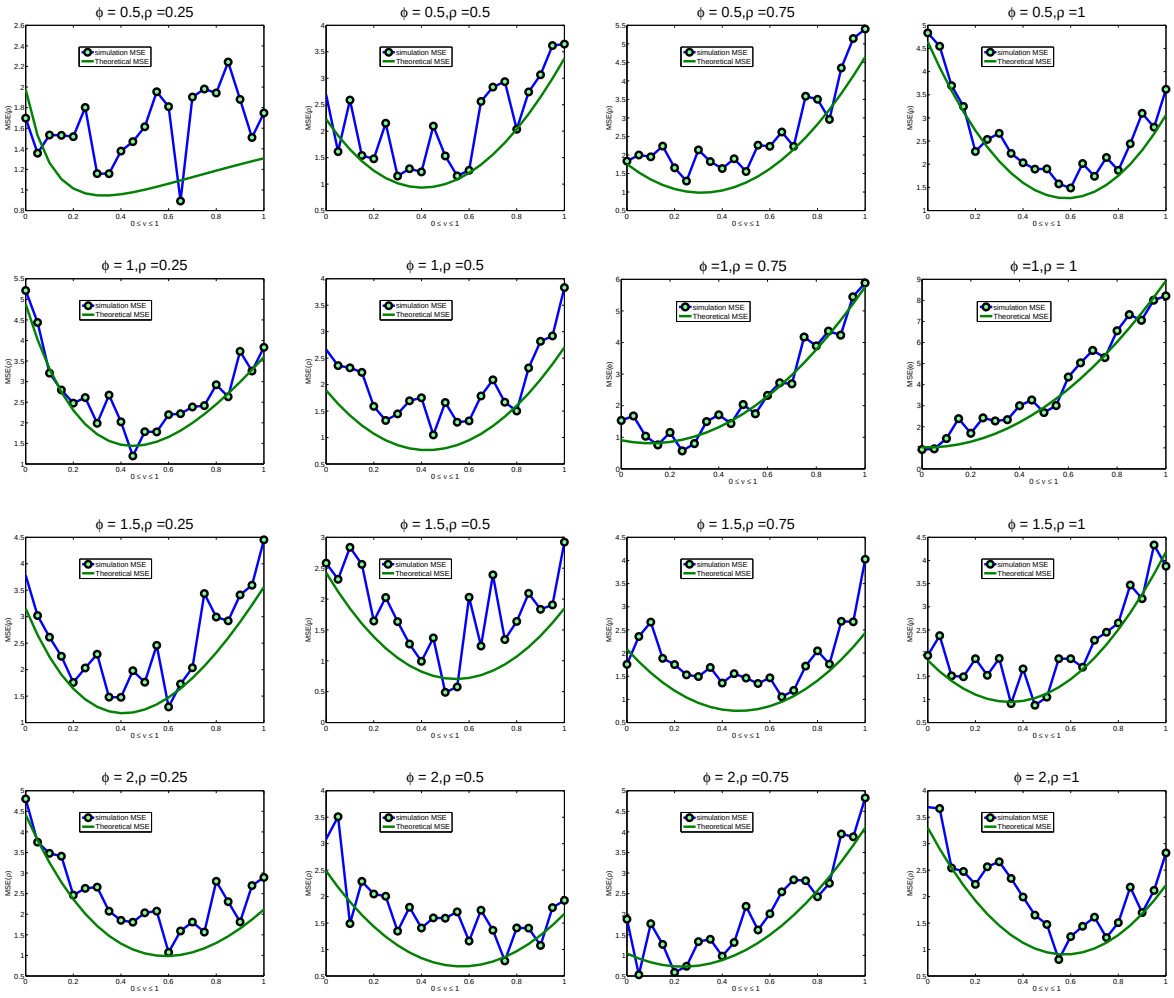

Figure 4.6: When R is long range dependence. Covariance function $K^l = \phi \exp(-d/\rho), l = 1, 2$. Plots of $MSE(\hat{\rho}_{\hat{R}})$ vs. ν, where $\hat{R}_\nu = \nu S + (1-\nu)S^*, \nu = 0, 0.05, 0.1, \ldots, 1$. Solid line: theoretical result by Theorem 4.3.1, Circled line: simulation results (100 iterations).

