Bayesian and Frequentist Methods for Approximate Inference in Generalized Linear Mixed Models

Evangelos A. Evangelou

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

Chapel Hill
2009

Approved by,

Richard L. Smith, Advisor

Zhengyuan Zhu, Advisor

Chuanshu Ji, Committee Member

Bahjat F. Qaqish, Committee Member

Haipeng Shen, Committee Member

Abstract

EVANGELOS A. EVANGELOU: Bayesian and Frequentist Methods for Approximate Inference in Generalized Linear Mixed Models
(Under the direction of Richard L. Smith and Zhengyuan Zhu)

Closed form expressions for the likelihood and the predictive density under the Generalized Linear Mixed Model setting are often nonexistent due to the fact that they involve integration of a nonlinear function over a high-dimensional space. We derive approximations to those quantities useful for obtaining results connected with the estimation and prediction from a Bayesian as well as from a frequentist point of view. Our asymptotic approximations work under the assumption that the sample size becomes large with a higher rate than the number of random effects.

The first part of the thesis presents results related to frequentist methodology. We derive an approximation to the log-likelihood of the parameters which, if maximized, gives estimates with low mean square error compared to other methods. Similar techniques are used for the prediction of the random effects where we propose an approximate predictive density from the Gaussian family of densities. Our simulations show that the predictions obtained using our method is comparable to other computationally intensive methods. Focus is given toward the analysis of spatial data where, as an example, the analysis of the rhizoctonia root rot data is presented.

The second part of the thesis is concerned with the Bayesian prediction of the random effects. First, an approximation to the Bayesian predictive distribution function is derived which can be used to obtain prediction intervals for the random effects without the use of Monte Carlo methods. In addition, given a prior for the covariance parameters of the random effects we derive approximations to the coverage probability bias and the Kullbak-Leibler divergence of the predictive distribution constructed using that prior. A simulation study is performed where we compute these quantities for different priors to select the prior with the smallest coverage probability bias and Kullbak-Leibler divergence.

Acknowledgments

I would like to acknowledge particularly my advisors, Richard Smith and Zhengyuan Zhu for their guidance, advice, and financial support. Special thanks go to all my teachers for providing me with education and inspiration.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

A statistical model is a tool for describing random phenomena in terms of mathematical equations. Although such models rarely manage to capture the phenomenon exactly, they are still useful in the sense that they provide a way of understanding the phenomenon under study. Here we focus on a specific model that is general enough to find many applications such as in medical experiments (ex. 6.2 in Breslow and Clayton, 1993), genetics (ex. 6.6 in Breslow and Clayton, 1993), environmental sciences (ex. 1 in Diggle et al., 1998), epidemiology (Zhang, 2002) among others.

The Generalized Linear Mixed Model (GLMM) is a type of model that is general enough to be used for modeling data from discrete as well as continuous distributions and to allow for different sources of variability in the mean response. The first feature is achieved by specifying the mean of the response as a function (usually nonlinear) of some explanatory variables. The second feature is achieved by modeling the mean of the response as a function of random variables called the random effects.

An example

Consider the following example adapted from Zhang (2002). In this example, a plantation of wheat and barley suffers from a disease called rhizoctonia that attaches to the roots of plants and hinders the absorption of water and nutrients by them. In order to be able to apply sufficient treatment to the plantation, we would like to construct a map showing the severity of the disease in the whole area. To achieve that, a sample of 15 plants where each plant has multiple roots was collected at 100 different locations and the number of roots and infected roots were counted. The data are shown in Figure 1.1 where we can see the number of infected

roots out of the total number of roots observed at each location.

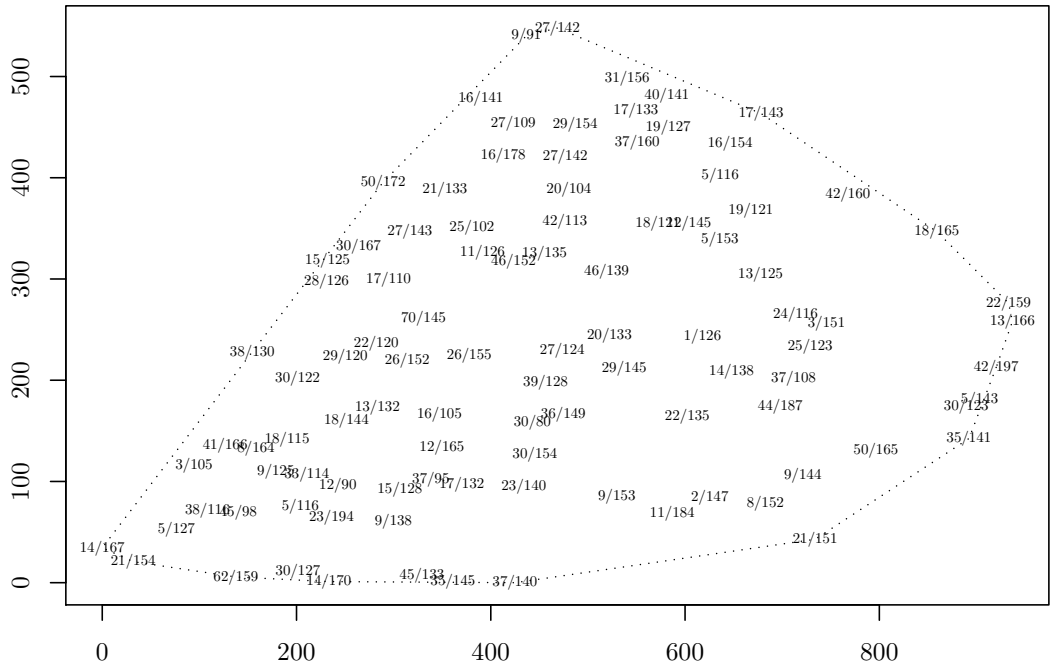

Figure 1.1: Figure showing the infected number of roots out of the total number of roots observed at the different locations.

The nature of the data does not allow us to model them as continuous variables but instead they should be modeled as binomial where the probability of observing an infected root is higher where the disease is more severe. In addition, since the data depend on the locations that were drawn from, a separate sampling involving data from different locations would result to a different disease mapping, therefore, the effect of the disease intensity at each location should be considered random in order to account for the variability due to sampling, hence a random effect. Furthermore, it is natural to assume that observations at nearby locations are highly correlated compared to observations from locations that are far apart.

In this example we are interested in estimating any parameters associated with the probability of an infected root as well as any parameters associated with the correlation structure of the random effects. Furthermore, a disease map should be constructed by predicting the random effects associated with the probability of an infection.

Two questions related to Generalized Linear Mixed Models with statistical interest are (i) estimating the parameters of the model and (ii) predicting the random effects. Methods exist

for solving these questions, suitable for each problem. They can be split into two categories, those that calculate the likelihood and the predictive density by simulations and those that evaluate those quantities by approximating them. On the one hand, simulation based methods become very accurate if the simulation is carried to a large size, but that is sometimes too computationally intensive. On the other hand, approximation methods are faster but the error of the approximation can be significant if the sample size is not sufficiently large, which makes these methods biased.

This thesis is focused on results related to asymptotic methods. Approximate techniques are used to derive an approximation to the likelihood with small error when the sample size is large. Parameter estimation is then performed by maximizing the proposed approximation. We compare the proposed method with other methods by simulations and find that it performs very well compared to the other methods.

The same techniques are used for approximating the predictive likelihood. These lead us to the definition of an approximate plug-in predictive density which has the Normal density. Prediction intervals for the random effects are constructed by obtaining the necessary quantiles of this approximate distribution. This method is also compared to other existing methods and shows good performance compared to them.

From a Bayesian point of view, estimation and prediction is performed by drawing random samples from the posterior distributions of the parameters and the random effects given the sample. We derive a similar approximation to the Bayesian predictive distribution function and show how our approximation can be used to derive corrections to the Bayesian (and frequentist) prediction intervals using random sampling. Furthermore, we propose criteria for assessing the performance of the prior based on approximations to the coverage probability bias and Kullback-Leibler divergence.

The rest of the thesis is organized as follows. In chapter 2 we review the literature in Generalized Linear Mixed Models and related topics. The model under study in this thesis is defined in chapter 3 where we also derive some useful results that will be used later in the thesis. Chapter 4 deals with issues related to the maximum likelihood estimation and chapter 5 contains similar results in the case of the predictive density. An example on how the theory in the previous chapters can be applied is presented in chapter 6. In chapter 7

we derive approximations to the Bayesian prediction quantiles. We also show how the bias of the Bayesian predictive density and the Kullabck-Leibler divergence can be approximated and describe how these quantities can be computed. A comparison of several Bayesian predictive densities is performed constructed under different priors by comparing the coverage probability bias and Kullback-Leibler divergence and selecting the one with the smallest of these.

CHAPTER 2

Background

2.1 Generalized Linear Mixed Models

Statistical models are generally used to explain the variability in the values of one variable, the *response*, in terms of the values of other variables, the *predictors*. From a statistician's point of view, the challenges of any model fitting are developing a methodology for estimating the parameters of the model, as well as predicting any unobserved random variables.

The simplest model, and the most widely studied, is the *Linear Model*. Its main assumptions are that (i) the observations are independent, (ii) the mean equals a linear combination of the predictors, and (iii) the variance of the response is constant for every observation. An additional fourth assumption is sometimes made, that (iv) the observations are a sample from the Normal distribution. Procedures for fitting linear models which are very easy to implement have been developed. However, the above assumptions are not always satisfied, therefore the use of more general models is necessary. Such models include the *Generalized Linear Model* (GLM) and the *Generalized Linear Mixed Model* (GLMM).

The GLM generalizes on the assumptions (ii), (iii), and (iv) above in the following way: (ii') a monotone transformation of the mean equals a linear combination of the predictors, (iii') the variance is a function of the mean of each observation, and (iv') the distribution of the observations is a member of the exponential family. More specifically, suppose that we have the sample

$$y_1, y_2, \ldots, y_n$$

which we want to model with respect to a set of covariates

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$$

Our data consist of a realization of the random variables $Y_1, \ldots Y_n$ which according to assumption (iv'), the ith variable follows a distribution with probability density/mass function

$$f(y; \theta, \omega) = \exp[\omega^{-1}\{y\,\theta - b(\theta)\} + c(y, \omega)] \tag{2.1}$$

for known functions b and c, and with *canonical parameter* θ and *dispersion parameter* ω. Examples of distributions with density/mass function (2.1) are the Normal, Poisson, Binomial, and Gamma distributions. For certain members of this family of distributions, the dispersion parameter ω is known and is in general of less importance. The parameter θ on the other hand, plays an important role, especially because of its relationship with the mean μ of the distribution by the equation

$$\mu = b'(\theta) \tag{2.2}$$

and needs to be estimated from the data. Furthermore, the variance, as a function of θ equals

$$v(\theta) = \omega\, b''(\theta) \tag{2.3}$$

The function b is called the *cumulant function* and v is called the *variance function* (McCullagh and Nelder, 1999).

A linear combination $\eta_i = \boldsymbol{x}_i^\top \beta$ of the response is called the *linear predictor* and according to assumption (ii'), for some strictly increasing function g, called the *link function*,

$$g(\mu_i) = \eta_i \tag{2.4}$$

In the special case of the linear model, g is the identity function.

Note that by (2.2) and (2.4), a one-to-one relationship exists between θ_i and η_i. A natural choice would be to choose g such that $\theta_i = \eta_i$ in which case g is called the *canonical link* function. Canonical links have the property that the sufficient statistic for β is $X^\top \boldsymbol{y}$, X being

the matrix with rows $\boldsymbol{x}_i$'s and $\boldsymbol{y}$ the vector of y_i's.

The GLMM introduces a further generalization of the linear model by relaxing the assumption (i) that the observations are independent and assuming (i') that they are *conditionally independent* given an unobserved random variable $\boldsymbol{Z}$ called the *random effect*. In this case, the linear predictor is written as

$$\eta_i = \boldsymbol{x}_i^\top \beta + \boldsymbol{u}_i^\top \boldsymbol{Z} \tag{2.5}$$

for known $\boldsymbol{u}_i$. A usual assumption of the distribution of the random variable $\boldsymbol{Z}$ is that is Normal with mean zero and some variance $\Sigma = \Sigma(\gamma)$ that is parameterized by an unknown parameter γ, called the *variance components*, that is

$$\boldsymbol{Z} \sim \mathrm{N}(\boldsymbol{0}, \Sigma) \tag{2.6}$$

In applications, assuming that the correct model is being used, two main questions arise with statistical interest:

1. How to estimate the parameters β and γ, and
2. how to predict the random effects $\boldsymbol{Z}$.

The solution to the first question is related to the notion of the likelihood while the solution to the second question with the predictive distribution.

2.1.1 Maximum Likelihood Estimation

The most broadly used method for estimation in parametric models is that of maximum likelihood estimation (MLE) originally proposed by Fisher (1912). The idea is to choose the value of the parameters that maximizes the joint density of the observations, called the *likelihood*. Under fairly weak regularity conditions the estimate obtained is asymptotically unbiased and asymptotically efficient as the sample size increases.

In linear models, the calculations for deriving the MLE can be done analytically and closed form expressions for the estimators of the parameters exist. This is not the case for the GLM and the GLMM so numerical methods are used instead. The MLE for GLM is obtained by applying

the Iterated Weighted Least Squares algorithm (IWLS) as shown in Nelder and Wedderburn (1972).

For the GLMM, the log-likelihood of the parameters (β, γ) given the observations $\boldsymbol{y}$ can be written (up to a constant), using (2.1) as

$$\ell(\beta, \gamma|\boldsymbol{y}) = -\frac{1}{2}\log|\Sigma(\gamma)| + \log\int \exp\left\{\sum_i y_i\,\theta_i(\beta, \boldsymbol{z}) - \sum_i b\big(\theta_i(\beta, \boldsymbol{z})\big) - \frac{1}{2}\boldsymbol{z}^\intercal\Sigma^{-1}(\gamma)\boldsymbol{z}\right\}\mathrm{d}\boldsymbol{z} \tag{2.7}$$

where $\theta_i(\beta, \boldsymbol{z})$ is the expression of θ as a function of $(\beta, \boldsymbol{z})$ obtained from (2.2), (2.4), and (2.5).

To understand the difficulties in calculating the MLE let's consider the simple example where there is only one random effect $z \sim \mathrm{N}(0, \gamma)$, β is just an intercept term, and for each i, $y_i|z \sim \mathrm{Po}(e^{\beta+z})$. Then the likelihood of (β, γ) equals

$$L(\beta, \gamma|\boldsymbol{y}) = e^{\beta\sum y_i}e^{-\frac{1}{2}\log\gamma}\int \exp\left\{z\sum y_i - ne^{\beta+z} - \frac{1}{2\gamma}z^2\right\}\mathrm{d}z \tag{2.8}$$

The integral in (2.8) does not have a closed form solution which makes it difficult to calculate the likelihood accurately. Evaluation of (2.8) is possible using numerical techniques such as Gauss-Hermite quadrature but as the dimension of z becomes larger, these methods become unreliable. More advanced methods have been developed for estimation in GLMM. These include simulation based methods, as in McCulloch (1997) and Booth and Hobert (1999), and approximation methods as in Breslow and Clayton (1993) and Shun and McCullagh (1995).

Approximate Likelihood

The idea of the approximate methods is to approximate the log-likelihood in (2.7) by replacing the integrand with an approximation of it which can be integrated analytically. Typically the approximation is performed by applying Taylor expansion to the exponent of the integrand in (2.7) around the point $\tilde{\boldsymbol{z}}$ that the exponent is maximized. This method of approximating integrals is known as *Laplace approximation* (Barndorff-Nielsen and Cox, 1989, sec 3.3). The estimates are then obtained by maximizing the approximate likelihood.

Penalized Quasi Likelihood

Breslow and Clayton (1993) expanded the exponent of the integrand up to a polynomial of

second degree and suggested using an IWLS algorithm for obtaining the MLE. Their method is outlined below:

Let

$$\psi(\boldsymbol{y}, \boldsymbol{z}; \beta, \gamma) = -\sum_i y_i \, \theta_i + \sum_i b\big(\theta_i\big) + \frac{1}{2}\boldsymbol{z}^{\intercal}\Sigma^{-1}\boldsymbol{z} \tag{2.9}$$

i.e. the log-likelihood is $\ell(\beta, \gamma|\boldsymbol{y}) = -\frac{1}{2}\log|\Sigma| + \log\int \exp\{-\psi(\boldsymbol{y}, \boldsymbol{z}; \beta, \gamma)\}\,\mathrm{d}\boldsymbol{z}$.

Let

$$\tilde{\boldsymbol{z}} = \underset{\boldsymbol{z}}{\operatorname{argmin}}\, \psi(\boldsymbol{y}, \boldsymbol{z}; \beta, \gamma). \tag{2.10}$$

Note that $\tilde{\boldsymbol{z}}$ is a function of $(\boldsymbol{y}; \beta, \gamma)$. Then by Laplace approximation (eq. 4 in Breslow and Clayton, 1993)

$$\ell(\beta, \gamma|\boldsymbol{y}) \approx -\frac{1}{2}\log|\Sigma| - \frac{1}{2}\log|\psi_{\boldsymbol{zz}}(\boldsymbol{y}, \tilde{\boldsymbol{z}}; \beta, \gamma)| - \psi(\boldsymbol{y}, \tilde{\boldsymbol{z}}; \beta, \gamma) \tag{2.11}$$

where $\psi_{\boldsymbol{zz}}(\boldsymbol{y}, \tilde{\boldsymbol{z}}; \beta, \gamma)$ denotes the matrix of the second order derivatives of ψ with respect to the components of $\boldsymbol{z}$ evaluated at $\tilde{\boldsymbol{z}}$. For fixed γ, Breslow and Clayton (1993) noted that the term $\log|\psi_{\boldsymbol{zz}}(\boldsymbol{y}, \tilde{\boldsymbol{z}}; \beta, \gamma)|$ varies very little with respect to β, therefore the estimate of β is obtained by maximizing $-\psi(\boldsymbol{y}, \tilde{\boldsymbol{z}}; \beta, \gamma)$, which is of the form of the log-likelihood for GLM. Hence, the IWLS algorithm which is used for obtaining the MLE $\hat{\beta}$ in GLM can be used for estimating β. Substituting $\hat{\beta}$ into (2.11) and noting that the profile likelihood for γ has the form of the likelihood if the data were following a Normal distribution, the estimate $\hat{\gamma}$ is obtained by REML. The obvious procedure to follow is first to set a starting value for γ, estimate β and use $\hat{\beta}$ to estimate γ. Then $\hat{\gamma}$ is used to obtain a new $\hat{\beta}$ which is in turn used to re-estimate γ. These steps are repeated until the algorithm converges.

From computational point of view, apart from the challenge of finding $\tilde{z}$ when its dimension is large, the algorithms suggested are easy to implement. Breslow and Clayton (1993) provided several examples that their method applies but they noted that it doesn't perform well in certain examples, e.g when used to analyze binary clustered data. Breslow and Lin (1995) and Lin and Breslow (1996) improved the method of Breslow and Clayton (1993) by including higher order terms in the Taylor expansion of ψ but even so there are cases where the estimates are highly biased. This is the case when the dimension of the random effects is comparable with the

sample size.

Direct Laplace Approximation

Shun and McCullagh (1995) looked at problems where the dimension of the random effects increases with the sample size. This assumption is necessary for the variance components to be estimated consistently but under this framework it is not clear if the remainder term in the classical Laplace approximation is bounded. In their paper Shun and McCullagh derived a formula that takes this into account by grouping terms according to their asymptotic order, an application of which can be found in Shun (1997).