and we get the resulting MSPE for $\hat{Z}$ is

$$E\{(Z_0 - \hat{\lambda}^T Y^t)^2\} = c - 2\hat{\lambda}^T \tau_0 + \hat{\lambda}^T K_{tt} \hat{\lambda}. \tag{4.25}$$

To calculate the right hand side of equations (4.25) and (4.26), we need to expand it in powers of ϵ_n, by using the following derivative expressions

$$\begin{aligned}
\tfrac{\partial K_{tt}^{-1}}{\partial \epsilon_n} &= -K_{tt}^{-1} \tfrac{\partial K_{tt}}{\partial \epsilon_n} K_{tt}^{-1}, \\
&= -K_{tt}^{-1} A K_{tt}^{-1}, \qquad (4.26) \\
\tfrac{\partial^2 K_{tt}^{-1}}{\partial \epsilon_n^2} &= 2K_{tt}^{-1} \tfrac{\partial K_{tt}}{\partial \epsilon_n} K_{tt}^{-1} \tfrac{\partial K_{tt}}{\partial \epsilon_n} K_{tt}^{-1} - K_{tt}^{-1} \tfrac{\partial^2 K_{tt}}{\partial \epsilon_n^2}, \\
&= 2K_{tt}^{-1} A K_{tt}^{-1} A K_{tt}^{-1}, \qquad (4.27) \\
\tfrac{\partial}{\partial \epsilon_n}(X_0^T K_{tt}^{-1} X_0) &= -X_0^T K_{tt}^{-1} A K_{tt}^{-1} X_0, \qquad (4.28) \\
\tfrac{\partial}{\partial \epsilon_n}(X_0^T K_{tt}^{-1} X_0)^{-1} &= (X_0^T K_{tt}^{-1} X_0)^{-1} X_0^T K_{tt}^{-1} A K_{tt}^{-1} X_0 (X_0^T K_{tt}^{-1} X_0)^{-1}, \qquad (4.29) \\
\tfrac{\partial^2}{\partial \epsilon_n^2}(X_0^T K_{tt}^{-1} X_0)^{-1} &= 2(X_0^T K_{tt}^{-1} X_0)^{-1} X_0^T K_{tt}^{-1} A K_{tt}^{-1} X_0 (X_0^T K_{tt}^{-1} X_0)^{-1} X_0^T K_{tt}^{-1} A K_{tt}^{-1} X_0 \cdot \\
&\quad (X_0^T K_{tt}^{-1} X_0)^{-1}, \qquad (4.30)
\end{aligned}$$

and so on. Besides, we can write $\hat{\lambda}$ as follows,

$$\begin{aligned}
\hat{\lambda} &= \lambda + \epsilon_n \cdot \tfrac{\partial \lambda}{\partial \epsilon_n} + O_p(\tfrac{1}{n}) \\
&= \lambda + \epsilon_n \cdot [\tau_0 \tfrac{\partial K_{tt}^{-1}}{\partial \epsilon_n} + \tfrac{\partial K_{tt}^{-1}}{\partial \epsilon_n} X_0 (X_0^T \hat{K}_{tt}^{-1} X_0)^{-1} x_0 \\
&+ K_{tt}^{-1} X_0 \tfrac{\partial}{\partial \epsilon_n} (X_0^T K_{tt}^{-1} X_0)^{-1} x_0 - \tfrac{\partial K_{tt}^{-1}}{\partial \epsilon_n} X_0 (X_0^T \hat{K}_{tt}^{-1} X_0)^{-1} X_0^T \hat{K}_{tt}^{-1} \tau_0 \\
&- \hat{K}_{tt}^{-1} X_0 \tfrac{\partial}{\partial \epsilon_n} (X_0^T K_{tt}^{-1} X_0)^{-1} X_0^T \hat{K}_{tt}^{-1} \tau_0 - \hat{K}_{tt}^{-1} X_0 (X_0^T \hat{K}_{tt}^{-1} X_0)^{-1} X_0^T \tfrac{\partial K_{tt}^{-1}}{\partial \epsilon_n} \tau_0]. \qquad (4.31)
\end{aligned}$$

Plugging (4.27) and (4.30) into equation (4.32), we can get the expression of $\hat{\lambda}$ and then substitute into equation (4.26). Similarly to the simulation work we have done in Section 4.3.2, we can simulate some data and check the effect of $\hat{R}_\nu$ (with different ν) on the kriging predictor $\hat{Z} = \hat{\lambda}^T Y^t$ and the kriging prediction error $\hat{\sigma}^2$.

4.5 Summary

In this Chapter, we consider estimating the unstructured covariance matrix in model (4.6). We study two different estimators $\hat{R}$'s and their effect on the estimator for θ, and will investigate how they affect the kriging predictor and kriging prediction error as well. Direct application of the estimator of covariance matrix leads to the "plug-in" approach for $\hat{\hat{\theta}}, \hat{Z}, MSPE(\hat{Z})$. To derive asymptotic distribution of the parameter estimator $\hat{\hat{\theta}}$ and the prediction error for an unobserved variable Z, we need to specify $A = \frac{1}{\epsilon_n}(X^T X)^{-1}(\hat{R} - R)$. We consider S (sample covariance matrix) or $\hat{R}_\nu$ (a convex linear combination of the sample covariance and diagonal matrix) here.

For the unstructured measurement error matrix, we could estimate it by the sample covariance $S = n^{-1} Y^T P Y,$ with $P = I - X(X^T X)^{-1} X^T$, or some shrinkage estimator, for instance, $\hat{R}_\nu = \nu S + (1-\nu) S^*, \nu \in [0, 1)$. We are interested in estimating θ and β based on the restricted log likelihood function $\ell(\theta; \hat{B}, R)$, and also the spatial predictions for α^j.

For fixed K and R, we compare the performances of $\hat{\theta}_{\hat{R}}$ ($\hat{R} = S$ or $\hat{R}_\nu$) by computing their mean squared errors using equations (4.18) - (4.22), simultaneously we can use the latter to select the tuning parameters ν. From Figure 4.1 - 4.6 we can see clearly that there is substantial improvement from the sample covariance to the shrinkage estimator, in the sense of the MSE of resulting covariance parameter estimator $\hat{\theta}$. The next analysis should be to find if there is any

explicit expression (or feasible algorithm) of ν in terms of $K, R,$ and $\hat{R}$. Also, we already have some preliminary results for the kring prediction error as well, similar things could be done for predicting α^j.

Up to now, we only consider the comparison the sample covariance with the linear combination, without comparing with any other alternative covariance estimators. On the other hand, we have already assumed the model is spatial-temporal data, so that any results from our work could be easily applied to either time series or spatial modeling. Both of these remain possible topics for future research.

Chapter 5

Summary and Comments

In this chapter, we summarize all the contributes we have already made, and propose some interesting problems for the future research.

5.1 Summary of the finished work

5.1.1 Asymptotic Comparisons of Predictive Densities for Dependent observations

In Chapter 2, we used the asymptotic expansion of the KL divergence as the main tool to compare different predictive distributions with dependent observations, and derived some explicit results for one-way random effect models.

1. In Section 2.3, we have derived the Laplace expansion of the expected difference between the KL divergence of both the REML-based estimative density and the Bayesian predictive density. We have also defined "second-order KL REML-dominance" and given explicit conditions for a prior to be second-order KL REML-dominant.

2. In Section 2.4, we used a specific mixed effect model as an example to illustrate our results. Under this model, we have proposed a class of improper priors and given specific conditions

for the corresponding predictive distributions to be second-order KL REML-dominant. We also showed that the commonly used Jeffreys prior does not lead to second-order KL REML-dominance.

We have already derived the methodology for asymptotic approximation of integration of KL-divergences, which can be applied to more general mixed effects models, for instance the spatial linear models commonly used in geostatistics. However the condition for a "second-order KL REML-dominant" prior in these models is more difficult to get.

5.1.2 Applications regarding the temporal (or spatial) AR(1) models

We indicated how to predict the conditional density on the AR(1) process in the Bayesian framework by introducing some noninformative priors, all of which are second-order KL REML-dominant. We also compared among the three candidate priors, of which the result indicates that in the asymptotic sense the reference prior is superior to the other two, in both the temporal and spatial AR(1) models. We simulate data for both the time series and spatial AR(1) models, which also validates the result when the sample size is moderately large. We also noticed that, when the sample size is very small, the simulation numerical values differ substantially from the asymptotic second-order approximation, possibly due to the biases of higher-order with respect to averaged KL divergence. In particular, the smaller the sample size is, the greater the influence of higher-order biases is. Tanaka and Komaki (2008) considered the comparison of priors in a different way: they evaluated the performance of the resulting Bayesian spectral densities and compare their expected KL divergences and suggested a superharmonic prior for the AR(2) case.