More specifically, they noted that a suitable approximation to the log-likelihood in (2.7) is

$$-\tilde{\psi} - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\log|\tilde{\psi}_{\boldsymbol{zz}}| + \frac{1}{8}\sum_{ijkl}\tilde{\psi}_{ijkl}\tilde{\psi}^{ij}\tilde{\psi}^{kl} - \frac{1}{8}\sum_{ijklmt}\tilde{\psi}_{ijk}\tilde{\psi}_{lmt}\tilde{\psi}^{ij}\tilde{\psi}^{kl}\tilde{\psi}^{mt} - \frac{1}{12}\sum_{ijklmt}\tilde{\psi}_{ijk}\tilde{\psi}_{lmt}\tilde{\psi}^{il}\tilde{\psi}^{jm}\tilde{\psi}^{kt} \quad (2.12)$$

Here, the subscripts on ψ denote differentiation with respect to the elements of $\boldsymbol{z}$, the superscripts denote the inverse elements of the Hessian of ψ, $\psi_{\boldsymbol{zz}}$ and the tilde means that the function ψ and its derivatives are evaluated at the $\tilde{\boldsymbol{z}}$ defined in (2.10). In addition, the indices in the summations range over the dimension of $\boldsymbol{z}$. The expression in (2.12) is maximized simultaneously over (β, γ) to obtain the corresponding estimates.

It should be noted here that although their method performs well for small sample sizes, it becomes slow even for moderate samples because it involves the summation of many terms. In fact, Shun in his application excluded some non-small terms from the likelihood to speed up the convergence of the algorithm. Later, Noh and Lee (2007) found a way to include these terms when the design matrix of the random effects has many zeros, which is the case of the crossed random effects.

Although these approximation methods perform very well in many situations, the estimates obtained are biased because of the error in the Taylor expansion. Of course the bias can become smaller by including more terms in the expansion, as in Raudenbush et al. (2000), but on the other hand, the approximations become harder to evaluate.

Simulation based methods

The idea behind simulation based methods is to evaluate any integrals that occur by simulating random numbers from an appropriate distribution. Under some regularity conditions, if $X_1, \ldots, X_N$ is an i.i.d. sequence and $h(x)$ is some function then, by the Law of Large Numbers, $\mathbb{E}(h(X_1)) \approx N^{-1} \sum h(X_i)$.

Note that the log-likelihood in (2.7) can be expressed as $\mathbb{E}(f(\boldsymbol{y}|\boldsymbol{Z}))$ where the expectation is taken over the distribution of $\boldsymbol{Z}$ which is $\mathrm{N}(\boldsymbol{0}, \Sigma)$. A rather naive method to evaluate this expectation would be to simulate a large number of random effects from (2.6), plug them into $f(\boldsymbol{y}|\boldsymbol{z})$ and then average over the simulations. Unfortunately this simple method fails if the sample size is large, and that is because the conditional density $f(\boldsymbol{y}|\boldsymbol{z})$ would be so small that is numerically indistinguishable from 0.

Three suggestions were proposed by McCulloch (1997) with the key point of using the Metropolis-Hastings algorithm to simulate from the distribution of $\boldsymbol{Z}|\boldsymbol{y}$.

Monte Carlo EM

The EM algorithm is an iterative method for maximizing the likelihood in the presence of latent variables. In each iteration the parameters are chosen such that they maximize the expectation $\mathbb{E}(\log f(\boldsymbol{y}, \boldsymbol{Z})|\boldsymbol{y})$ over the distribution of $\boldsymbol{Z}|\boldsymbol{y}$ with parameters taken from the previous iteration. The parameters are updated until convergence of the estimates.

In GLMM β is updated by maximizing

$$\mathbb{E}(\log f(\boldsymbol{y}|\boldsymbol{Z}; \beta)|\boldsymbol{y}) \tag{2.13}$$

and γ is updated by maximizing

$$\mathbb{E}(\log f(\boldsymbol{Z}; \gamma)|\boldsymbol{y}) \tag{2.14}$$

Since none of (2.13) or (2.14) can be evaluated analytically, McCulloch (1997) suggests using the Metropolis-Hastings algorithm to simulate a sample from the distribution of $\boldsymbol{Z}|\boldsymbol{y}$ and evaluate the expectations numerically. On the other hand, some testing has to be made on whether the sample produced is approximately an i.i.d. sample from the target distribution. Alternatively Booth and Hobert (1999) suggested two other algorithms that don't require the simulation of

a Markov Chain, and hence are more efficient than the one proposed by McCulloch.

Monte Carlo Newton-Raphson

In another point of view, instead of trying to maximize the log-likelihood, McCulloch (1997) proposes estimating the parameters by setting the scores equal to 0. In this case, an estimate for β is obtained by solving

$$\mathbb{E}\left(\frac{\partial}{\partial \beta} \log f(\boldsymbol{y}|\boldsymbol{Z};\beta) \middle| \boldsymbol{y} \right) = 0 \tag{2.15}$$

and for γ by solving

$$\mathbb{E}\left(\frac{\partial}{\partial \gamma} \log f(\boldsymbol{Z};\gamma) \middle| \boldsymbol{y} \right) = 0 \tag{2.16}$$

The left hand side of (2.16) has an analytical expression and it can be solved easily. On the other hand (2.15) is not so easy to solve but it is evaluated numerically using the same algorithms as with the Monte Carlo EM case.

Simulated Maximum Likelihood

The idea of Simulated Maximum Likelihood is as follows. Suppose there exists a random variable $\boldsymbol{Z}^*$ with density function f^* such that $f^*(\boldsymbol{z};\beta,\gamma) > 0$ whenever $f(\boldsymbol{z}|\boldsymbol{y};\beta,\gamma) > 0$. Then the log-likelihood can be written as

$$\ell(\beta,\gamma|\boldsymbol{y}) = \log \mathbb{E}\left(\frac{f(\boldsymbol{y},\boldsymbol{Z}^*;\beta,\gamma)}{f^*(\boldsymbol{Z}^*;\beta,\gamma)} \right) \tag{2.17}$$

where the expectation is taken over the distribution of $\boldsymbol{Z}^*$. The expectation is then calculated numerically by simulating from the distribution of $\boldsymbol{Z}^*$ and averaging. We noted earlier that evaluating the log-likelihood by naive Monte Carlo integration is not possible but here if f^* is chosen appropriately then the simulated likelihood should be possible to evaluate. (In fact, the best choice for $f^*(z)$ is $f(\boldsymbol{z}|\boldsymbol{y})$.) To this end, McCulloch (1997) did not give a clear answer as to what to choose for f^* and in his example he chose $f^*(z;\beta,\gamma) = f(z;\gamma)$ which is of course unknown, but even in this case his method did not perform as well as the MCEM or MCNR methods.

2.1.2 Prediction

We now review methods for predicting the random effects themselves or other random variables that are correlated with them such that conditioned on the random effects, they are independent on the observations. We use the symbol Z_0 for the variable that we want to predict given a sample $\boldsymbol{Y}$.

Best Prediction

The best prediction for the random variable Z_0 is defined as the random variable $\hat{Z}_0 = \hat{Z}_0(\boldsymbol{Y})$ such that the mean square prediction error $\mathbb{E}\{(\hat{Z}_0 - Z_0)^2\}$ is minimized. This reduces to $\hat{Z}_0 = \mathbb{E}(Z_0|\boldsymbol{Y})$ which also depends on the parameters of the model. If the parameters were known, the best predictor can be calculated either by using the Metropolis-Hastings algorithm as was proposed by McCulloch (1997) or any of the two sampling methods of Booth and Hobert (1999), or approximate the expectation using Laplace approximation (see Vidoni, 2006, for the case of independent random effects). Of course, in most applications the parameters are not known. A way to overcome this problem is to replace them with a good estimate, obtained using one of the methods described in the previous section. This is known as the plug-in approach.

Plug-in approach

A method for obtaining prediction intervals is by constructing the *predictive density*, that is the conditional density of the variable we want to predict given the observations: $f(z_0|\boldsymbol{y};\beta,\gamma)$. It can be expressed as

$$f(z_0|\boldsymbol{y};\beta,\gamma) = \frac{\int f(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{y}|\boldsymbol{z};\beta) f(\boldsymbol{z};\gamma)\,\mathrm{d}\boldsymbol{z}}{\int f(\boldsymbol{y}|\boldsymbol{z};\beta) f(\boldsymbol{z};\gamma)\,\mathrm{d}\boldsymbol{z}} \tag{2.18}$$

Similarly, the predictive distribution function is written as

$$F(z_0|\boldsymbol{y};\beta,\gamma) = \frac{\int F(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{y}|\boldsymbol{z};\beta) f(\boldsymbol{z};\gamma)\,\mathrm{d}\boldsymbol{z}}{\int f(\boldsymbol{y}|\boldsymbol{z};\beta) f(\boldsymbol{z};\gamma)\,\mathrm{d}\boldsymbol{z}} \tag{2.19}$$

As we mentioned earlier, (2.18) and (2.19) depend on the parameters which are unknown. The *plug-in predictive density* is constructed by replacing the parameters with their estimates based on the sample $\boldsymbol{y}$ and is a rather simple to construct but on the other hand it has been criticized

as failing to take into account the uncertainty in estimating the parameters and assumes that they are the true values.

Barndorff-Nielsen and Cox (1996) made some suggestions on correcting the plug-in approach, although these have not been applied to GLMM. They considered the fact that the plug-in and the true predictive distribution are related asymptotically by

$$F(z_0|\boldsymbol{y};\hat{\beta},\hat{\gamma}) = F(z_0|\boldsymbol{y};\beta,\gamma) + k^{-1}D(z_0|\boldsymbol{y};\beta,\gamma) + O(k^{-3/2}) \tag{2.20}$$

as $k \to \infty$ where k is some constant related with the sample size and D is a known expression. Let z_α be the quantiles obtained by inverting the true predictive distribution and $\hat{z}_\alpha$ be the quantiles obtained by inverting the plug-in predictive distribution. Their first suggestion was to estimate z_α by $\hat{z}_{\alpha_1}$ where $\alpha_1 = \alpha - k^{-1}D(\hat{z}_\alpha|\boldsymbol{y};\hat{\beta},\hat{\gamma})$ in which case $F(\hat{z}_{\alpha_1}|\boldsymbol{y};\hat{\beta},\hat{\gamma}) = \alpha + O(k^{-3/2})$ instead of $O(k^{-1})$ that would have been if we used just $\hat{z}_\alpha$. We should note though, that if the correction is too large, then α_1 might not be between 0 and 1. Their second suggestion corrects $\hat{z}_\alpha$ directly by using $\hat{z}_\alpha - k^{-1}\frac{D(\hat{z}_\alpha|\boldsymbol{y};\hat{\beta},\hat{\gamma})}{f(\hat{z}_\alpha|\boldsymbol{y};\hat{\beta},\hat{\gamma})}$ as an estimator for z_α which doesn't have the disadvantage of the first method and gives the same order of accuracy. Furthermore, they derived an approximation to the predictive density of order $O(k^{-3/2})$ given by

$$f(z_0|\boldsymbol{y};\beta,\gamma) = (1 + \hat{r}'(z_0))f(z_0 + \hat{r}(z_0)|\boldsymbol{y};\hat{\beta},\hat{\gamma}) \tag{2.21}$$

where $\hat{r}(z_0) = D(z_0|\boldsymbol{y};\hat{\beta},\hat{\gamma})/f(z_0|\boldsymbol{y};\beta,\gamma)$ and $\hat{r}'(z_0) = (\mathrm{d}/\mathrm{d}z_0)\hat{r}(z_0)$.

2.1.3 Bayesian solution

We describe the Bayesian approaches to GLMM in a different section since from the Bayesian point of view, there is no distinction between estimation and prediction.

A major concern regarding Bayesian analysis is the choice of prior distribution for the parameters. On this subject, we note the papers by Zeger and Karim (1991) where they discuss Gibbs sampling in longitudinal GLMM, Karim and Zeger (1992) for similar ideas in crossed random effects models, Berger et al. (2001) and Diggle et al. (1998) discuss prior selection in spatial GLMM while Natarajan and Kass (2000) provide priors for the covariance matrix in a

general setting.

Bayesian Monte Carlo Methods

After the prior is selected, say $\beta \sim \pi(\beta)$ and independently $\gamma \sim \pi(\gamma)$, one proceeds by calculating confidence regions from the posterior distributions $\pi(\beta|\boldsymbol{y})$, $\pi(\gamma|\boldsymbol{y})$, and $f(z_0|\boldsymbol{y})$. The Monte Carlo methods proceed by constructing the conditional distributions $\pi(\beta|\boldsymbol{y}, \boldsymbol{z})$, $\pi(\gamma|\boldsymbol{z})$, $f(\boldsymbol{z}|\boldsymbol{y}, \beta, \gamma)$, and $f(z_0|\boldsymbol{z}, \gamma)$. As in general, these conditional distributions cannot be obtained analytically (besides $f(z_0|\boldsymbol{z}, \gamma)$ which is Normal), Markov Chain Monte Carlo techniques, such as Gibbs sampler and Metropolis-Hastings, are used to simulate from them which sometimes can be burdensome (see Clayton, 1996). In a recent paper, Fan et al. (2008) proposed a Sequential Monte Carlo algorithm for simulating from the posterior distributions which is faster than MCMC, though it does not completely avoid the simulation of Markov Chains.

Integrated Nested Laplace Approximation

Alternatively, Rue et al. (2009), suggest a new methodology based on Laplace approximation. Writing $\theta = (\beta, \gamma)$, they express the posterior as

$$\pi(\theta|\boldsymbol{y}) \propto \frac{f(\boldsymbol{y}, \boldsymbol{z}; \theta)\pi(\theta)}{f(\boldsymbol{z}|\boldsymbol{y}; \theta)} \tag{2.22}$$

and the Bayesian predictive density as

$$f(\boldsymbol{z}|\boldsymbol{y}) = \int f(\boldsymbol{z}|\boldsymbol{y}; \theta)\pi(\theta|\boldsymbol{y})\, \mathrm{d}\theta \tag{2.23}$$

The idea is to replace the denominator of (2.22) with its Normal approximation using first order Laplace approximation, centered around the point $\hat{\boldsymbol{z}} = \text{argmax}_{\boldsymbol{z}}\, f(\boldsymbol{y}, \boldsymbol{z}; \theta)$. Denoting the aforementioned approximation by $\tilde{f}_G(\boldsymbol{z}|\boldsymbol{y}; \theta)$ define

$$\tilde{\pi}(\theta|\boldsymbol{y}) \propto \frac{f(\boldsymbol{y}, \hat{\boldsymbol{z}}; \theta)\pi(\theta)}{\tilde{f}_G(\hat{\boldsymbol{z}}|\boldsymbol{y}; \theta)} \tag{2.24}$$

Equation (2.24) is an approximation to the posterior of θ which can be used to substitute the second term of the integrand in (2.23). The first term of the integrand is also computed by a separate application of Laplace approximation. Finally, the integration in (2.23) is performed

numerically by Gauss-Hermite quadrature.

This approach provides a quick and accurate way of obtaining approximations to the conditional distributions of interest and therefore obtain accurate predictions to the parameters and the random effects. On the other hand, the computational advantages of this method are in effect when the inverse covariance matrix of $f(\boldsymbol{z}|\gamma)$ is sparse and when the number of parameters is small.

Bayesian Predictive Distribution Function

In analogy to the plug-in predictive distribution, we note here the Bayesian predictive distribution function

$$F(z_0|\boldsymbol{y}) = \frac{\iiint F(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{y}|\boldsymbol{z};\beta) f(\boldsymbol{z};\gamma)\pi(\beta)\pi(\gamma)\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\beta\,\mathrm{d}\gamma}{\iiint f(\boldsymbol{y}|\boldsymbol{z};\beta) f(\boldsymbol{z};\gamma)\pi(\beta)\pi(\gamma)\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\beta\,\mathrm{d}\gamma} \tag{2.25}$$

Several authors have argued for the Bayesian predictive distribution against the plug-in one in the sense that it naturally takes into account the uncertainty in the parameters by assigning the prior distribution (see Geisser, 1993), nevertheless this should not always be conceded as Smith (1999) mentions examples where the plug-in predictive density performs better.

2.2 Modeling Geostatistical Data with GLMM

In this section we review methods for analyzing geostatistical data, that is, correlated data observed over a continuous spatial domain $\mathbb{S}$ whose correlation arises through their dependence on an unobserved spatial process $\mathcal{Z} = \{Z(s), s \in \mathbb{S}\}$. ($s$ here is an index representing the individual elements of $\mathbb{S}$.) In practice, only a finite sample $Y_1, \ldots, Y_n$ is observed corresponding to the sampling sites $s_1, \ldots, s_n$ and the objective is to predict the spatial process $\mathcal{Z}$ over the whole domain $\mathbb{S}$. In practice this is done by predicting $Z(s)$ on a fine grid that covers $\mathbb{S}$.

2.2.1 Gaussian Random Fields

A *Random Field* over the spatial domain $\mathbb{S}$ is a process $\mathcal{Z} = \{Z(s,\omega), s \in \mathbb{S}, \omega \in \Omega\}$ such that for fixed s, $Z(s,\omega)$ is a random variable on a probability space, say $(\Omega, \mathcal{A}, \Pr)$ while for fixed ω, $Z(s,\omega)$ is a realization of a stochastic process indexed by s.

The mean $\mu(s)$ and covariance $c(s, s')$ of the random field $\mathcal{Z}$ are defined on $\mathbb{S}$ and $\mathbb{S}^2$ respectively in the following way

$$\mu(s) = \int_{\Omega} Z(s, \omega) \, \mathrm{d} \Pr(\omega) \tag{2.26}$$

$$c(s, s') = \int_{\Omega} \{Z(s, \omega) - \mu(s)\}\{Z(s', \omega) - \mu(s')\} \, \mathrm{d} \Pr(\omega) \tag{2.27}$$

For simplicity, hereafter, we will avoid writing explicitly the second argument ω when we describe Random Fields.

Note that the definitions in (2.26) and (2.27) can be written as $\mu(s) = \mathbb{E}(Z(s))$ and $c(s, s') = \mathrm{Cov}(Z(s), Z(s'))$

Homogeneous Gaussian Random Field

In this section, we present a class of Random Fields, the *isotropic Gaussian Random Field*, that is usually assumed in order to assist with the inference regarding the probability measure Pr.

The random field $\mathcal{Z}(s)$ with mean $\mu(s)$ and covariance $c(s, s')$ is called *weakly stationary* if for all $s, s' \in \mathbb{S}$ the following three conditions hold:

1. $\mathbb{E}(Z(s)) = \mathbb{E}(Z(s'))$,
2. $\mathrm{Cov}(Z(s), Z(s)) < \infty$, and
3. $\mathrm{Cov}(Z(s), Z(s'))$ can be expressed as a function of $s - s'$.

A weaker form of stationarity is a process that is *intrinsically stationary* defined by replacing condition 3 above by

3'. $\mathrm{Var}(Z(s) - Z(s'))$ can be expressed as a function of $s - s'$.

If in addition to 1, 2 and 3', the following condition holds

4. $\mathrm{Var}(Z(s) - Z(s'))$ can be expressed as a function of $\|s - s'\|$.

then the process is called *isotropic*. If all conditions 1, 2, 3, and 4 hold then the process is called *homogeneous*.

A *Gaussian Random Field* is the Random Field $\mathcal{Z}$ on $\mathbb{S}$ where for every subset $\{s_1, \ldots, s_k\}$ of $\mathbb{S}$, the joint distribution of $(Z(s_1), \ldots, Z(s_k))$ is Gaussian. In connection with the previous definitions the distribution of a homogeneous Gaussian Random Field is characterized by its mean μ, which without loss of generality we will assume that $\mu = 0$, and by its covariance function $c(d_{ii'}) = c(s_i, s_{i'})$ where $d_{ii'} = \|s_i - s_{i'}\|$.

Intuitively, the probability measure underlining a stationary Gaussian Random Field is invariant under parallel transition of the co-ordinate system while for an isotropic Random Field it is invariant under rotations of the co-ordinate system.

Here we only consider zero-mean homogeneous Gaussian Random Fields.