5.1.3 Estimation and Prediction with Errors in Covariances

For shrinkage estimator of the covariance matrix, and its effect on the estimation of covariance function parameter and the kriging performance, we have the following results in Chapter 4.

1. In Section 4.1 we consider a general model with a semiparametric covariance matrix V. We study two estimators for R: S as the sample covariance of R, and a linear shrinkage estimator $\hat{R}_\nu = \nu S + (1-\nu)S^*, \nu \in [0,1]$, where S^* denotes the diagonal estimator of R. For the parameter θ, an empirical REML estimator based on estimator of R and $\hat{B}$ has been proposed. The bias, $\hat{\theta}^i - \theta$, asymptotically depends on the first and second order moments of the first order derivative of the restricted log-likelihood function, and the first order moments of the second order derivative. We have derived the asymptotic expression for these moments, leading to the computation of $MSE(\hat{\theta}_{\hat{R}})$ with $\hat{R} = S$ or $\hat{R}_\nu$. For fixed exponential structure of K and three different types of R (random, exponential and long range dependence, respectively), we have done some simulation work and the numerical results agrees with theoretical values very well, indicating that in most cases, some $\hat{R}_\nu$ with certain ν will result in the smallest MSE of $\hat{\theta}$.

2. Preliminary results for predicting α^j and getting the mean squared prediction error in kriging has been given in section 4.4. If we replace R by $\hat{R}_\nu$ or S, we can get the empirical kriging predictor and its mean squared prediction error, depending on K, R and $\hat{R}$.

5.2 Future work

In the course of this research, additional future work has become apparent. There are many potential future research problems related to comparisons of densities, estimation for covariance

function parameter and prediction methods. Here we propose some of them as my future research topics.

5.2.1 Asymptotic Expansions for KL divergence

1. Asymptotic study of the predictive densities for dependent observations is expected to give many interesting and insightful results. In the future, we plan to consider spatial sampling design in the context of spatial linear model. We will consider a design criterion by use of the Kullback-Leibler divergence between the true density and REML plug-in density or Bayesian predictive density, with respect to the block predictor. We plan to derive the optimal design criterion by applying the asymptotic approximation to the KL divergence to the second order. This gives some explicit form for the integration of Kullback-Leibler divergences, and reduces the computation workload, when searching for optimal design.

2. Garcia-Donato and Sun (2007) discussed objective priors for hypothesis testing for one-way random effects models, and derived the divergence based (DB) prior and the intrinsic prior. Their work is related with ours, while their emphasis is on the use of these priors to develop consistent objective Bayesian factors, which is different from ours. We can check whether their priors are also second order KL REML-dominant, and whether some of their priors can dominate others in the sense of second order KL divergence, by the asymptotic expressions we derived for KL divergence (2.26) - (2.28). This could be future research topic beyond thesis work.

5.2.2 Applications to the regression models with temporal or spatial correlated error

In Chapter 3, our attentions is mainly paid to the AR(1) model, a special case of AR(p) processes. For the general AR(p) model, especially when the order is moderately large, there is another way to estimate the model: utilizing the Bayesian estimation of the spectral density of the model, e.g. Tanaka and Komaki (2005) shows that in i.i.d. case the Bayesian estimation of spectral densities based on a superharmonic prior (if exists) asymptotically dominate those based on the Jeffreys prior, using the asymptotic expansion of the risk difference. Tanaka and Komaki (2008) focuses on the AR(2) process and propose an explicit form of such a superharmonic prior. We could also consider the AR(2) model from the viewpoint of its predictive densities and make comparisons of different methods.

On the other hand, the moving average (MA) model is also one of the most important models in data analysis. The MA models are completely different from the AR models as a stochastic process and in the information geometrical viewpoint, they are known to have different structures. We can also consider the ARMA model for the most general situation.

5.2.3 Estimation and Prediction with Errors in Covariances

We could continue the following work as an extension of Chapter 4:

1. In Chapter 4, we compare the empirical REML estimator of parameter θ based on sample covariance matrix S and the one based on shrinkage estimator $\hat{R}_\nu$, in the sense of the resulting MSE of $\hat{\theta}$ (for specific exponential K, and three kinds of R: random, exponential and long range dependence, see details on page in Section 4.3). Also we can use the first one to select the tuning parameter ν. We would consider the underlying connection

between the choice of ν and the structure of K, R and $\hat{R}_\nu$, and propose certain algorithm to select the optimal tuning parameter.

2. Similar work as above could be done for predicting α^j in kriging. In this way we can investigate the effect of $\hat{R}$ ($\hat{R}_\nu$ or S) on kriging prediction and the Mean Squared Prediction Error (MSPE).

3. Besides the linear combination $\hat{R}_\nu$, we will try some other estimators for R, for instance the Stein-shrinkage estimator (Daniels and Kass (2001)), to check if there will be substantial improvement from the sample covariance, in the sense of MSE($\hat{\hat{\theta}}$) or empirical kriging prediction MSE.

Besides, certain related future research is as follows:

1. Model (4.6) is assumed to be stationary. We would like to estimate the covariance matrix without assumption of stationarity, or for non-stationary models, and also check its effect on parameter estimation and kriging predictions.

2. Illuminated by Section 4.1, we can consider systematic comparison of the Bayesian methods and the regularized REML as future research work, though the computation cost of the Bayesian methods is much more now.

Bibliography

Aitchison, J. (1975), "Goodness of prediction fit," *Biometrika*, 62, 547.

Amari, S. (1987), "Differential geometry of a parametric family of invertible linear systems - Riemannian metric, dual affine connections, and divergence," *Mathematical System Theory*, 20, 53 – 82.

Azzalini, A. (1984), "Estimation and hypothesis testing for collection of autoregressive time series," *Biometrika*, 71, 85 – 90.

Barndorff-Nielsenn, O. (1983), "On a formula for the distribution of the maximum likelihood estimator," *Biometrika*, 70, 343 – 365.

Bartlett, M. S. (1975), *The Statistical Analysis of Spatial Pattern*, Chapman and Hall, London.

Basu, S. and Reinsel, G. C. (1993), "Properties of the Spatial Unilateral First-Order ARMA Model," *Advances in Applied Probability*, 25, 631–648.

— (1994), "Regression Models with Spatially Correlated Errors," *J. Amer. Statist. Assoc.*, 89, 88 – 99.

Bayes, T. R. (1763), "Essay towards solvng a problem in the doctrine of chances," *Philosophical Transactions*, 53, 370–418.

Berger, J. and Bernardo, J. M. (1992), *On the develpment of the reference prior method. In J.M. Bernardo, J.O. Berger, D.V. Lindley A.F.M. Smith (eds.) Baysian Statistics 4*, London: Oxford University Press.

Berger, J. O., de Oliveira, V., and Sanso, B. (2001), "Objective Bayesian Analysis of Spatially Correlated Data," *Journal of the American Statistical Association*, 96, 1361–1374.

Berger, J. O. and Yang, R. (1994), "Noninformative Priors and Bayesian Testing for the AR(1) Model," *Econometric Theory*, 10, 461 – 482.

Bernardo, J. M. (1979), "Reference Posterior Distributions for Bayesian Inference," *Journal of Royal Statistical Society B*, 41, 113 – 147.

Besag, J., York, J., and Mollie, A. (1991), "Bayesian image restoration, with two applications in spatial statistics (with discussion)," *Annals of the Institute of Statistical Mathematics*, 43, 1–20.

Bleistein, N. and Handelsman, R. A. (1986), *Asymptotic Expansions of Integrals*, Courier Dover.

Box, G. E. P. and Jenkins, G. M. (1976), *Time Series Analysis: Forecasting and Control*, San Francisco: Holden Day.

Brockwell, P. J. and Davis, R. A. (1991), *Time series: theory and methods*, Springer.

Brooks, S. P. and Gelman, A. (1997), "General methods for monitoring convergence of iterative simulations," *Journal of Computational and Graphical Statistics*, 7, 434–455.

Chen, C. (1979), "Bayesian inference for a normal dispersion matrix and its application to stochastic multiple regression analysis," *J. R. Statist. Soc. B*, 41, 235 – 248.

Cliff, A. D. and Ord, J. K. (1981), *Spatial Processes: Models and Applications*, Pion, London.

Cooper, D. M. and Thompson, R. (1977), "A note on the estimation of the parameters of autoregressive-moving average process," *Biometrika*, 64, 625 – 628.

Cressie, N. (1992), "REML estimation in empirical Bayes smoothing of census undercount," *Survey Methodology*, 18, 75 – 94.

Cullis, B. R. and Gleeson, A. C. (1991), "Spatial analysis of Field experiments: an extension to two dimensions," *Biometrics*, 47, 1449 – 1460.