Covariance Structure

A common practice is to assume that the covariance function $c(\cdot)$ is parameterized by a few parameters γ that have a reasonable interpretation regarding the structure of the covariance between two sampling sites. Some forms of the covariance function $c(\cdot)$ are

1. Exponential:
$$c(d) = \begin{cases} \gamma_1 + \gamma_2, & \text{if } d = 0 \\ \gamma_2 \exp\{-d/\gamma_3\}, & \text{if } d > 0 \end{cases} \tag{2.28}$$

2. Gaussian
$$c(d) = \begin{cases} \gamma_1 + \gamma_2, & \text{if } d = 0 \\ \gamma_2 \exp\{-(d/\gamma_3)^2\}, & \text{if } d > 0 \end{cases} \tag{2.29}$$

3. Spherical:
$$c(d) = \begin{cases} \gamma_1 + \gamma_2, & \text{if } d = 0 \\ \gamma_2\{1 - 1.5(d/\gamma_3) + 0.5(d/\gamma_3)^3\}, & \text{if } 0 < d < \gamma_3 \\ 0, & \text{if } d \geq \gamma_3 \end{cases} \tag{2.30}$$

4. Matérn:
$$c(d) = \begin{cases} \gamma_1 + \gamma_2, & \text{if } d = 0 \\ \gamma_2 \dfrac{\Gamma(\nu)}{2^{\nu-1}} \left(\dfrac{2\sqrt{\nu}\, d}{\gamma_3} \right)^{\nu} \mathcal{K}_\nu(2\sqrt{\nu}\, d/\gamma_3) & \text{if } d > 0 \end{cases} \tag{2.31}$$

γ_1 is called the *nugget* and is interpreted as the variability due to measurement error and microscale variation (i.e. the part of the variation that cannot be estimated because there are no two sampling sites close enough to indicate it). γ_2 is called the *partial sill* and is interpreted as the variance of the random field if there was no nugget. γ_3 is called the *range* and is interpreted as the distance at which the correlation function reaches 0 for the spherical form, or 5% of the partial sill for the other forms. In (2.31), ν is called the smoothness parameter and $\mathcal{K}_\nu$ corresponds to the modified Bessel function of the second kind of order ν (Abramowitz and Stegun, 1964).

The Gaussian Geostatistical Model

Suppose a Gaussian Random Field $\mathcal{Z}$ on $\mathbb{S}$. In practice, we sample at $s_1, \ldots, s_k$ and observe $\boldsymbol{Y} = (Y_1, \ldots, Y_k)$ whose mean is affected by a set of covariates X. For simplicity, let $\boldsymbol{Z} = (Z_1, \ldots, Z_k)$, $Z_i = Z(s_i)$, $i = 1, \ldots, k$ so that $\boldsymbol{Z} \sim \mathrm{N}_k(\boldsymbol{0}, \Sigma)$ where the (i, i') element of Σ is $c(\|s_i - s_{i'}\|)$, parameterized by γ.

The simplest type of relationship we can have between the observations and the covariates is to express the mean of $\boldsymbol{Y}$ as a linear combination of the covariates:

$$\boldsymbol{Y} = X\beta + \boldsymbol{Z} \tag{2.32}$$

The model in (2.32) implies that the distribution of $\boldsymbol{Y}$ is k-dimensional Gaussian with mean $X\beta$ and covariance matrix Σ and in this case estimation of β and γ can be performed by either weighted least squares, maximum likelihood, or restricted maximum likelihood.

Several tools under the name *kriging* have been developed for predicting the random variable Z_0 associated with the sampling site s_0 under the model (2.32). The original work on this subject was done by Krige (1951) and later developed by Matheron in a series of papers and books (e.g. Matheron, 1962, 1963). For a collection of these methods see Cressie (1993).

2.2.2 Spatial GLMM

In many applications, for example when the observations involve counts, the Gaussian Geostatistical model (2.32) is not appropriate. Diggle et al. (1998) extended the Gaussian geosta-

tistical model to include parametric families depending only on the mean of the spatial process in the same way that the classical linear model is extended to the generalized linear mixed model. In other words they define the linear predictor by

$$\boldsymbol{\eta} = X\beta + \boldsymbol{Z} \tag{2.33}$$

and they assumed that the conditional distribution of the observations $\boldsymbol{Y}$ given the random field $\boldsymbol{Z}$ is a member of the exponential family as defined in (2.1). In addition, all the assumptions regarding GLMM also hold for the spatial GLMM and as a consequence, all methods that have been developed for the analysis of GLMM can also be used for this model. Here we present a few from the literature.

Bayesian MCMC

As we mentioned earlier, when we want to predict the random effect Z_0 under GLMM, the Bayesian approach has the advantage that it incorporates the unknowing of the parameters into the prior. In their paper Diggle et al. (1998) assigned independent uniform priors with bounded support for the components of β and γ and used Metropolis-Hastings to simulate from the posterior distribution of γ while random samples from the posteriors of β and Z_0 were obtained by simulating from the normal distribution. Later on, Christensen et al. (2000) and Christensen and Waagepetersen (2002) commented on the use of Langevin-Hastings algorithm for the MCMC simulation as it leads to better convergence and mixing properties than the Metropolis-Hastings algorithm. They also suggested the use of non-informative flat priors for the components of β and non-informative inverse Gamma for the partial sill γ_2. For the range parameter γ_3, the use of improper prior results to improper posterior, hence they used uniform proper prior in the first paper and exponential prior in the second. The estimation of the nugget γ_1 was not considered.

Monte Carlo EM Gradient

In connection with the MCEM idea of McCulloch (1997), Zhang (2002) uses a Metropolis-Hastings algorithm to simulate from the distribution of $\boldsymbol{Z}|\boldsymbol{y}$ and calculate the expectations (2.13) and (2.14) needed for the E step but instead of maximizing those expectations at the M step, he applies a one-step Newton-Raphson update. This way he avoids performing full

maximization of the expectations which can be cumbersome. For prediction, he noted that the best predictor is expressed in terms of the expectations of the random effects at the sampling sites conditioned on the observations, i.e.

$$\mathbb{E}(Z_0|\boldsymbol{y}) = \sum_{i=1}^{k} w_i \mathbb{E}(Z_i|\boldsymbol{y}) \tag{2.34}$$

where the vector of w_i's, $\boldsymbol{w} = \mathbf{c}^\top \Sigma^{-1}$, $\mathbf{c} = \mathrm{Cov}(\boldsymbol{Z}, Z_0)$. Therefore, after the covariance parameters are estimated, the best predictor for Z_0 is obtained by evaluating (2.34) for each iteration of $\boldsymbol{Z}|\boldsymbol{y}$ and then averaging over all simulations.

Simulated Maximum Likelihood

Christensen (2004) demonstrated the use of Simulated MLE for the spatial GLMM. He wrote the likelihood function in the form of (2.17) and used simulation to approximate it. As we mentioned earlier, the best choice for f^* is $f(\boldsymbol{z}|\boldsymbol{y}; \beta, \gamma)$ which depends on the unknown parameters. For this reason the author suggested simulating from $f(\boldsymbol{z}|\boldsymbol{y}; \beta_0, \gamma_0)$ for some fixed values (β_0, γ_0) which are believed to be close to the true ones and which are updated after a few iterations.

2.3 Contribution of the thesis

This thesis contributes on topics related to Bayesian and frequentist methods for the analysis of Generalized Linear Mixed Models. The basic idea is to use asymptotic expansions to compute quantities of interest such as the likelihood or the predictive density. A key role is played by Laplace approximation, a method for approximating integrals, where an approximation of this type, suitable for our case, is derived. The particular feature of our approximation is that it allows the dimension of the variable that is integrated to increase to infinity, a necessary assumption for estimation and prediction for these models.

For parameter estimation, we propose an approximate likelihood method. There are several advantages when estimating the parameters this way. First, as our simulations show, the estimates obtained have low bias and mean square error and second, our method is fast to compute. Some details are also given on correcting the bias of the approximation.

We also suggest an approximate prediction method based on the same idea. We apply a high order approximation to the predictive density and replace the unknown parameters with their estimates, what is known as the plug-in approach. The proposed density belongs to the Gaussian family and prediction intervals can be easily computed. We present a simulation study where we compare our method with other approximate and simulation based methods and show that our method has similar performance with the other methods with less computation time.

From Bayesian perspective, we investigate the issue of prior selection. We derive approximations to the coverage probability bias and Kullback-Leibler divergence of the Bayesian predictive density constructed under different priors. These are computed for different simulations and for possible choices for priors that are found in the literature. We make selection based on the criterion of minimum coverage probability bias and minimum Kullback-Leibler divergence. Our approximation to the coverage probability bias agrees mostly to the one obtained by simulations but has smaller variation. We find that the best choices are uniform priors for the nugget and range parameter while for the partial sill we recommend either uniform prior or inverse gamma. Using exponential prior for the range, results to higher coverage probability bias unless the true value is close to the mean of the exponential prior.

CHAPTER 3

General Results

3.1 Model and Notation

Throughout this document we use indices to denote components, derivatives and summations. For the last purpose, any index that appears in an expression as a subscript and as a superscript, a summation over all possible values of that index is implicit. For this reason we will denote the components of a vector sometimes by subscripts and sometimes by superscripts. For example, the components of the three dimensional vector $\boldsymbol{x}$ will be written as x_1, x_2 and x_3 or as x^1, x^2 and x^3 depending on the expression i.e. $x_i x^i = x^i x_i = \sum_{i=1}^{3} (x_i)^2$ but $x^i x^i$ is the square of the ith element of $\boldsymbol{x}$: $(x_i)^2$. The (i,j) component of a matrix A will be written as a_{ij} and its inverse (when exists) will have components a^{ij}. It is also convenient to enclose any set of indices within square brackets to denote the sum over all partitions of those indices of the products of the corresponding arrays and a number withing square brackets to denote the different permutations of indices for the corresponding partition, i.e. $\boldsymbol{x}_{[ijk]} = x_{ijk} + [3] x_{ij} x_k + x_i x_j x_k = x_{ijk} + x_{ij} x_k + x_{ik} x_j + x_{jk} x_i + x_i x_j x_k$.

For any real function $f(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^k$, its derivative with respect to the i^{th} component of $\boldsymbol{x}$ is denoted by a subscript i.e. $f_i(\boldsymbol{x}) \coloneqq \frac{\partial f(\boldsymbol{x})}{\partial x_i}$ and $f_{ij}(\boldsymbol{x}) \coloneqq \frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}$. Furthermore, $f_{\boldsymbol{x}}$ is the gradient of f and $f_{\boldsymbol{xx}}$ is the Hessian matrix. Based on our notation on matrix inversion, f^{ij} is the (i,j) element of $f_{\boldsymbol{xx}}^{-1}$: the inverse of the Hessian matrix. Finally, when we refer to the probability density/mass function of a random variable, we will use the generic symbol $f(\cdot;\cdot)$ with the random variables written at the left of the semicolon and the parameters at the right, i.e. $f(x;\theta)$ is the density/mass of X depending on parameter θ and $f(x|y;\theta)$ is the conditional density of $X|Y$. In a similar fashion, we will write $F(\cdot;\cdot)$ for the cumulative distribution

function. An exception will be made when the distribution is Gaussian, in which case we will use the letters ϕ and Φ for the pdf and cdf respectively.

The vector of the response variable is denoted by $\boldsymbol{Y}$ with components $\{Y_{il}\, i = 1, \ldots, k, l = 1, \ldots, n_i\}$ repeatedly sampled at k different locations within a spatial domain $\mathbb{S}$. We assume the existence of an unobserved homogeneous Gaussian random field $\mathcal{Z}$ over the whole spatial region $\mathbb{S}$ such that conditioned on $\mathcal{Z}$ the observations are independent. We denote by $\boldsymbol{Z}$ the k-dimensional vector that consists of the components of $\mathcal{Z}$ that correspond to the sampled sites and we refer to it as the random effects. Furthermore, the mean $\mu_i = \mathbb{E}(Y_{il}|Z_i) = b(\theta_i)$ for some known differentiable function b, called the cumulant function, such that b' is strictly increasing, and variance $v_i(\mu_i)$ where v_i is a known function called the variance function (McCullagh and Nelder, 1999). The parameter θ_i relates to the linear predictor $\eta_i = \boldsymbol{x}_i^\top \beta + Z_i$ through the relationship $\mu_i = b(\theta_i) = g^{-1}(\eta_i)$ for some function g called the link function. In our asymptotic analysis we consider the case in which k and n_i increase to infinity with the n_i's having the same order, $\min\{n_i\} = O(n)$ but k is increasing in a lower rate, $k/n \to 0$.

We assume that the joint distribution of the random field $\mathcal{Z}$ is Normal with mean $\mathbf{0}$ and covariance matrix parameterized by γ, i.e. for the random effects $\boldsymbol{Z}$ we have,

$$\boldsymbol{Z} \sim \mathrm{N}_k(\mathbf{0}, \Sigma(\gamma)) \tag{3.1}$$

Conditioned on $\boldsymbol{Z}$, the density of $\boldsymbol{Y}$ has the form

$$f(\boldsymbol{y}|\boldsymbol{z}; \beta) = \exp\left\{ \sum_{i=1}^{k} y_i(\boldsymbol{x}_i^\top \beta + z_i) - \sum_{i=1}^{k} n_i b(\boldsymbol{x}_i^\top \beta + z_i) + \sum_{i=1}^{k} c(y_i) \right\} \tag{3.2}$$

for known functions b and c, where $y_i = \sum_{j=1}^{n_i} y_{ij}$. Note that the form of the density implies conditional independence among the observations given the corresponding random effects. Although in (3.2) we implicitly used the canonical link for the distribution of $\boldsymbol{y}$, the results that follow don't necessarily require this restriction.

We are interested in estimating the parameters (β, γ) of the model as well as predicting a component Z_0 of $\mathcal{Z}$ that corresponds to an unsampled site.

Under our model, the likelihood for the parameters is

$$\ell(\beta, \gamma; \boldsymbol{y}) = \int f(\boldsymbol{y}|\boldsymbol{z}; \beta)\phi(\boldsymbol{z}; \gamma)\,\mathrm{d}\boldsymbol{z} \tag{3.3}$$

which does not have an analytic expression. In addition, writing the distribution function of $Z_0|\boldsymbol{Z}$ as

$$\Phi(z_0|\boldsymbol{z}; \gamma) = \Phi\left(\frac{z_0 - \mu}{\tau}\right) \tag{3.4}$$

where $\Phi(\cdot)$ denotes the standard normal distribution function and

$$\mu = \mathbf{c}^\top \Sigma^{-1} \boldsymbol{z} \tag{3.5}$$

$$\tau^2 = \sigma_0^2 - \mathbf{c}^\top \Sigma^{-1} \mathbf{c} \tag{3.6}$$

$$\sigma_0^2 = \mathrm{Var}(Z_0) \tag{3.7}$$

$$\mathbf{c} = \mathrm{Cov}(\boldsymbol{Z}, Z_0) \tag{3.8}$$

the predictive distribution function for Z_0 given the data $\boldsymbol{Y}$ is

$$F(z_0|\boldsymbol{y}; \beta, \gamma) = \frac{\int \Phi(z_0|\boldsymbol{z}; \gamma) f(\boldsymbol{y}|\boldsymbol{z}; \beta)\phi(\boldsymbol{z}; \gamma)\,\mathrm{d}\boldsymbol{z}}{\int f(\boldsymbol{y}|\boldsymbol{z}; \beta)\phi(\boldsymbol{z}; \gamma)\,\mathrm{d}\boldsymbol{z}} \tag{3.9}$$

Similarly, the predictive density is written as

$$f(z_0|\boldsymbol{y}; \beta, \gamma) = \frac{\int \phi(z_0|\boldsymbol{z}; \gamma) f(\boldsymbol{y}|\boldsymbol{z}; \beta)\phi(\boldsymbol{z}; \gamma)\,\mathrm{d}\boldsymbol{z}}{\int f(\boldsymbol{y}|\boldsymbol{z}; \beta)\phi(\boldsymbol{z}; \gamma)\,\mathrm{d}\boldsymbol{z}} \tag{3.10}$$

Neither (3.9) nor (3.10) have an analytic expression and hence exact prediction intervals cannot be calculated exactly.

Before we reach the point of proposing methods for obtaining maximum likelihood estimates and prediction intervals, we derive a formula that allows us to approximate the likelihood and the predictive likelihood when the sample size is large.

3.2 Asymptotic Expansions of Integrals

3.2.1 Modified Laplace approximation

Shun and McCullagh (1995) proposed a modification of Laplace's approximation that can be used for evaluating integrals of the form

$$I_1 = \int e^{-g(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z} \tag{3.11}$$

where $g = O(n)$. Assuming that g has a unique minimum at $\hat{\boldsymbol{z}}$, Shun and McCullagh suggest an expansion of the integral around that minimum. They derive the identities

$$\log I_1 = -\hat{g} - \frac{1}{2}\log\left|\frac{\hat{g}_{\boldsymbol{zz}}}{2\pi}\right| + \sum_{m=1}^{\infty}\sum_{\substack{P,Q\\ P\vee Q=1}}\frac{(-1)^t}{(2m)!}\hat{g}_{p_1}\ldots\hat{g}_{p_t}\hat{g}^{q_1}\cdots\hat{g}^{q_m} \tag{3.12}$$

$$I_1 = e^{-\hat{g}}\left|\frac{\hat{g}_{\boldsymbol{zz}}}{2\pi}\right|^{-1/2}\sum_{m=1}^{\infty}\sum_{P,Q}\frac{(-1)^t}{(2m)!}\hat{g}_{p_1}\ldots\hat{g}_{p_t}\hat{g}^{q_1}\cdots\hat{g}^{q_m} \tag{3.13}$$

where the second sum in each of (3.12) and (3.13) is over all partitions P, Q such that $P = p_1|\ldots|p_t$ is a partition of $2m$ indices into t blocks, each of size 3 or more and $Q = q_1|\ldots|q_m$ is a partition of the same indices into m blocks, each of size 2. $P \vee Q = 1$ means that the union of the graphs produced by joining elements in the same block of the two partitions is connected e.g. $Q = i_1i_2|i_3i_4$ is connected with $P_1 = i_1|i_2i_3|i_4$ but not with $P_2 = i_1|i_2|i_3i_4$ (see Figure 3.1). The summation over all the possible values of the $2m$ indices is also implicit.

Figure 3.1: Connected partitions Q and P_1 (left) and unconnected Q and P_2 (right)

These formulae require expressing the integrand in a fully exponential form while for the results here we require the integrand to be written in the standard form (Tierney et al., 1989).

In our approach, we consider the following integral:

$$I_2 = \int \exp\{-g(\boldsymbol{z})\} \times f(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} \tag{3.14}$$

where f is not necessarily positive. Suppose that $\boldsymbol{z} \in \mathbb{R}^k$, $g(\boldsymbol{z}) = O(n)$ has a minimum at $\mathbf{0}$ and f and its derivatives are $o(n)$. We will develop a formula for the approximation of (3.14).

Taylor expansion of g around $\mathbf{0}$ gives

$$g(\boldsymbol{z}) = \hat{g} + \frac{1}{2!} z^{i_1} z^{i_2} \hat{g}_{i_1 i_2} + \frac{1}{3!} z^{i_1} z^{i_2} z^{i_3} \hat{g}_{i_1 i_2 i_3} + \frac{1}{4!} z^{i_1} z^{i_2} z^{i_3} z^{i_4} \hat{g}_{i_1 i_2 i_3 i_4} + \dots \tag{3.15}$$

where the subscripts under g imply differentiation with respect to the indicated component of $\boldsymbol{z}$ and the hats imply that the function or its derivatives are evaluated at $\mathbf{0}$. The indices range from 1 to k and the sums are over all indices. Let $\hat{g}_{\boldsymbol{zz}}$ denote the hessian matrix of g evaluated at $\mathbf{0}$.