Daniels, M. and Kass, R. (1999), "Nonconjugate bayesian estimation of covariance ma- trices and ist use in hierarchical models," *J. Am. Statist. Assoc.*, 94, 1254 – 1263.

Daniels, M. J. and Cressie, N. A. C. (1999), "A hierarchical approach to covariance function estimation for time series," *Journal of Time Series Analysis*, 22, 253–266.

Daniels, M. J. and Kass, R. E. (2001), "Shrinkage Estimators for Covariance Matrices," *Biometrics*, 57, 1173 – 1184.

Daniels, M. J. and Pourahmadi, M. (2002), "Bayesian analysis of covariance matrices and

dynamic models for longitudinal data," *Biometrika*, 89, 553–566.

Ferreira, M. A. R. and Oliveira, V. D. (2007), "Bayesian reference analysis for Gaussian Markov random fields," *Journal of Multivariate Analysis*, 98, 789–812.

Garcia-Donato, G. and Sun, D. (2007), "Objective priors for hypothesis testing in one-way random effects models," *The Canadian Journal of Statistics*, 35, 303 – 320.

Gelman, A. and Rubin, D. B. (1992), "Inference from iterative simulation using multiple sequences," *Statistical Science*, 7, 457 – 511.

George, E. I., Liang, F., and Xu, X. (2006), "Improved Minimax Predictive Densities under Kullback-Leibler Loss," *The Annals of Statistics*, 34, 78 – 91.

Geweke, J. (1992), "Evaluating the accuracy of sampling-based approaches to calculating posterior moments," *Bayesian Statistics*, 4, 1 – 30.

Gleeson, A. C. and Cullis, B. R. (1987), "Residual maximum likelihood (REML) estimation of a neighbour model for field experiments," *Biometrics*, 43, 277 – 287.

Green, P. J. (1985), "Linear models for field trials, smoothing and cross-validation," *Biometrika*, 72, 527 – 538.

Haff, L. R. (1977), "Minimax estimators for a multinormal precision matrix," *J. Multivariate Anal*, 7, 374 – 385.

— (1980), "Empirical bayes estimation of the multivariate normal covariance matrix," *Ann. Statist.*, 8, 586 – 597.

— (1991), "The variational form of certain bayes estimators," *Ann. Statist.*, 19, 1163 – 1190.

Hartigan, J. A. (1998), "The maximum likelihood prior," *The Annals of Statistics*, 26, 2083–2103.

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

— (1977), "Maximum likelihood approaches to variance components estimation and related

problems," *J. Amer. Statist. Assoc*, 72, 320 – 340.

Hobert, J. P. and Casella, G. (1996), "The effect of improper priors on Gibbs Sampling in Hierarchical Linear Mixed Models," *Jounal of the American Statistical Association*, 91, 1461 – 1473.

Holland, D. M., Oliveira, V. D., Cox, L. H., and Smith, R. L. (2000), "Estimation of regional trends in sulfur dioxide over the eastern United States," *Environmetrics*, 11, 373–393.

Huang, J. Z., Liu, L., and Liu, N. (2007), "Estimation of Large Covariance Matrices of Longitudinal Data With Basis Function Approximations," *Journal of Computational and Graphical Statistics*, 16, 189–209.

James, W. and Stein, C. (1961), "Estimation with quadratic loss. In Jerzy Neyman, editor," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statis- tics and Probability*, 1, 361 – 379.

Jeffreys, H. (1961a), *Theory of Probability*, London: Oxford University Press.

— (1961b), *Theory of Probability, [1937]*, London: Oxford University Press.

Khuri, A. I. and Sahai, H. (1985), "Variance components analysis: a selective literature survey," *Internat. Statist. Rev*, 53, 279 – 300.

Komaki, F. (1996), "On asymptotic properties of predictive distributions," *Biometrika*, 83, 299–313.

— (2006), "Shrinkage Priors for Bayesian Prediction," *The Annals of Statistics*, 34, 808 – 819.

Laird, N. M. and Ware, J. M. (1982), "Random effects models for longitudinal data," *Biometrics*, 38, 963 – 974.

Laplace, P. (1812), *Theorie Analytique des Probabilites*, Paris: Courcier.

Ledoit, O. and Wolf, M. (2004), "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, 88, 365–411.