A similar expansion of f around the same point gives

$$f(z) = \hat{f} + \hat{f}_{j_1} z^{j_1} + \frac{1}{2} \hat{f}_{j_1 j_2} z^{j_1} z^{j_2} + \dots \tag{3.16}$$

Thus

$$\begin{aligned}
I_2 &= e^{-\hat{g}} \int e^{-\frac{1}{2}\boldsymbol{z}^{\mathsf{T}} \hat{g}_{\boldsymbol{zz}} \boldsymbol{z}} \exp\left\{ -\frac{1}{3!} \hat{g}_{i_1 i_2 i_3} z^{i_1} z^{i_2} z^{i_3} - \frac{1}{4!} \hat{g}_{i_1 i_2 i_3 i_4} z^{i_1} z^{i_2} z^{i_3} z^{i_4} - \dots \right\} \\
&\qquad \times \left(\hat{f} + \hat{f}_{j_1} z^{j_1} + \frac{1}{2} \hat{f}_{j_1 j_2} z^{j_1} z^{j_2} + \dots \right) \mathrm{d}\boldsymbol{z} \\
&= e^{-\hat{g}} \int e^{-\frac{1}{2}\boldsymbol{z}^{\mathsf{T}} \hat{g}_{\boldsymbol{zz}} \boldsymbol{z}} \left(1 - \frac{1}{3!} \hat{g}_{[i_1 i_2 i_3]} z^{i_1} z^{i_2} z^{i_3} - \frac{1}{4!} \hat{g}_{[i_1 i_2 i_3 i_4]} z^{i_1} z^{i_2} z^{i_3} z^{i_4} - \dots \right) \\
&\qquad \times \left(\hat{f} + \hat{f}_{j_1} z^{j_1} + \frac{1}{2} \hat{f}_{j_1 j_2} z^{j_1} z^{j_2} + \dots \right) \mathrm{d}\boldsymbol{z} \\
&= e^{-\hat{g}} \left| \frac{\hat{g}_{\boldsymbol{zz}}}{2\pi} \right|^{-1/2} \mathbb{E}\left[\left(1 - \frac{1}{3!} \hat{g}_{[i_1 i_2 i_3]} Z^{i_1} Z^{i_2} Z^{i_3} - \frac{1}{4!} \hat{g}_{[i_1 i_2 i_3 i_4]} Z^{i_1} Z^{i_2} Z^{i_3} Z^{i_4} - \dots \right) \right. \\
&\qquad \left. \times \left(\hat{f} + \hat{f}_{j_1} Z^{j_1} + \frac{1}{2} \hat{f}_{j_1 j_2} Z^{j_1} Z^{j_2} + \dots \right) \right]
\end{aligned}$$

where $\boldsymbol{Z}$ is a normally distributed random variable with mean $\mathbf{0}$ and covariance matrix $\hat{g}_{\boldsymbol{zz}}^{-1}$.

Then,

$$I_2 = e^{-\hat{g}} \left| \frac{\hat{g}_{\boldsymbol{zz}}}{2\pi} \right|^{-1/2} \sum_{r \in \{0,3,4,\ldots\}} \sum_{s=0}^{\infty} (-1)^r \frac{1}{r!s!} \hat{g}_{[i_1 \ldots i_r]} \hat{f}_{j_1 \ldots j_s} \mathbb{E}\left[Z^{i_1} \cdots Z^{i_r} \cdot Z^{j_1} \cdots Z^{j_s} \right]$$

where we make the convention if $r = 0$ then $\hat{g}_{[i_1 \ldots i_r]} = 1$, if $s = 0$ then $\hat{f}_{j_1 \ldots j_s} = \hat{f}$ and if $r = s = 0$ then $\mathbb{E}\left[Z^{i_1} \cdots Z^{i_r} \cdot Z^{j_1} \cdots Z^{j_s} \right] = 1$

Using equation (2.8) from McCullagh (1987), I_2 becomes

$$I_2 = e^{-\hat{g}} \left| \frac{\hat{g}_{\boldsymbol{zz}}}{2\pi} \right|^{-1/2} \sum_{m=0}^{\infty} \sum_{s=0}^{2m} \sum_{P,Q} \frac{(-1)^t}{(2m)!} \hat{f}_{j_1 \ldots j_s} \hat{g}_{p_1} \cdots \hat{g}_{p_t} \hat{g}^{q_1} \cdots \hat{g}^{q_m} \tag{3.17}$$

where P is a partition of $2m - s$ indices into t blocks each of size 3 or more and Q is a partition of the same indices together with $\{j_1, \ldots, j_s\}$ into m blocks of size 2. It is not required for P and Q to be connected.

In the special case where $f(\boldsymbol{z}) > 0$, say $f(\boldsymbol{z}) = \exp\{h(\boldsymbol{z})\}$, then from (3.17),

$$\log I_2 = -\hat{g} + \hat{h} - \frac{1}{2} \log \left| \frac{1}{2\pi} \hat{g}_{\boldsymbol{zz}} \right| + \sum_{m=1}^{\infty} \frac{1}{(2m)!} \sum_{\substack{P,Q \\ P \vee Q = 1}} \chi_{p_1} \cdots \chi_{p_t} \cdot \hat{g}^{q_1} \cdots \hat{g}^{q_m} \tag{3.18}$$

where

$$\chi_{i_1 \cdots i_s} = \begin{cases} \hat{h}_{i_1 \cdots i_s} & \text{if } s \leq 2 \\ \hat{h}_{i_1 \cdots i_s} - \hat{g}_{i_1 \cdots i_s} & \text{if } s \geq 3 \end{cases}$$

3.2.2 Approximation to the ratio of two integrals

In the following sections we will need to approximate ratios of integrals e.g. when we want to approximate conditional densities. Suppose we want to approximate

$$\frac{I_2}{I_1} = \frac{\int \exp\{-g(\boldsymbol{z})\} \times f(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}}{\int e^{-g(\boldsymbol{z})} \, \mathrm{d}\boldsymbol{z}} \tag{3.19}$$

Using equations (3.13) and (3.17)

$$\frac{I_2}{I_1} = \frac{\sum_{m=1}^{\infty} \sum_{s=0}^{2m} \sum_{P,Q} \frac{(-1)^t}{(2m)!} \hat{f}_{j_1 \ldots j_s} \hat{g}_{p_1} \cdots \hat{g}_{p_t} \hat{g}^{q_1} \cdots \hat{g}^{q_m}}{\sum_{m=1}^{\infty} \sum_{P,Q} \frac{(-1)^t}{(2m)!} \hat{g}_{p_1} \cdots \hat{g}_{p_t} \hat{g}^{q_1} \cdots \hat{g}^{q_m}}$$

$$= \sum_{m=1}^{\infty} \sum_{s=1}^{2m} \sum_{P,Q} \frac{(-1)^t}{(2m)!} \hat{f}_{j_1 \ldots j_s} \hat{g}_{p_1} \cdots \hat{g}_{p_t} \hat{g}^{q_1} \cdots \hat{g}^{q_m} \tag{3.20}$$

As a demonstration of (3.20) suppose $kn^{-1} \to 0$, and that f and its derivatives are $O(1)$ as $k \to \infty$. In addition g and its derivatives are $O(n)$ when the differentiation is performed with respect to the same component of $\boldsymbol{z}$, otherwise they are $O(1)$. As we will show later in Lemma 1, the inverse Hessian matrix of g is $O(n^{-1})$ at the diagonal and $O(n^{-2})$ at the off diagonal elements as $k \to \infty$. This a typical situation which we encounter in the subsequent sections. The numerator of (3.19) is approximated by

$$\hat{f} - \frac{1}{8}\hat{f}\hat{g}_{i_1 i_2 i_3 i_4}\hat{g}^{i_1 i_2}\hat{g}^{i_3 i_4} + \frac{1}{8}\hat{f}\hat{g}_{i_1 i_2 i_3}\hat{g}_{i_4 i_5 i_6}\hat{g}^{i_1 i_2}\hat{g}^{i_3 i_4}\hat{g}^{i_5 i_6} + \frac{1}{12}\hat{f}\hat{g}_{i_1 i_2 i_3}\hat{g}_{i_4 i_5 i_6}\hat{g}^{i_1 i_4}\hat{g}^{i_2 i_5}\hat{g}^{i_3 i_6}$$
$$- \frac{1}{2}\hat{f}_{i_1}\hat{g}_{i_2 i_3 i_4}\hat{g}^{i_1 i_2}\hat{g}^{i_3 i_4} + \frac{1}{2}\hat{f}_{j_1 j_2}\hat{g}^{j_1 j_2} + O(n^{-1} \vee k^2 n^{-2}) \tag{3.21}$$

where besides the first term: $\hat{f}$, all the other terms in (3.21) are $O(k\, n^{-1})$. A similar expansion exists for the denominator by replacing f in (3.21) by 1. Thus (3.20) becomes after we take $\hat{f}$ as a common factor

$$\begin{aligned}
\frac{I_2}{I_1} = \hat{f} \Bigg(& 1 - \frac{1}{8}\hat{g}_{i_1 i_2 i_3 i_4}\hat{g}^{i_1 i_2}\hat{g}^{i_3 i_4} + \frac{1}{8}\hat{g}_{i_1 i_2 i_3}\hat{g}_{i_4 i_5 i_6}\hat{g}^{i_1 i_2}\hat{g}^{i_3 i_4}\hat{g}^{i_5 i_6} + \frac{1}{12}\hat{g}_{i_1 i_2 i_3}\hat{g}_{i_4 i_5 i_6}\hat{g}^{i_1 i_4}\hat{g}^{i_2 i_5}\hat{g}^{i_3 i_6} \\
& - \frac{1}{2}\frac{\hat{f}_{i_1}}{\hat{f}}\hat{g}_{i_2 i_3 i_4}\hat{g}^{i_1 i_2}\hat{g}^{i_3 i_4} + \frac{1}{2}\frac{\hat{f}_{j_1 j_2}}{\hat{f}}\hat{g}^{j_1 j_2} + O(k^2 n^{-1}) \Bigg) \\
& \times \Bigg(1 - \frac{1}{8}\hat{g}_{i_1 i_2 i_3 i_4}\hat{g}^{i_1 i_2}\hat{g}^{i_3 i_4} + \frac{1}{8}\hat{g}_{i_1 i_2 i_3}\hat{g}_{i_4 i_5 i_6}\hat{g}^{i_1 i_2}\hat{g}^{i_3 i_4}\hat{g}^{i_5 i_6} \\
& \quad + \frac{1}{12}\hat{g}_{i_1 i_2 i_3}\hat{g}_{i_4 i_5 i_6}\hat{g}^{i_1 i_4}\hat{g}^{i_2 i_5}\hat{g}^{i_3 i_6} + O(n^{-1}) \Bigg)^{-1}
\end{aligned} \tag{3.22}$$

Employing the identity $(1-\epsilon)^{-1} = 1 + \epsilon + O(\epsilon^2)$ we have in (3.22) after canceling between the numerator and the denominator

$$\begin{aligned}
\frac{I_2}{I_1} &= \hat{f} \left(1 - \frac{1}{2}\frac{\hat{f}_{i_1}}{\hat{f}}\hat{g}_{i_2 i_3 i_4}\hat{g}^{i_1 i_2}\hat{g}^{i_3 i_4} + \frac{1}{2}\frac{\hat{f}_{j_1 j_2}}{\hat{f}}\hat{g}^{j_1 j_2} + O(k^2 n^{-1}) \right) \\
&= \hat{f} - \frac{1}{2}\hat{f}_{i_1}\hat{g}_{i_2 i_3 i_4}\hat{g}^{i_1 i_2}\hat{g}^{i_3 i_4} + \frac{1}{2}\hat{f}_{j_1 j_2}\hat{g}^{j_1 j_2} + O(n^{-1})
\end{aligned} \tag{3.23}$$

CHAPTER 4

Likelihood Methods

Define

$$\ell(\beta, \gamma | \boldsymbol{y}, \boldsymbol{z}) = \log f(\boldsymbol{y}|\boldsymbol{z}; \beta) + \log f(\boldsymbol{z}; \gamma) \tag{4.1}$$

to be the log-likelihood when the *complete* dataset $(\boldsymbol{y}, \boldsymbol{z})$ is observed. Then, the likelihood based only on $\boldsymbol{y}$ is defined by integrating over the unobserved random effects:

$$\ell(\beta, \gamma | \boldsymbol{y}) = \log \int \exp\{\ell(\beta, \gamma | \boldsymbol{y}, \boldsymbol{z})\} \, \mathrm{d}\boldsymbol{z} \tag{4.2}$$

which is of the form (3.11). Joint maximization of (4.2) with respect to (β, γ) results to the Maximum Likelihood Estimates for those parameters. In order to be able to derive the order of the asymptotic approximations, we need to know the order of the elements of $\ell^{-1}_{\boldsymbol{zz}}(\beta, \gamma | \boldsymbol{y}, \boldsymbol{z})$, the inverse Hessian matrix of the log-likelihood of the complete data. The following lemma gives the answer

Lemma 1. *If $k = o(n)$ then the diagonal elements of $\ell^{-1}_{\boldsymbol{zz}}(\beta, \gamma | \boldsymbol{y}, \boldsymbol{z})$ are $O(n^{-1})$ and the off diagonal are $O(n^{-2})$.*

Proof. Keeping only the terms that depend on $\boldsymbol{z}$, the Hessian of the complete log-likelihood has the form

$$\ell_{\boldsymbol{zz}} = n\, D - \Sigma^{-1}$$

where D is diagonal with elements of order $O(1)$ while Σ^{-1} has elements of order $O(1)$ and the dimension of these matrices is $k \times k$. Then, using the identity

$$(I - \varepsilon\, A)^{-1} = I + \varepsilon\, A + \varepsilon^2 A^2 + \varepsilon^3 A^3 + \ldots$$

we have

$$\begin{aligned}
\ell_{\boldsymbol{zz}}^{-1} &= (n\,D - \Sigma^{-1})^{-1} \\
&= n^{-1}D^{-1}\{I - n^{-1}(D\Sigma)^{-1}\}^{-1} \\
&= n^{-1}D^{-1}\{I + n^{-1}(D\Sigma)^{-1} + O(k\,n^{-2})\} \\
&= n^{-1}D^{-1} + n^{-2}(D\Sigma D)^{-1} + O(k\,n^{-3})
\end{aligned}$$

where we can see that the diagonal elements of $\ell_{\boldsymbol{zz}}^{-1}(\beta,\gamma|\boldsymbol{y},\boldsymbol{z})$ are $O(n^{-1})$ and the off diagonal are $O(n^{-2})$. □

4.1 The Conditional Distribution of the Random Effects

We derive an approximation to the conditional distribution of the random effects $\boldsymbol{Z}$ given the observations $\boldsymbol{Y}$ and the parameters (β,γ) by approximating the cumulant generating function of the distribution in question. We first start by approximating the moment generating function.

Let $\ell(\boldsymbol{z})$ be the complete likelihood defined in (4.1). For given $(\boldsymbol{y},\beta,\gamma)$, $\hat{\boldsymbol{z}} = \operatorname{argmax}_{\boldsymbol{z}}\,\ell(\boldsymbol{z})$ and denote the evaluation of $\ell(\boldsymbol{z})$ and its derivatives at $\hat{\boldsymbol{z}}$ by a hat over the corresponding function. Then

$$\begin{aligned}
\mathbb{E}(e^{\boldsymbol{t}^\top\boldsymbol{Z}}|\boldsymbol{Y}) &= \frac{\int e^{\boldsymbol{t}^\top\boldsymbol{z}}e^{\ell(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z}}{\int e^{\ell(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z}} \\
&= e^{\boldsymbol{t}^\top\hat{\boldsymbol{z}}} - \frac{1}{2}e^{\boldsymbol{t}^\top\hat{\boldsymbol{z}}}t_{i_1}t_{i_2}\hat{\ell}^{i_1i_2} + \frac{1}{2}e^{\boldsymbol{t}^\top\hat{\boldsymbol{z}}}t_{i_1}\hat{\ell}_{i_2i_2i_2}\hat{\ell}^{i_1i_2}\hat{\ell}^{i_2i_2} + O(k^2n^{-2}) \\
&= e^{\boldsymbol{t}^\top\hat{\boldsymbol{z}}}\left(1 - \frac{1}{2}t_{i_1}t_{i_2}\hat{\ell}^{i_1i_2} + \frac{1}{2}t_{i_1}\hat{\ell}_{i_2i_2i_2}\hat{\ell}^{i_1i_2}\hat{\ell}^{i_2i_2} + O(k^2n^{-2})\right) \qquad (4.3)
\end{aligned}$$

Therefore, by taking logarithms on (4.3) and using the fact that $\log(1+\epsilon) \approx \epsilon + o(\epsilon^2)$, we observe that the cumulant generating function of the conditional distribution of the random effects matches up to order $O(k^2n^{-2})$ the one of a k-dimensional Normally distributed random variable with mean whose ith element is $\hat{z}_i + \frac{1}{2}\hat{\ell}_{i_1i_1i_1}\hat{\ell}^{i_1i_1}\hat{\ell}^{i_1i}$ and covariance matrix whose (i_1,i_2) element is $-\hat{\ell}^{i_1i_2}$.

4.2 Fisher Information Matrix

For the Gaussian model, Mardia and Marshall (1984) showed that the MLE is consistent and asymptotically Normal with covariance matrix having block diagonal form with two blocks, one corresponding to the inverse Fisher information for β and one to the inverse Fisher information for γ. We show a similar result for our model and derive the asymptotic form of the Fisher information matrix.

Let

$$h(\boldsymbol{y}|\boldsymbol{z};\beta) = -\log f(\boldsymbol{y}|\boldsymbol{z};\beta) \tag{4.4}$$

$$h(\boldsymbol{z};\gamma) = -\log f(\boldsymbol{z};\gamma) \tag{4.5}$$

$$h(\boldsymbol{y},\boldsymbol{z};\beta,\gamma) = h(\boldsymbol{y}|\boldsymbol{z};\beta) + h(\boldsymbol{z};\gamma) \tag{4.6}$$

Then, the log-likelihood for the observed data is written as

$$\ell(\beta,\gamma|\boldsymbol{y}) = \log \int e^{-h(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)}\,\mathrm{d}\boldsymbol{z} \tag{4.7}$$

so, up to first order,

$$\frac{\partial}{\partial\beta_m}\ell(\beta,\gamma|\boldsymbol{y}) = -\frac{\int h_m(\boldsymbol{y}|\boldsymbol{z};\beta)e^{-h(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)}\,\mathrm{d}\boldsymbol{z}}{\int e^{-h(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)}\,\mathrm{d}\boldsymbol{z}} \approx -h_m(\boldsymbol{y}|\hat{\boldsymbol{z}};\beta) \tag{4.8}$$

$$\frac{\partial}{\partial\gamma_j}\ell(\beta,\gamma|\boldsymbol{y}) = -\frac{\int h_j(\boldsymbol{z};\gamma)e^{-h(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)}\,\mathrm{d}\boldsymbol{z}}{\int e^{-h(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)}\,\mathrm{d}\boldsymbol{z}} \approx -h_j(\hat{\boldsymbol{z}};\gamma) \tag{4.9}$$

where

$$\hat{\boldsymbol{z}} = \underset{\boldsymbol{z}}{\operatorname{argmin}}\, h(\boldsymbol{y},\boldsymbol{z};\beta,\gamma) \tag{4.10}$$

We can show (see Appendix) that $\hat{\boldsymbol{z}}(\boldsymbol{Y})$ converges to $\boldsymbol{Z}$ in probability as $k \to \infty$, therefore, as $k \to \infty$, $(\partial/\partial\beta_m)\ell(\beta,\gamma|\boldsymbol{Y}) \xrightarrow{p} -h_m(\boldsymbol{Y}|\boldsymbol{Z};\beta)$ and $(\partial/\partial\gamma_j)\ell(\beta,\gamma|\boldsymbol{Y}) \xrightarrow{p} -h_j(\boldsymbol{Z};\gamma)$.

The asymptotic expression for the likelihood for β is the same as the likelihood for GLM with a total of $\sum n_i$ observations, hence the bias of the MLE for β has order $O((\sum n_i)^{-1/2})$. Similarly, the asymptotic expression for the likelihood of γ is the same as the one in the Gaussian case, therefore, according to the result of Mardia and Marshall, the bias of the MLE for γ has

order $O(k^{-1/2})$. Furthermore, since

$$\mathbb{E}\{h_m(\boldsymbol{Y}|\boldsymbol{Z};\beta)\, h_j(\boldsymbol{Z};\gamma)\} = \mathbb{E}[\mathbb{E}\{h_m(\boldsymbol{Y}|\boldsymbol{Z};\beta)|\boldsymbol{Z}\}h_j(\boldsymbol{Z};\gamma)] = 0,$$

the estimates for β and γ are asymptotically uncorrelated.

Let

$$\ell_{j_1 j_2}(\beta,\gamma|\boldsymbol{Y}) = \frac{\partial}{\partial\gamma_{j_1}\gamma_{j_2}}\ell(\beta,\gamma|\boldsymbol{Y}) \tag{4.11}$$

and κ_{j_1,j_2} be the (j_1,j_2) element of the Fisher information matrix. Then, it can be shown that $\kappa_{j_1,j_2} = -\mathbb{E}\{\ell_{j_1 j_2}(\beta,\gamma|\boldsymbol{Y})\}$ where

$$\begin{aligned}\ell_{j_1 j_2}(\beta,\gamma|\boldsymbol{Y}) = &-\frac{\int h_{j_1 j_2}(\boldsymbol{z};\gamma)e^{-h(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)}\,\mathrm{d}\boldsymbol{z}}{\int e^{-h(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)}\,\mathrm{d}\boldsymbol{z}} + \frac{\int h_{j_1}(\boldsymbol{z};\gamma)h_{j_2}(\boldsymbol{z};\gamma)e^{-h(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)}\,\mathrm{d}\boldsymbol{z}}{\int e^{-h(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)}\,\mathrm{d}\boldsymbol{z}} \\ &- \frac{\int h_{j_1}(\boldsymbol{z};\gamma)e^{-h(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)}\,\mathrm{d}\boldsymbol{z}}{\int e^{-h(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)}\,\mathrm{d}\boldsymbol{z}} \cdot \frac{\int h_{j_2}(\boldsymbol{z};\gamma)e^{-h(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)}\,\mathrm{d}\boldsymbol{z}}{\int e^{-h(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)}\,\mathrm{d}\boldsymbol{z}}\end{aligned} \tag{4.12}$$

Applying (3.20), an asymptotic expansion of (4.12) around $\hat{\boldsymbol{z}}$ is

$$\begin{aligned}\ell_{j_1 j_2} = &-\left(\hat{h}_{j_1 j_2} + \frac{1}{2}\hat{h}_{i_1 i_2 j_1 j_2}\hat{h}^{i_1 i_2} - \frac{1}{2}\hat{h}_{i_1 j_1 j_2}\hat{h}_{i_2 i_2 i_2}\hat{h}^{i_1 i_2}\hat{h}^{i_2 i_2} + O(k^2 n^{-2})\right) \\ &+ \left(\hat{h}_{j_1}\hat{h}_{j_2} + \frac{1}{2}(\hat{h}_{i_1 i_2 j_1}\hat{h}_{j_2} + 2\hat{h}_{i_1 j_1}\hat{h}_{i_2 j_2} + \hat{h}_{j_1}\hat{h}_{i_1 i_2 j_2})\hat{h}^{i_1 i_2}\right. \\ &\quad \left. - \frac{1}{2}(\hat{h}_{i_1 j_1}\hat{h}_{j_2} + \hat{h}_{j_1}\hat{h}_{i_1 j_2})\hat{h}_{i_2 i_2 i_2}\hat{h}^{i_1 i_2}\hat{h}^{i_2 i_2} + O(k^2 n^{-2})\right) \\ &- \left(\hat{h}_{j_1} + \frac{1}{2}\hat{h}_{i_1 i_2 j_1}\hat{h}^{i_1 i_2} - \frac{1}{2}\hat{h}_{i_1 j_1}\hat{h}_{i_2 i_2 i_2}\hat{h}^{i_1 i_2}\hat{h}^{i_2 i_2} + O(k^2 n^{-2})\right) \\ &\quad \times \left(\hat{h}_{j_2} + \frac{1}{2}\hat{h}_{i_1 i_2 j_2}\hat{h}^{i_1 i_2} - \frac{1}{2}\hat{h}_{i_1 j_2}\hat{h}_{i_2 i_2 i_2}\hat{h}^{i_1 i_2}\hat{h}^{i_2 i_2} + O(k^2 n^{-2})\right)\end{aligned} \tag{4.13}$$

After some cancellation in (4.13), we approximate the (j_1,j_2) element of the information matrix by

$$\begin{aligned}&\hat{h}_{j_1 j_2} + \frac{1}{2}\hat{h}_{i_1 i_2 j_1 j_2}\hat{h}^{i_1 i_2} - \frac{1}{2}\hat{h}_{i_1 j_1 j_2}\hat{h}_{i_2 i_2 i_2}\hat{h}^{i_1 i_2}\hat{h}^{i_2 i_2} - \hat{h}_{i_1 j_1}\hat{h}_{i_2 j_2}\hat{h}^{i_1 i_2} \\ &\qquad + \frac{1}{4}(\hat{h}_{i_1 i_2 j_1}\hat{h}^{i_1 i_2} - \hat{h}_{i_1 j_1}\hat{h}_{i_2 i_2 i_2}\hat{h}^{i_1 i_2}\hat{h}^{i_2 i_2})(\hat{h}_{i_1 i_2 j_2}\hat{h}^{i_1 i_2} - \hat{h}_{i_1 j_2}\hat{h}_{i_2 i_2 i_2}\hat{h}^{i_1 i_2}\hat{h}^{i_2 i_2})\end{aligned} \tag{4.14}$$

where the error of the approximation is $O(k^2 n^{-2})$.

4.3 Approximation to the Likelihood

Write (4.2) as

$$\ell(\beta,\gamma|\boldsymbol{y}) = \int \exp\{-h(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)\}\,\mathrm{d}\boldsymbol{z} \tag{4.15}$$

and define $\hat{\boldsymbol{z}}$ by (4.10). Then by (3.12), ignoring the terms that don't depend on the parameters,

$$\begin{aligned}\ell(\beta,\gamma;\boldsymbol{y}) = &-\hat{h} - \frac{1}{2}\log|h_{\boldsymbol{zz}}| \\ &- \frac{1}{8}\hat{h}_{iiii}\hat{h}^{ii}\hat{h}^{ii} + \frac{1}{12}\hat{h}_{iii}\hat{h}_{iii}\hat{h}^{ii}\hat{h}^{ii}\hat{h}^{ii} + \frac{1}{8}\hat{h}_{i_1i_1i_1i_1}\hat{h}_{i_2i_2i_2i_2}\hat{h}^{i_1i_1}\hat{h}^{i_2i_2}\hat{h}^{i_1i_2} + O(k\,n^{-2})\end{aligned} \tag{4.16}$$

where the functions in the right hand side are evaluated at $\hat{\boldsymbol{z}}$.

The terms $\hat{h}_{iiii}\hat{h}^{ii}\hat{h}^{ii}$ and $\hat{h}_{iii}\hat{h}_{iii}\hat{h}^{ii}\hat{h}^{ii}\hat{h}^{ii}$ appearing in (4.16) have order $O(k\,n^{-1})$ and the term $\hat{h}_{i_1i_1i_1}\hat{h}_{i_2i_2i_2}\hat{h}^{i_1i_1}\hat{h}^{i_2i_2}\hat{h}^{i_1i_2}$ has order $O(k^2\,n^{-2})$. The remainder terms which are excluded from (4.16), such as $h_{iiiiii}h^{ii}h^{ii}h^{ii}$ and $h_{iiii}h_{iiii}h^{ii}h^{ii}h^{ii}h^{ii}$, have order $O(k\,n^{-2})$.

On the other hand, if k is too large, obtaining $\hat{\boldsymbol{z}}$ accurately can be numerically challenging. A second approach would be to write the likelihood in the form of (3.14) i.e.

$$L(\beta,\gamma|\boldsymbol{y}) = \int \phi(\boldsymbol{z};\gamma)\exp\{-h(\boldsymbol{y}|\boldsymbol{z};\beta)\}\,\mathrm{d}\boldsymbol{z} \tag{4.17}$$

where now $h(\boldsymbol{y}|\boldsymbol{z};\beta) = -\sum y_i\eta_i + \sum n_i b(\eta_i)$. Then, letting $\hat{\boldsymbol{z}} = \operatorname{argmin}_{\boldsymbol{z}} h(\boldsymbol{y}|\boldsymbol{z};\beta)$, we have the following approximation

$$\begin{aligned}\ell(\beta,\gamma|\boldsymbol{y}) = &-\frac{1}{2}\log|\Sigma| - \frac{1}{2}\hat{\boldsymbol{z}}^{\intercal}\Sigma^{-1}\hat{\boldsymbol{z}} - \frac{1}{2}\log\left|\hat{h}_{\boldsymbol{zz}} + \Sigma^{-1}\right| + \frac{1}{2}\frac{\hat{\phi}_{i_1}}{\hat{\phi}}\frac{\hat{\phi}_{i_2}}{\hat{\phi}}\hat{v}^{i_1i_2} \\ &- \frac{1}{2}\frac{\hat{\phi}_{i_1}}{\hat{\phi}}\hat{h}_{i_2i_2i_2}\hat{v}^{i_1i_2}\hat{v}^{i_2i_2} - \frac{1}{8}\hat{h}_{iiii}\hat{v}^{ii}\hat{v}^{ii} + \frac{5}{24}\hat{h}_{i_1i_1i_1}\hat{h}_{i_2i_2i_2}\hat{v}^{i_1i_1}\hat{v}^{i_1i_2}\hat{v}^{i_2i_2} + O(k\,n^{-2})\end{aligned} \tag{4.18}$$

where $v^{i_1i_2}$ is the (i_1,i_2) element of $(\hat{h}_{\boldsymbol{zz}} + \Sigma^{-1})^{-1}$. Comparing (4.18) to (4.16), both have the same asymptotic order. The main advantage of (4.18) is that it performs faster because it doesn't require the use of an optimization algorithm for finding $\hat{z}$ and it can be consider an alternative to other computational intensive methods such as MCMC. On the other hand, there is no guarantee that it actually has a maximum as the high order correction terms sometimes become too large and the resulted estimates become biased. A bias corrected estimate can be

calculated in this using bootstrap methods, as described in the next section.

4.3.1 Bootstrap bias correction and bootstrap variance

Here we apply the methodology for bias correction via bootstrapping as described in Chapter 10 of Efron and Tibshirani (1993).

Denote by $(\hat{\beta}, \hat{\gamma})$ the maximizer of (4.18). The bootstrap estimate of the bias is obtained as follows: First we independently draw B realizations from the underlying model with true parameters $(\hat{\beta}, \hat{\gamma})$. For each realization $r \in \{1, \ldots, B\}$ we estimate the parameters of the model by the approximate likelihood method. Let $(\hat{\beta}_r^*, \hat{\gamma}_r^*)$ denote the estimate from the rth realization and define

$$\hat{\beta}^* = B^{-1} \sum_{r=1}^{B} \hat{\beta}_r^*$$
$$\hat{\gamma}^* = B^{-1} \sum_{r=1}^{B} \hat{\gamma}_r^*$$

Then an estimate for the bias of the approximate likelihood is given by

$$\widehat{\text{bias}}(\hat{\beta}) = \hat{\beta}^* - \hat{\beta}$$
$$\widehat{\text{bias}}(\hat{\gamma}) = \hat{\gamma}^* - \hat{\gamma}$$

and a bootstrap bias correction to our original estimates is

$$\hat{\beta} - \widehat{\text{bias}}(\hat{\beta})$$
$$\hat{\gamma} - \widehat{\text{bias}}(\hat{\gamma})$$

The bootstrap realizations can also be used to approximate the variance of our estimates:

$$\widehat{\text{var}}(\hat{\beta}) = (B-1)^{-1} \sum_{r=1}^{B} (\hat{\beta}_r^* - \hat{\beta}^*)^2$$
$$\widehat{\text{var}}(\hat{\gamma}) = (B-1)^{-1} \sum_{r=1}^{B} (\hat{\gamma}_r^* - \hat{\gamma}^*)^2$$

As Efron and Tibshirani (1993) point out, bootstrap bias correction is not always effective. In general, the variability of the estimates is increased, so they recommend applying the bootstrap bias correction if the estimate of the bias is at least of the same magnitude as the estimate of the variance.

4.3.2 Assessing the error of the approximation

Let

$$p(\boldsymbol{z}) = \hat{h} + \frac{1}{2}\hat{h}_{i_1 i_2}(z^{i_1} - \hat{z}^{i_1})(z^{i_2} - \hat{z}^{i_2}) + \frac{1}{3!}\hat{h}_{i_1 i_2 i_3}(z^{i_1} - \hat{z}^{i_1})(z^{i_2} - \hat{z}^{i_2})(z^{i_3} - \hat{z}^{i_3}) + \frac{1}{4!}h_{i_1 i_2 i_3 i_4}(z^{i_1} - \hat{z}^{i_1})(z^{i_2} - \hat{z}^{i_2})(z^{i_3} - \hat{z}^{i_3})(z^{i_4} - \hat{z}^{i_4}) \quad (4.19)$$

i.e. the first four terms in the Taylor expansion of $h(\boldsymbol{z})$ around $\hat{\boldsymbol{z}}$ where $h(\boldsymbol{z})$ is given by either (4.4) or (4.6) according to whether (4.16) or (4.18) is used. Define the remainder of this expansion by

$$r(\boldsymbol{z}) = h(\boldsymbol{z}) - p(\boldsymbol{z}) \quad (4.20)$$

where the functions h, p, and r implicitly depend on $(\beta, \gamma, \boldsymbol{y})$. In the following we will assume that (4.16) is used to approximate the likelihood but similar corrections can be carried when (4.18) is used.

We have, after transforming to $\hat{\boldsymbol{z}} = \boldsymbol{0}$,

$$\begin{aligned}
\int e^{-p(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z} &= e^{-\hat{h}} \int \exp\{-\frac{1}{2}\boldsymbol{z}^{\mathsf{T}}\hat{h}_{\boldsymbol{zz}}\boldsymbol{z}\} \exp\{-\frac{1}{3!}\hat{h}_{i_1 i_2 i_3} z^{i_1} z^{i_2} z^{i_3} - \frac{1}{4!}\hat{h}_{i_1 i_2 i_3 i_4} z^{i_1} z^{i_2} z^{i_3} z^{i_4}\}\,\mathrm{d}\boldsymbol{z} \\
&\approx e^{-\hat{h}} \int \exp\{-\frac{1}{2}\boldsymbol{z}^{\mathsf{T}}\hat{h}_{\boldsymbol{zz}}\boldsymbol{z}\} \left\{1 - \frac{1}{3!}\hat{h}_{i_1 i_2 i_3} z^{i_1} z^{i_2} z^{i_3} - \frac{1}{4!}\hat{h}_{i_1 i_2 i_3 i_4} z^{i_1} z^{i_2} z^{i_3} z^{i_4} \right. \\
&\qquad \left. + \frac{1}{2}\left(\frac{1}{3!}\hat{h}_{i_1 i_2 i_3} z^{i_1} z^{i_2} z^{i_3} + \frac{1}{4!}\hat{h}_{i_1 i_2 i_3 i_4} z^{i_1} z^{i_2} z^{i_3} z^{i_4}\right)^2 \right\} \mathrm{d}\boldsymbol{z} \\
&= e^{-\hat{h}} \left|\frac{1}{2\pi}\hat{h}_{\boldsymbol{zz}}\right|^{-1/2} \left(1 - \frac{1}{8}\hat{h}_{i_1 i_2 i_3 i_4}\hat{h}^{i_1 i_2}\hat{h}^{i_3 i_4} + \frac{1}{8}\hat{h}_{i_1 i_2 i_3}\hat{h}_{i_4 i_5 i_6}\hat{h}^{i_1 i_2}\hat{h}^{i_3 i_4}\hat{h}^{i_5 i_6} \right. \\
&\quad \left. + \frac{1}{12}\hat{h}_{i_1 i_2 i_3}\hat{h}_{i_4 i_5 i_6}\hat{h}^{i_1 i_4}\hat{h}^{i_2 i_5}\hat{h}^{i_3 i_6} + \frac{1}{128}\hat{h}_{i_1 i_2 i_3 i_4}\hat{h}_{i_5 i_6 i_7 i_8}\hat{h}^{i_1 i_2}\hat{h}^{i_3 i_4}\hat{h}^{i_5 i_6}\hat{h}^{i_7 i_8}\right.
\end{aligned}$$

$$+\frac{1}{16}\hat{h}_{i_1i_2i_3i_4}\hat{h}_{i_5i_6i_7i_8}\hat{h}^{i_1i_2}\hat{h}^{i_3i_5}\hat{h}^{i_4i_6}\hat{h}^{i_7i_8}+\frac{1}{48}\hat{h}_{i_1i_2i_3i_4}\hat{h}_{i_5i_6i_7i_8}\hat{h}^{i_1i_5}\hat{h}^{i_2i_6}\hat{h}^{i_3i_7}\hat{h}^{i_4i_8}\Big) \tag{4.21}$$

which by (4.16), is equal to the likelihood $L(\beta,\gamma|\boldsymbol{y})$ up to order $O(k\,n^{-2})$. Denote the right hand side of (4.21) by $\hat{L}_0(\beta,\gamma|\boldsymbol{y})$ and

$$(\hat{\beta}_0,\hat{\gamma}_0)=\underset{(\beta,\gamma)}{\operatorname{argmax}}\,\hat{L}_0(\beta,\gamma|\boldsymbol{y}) \tag{4.22}$$

Now write

$$\begin{aligned}L(\beta,\gamma|\boldsymbol{y})&=\int e^{-h(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z}\\&=\int e^{-p(\boldsymbol{z})-r(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z}\\&=\int e^{-p(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z}\int\frac{e^{-p(\boldsymbol{z})}}{\int e^{-p(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z}}e^{-r(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z}\end{aligned} \tag{4.23}$$

The first term in (4.23), $\int e^{-p(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z}$, is what we propose in (4.16), hence

$$\int e^{-h(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z}=\int e^{-p(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z}\big(1+O(k\,n^{-2})\big)$$

On the other hand,

$$\frac{e^{-p(\boldsymbol{z})}}{\int e^{-p(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z}}\approx\frac{e^{-h(\boldsymbol{z})}}{\int e^{-h(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z}}=f(\boldsymbol{z}|\boldsymbol{y};\beta,\gamma)$$

so, by writing the second integral in (4.23) as

$$R(\beta,\gamma|\boldsymbol{y})=\int\frac{e^{-p(\boldsymbol{z})}}{\int e^{-p(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z}}e^{-r(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z}\approx\int f(\boldsymbol{z}|\boldsymbol{y};\beta,\gamma)e^{-r(\boldsymbol{z}|\boldsymbol{y};\beta,\gamma)}\,\mathrm{d}\boldsymbol{z} \tag{4.24}$$

we obtain an expression of the error of the approximation.