MacNab, Y. C. (2003), "Hierarchical Bayesian Modeling of Spatially Correlated Health Service

Outcome and Utilization Rates," *Biometrics*, 59, 305–316.

Mardia, K. V. and Marshall, R. J. (1984), "Maximum likelihood estimation of models for residual covariance in spatial regression," *Biometrika*, 71, 135–146.

Martin, R. J. (1990), "The use of time-series models and methods in the analysis of agricultural Field trials," *Communications in statistics. Theory and methods*, 19, 55 – 81.

Mitchell, A. (1967), "Discussion of paper by I.J. Good." *Journal of Royal Statistical Society B*, 29, 423 – 424.

Monette, G., Fraser, D. A. S., and Ng, K. W. (1984), *Marginalization, likelihood, and structural models*, P.R. Krishnaiah (eds.) Mulitivariate Analysis VI. Amsterdam: North-Holland.

Murray, G. D. (1977), "A note on the estimation of probability density functions," *Biometrika*, 64, 150–2.

Patterson, H. D. and Thompson, R. (1971), "Recovery of interblock information when block sizes are unequal," *Biometrika*, 58, 545 – 554.

Phillips, P. C. B. (1991), "To criticize the critics: An objective Bayesian analysis of stochastic trends," *Journal of Applied Econometrics*, 6, 333 – 364.

Rajaratnam, B., Massam, H., and Carvalho, C. M. (2008), "Flexible covariance estimation in graphical Gaussian models," *Ann. Statist*, 36, 2818–2849.

Ren, C., Sun, D., and Dey, D. K. (2006), "Bayesian and frequentist estimation and prediction for exponential distributions," *Jounal of Statistical Planning and Inference*, 136, 2873 – 2897.

Ripley, B. D. (1981), *Spatial Statistics*, Wiley, New York.

Robinson, D. L. (1987), "Estimation and use of variance components," *The Statistician*, 36, 3 – 14.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992), *Variance Components*, Wiley, New York.

Smith, R. L. and Zhu, Z. (2004), "Asymptotic Theory for Kriging with Estimated Parameters

and Its Application to Network Design," *Techinical Report*.

Speed, T. P. (1991), "Comment on "That BLUP is a good thing — The estimation of random effects" by G. K. Robinson," *Statist. Sci*, 6, 42 – 44.

Stein, C. (1956), "Some problems in multivariate analysis," *Standford University Technical Report*, 6.

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

Sun, D., Tsutakawa, R. K., Kim, H., and He, Z. (2000), "Spatio-temporal interaction with disease mapping," *Statistics in Medicine*, 19, 2015–2035.

Tanaka, F. and Komaki, F. (2005), "Asymptotic Expansion of the Risk Difference of the Bayesian Spectral Density in the ARMA model," *Mathematical Engineering Technical Reports, University of Tokyo*, 31, 1 – 21.

— (2008), "A superharmonic Prior for the Autoregressive Process of the Second Order," *Journal of Time Series Analysis*, 29, 444 – 452.

Thompson, W. A. J. (1962), "The problem of negative estimates of variance components," *Ann. Math. Statist*, 33, 273 – 289.

Verbyla, A. P. (1990), "A conditional derivation of residual maximum likelihood," *Austral. J. Statist*, 32, 227 – 230.

Vidoni, P. (1995), "A simple predictive density based on the p^*-formula," *Biometrika*, 82, 855 – 863.

Wahba, G. (1990), *Spline Models for Observational Data*, SIAM, Philadelphia.

Wu, W. and Pourahmadi, M. (2003), "Nonparametric estimation of large covariance matrices of longitudinal data," *Biometrika*, 90, 831–844.

Ye, K. Y. and Berger, J. (1991), "Noninformative priors for inference in exponential regression models," *Biometrika*, 78, 645 – 656.

Zellner, A. (1971), *An introduction to Bayesian Inference in Econometrics*, New York: Wiley.

— (1977), *Maximal data information prior distributions*, A. Aykac (eds.), New Methods in the Application of Bayesian Methods. Amsterdam: North-Holland.

Zhu, Z. and Liu, Y. (2007), "Estimating Spatial Covariance using Penalized Likelihood with Weighted L_1*Penalty*," *Technical Report UNC/STOR/07/14*.

Zimmerman, D. L. and Harville, D. A. (1991), "A random Field approach to the analysis of Field-plot experiments and other spatial experiments," *Biometrics*, 47, 223 – 239.