There are two ways we can evaluate the right hand side of (4.24), both of which use the result from section 4.1, namely that $f(\boldsymbol{z}|\boldsymbol{y};\beta,\gamma)$ is approximately the Normal density. The first method is by importance sampling using the Normal approximation to $f(\boldsymbol{z}|\boldsymbol{y};\beta,\gamma)$ as the

importance density. Let $\hat{R}_2(\beta, \gamma|\boldsymbol{y})$ be the remainder approximated this way and

$$\hat{L}_2(\beta, \gamma|\boldsymbol{y}) = \hat{L}_0(\beta, \gamma|\boldsymbol{y})\hat{R}_2(\beta, \gamma|\boldsymbol{y}) \tag{4.25}$$

The approximate likelihood estimates are obtained by maximizing (4.25) with respect to the parameters, i.e.

$$(\hat{\beta}_2, \hat{\gamma}_2) = \underset{(\beta,\gamma)}{\operatorname{argmax}}\, \hat{L}_2(\beta, \gamma|\boldsymbol{y}). \tag{4.26}$$

Davis and Rodriguez-Yam (2005) applied this method to correct the Laplace approximation under a different setting and found that $(\hat{\beta}_2, \hat{\gamma}_2)$ has smaller bias compared to $(\hat{\beta}_0, \hat{\gamma}_0)$. On the other hand, random sampling has to be performed at every function evaluation of the optimization algorithm used to maximize (4.25) which can be time consuming. For this reason Davis and Rodriguez-Yam (2005) suggest a first order approximation to the logarithm of (4.24) by applying Taylor expansion around $(\hat{\beta}_0, \hat{\gamma}_0)$, i.e.

$$\log R(\beta, \gamma|\boldsymbol{y}) \approx \log R(\hat{\beta}_0, \hat{\gamma}_0|\boldsymbol{y}) + \big((\beta, \gamma) - (\hat{\beta}_0, \hat{\gamma}_0)\big)^{\mathsf{T}} \nabla \log R(\hat{\beta}_0, \hat{\gamma}_0|\boldsymbol{y}) \tag{4.27}$$

where $\nabla \log R(\hat{\beta}_0, \hat{\gamma}_0|\boldsymbol{y})$ denotes the gradient of $\log R(\beta, \gamma|\boldsymbol{y})$ with respect to (β, γ) evaluated at $(\hat{\beta}_0, \hat{\gamma}_0)$. As the quantities in the right hand side of (4.27) are unknown, they replaced $\log R(\hat{\beta}_0, \hat{\gamma}_0|\boldsymbol{y})$ by $\log R_2(\hat{\beta}_0, \hat{\gamma}_0|\boldsymbol{y})$ and $\nabla \log R(\hat{\beta}_0, \hat{\gamma}_0|\boldsymbol{y})$ by

$$\frac{\log R_2(\hat{\beta}_0 + \delta, \hat{\gamma}_0 + \delta|\boldsymbol{y}) - \log R_2(\hat{\beta}_0, \hat{\gamma}_0|\boldsymbol{y})}{\delta}$$

for sufficiently small δ. This way the importance sampling approximation to the remainder is carried only once which significantly improves the speed of the algorithm. On the other hand, since $R_2(\hat{\beta}_0 + \delta, \hat{\gamma}_0 + \delta|\boldsymbol{y})$ and $R_2(\hat{\beta}_0, \hat{\gamma}_0|\boldsymbol{y})$ are computed by simulation and for small δ, it is not clear how accurate the numerical differentiation is. Here we suggest a different way of approximating $\nabla \log R(\hat{\beta}_0, \hat{\gamma}_0|\boldsymbol{y})$ which is also based on importance sampling. The idea is to write

$$\nabla R(\hat{\beta}_0, \hat{\gamma}_0|\boldsymbol{y}) = \int f(\boldsymbol{z}|\boldsymbol{y}; \hat{\beta}_0, \hat{\gamma}_0) e^{-r(\boldsymbol{z}|\boldsymbol{y}; \hat{\beta}_0, \hat{\gamma}_0)} \Big(\nabla \log f(\boldsymbol{z}|\boldsymbol{y}; \hat{\beta}_0, \hat{\gamma}_0) - \nabla r(\boldsymbol{z}|\boldsymbol{y}; \hat{\beta}_0, \hat{\gamma}_0) \Big) \, \mathrm{d}\boldsymbol{z} \tag{4.28}$$

which can be evaluated using the same importance sample that was used for the evaluation of $\hat{R}_2(\hat{\beta}_0, \hat{\gamma}_0|\boldsymbol{y})$. Therefore, denoting by $\nabla\hat{R}_1(\hat{\beta}_0, \hat{\gamma}_0|\boldsymbol{y})$ the approximation to (4.28) computed this way and by

$$\hat{R}_1(\beta, \gamma|\boldsymbol{y}) = \exp\left\{\log \hat{R}_2(\hat{\beta}_0, \hat{\gamma}_0|\boldsymbol{y}) + \left((\beta, \gamma) - (\hat{\beta}_0, \hat{\gamma}_0)\right)^{\intercal} \nabla\hat{R}_1(\hat{\beta}_0, \hat{\gamma}_0|\boldsymbol{y})/\hat{R}_2(\hat{\beta}_0, \hat{\gamma}_0|\boldsymbol{y})\right\} \quad (4.29)$$

we define

$$\hat{L}_1(\beta, \gamma|\boldsymbol{y}) = \hat{L}_0(\beta, \gamma|\boldsymbol{y})\hat{R}_1(\beta, \gamma|\boldsymbol{y}) \quad (4.30)$$

and

$$(\hat{\beta}_1, \hat{\gamma}_1) = \underset{(\beta,\gamma)}{\operatorname{argmax}}\, \hat{L}_1(\beta, \gamma|\boldsymbol{y}) \quad (4.31)$$

The estimates $(\hat{\beta}_1, \hat{\gamma}_1)$ can be used as alternative to the more computational intensive estimates $(\hat{\beta}_2, \hat{\gamma}_2)$ for correcting the error of the approximation to the likelihood.

4.3.3 Example: Binomial Spatial Data

Suppose we observe n_i Bernoulli random variables at location i with a total of k locations and we use the canonical link $g(\mu) = \log\{\mu/(1-\mu)\}$. Then

$$\log f(\boldsymbol{y}|\boldsymbol{z}; \beta, \gamma) = -h(\boldsymbol{y}|\boldsymbol{z}; \beta, \gamma) = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}\eta_i - \sum_{i=1}^{k} n_i \log(1 + e^{\eta_i})$$

where $\eta_i = x_i\beta + z_i$. It's easy to see that $h(\boldsymbol{y}|\boldsymbol{z}; \beta, \gamma)$ has a maximum at $\hat{z}_i = \log(\bar{y}_{i\cdot}/(1-\bar{y}_{i\cdot})) - x_i\beta$ for $\bar{y}_{i\cdot}$ being the average of the y_{ij}'s for fixed i. Two remarks on this

1. In rare cases we might have $\bar{y}_{i\cdot} = 0$ or 1 causing problems in finding $\hat{z}_i$. Therefore when a 0 occurs we take it as it is equal to $.5/n_i$ and when 1 occurs to $1 - .5/n_i$.

2. The calculations can be simplified by noting that when we evaluate $h(\boldsymbol{y}|\boldsymbol{z}; \beta, \gamma)$ and its derivatives at $\hat{\boldsymbol{z}}$ the terms that depend on β cancel i.e. h and its derivatives are constants in the likelihood.

Assuming that the n_i's increase to infinity with the same order, say $\min\{n_i\} = O(n)$, then the order of h is $O(k\,n)$ and its derivatives with respect to $\boldsymbol{z}$ is $O(n)$ if the differentiation is

with respect to one component and 0 if the differentiation is with respect to different components. The order of ϕ_i in (4.18) depends on the form of Σ: here we assume increasing domain asymptotics in the spirit of Mardia and Marshall (1984), in which case the order is $O(1)$. By Lemma 1 we have that the diagonal elements of $(\hat{h}_{\boldsymbol{zz}} + \Sigma^{-1})^{-1}$ have order $O(n^{-1})$ while the off-diagonal elements have order $O(n^{-2})$. Thus, an asymptotic expansion to the log-likelihood (up to a constant) of order $O(k\, n^{-2})$ obtained from (4.18) where

$$\begin{aligned}
\hat{h}_{\boldsymbol{zz}} &= \operatorname{diag}\left\{ y_{i\cdot}\left(1 - \frac{y_{i\cdot}}{n_i}\right)\right\} \\
\hat{h}_{iii} &= y_{i\cdot}\left(1 - \frac{y_{i\cdot}}{n_i}\right)\left(1 - 2\frac{y_{i\cdot}}{n_i}\right) \\
\hat{h}_{iiii} &= y_{i\cdot}\left(1 - \frac{y_{i\cdot}}{n_i}\right)\left(1 - 6\frac{y_{i\cdot}}{n_i} + 6\frac{y_{i\cdot}^2}{n_i^2}\right)
\end{aligned}$$

4.3.4 Simulations

We compare the performance of the second order Laplace approximation to the likelihood (LA 2) as given by (4.18) with three other methods: first by taking the logit transformation to the observed probabilities and then performing maximum likelihood on the transformed data assuming they follow a normal distribution (trans.), the first order Laplace approximation (LA 1), and the MCMC method from the `R` package `geoRglm` (R Development Core Team, 2008; Christensen and Ribeiro, 2002) which computed the likelihood by drawing random samples from the distribution $f(\boldsymbol{z}|\boldsymbol{y}; \beta, \gamma)$. The transformation and the first order Laplace approximation are lower order approximations to the log-likelihood and they consist of the first two and the first three terms of (4.18) respectively.

We randomly choose $k = 50$ locations on a 10×10 grid (Figure 4.1) from where we simulate 1000 realizations of a Gaussian random field $\boldsymbol{Z}$ with exponential covariance function given by

$$\mathcal{C}(\gamma) = \begin{cases} \gamma_1 + \gamma_2 & \text{if } d_{ij} = 0 \\ \gamma_2 \exp\left\{-d_{ij}/\gamma_3\right\} & \text{if } d_{ij} \neq 0 \end{cases}$$

with parameters $\gamma = (0.2, 2.0, 4.0)$ corresponding to nugget, partial sill and range and d_{ij} is the euclidean distance between the locations i and j. The linear predictor at location i is given by

$\eta_i = x_i\beta + z_i$ where x_i is the ith row of the $k \times 2$ matrix X with the first column equal to 1 and the second equal to the first co-ordinate of the locations, and $\beta = (-1.0, 0.2)$. We consider estimation when the nugget is unknown and when it's not. Conditioned on the random field we repeatedly generate 1000 binomial observations with parameters $n_i = 60$ in the case where the nugget is known and $n_i = 200$ when the nugget is unknown, and $p_i = e^{\eta_i}/(1 + e^{\eta_i})$ at location i. The reason for the larger sample in the unknown-nugget case was because for some simulations, the profile likelihood for the nugget didn't have a maximum for $n_i = 60$ for the approximation methods. Starting values for the fixed effects are obtained by fitting a GLM on the observations and then applying variogram estimation on the residuals to obtain starting values for the covariance parameters. As indicated by Stein (1999, sec. 6.2) the parameters γ_2 and γ_3 are not identifiable in infill asymptotics but their ratio is, and since our simulation design corresponds more closely to infill asymptotics than increasing domain asymptotics, we also provide an estimate of $\log(\gamma_3/\gamma_2)$. We compare the estimates from each method with the corresponding REML estimates obtained assuming that the random effects were actually observed. This comparison is useful because frequently the estimated variogram matches the data better than the true variogram. The programs were executed on a computer with 3.20 GHz processor and 1 Gb RAM.

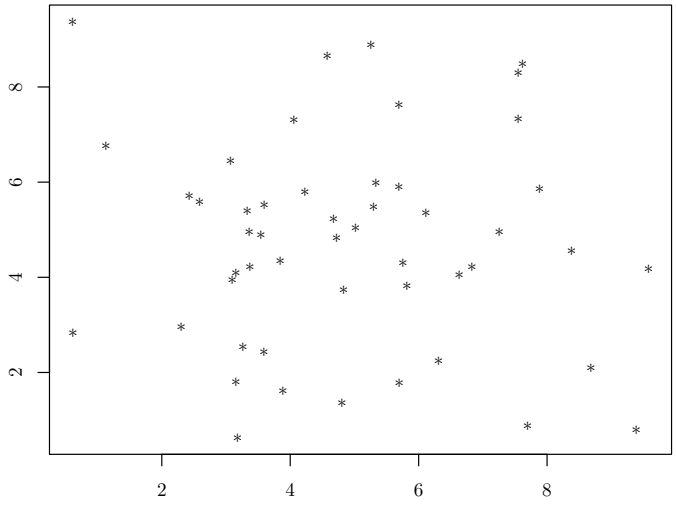

Figure 4.1: Observed locations for estimation.

The results of the simulations are summarized in Tables 4.1 and 4.2. The estimates obtained using the MCMC method have the least bias and are closer to the estimated values when

the random effects are observed. The second order Laplace approximation (LA 2) performs generally well in terms of mean square error. In fact, it has the smallest mean square error among all methods for the estimation of β. Also it is generally close to the REML estimates obtained if the random effects were observed, again with the smallest difference for the estimation of β. MCMC has low bias for the estimation of the fixed effects but takes much longer.

		β_0	β_1	γ_2	γ_3	$\log(\gamma_3/\gamma_2)$	time (sec)
trans.	Bias	−0.0180	0.0068	−0.7049	−2.5494	−0.6873	864
	S.D.	0.03712	0.00548	0.02179	0.03592	0.01750	
	M.S.E.	1.3779	0.0298	0.9716	7.7899	0.8619	
	M.S. diff. REML	0.1114	0.0020	3.1432	27.3643	0.6675	
LA 1	Bias	−0.0233	0.0077	−0.5925	−2.7741	−0.9493	887
	S.D.	0.03772	0.00550	0.02142	0.02802	0.01337	
	M.S.E.	1.4233	0.0303	0.8098	8.4805	1.3305	
	M.S. diff. REML	0.1351	0.0023	2.9635	30.6541	1.1238	
LA 2	Bias	0.0690	−0.0103	−1.0102	−1.9184	−0.0223	807
	S.D.	0.03323	0.00498	0.02001	0.05557	0.03473	
	M.S.E.	1.1093	0.0249	1.4208	6.7681	0.2841	
	M.S. diff. REML	0.0777	0.0015	3.8777	19.4539	0.1450	
MCMC	Bias	0.0140	−0.0007	−0.9837	−2.2573	−0.2025	27682
	S.D.	0.03539	0.00533	0.01298	0.06462	0.03270	
	M.S.E.	1.2523	0.0284	1.1362	9.2705	0.2548	
	M.S. diff. REML	0.0977	0.0017	4.4771	26.5111	0.1412	

Table 4.1: Simulation results for comparing the performance of each method when the nugget (γ_1) is known. The fourth row from each method corresponds to the mean square of the differences from the REML estimation assuming that the random effects are observed.

		β_0	β_1	γ_1	γ_2	γ_3	$\log(\gamma_3/\gamma_2)$	time (sec)
trans.	Bias	0.0105	3.611×10^{-3}	-0.0222	-0.7415	-2.0705	-1.5519	1725
	S.D.	0.03607	0.00531	0.00525	0.02180	0.04615	0.03573	
	M.S.E.	1.3009	0.02825	0.0280	1.0249	6.4164	3.6852	
	M.S. diff. REML	0.1384	1.319×10^{-3}	0.0148	9.3500	50.3339	0.7777	
LA 1	Bias	0.0088	3.667×10^{-3}	0.0059	-0.7315	-2.1090	-1.5723	1655
	S.D.	0.03619	0.00532	0.00570	0.02186	0.04419	0.03674	
	M.S.E.	1.3108	0.02833	0.0325	1.0129	6.4011	3.8221	
	M.S. diff. REML	0.1475	1.416×10^{-3}	0.0192	9.3344	50.8865	0.8914	
LA 2	Bias	0.0699	-7.143×10^{-3}	-0.0702	-0.8382	-2.0225	-1.4638	1748
	S.D.	0.03494	0.00511	0.00509	0.02042	0.05260	0.03841	
	M.S.E.	1.0650	0.02521	0.0276	1.1130	6.9512	3.6183	
	M.S. diff. REML	0.1161	1.209×10^{-3}	0.0167	9.5095	48.4351	1.1570	
MCMC	Bias	0.0269	-0.6667×10^{-3}	-0.0459	-0.7738	-2.0369	-1.7173	81398
	S.D.	0.03529	0.00521	0.00539	0.02593	0.05453	0.02402	
	M.S.E.	1.2459	0.02717	0.0312	1.2712	7.1227	3.5262	
	M.S. diff. REML	0.1345	1.267×10^{-3}	0.0197	9.4852	50.2075	0.6742	

Table 4.2: Simulation results for comparing the performance of each method when the nugget (γ_1) is unknown. The fourth row from each method corresponds to the mean square of the differences from the REML estimation assuming that the random effects are observed.

CHAPTER 5

Prediction Methods

Let's consider now the issue of predicting the random effect, Z_0 say, at location s_0 based on the observations $y_1, \ldots, y_k$ at locations $s_1, \ldots, s_k$.

The joint density of $(\boldsymbol{Z}, Z_0)$ is written as

$$(\boldsymbol{Z}, Z_0)^\mathsf{T} \sim \mathrm{N}_{k+1}\left[(\mathbf{0}, 0)^\mathsf{T}, \begin{pmatrix} \Sigma & \mathbf{c} \\ \mathbf{c}^\mathsf{T} & \sigma_0^2 \end{pmatrix}\right] \tag{5.1}$$

where the terms in the covariance matrix depend on the variance components γ.

We write the density of $Z_0|\{\boldsymbol{Z} = \boldsymbol{z}\}$ as

$$\phi(z_0|\boldsymbol{z};\gamma) = \tau^{-1}\phi\left(\frac{z_0 - \mu}{\tau}\right) \tag{5.2}$$

where $\mu = \mathbf{c}^\mathsf{T}\Sigma^{-1}\boldsymbol{z}$ and $\tau^2 = \sigma_0^2 - \mathbf{c}^\mathsf{T}\Sigma^{-1}\mathbf{c}$.

An important role for prediction is played by the *predictive density* $f(z_0|\boldsymbol{y};\beta,\gamma)$. Using the fact that conditioned on $\boldsymbol{Z}$, the random effect at location Z_0 is independent of the observations at the other locations $\boldsymbol{Y}$, we write

$$f(z_0|\boldsymbol{y};\beta,\gamma) = \frac{\int \phi(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)\,\mathrm{d}\boldsymbol{z}}{\int f(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)\,\mathrm{d}\boldsymbol{z}} \tag{5.3}$$

For the same reason that the likelihood cannot be expressed analytically, the predictive density does not have a closed form expression.

In connection to the predictive density, we write the *predictive distribution function*

$$F(z_0|\boldsymbol{y};\beta,\gamma) = \frac{\int \Phi(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)\,\mathrm{d}\boldsymbol{z}}{\int f(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)\,\mathrm{d}\boldsymbol{z}} \tag{5.4}$$

As is indicated by the expressions in (5.3) and (5.4), the predictive density depends on the unknown parameters β and γ. A way to overcome this is to replace the unknown parameters with some consistent estimates $(\hat{\beta},\hat{\gamma})$. The predictive density constructed by this method is called the *plug-in predictive density*.

5.1 Plug-in Predictive Density

Suppose that based on the sample $\boldsymbol{y} = (y_1,\ldots,y_k)^\intercal$ drawn from the sampling sites $s_1,\ldots,s_k$, we estimate the parameter β by $\hat{\beta}$, and γ by $\hat{\gamma}$. The plug-in predictive density is given by

$$f(z_0|\boldsymbol{y};\hat{\beta},\hat{\gamma}) = \frac{\int \phi(z_0|\boldsymbol{z};\hat{\gamma}) f(\boldsymbol{y},\boldsymbol{z};\hat{\beta},\hat{\gamma})\,\mathrm{d}\boldsymbol{z}}{\int f(\boldsymbol{y},\boldsymbol{z};\hat{\beta},\hat{\gamma})\,\mathrm{d}\boldsymbol{z}} \tag{5.5}$$

and similarly, the *plug-in predictive distribution function* by

$$F(z_0|\boldsymbol{y};\hat{\beta},\hat{\gamma}) = \frac{\int \Phi(z_0|\boldsymbol{z};\hat{\gamma}) f(\boldsymbol{y},\boldsymbol{z};\hat{\beta},\hat{\gamma})\,\mathrm{d}\boldsymbol{z}}{\int f(\boldsymbol{y},\boldsymbol{z};\hat{\beta},\hat{\gamma})\,\mathrm{d}\boldsymbol{z}} \tag{5.6}$$

We now proceed to construct an approximation to the predictive distribution of Z_0 using similar techniques with the approximation to the likelihood function. Write

$$f(z_0|\boldsymbol{y};\hat{\beta},\hat{\gamma}) = \frac{\int \phi(z_0|\boldsymbol{z};\hat{\gamma}) \exp\{-h(\boldsymbol{y},\boldsymbol{z};\hat{\beta},\hat{\gamma})\,\mathrm{d}\boldsymbol{z}}{\int \exp\{-h(\boldsymbol{y},\boldsymbol{z};\hat{\beta},\hat{\gamma})\,\mathrm{d}\boldsymbol{z}} \tag{5.7}$$

and define $\hat{\boldsymbol{z}} = \operatorname{argmin} h(\boldsymbol{y},\boldsymbol{z};\hat{\beta},\hat{\gamma})$. Then, by (3.20)

$$f(z_0|\boldsymbol{y};\hat{\beta},\hat{\gamma}) = \hat{\phi} - \frac{1}{2}\hat{\phi}_{i_1}\hat{h}_{i_2i_2i_2}\hat{h}^{i_1i_2}\hat{h}^{i_2i_2} + \frac{1}{2}\hat{\phi}_{i_1i_2}\hat{h}^{i_1i_2} + O(k\,n^{-2}) \tag{5.8}$$

Taking $\hat{\phi}$ as a common factor in (5.8) we write

$$f(z_0|\boldsymbol{y};\hat{\beta},\hat{\gamma}) = \hat{\phi}\left(1 - \frac{1}{2}\frac{\hat{\phi}_{i_1}}{\hat{\phi}}\hat{h}_{i_2i_2i_2}\hat{h}^{i_1i_2}\hat{h}^{i_2i_2} + \frac{1}{2}\frac{\hat{\phi}_{i_1i_2}}{\hat{\phi}}\hat{h}^{i_1i_2} + O(k\,n^{-2})\right)$$

and then logarithms on each side,

$$\log f(z_0|\boldsymbol{y};\hat{\beta},\hat{\gamma}) = \log\hat{\phi} - \frac{1}{2}\frac{\hat{\phi}_{i_1}}{\hat{\phi}}\hat{h}_{i_2 i_2 i_2}\hat{h}^{i_1 i_2}\hat{h}^{i_2 i_2} + \frac{1}{2}\frac{\hat{\phi}_{i_1 i_2}}{\hat{\phi}}\hat{h}^{i_1 i_2} + O(k\,n^{-2}) \tag{5.9}$$

where

$$\begin{aligned}
\log\hat{\phi} &= -\frac{1}{2}\log(2\pi\,\hat{\tau}) - \frac{1}{2}\left(\frac{z_0-\hat{\mu}}{\hat{\tau}}\right)^2 \\
\frac{\hat{\phi}_{i_1}}{\hat{\phi}} &= \hat{\tau}^{-1}\hat{\mu}_i\left(\frac{z_0-\hat{\mu}}{\hat{\tau}}\right) \\
\frac{\hat{\phi}_{i_1 i_2}}{\hat{\phi}} &= \hat{\tau}^{-2}\hat{\mu}_{i_1}\hat{\mu}_{i_2}\left\{\left(\frac{z_0-\hat{\mu}}{\hat{\tau}}\right)^2 - \delta_{i_1,i_2}\right\}
\end{aligned}$$

$\delta_{i_1,i_2} = 1$ if $i_1 = i_2$ and 0 otherwise.

Notice that the terms in the right hand side of (5.9) consist of a polynomial of second degree with respect to z_0, which suggests that the predictive density constructed by including the higher order terms in the right hand side of (5.9) is Normal. Consequently, we define the *second order corrected plug-in predictive density* by

$$\hat{f}(z_0|\boldsymbol{y}) = \exp\left\{\log\hat{\phi} - \frac{1}{2}\frac{\hat{\phi}_{i_1}}{\hat{\phi}}\hat{h}_{i_2 i_2 i_2}\hat{h}^{i_1 i_2}\hat{h}^{i_2 i_2} + \frac{1}{2}\frac{\hat{\phi}_{i_1 i_2}}{\hat{\phi}}\hat{h}^{i_1 i_2}\right\} \tag{5.10}$$

Notice that the coefficient of z_0^2 is

$$-\frac{1}{2\hat{\tau}^2}\left(1 - \hat{\tau}^{-2}\mu_{i_1}\mu_{i_2}\hat{h}^{i_1 i_2}\right) \tag{5.11}$$

Since $\hat{h}_{\boldsymbol{zz}}$ is evaluated at $\hat{\boldsymbol{z}}$, it is positive definite, hence, so is $\hat{h}_{\boldsymbol{zz}}^{-1}$, therefore $\mu_{i_1}\mu_{i_2}\hat{h}^{i_1 i_2} > 0$ so the higher order correction always has bigger variance than the first order Laplace approximation. On the other hand, it might happen that $\hat{\tau}^{-3}\mu_{i_1}\mu_{i_2}\hat{h}^{i_1 i_2} > 1$, hence (5.10) cannot be defined because (5.11) becomes positive, in which case a modification as we explain below can be used. Note though that since $\mu_{i_1}\mu_{i_2}\hat{h}^{i_1 i_2} = O(k\,n^{-1})$, then the coefficient of z_0^2 should be negative if the sample size is sufficiently large.

The mean of (5.10) is

$$\hat{\mu}_c = \hat{\mu} - \frac{1}{2}\hat{\tau}^{-1}(\hat{\mu}_{i_1}\hat{h}_{i_2 i_2 i_2}\hat{h}^{i_1 i_2}\hat{h}^{i_2 i_2})\,(1 - \hat{\tau}^{-2}\hat{\mu}_{i_1}\hat{\mu}_{i_2}\hat{h}^{i_1 i_2})^{-1} \tag{5.12}$$

and its variance is

$$\hat{\sigma}_c^2 = \hat{\tau}^2(1 - \hat{\tau}^{-2}\hat{\mu}_{i_1}\hat{\mu}_{i_2}\hat{h}^{i_1 i_2})^{-1} \tag{5.13}$$

therefore, the α-quantile of the distribution of $Z_0|\{\boldsymbol{Y} = \boldsymbol{y}\}$ is estimated by

$$\hat{z}_\alpha = \hat{\mu}_c + \hat{\sigma}_c\Phi^{-1}(\alpha) \tag{5.14}$$

where $\Phi^{-1}(\alpha)$ is the α-quantile of the standard Normal distribution.

By (5.8), $\hat{f}(z_0|\boldsymbol{y}) - f(z_0|\boldsymbol{y};\hat{\beta},\hat{\gamma}) = O(k\,n^{-2})$. On the other hand, since $\hat{\gamma} - \gamma = O(k^{-1/2})$ and $\hat{\beta} - \beta = O(k^{-1/2}n^{-1/2})$, then $\hat{f}(z_0|\boldsymbol{y}) - f(z_0|\boldsymbol{y};\beta,\gamma) = O(k^{-1/2})$, which implies $\hat{z}_\alpha - z_\alpha = O(k^{-1/2})$.

Making the approximation a proper density

As we mentioned above, it might happen that $\hat{\tau}^{-2}\hat{\mu}_{i_1}\hat{\mu}_{i_2}\hat{h}^{i_1 i_2} > 1$, in which case (5.10) is not an actual density since the quantity in (5.13) is negative. If $\hat{\tau}^{-2}\hat{\mu}_{i_1}\hat{\mu}_{i_2}\hat{h}^{i_1 i_2} < m$ for some $m > 0$, then the term $(1 - \hat{\tau}^{-2}\hat{\mu}_{i_1}\hat{\mu}_{i_2}\hat{h}^{i_1 i_2})^{-1}$ in (5.12) and (5.13) can be replaced by $(1 - m^{-1}\hat{\tau}^{-2}\hat{\mu}_{i_1}\hat{\mu}_{i_2}\hat{h}^{i_1 i_2})^{-m}$ without changing the order of the approximation.

Alternatively, since $\hat{\tau}^{-2}\hat{\mu}_{i_1}\hat{\mu}_{i_2}\hat{h}^{i_1 i_2} = O(k\,n^{-1})$, we can write $(1 - \hat{\tau}^{-2}\hat{\mu}_{i_1}\hat{\mu}_{i_2}\hat{h}^{i_1 i_2})^{-1} = 1 + \hat{\tau}^{-2}\hat{\mu}_{i_1}\hat{\mu}_{i_2}\hat{h}^{i_1 i_2} + O(k^2 n^{-2})$ which is always positive.

5.1.1 Plug-in corrections

Here we propose an adjustment for the bias of the predictive quantiles. Let $\hat{z}_\alpha$ be as in (5.14). The coverage probability of $\hat{z}_\alpha$ when the true parameters are $(\beta_\intercal, \gamma_\intercal)$ is

$$\alpha'(\boldsymbol{y};\beta_\intercal,\gamma_\intercal) = F(\hat{z}_\alpha|\boldsymbol{y};\beta_\intercal,\gamma_\intercal) = \frac{\int \Phi(\hat{z}_\alpha|\boldsymbol{z};\gamma_\intercal)\exp\{-h(\boldsymbol{y},\boldsymbol{z};\beta_\intercal,\gamma_\intercal)\}\,\mathrm{d}\boldsymbol{z}}{\int \exp\{-h(\boldsymbol{y},\boldsymbol{z};\beta_\intercal,\gamma_\intercal)\}\,\mathrm{d}\boldsymbol{z}} \tag{5.15}$$

By Laplace approximation around $\tilde{\boldsymbol{z}} = \text{argmin}_{\boldsymbol{z}}\, h(\boldsymbol{y}, \boldsymbol{z}; \beta_{\mathsf{T}}, \gamma_{\mathsf{T}})$ to the right hand side of (5.15), we have

$$\alpha'(\boldsymbol{y}; \beta_{\mathsf{T}}, \gamma_{\mathsf{T}}) = \tilde{\Phi} - \frac{1}{2}\tilde{\Phi}_{i_1}\tilde{h}_{i_2 i_2 i_2}\tilde{h}^{i_1 i_2}\tilde{h}^{i_2 i_2} + \frac{1}{2}\tilde{\Phi}_{i_1 i_2}\tilde{h}^{i_1 i_2} + O(k\, n^{-2}) \tag{5.16}$$

where $\tilde{\Phi} = \Phi(\hat{z}_\alpha|\tilde{\boldsymbol{z}}; \gamma_{\mathsf{T}})$ and similarly for the other quantities of (5.16). Define

$$\hat{\alpha}' = \alpha'(\boldsymbol{y}; \hat{\beta}, \hat{\gamma}) \tag{5.17}$$

Then the difference

$$\hat{z}_{\hat{\alpha}'} - \hat{z}_\alpha$$

can be used to estimate the bias of $\hat{z}_{\hat{\alpha}'}$. As we expect the quantiles $\hat{z}_{\hat{\alpha}'}$, $\hat{z}_\alpha$, and z_α to be relatively close to each other, the bias of $\hat{z}_{\hat{\alpha}'}$ can be used to adjust for the bias of $\hat{z}_\alpha$. Therefore, a bias adjustment to $\hat{z}_\alpha$ would be

$$\hat{z}'_\alpha = \hat{z}_\alpha - (\hat{z}_{\hat{\alpha}'} - \hat{z}_\alpha)$$

5.2 Simulations

Using the same simulated data as in Chapter 4, we compare our method for prediction with the Markov Chain EM Gradient (MCEMG) method from Zhang (2002), and the MCMC from Christensen and Ribeiro (2002). We also consider the no-correction plug-in, i.e. using only the first term of (5.9) as a predictive density. We choose 41 locations from the previously defined grid for prediction as shown in Figure 5.1. We only consider the case where the nugget is unknown. For the plug-in methods we use the parameter estimates obtained from our simulations while for the other two methods we choose the estimates obtained from the MCMC method. For the MCEMG and MCMC algorithms the total run length of a Markov chain was 37700 with burn in 200 and thinning 25, chosen such that both methods give sufficient random samples. The calculations for the plug-in and MCEMG methods were programmed in FORTRAN and for the MCMC method we used the `R` package `geoRglm` (Christensen and Ribeiro, 2002) which calls a routine written in C. A large part of the calculations required by the MCEMG and MCMC is due to random sampling from the conditional distribution of the random effects, the

first implementing the Metropolis-Hastings algorithm, while the second the Langevin-Hastings modification (see Christensen and Waagepetersen, 2002). A significant difference between the two is that for MCMC, the random effects at each location are updated simultaneously instead of updating a single component each time. Note that the length of the Markov chain is the same for these two methods, thus even though the computation time for MCMC could be different if it was implemented in FORTRAN, it is our belief that it won't make a notable difference when comparing its speed with the plug-in method, and we choose not to address this issue.

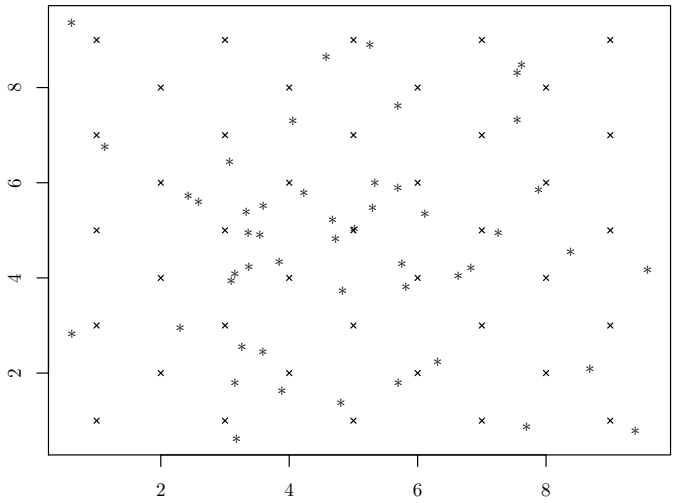

Figure 5.1: Locations for the simulations. The sampled locations are marked by $*$, and the locations for prediction by $\times$.

Using each method, we predict the random effects at the new locations z_{0i} and estimate the success probability at each location by $\hat{p}_i = e^{x_i\hat{\beta}+z_{0i}}/(1+e^{x_i\hat{\beta}+z_{0i}})$. We also obtain the conditional mean value for the random effects at the new locations assuming the random effects at the observed locations were observed. Then, the "true" probabilities at the new locations are calculated similarly by $p_i = e^{x_i\beta+z_{0i}}/(1+e^{x_i\beta+z_{0i}})$

Three scoring rules that appear in Gneiting and Raftery (2007) allow us to compare the different methods. They are defined in the following way: Let $U \sim \text{Bin}(n,p)$ and $p_j = \Pr(U = j) = \binom{n}{j}p^j(1-p)^{n-j}$ be the probability assigned to the event j, $j = 0,\ldots,n$ and $\hat{p}_j$ the corresponding estimate. If the event $\{U = i\}$ is observed then define

- The Brier score: $-\sum_{j=0}^{n}(\delta_{ij}-\hat{p}_j)^2$ where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise,

- the spherical score: $\hat{p}_i^{\alpha-1}\Big(\sum_{j=0}^{n}\hat{p}_j^{\alpha}\Big)^{-(\alpha-1)/\alpha}$ where $\alpha > 1$, and
- the logarithmic score: $\log \hat{p}_i$

The higher the score, the better the prediction.

Assuming that we make one draw at each location, there is probability p_i to get 1 and $1 - p_i$ to get 0. Then we can calculate the expected scoring rules at each location and for each simulation. The average over all locations and all simulations is compared for each method. In addition, for each simulation we take the Mahalanobis distance with covariance matrix equal to the conditional variance of the random effects. The average Mahalanobis distance over all simulations gives another measure for comparison. The results are in Table 5.1

Method	Brier	Spherical ($\alpha = 2$)	Logarithmic	Mah/bis dist.	time (sec)
uncorr. plug-in	−0.38125	0.78345	−0.56193	5.2934	76
2nd order plug-in	−0.38094	0.78363	−0.56156	5.0944	84
MCEMG	−0.38097	0.78361	−0.56160	5.1783	2795
MCMC	−0.40794	0.77028	−0.61392	3.1690	2768

Table 5.1: Measures for comparing different methods for prediction.

Regarding the three scoring rules, the corrected plug-in method seems to be doing slightly better than the other three methods and MCMC is the worse. On the other hand, MCMC ranks first when we compare the average Mahalanobis distance and the corrected plug-in is the second best. Also note that the uncorrected plug-in is always doing worse than the corrected plug-in. Clearly the two plug-in methods are much faster.

CHAPTER 6

An Application: The Rhizoctonia Disease

The rhizoctonia root rot is a disease that attaches on the roots of plants and hinders the process of absorbing water and nutrients. In this example examined by Zhang (2002), 15 plants were pulled from each of 100 randomly chosen locations in a farm and the number of crown roots and infected crown roots were counted. Similar to Zhang (2002), we assume constant mean and spherical covariance structure given by

$$\mathcal{C}(\gamma) = \begin{cases} \gamma_1 + \gamma_2 & \text{if } d_{ij} = 0 \\ \gamma_2 \left(1 - 1.5\frac{d_{ij}}{\gamma_3} + 0.5\left(\frac{d_{ij}}{\gamma_3}\right)^3\right) & \text{if } 0 < d_{ij} < \gamma_3 \\ 0 & \text{o.w.} \end{cases} \tag{6.1}$$

for the underlying Gaussian random field and treat the data as samples from binomial distribution. Because $\gamma_j > 0$, in our optimization we estimated the $\log \gamma_j$ and then exponentiated. Applying our method we obtain the estimates $\hat{\beta} = -1.677$ with standard error 0.0925 obtained by inverting the hessian matrix evaluated at the maximum, and $\hat{\gamma} = (0.405, 0.098, 145.69)$ with standard errors $(0.0804, 0.0808, 38.42)$ obtained using our approximate formula (4.14). Our estimates are close to Zhang's and fall within a 95% confidence interval constructed by our estimates but we find smaller standard error for the covariance parameters. The analysis using MCMC gives $\hat{\beta} = -1.721$ and $\hat{\gamma} = (0.4776, 0.108, 149.1)$ which also fall within our 95% confidence interval. The estimates using the first order Laplace approximation are $\hat{\beta} = -1.760$ and $\hat{\gamma} = (0.7074, 0.1241, 151.4)$.

It's important to know the severity of the disease at locations that we don't have an observation. Using (5.10) as an approximation to the predictive density we take the mode to be our

prediction. The resulting map, shown in Figure 6.1, has a very close pattern with the one in Zhang (2002) though the range of our predictions is narrower.

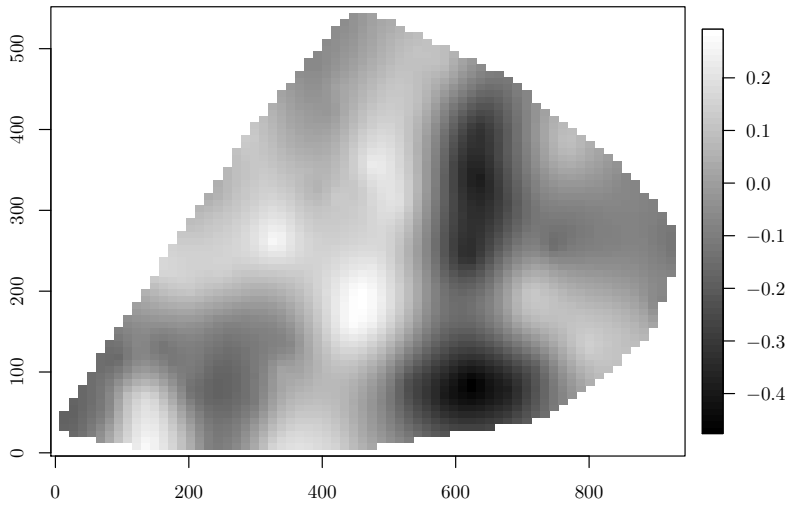

Figure 6.1: Map of the predicted random effects (disease intensity) using the second order corrected plug-in predictive density.

We compare the four methods: corrected and uncorrected plug-in (this paper), MCEMG (Zhang, 2002), and MCMC (Christensen and Ribeiro, 2002) by cross-validation. We predict the value of the random effect z_i at each of the $i = 1, \ldots, 100$ locations using each method and then calculate the binomial probability of what we actually observed given n_i and $\hat{p}_i = e^{\hat{\beta}+\hat{z}_i}/(1 + e^{\hat{\beta}+\hat{z}_i})$. The parameter values were taken from our estimates above for the plug-in and MCMC methods and from Zhang (2002) for the MCEMG method.

We use the same three scoring rules as in Chapter 5 for comparison. Figure 6.2 shows that the calculated scores at each location for the plug-in and MCEMG are very similar. The average scores over all the locations for each method are summarized in Table 6.1. MCEMG does slightly better than plug-in but MCMC is worse. The corrected plug-in is more accurate than the uncorrected. Again note the time difference between the methods.

Method	Brier	Spherical ($\alpha = 2$)	Logarithmic	Time (sec)
uncorrected plug-in	-1.0096	0.11198	-7.4450	20
corrected plug-in	-1.0086	0.11257	-7.2795	20
MCEMG	-1.0078	0.11400	-7.2254	651
MCMC	-1.0164	0.10133	-7.9378	703

Table 6.1: Measures of scoring for comparison of plug-in and MCEMG.

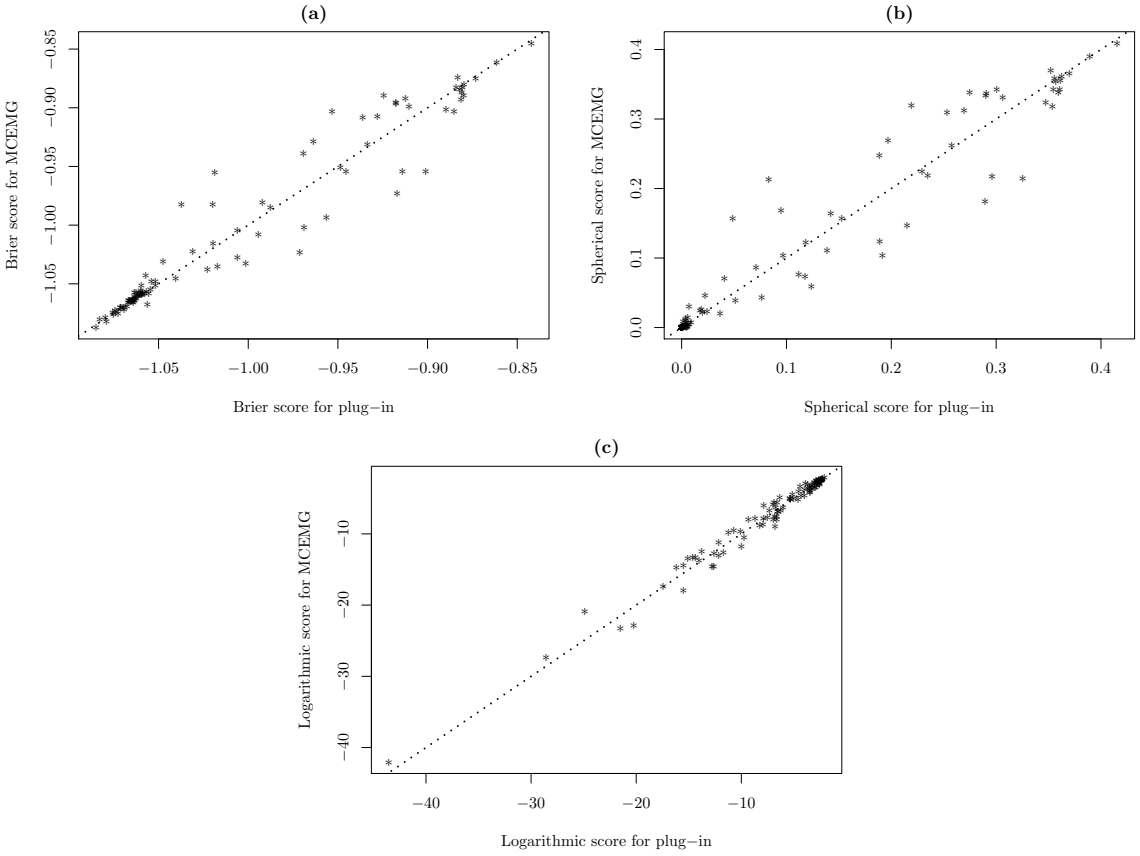

Figure 6.2: Plot showing the calculated scores for plug-in and MCEMG for cross-validation. The dotted line corresponds to the equation $y = x$. (a) Brier score, (b) Spherical score, (c) Logarithmic score.

CHAPTER 7

Bayesian Prediction

Here, we take the Bayesian approach in predicting the random effects in GLMM. We investigate the effect of the prior for γ on the prediction for Z_0.

7.1 Bayesian Predictive Distribution

We start by assigning priors for β and γ:

$$(\beta, \gamma) \sim \exp\{u(\beta) + r(\gamma)\} \tag{7.1}$$

and define the *Bayesian Predictive Distribution Function* by

$$\tilde{F}(z_0|\boldsymbol{y}) = \frac{\iiint \Phi(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{y},\boldsymbol{z};\beta,\gamma) \exp\{u(\beta)+r(\gamma)\}\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\beta\,\mathrm{d}\gamma}{\iiint f(\boldsymbol{y},\boldsymbol{z};\beta,\gamma) \exp\{u(\beta)+r(\gamma)\}\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\beta\,\mathrm{d}\gamma} \tag{7.2}$$

Let

$$\psi(\boldsymbol{y},\boldsymbol{z};\beta,\gamma) = -\log f(\boldsymbol{y},\boldsymbol{z};\beta,\gamma) - u(\beta) \tag{7.3}$$

Then (7.2) becomes

$$\tilde{F}(z_0|\boldsymbol{y}) = \frac{\iiint \Phi(z_0|\boldsymbol{z};\gamma) \exp\{-\psi(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)\} \exp\{r(\gamma)\}\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\beta\,\mathrm{d}\gamma}{\iiint \exp\{-\psi(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)\} \exp\{r(\gamma)\}\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\beta\,\mathrm{d}\gamma} \tag{7.4}$$

7.1.1 Expansion of the Bayesian Predictive Distribution

Write $\xi = (\boldsymbol{z}, \beta, \gamma)$ and define $\hat{\xi} = \operatorname{argmin}_{(\boldsymbol{z},\beta,\gamma)} \psi(\boldsymbol{y}, \boldsymbol{z}; \beta, \gamma)$. Then, (7.4) is written as

$$\tilde{F}(z_0|\boldsymbol{y}) = \frac{\int \Phi(\xi) \exp\{-\psi(\xi) + r(\xi)\} \, \mathrm{d}\xi}{\int \exp\{-\psi(\xi) + r(\xi)\} \, \mathrm{d}\xi} = \frac{I^{\mathsf{N}}}{I^{\mathsf{D}}} \tag{7.5}$$

Using (3.17) on each of I^{N} and I^{D},

$$I^{\mathsf{N}} = e^{-\hat{\psi}+\hat{r}} \left| \frac{\hat{\psi}_{\xi\xi}}{2\pi} \right|^{-1/2} \sum_{m=1}^{\infty} \sum_{s=0}^{2m} \sum_{\lambda=0}^{2m-s} \sum_{P,Q} \frac{(-1)^t}{(2m)!} \hat{\Phi}_{l_1 \ldots l_s} \cdot \hat{r}_{[j_1 \ldots j_\lambda]} \cdot \hat{\psi}_{p_1} \ldots \hat{\psi}_{p_t} \cdot \hat{\psi}^{q_1} \ldots \hat{\psi}^{q_m} \tag{7.6}$$

$$I^{\mathsf{D}} = e^{-\hat{\psi}+\hat{r}} \left| \frac{\hat{\psi}_{\xi\xi}}{2\pi} \right|^{-1/2} \sum_{m=1}^{\infty} \sum_{\lambda=0}^{2m} \sum_{P,Q} \frac{(-1)^t}{(2m)!} \hat{r}_{[j_1 \ldots j_\lambda]} \cdot \hat{\psi}_{p_1} \ldots \hat{\psi}_{p_t} \cdot \hat{\psi}^{q_1} \ldots \hat{\psi}^{q_m} \tag{7.7}$$

Hence, dividing (7.6) by (7.7), after some cancellations, we obtain

$$\tilde{F}(z_0|\boldsymbol{y}) = \sum_{m=1}^{\infty} \sum_{s=1}^{2m} \sum_{\lambda=0}^{2m-s} \sum_{P,Q} \frac{(-1)^t}{(2m)!} \hat{\Phi}_{l_1 \ldots l_s} \cdot \hat{r}_{[j_1 \ldots j_\lambda]} \cdot \hat{\psi}_{p_1} \ldots \hat{\psi}_{p_t} \cdot \hat{\psi}^{q_1} \ldots \hat{\psi}^{q_m} \tag{7.8}$$

(Notice the difference from (7.6) that s starts at 1.)

Note that (7.8) implies differentiation with respect to different variables with derivatives having different orders and this complicates the calculations. In general the indices in the two partitions range from 1 to $k + d$, d being the dimension of $(\beta^{\mathsf{T}}, \gamma^{\mathsf{T}})^{\mathsf{T}}$. In order to be able to evaluate (7.8) we need to know the order of the elements of $\hat{\psi}_{\xi\xi}$ when the differentiation is performed on the various components of ξ.

Order of the inverse Hessian of ψ

In Table 7.1 we show the asymptotic order of the derivatives of ψ with respect to the different components of ξ evaluated at $\hat{\xi}$.

Write

$$\psi_{\xi\xi} = \begin{pmatrix} \psi_{\boldsymbol{z}\boldsymbol{z}} & \psi_{\boldsymbol{z}\beta} & \psi_{\boldsymbol{z}\gamma} \\ \psi_{\beta\boldsymbol{z}} & \psi_{\beta\beta} & \psi_{\beta\gamma} \\ \psi_{\gamma\boldsymbol{z}} & \psi_{\gamma\beta} & \psi_{\gamma\gamma} \end{pmatrix}$$

For the inverse of $\psi_{\xi\xi}$, we make use of the following two lemmata.

$\cdot$ \ $\partial z_{i_2}^{m_2}$	$k \cdot n$	$1,\ m_1 = m_2 = 1;\ 0,$ o.w.	$\partial \beta_{j_2}^{l_2}$	
$\partial z_{i_1}^{m_1}$	$0,\ m_1 = 1;\ n,$ o.w.	n	$k \cdot n$	$\partial \gamma_{j_2}^{l_2}$
$\partial \beta_{j_1}^{l_1}$	$0,\ l_1 = 1;\ k \cdot n,$ o.w.	1	0	k
$\partial \gamma_{j_1}^{l_1}$	$0,\ l_1 = 1;\ k,$ o.w.			

Table 7.1: Table showing the order for different derivatives of ψ. The derivative is taken with respect to the components shown outside of the borders of the table and the $\cdot$ means that no derivative is taken.

Lemma 2. *Let A be an invertible 3×3 block matrix with blocks A_{ij}, $i, j = 1, 2, 3$ such that A_{11}, A_{22} and A_{33} are square matrices. Write A^{-1} as a 3×3 block matrix where its blocks A^{ij} have the same dimensions as the corresponding A_{ij} in A. Then,*

$$\begin{aligned} A^{11} = \big(&A_{11} - A_{12}(A_{22} - A_{23}A_{33}^{-1}A_{32})^{-1}(A_{21} - A_{23}A_{33}^{-1}A_{31}) \\ &- A_{13}(A_{33} - A_{32}A_{22}^{-1}A_{23})^{-1}(A_{31} - A_{32}A_{22}^{-1}A_{21})\big)^{-1} \end{aligned} \tag{7.9}$$

$$A^{12} = -A^{11}\left(A_{12}(A_{22} - A_{23}A_{33}^{-1}A_{32})^{-1} - A_{13}(A_{33} - A_{32}A_{22}^{-1}A_{23})^{-1}A_{32}A_{22}^{-1}\right) \tag{7.10}$$

The other block elements of A^{-1} are given similarly.

Proof. By multiplication. □

Lemma 3. *Let A, B be symmetric and invertible. Then*

$$(A - UB^{-1}U^{\intercal})^{-1} = A^{-1} + A^{-1}U(B - U^{\intercal}A^{-1}U)^{-1}U^{\intercal}A^{-1} \tag{7.11}$$

Proof. By multiplication (see Henderson et al., 1959, page 196). □

Using Lemma 2,

$$\begin{aligned} \psi^{\boldsymbol{zz}} &= \left(\psi_{\boldsymbol{zz}} - \psi_{\boldsymbol{z}\beta}\psi_{\beta\beta}^{-1}\psi_{\beta\boldsymbol{z}} - \psi_{\boldsymbol{z}\gamma}\psi_{\gamma\gamma}^{-1}\psi_{\gamma\boldsymbol{z}}\right)^{-1} \\ &= \left(\psi_{\boldsymbol{zz}} - \psi_{\boldsymbol{z}\beta}\psi_{\beta\beta}^{-1}\psi_{\beta\boldsymbol{z}}\right)^{-1}\left(I - \psi_{\boldsymbol{z}\gamma}\psi_{\gamma\gamma}^{-1}\psi_{\gamma\boldsymbol{z}}\left(\psi_{\boldsymbol{zz}} - \psi_{\boldsymbol{z}\beta}\psi_{\beta\beta}^{-1}\psi_{\beta\boldsymbol{z}}\right)^{-1}\right)^{-1} \end{aligned}$$

Applying (7.11) on the first parenthesis we get

$$\left(\psi_{\boldsymbol{zz}} - \psi_{\boldsymbol{z}\beta}\psi_{\beta\beta}^{-1}\psi_{\beta\boldsymbol{z}}\right)^{-1} = \psi_{\boldsymbol{zz}}^{-1} + \psi_{\boldsymbol{zz}}^{-1}\psi_{\boldsymbol{z}\beta}\left(\psi_{\beta\beta} - \psi_{\beta\boldsymbol{z}}\psi_{\boldsymbol{zz}}^{-1}\psi_{\boldsymbol{z}\beta}\right)^{-1}\psi_{\beta\boldsymbol{z}}\psi_{\boldsymbol{zz}}^{-1} \tag{7.12}$$

For the order of (7.12) remember that summation over the values of the vector $\boldsymbol{z}$ contributes an extra factor of order $O(k)$. Thus, by Lemma 1, the elements of $\psi_{\boldsymbol{zz}}^{-1}\psi_{\boldsymbol{z}\beta}$ have order $O(1)$ Therefore, the term in the parenthesis of the right hand side: $\psi_{\beta\beta} - \psi_{\beta\boldsymbol{z}}\psi_{\boldsymbol{zz}}^{-1}\psi_{\boldsymbol{z}\beta}$ has elements of order $O(k\,n)$. As a result, the diagonal elements of $\left(\psi_{\boldsymbol{zz}} - \psi_{\boldsymbol{z}\beta}\psi_{\beta\beta}^{-1}\psi_{\beta\boldsymbol{z}}\right)^{-1}$ are $O(n^{-1})$ and the off-diagonal are $O(k^{-1}n^{-1})$. In addition, the elements of $\psi_{\boldsymbol{z}\gamma}\psi_{\gamma\gamma}^{-1}\psi_{\gamma\boldsymbol{z}}\left(\psi_{\boldsymbol{zz}} - \psi_{\boldsymbol{z}\beta}\psi_{\beta\beta}^{-1}\psi_{\beta\boldsymbol{z}}\right)^{-1}$ are $O(k^{-1}n^{-1})$. Consequently, the diagonal elements of $\psi^{\boldsymbol{zz}}$ are $O(n^{-1})$ and the off-diagonal are $O(k^{-1}n^{-1})$.

For $\psi^{\boldsymbol{z}\beta}$ and $\psi^{\boldsymbol{z}\gamma}$ we have

$$\begin{aligned}\psi^{\boldsymbol{z}\beta} &= -\psi^{\boldsymbol{zz}}\psi_{\boldsymbol{z}\beta}\psi_{\beta\beta}^{-1}\\ \psi^{\boldsymbol{z}\gamma} &= -\psi^{\boldsymbol{zz}}\psi_{\boldsymbol{z}\gamma}\psi_{\gamma\gamma}^{-1}\end{aligned}$$

where we note that when we multiply $\psi^{\boldsymbol{zz}}\psi_{\boldsymbol{z}\beta}$ and $\psi^{\boldsymbol{zz}}\psi_{\boldsymbol{z}\gamma}$ we get elements of order $O(1)$ and $O(n^{-1})$ respectively. Thus the elements of $\psi^{\boldsymbol{z}\beta}$ and $\psi^{\boldsymbol{z}\gamma}$ are $O(k^{-1}n^{-1})$.

For the other blocks,

$$\begin{aligned}\psi^{\beta\beta} &= \left(\psi_{\beta\beta} - \psi_{\beta\boldsymbol{z}}(\psi_{\boldsymbol{zz}} - \psi_{\boldsymbol{z}\gamma}\psi_{\gamma\gamma}^{-1}\psi_{\gamma\boldsymbol{z}})^{-1}\psi_{\boldsymbol{z}\beta}\right)^{-1}\\ \psi^{\gamma\gamma} &= \left(\psi_{\gamma\gamma} - \psi_{\gamma\boldsymbol{z}}(\psi_{\boldsymbol{zz}} - \psi_{\boldsymbol{z}\beta}\psi_{\beta\beta}^{-1}\psi_{\beta\boldsymbol{z}})^{-1}\psi_{\boldsymbol{z}\gamma}\right)^{-1}\\ \psi^{\beta\gamma} &= \psi^{\beta\beta}\psi_{\beta\boldsymbol{z}}(\psi_{\boldsymbol{zz}} - \psi_{\boldsymbol{z}\gamma}\psi_{\gamma\gamma}^{-1}\psi_{\gamma\boldsymbol{z}})^{-1}\psi_{\boldsymbol{z}\gamma}\psi_{\gamma\gamma}^{-1}\end{aligned}$$

We have already derived the order of $(\psi_{\boldsymbol{zz}} - \psi_{\boldsymbol{z}\beta}\psi_{\beta\beta}^{-1}\psi_{\beta\boldsymbol{z}})^{-1}$. Applying the same technique we write

$$\left(\psi_{\boldsymbol{zz}} - \psi_{\boldsymbol{z}\gamma}\psi_{\gamma\gamma}^{-1}\psi_{\gamma\boldsymbol{z}}\right)^{-1} = \psi_{\boldsymbol{zz}}^{-1} + \psi_{\boldsymbol{zz}}^{-1}\psi_{\boldsymbol{z}\gamma}\left(\psi_{\gamma\gamma} - \psi_{\gamma\boldsymbol{z}}\psi_{\boldsymbol{zz}}^{-1}\psi_{\boldsymbol{z}\gamma}\right)^{-1}\psi_{\gamma\boldsymbol{z}}\psi_{\boldsymbol{zz}}^{-1} \tag{7.13}$$

In this case, the order of the elements of $\psi_{\gamma\boldsymbol{z}}\psi_{\boldsymbol{zz}}^{-1}\psi_{\boldsymbol{z}\gamma}$ is $O(k\,n^{-1})$ so the order of the elements of $\left(\psi_{\gamma\gamma} - \psi_{\gamma\boldsymbol{z}}\psi_{\boldsymbol{zz}}^{-1}\psi_{\boldsymbol{z}\gamma}\right)^{-1}$ is $O(k^{-1})$. Thus the second term at the right hand side of (7.13) has elements of order $O(k^{-1}n^{-2})$ which is lower than the order of the first term. Finally we

have that the diagonal elements of $\left(\psi_{\boldsymbol{zz}} - \psi_{\boldsymbol{z}\gamma}\psi_{\gamma\gamma}^{-1}\psi_{\gamma\boldsymbol{z}}\right)^{-1}$ are $O(n^{-1})$ and the off-diagonal are $O(n^{-2})$. After similar calculations we have

$$\psi_{j_1 j_2}^{\beta\beta} = O(k^{-1}n^{-1})$$
$$\psi_{j_1 j_2}^{\gamma\gamma} = O(k^{-1})$$
$$\psi_{j_1 j_2}^{\beta\gamma} = O(k^{-1}n^{-1})$$

7.1.2 Asymptotic approximation to the Bayesian predictive density

Note that the derivatives of $\Phi(\xi) = \Phi(z_0|\boldsymbol{z};\gamma)$ are all $O(1)$ and we also assume that the derivatives of $r(\gamma)$ are $O(1)$.

In (7.8), we can break each index to three different ranges, one for the dimension of $\boldsymbol{z}$, one for the dimension of β and one for γ. Then each block of the partitions P and Q has indices belonging in one of six cases: "only $\boldsymbol{z}$": $(\boldsymbol{zz})$, "only β": $(\beta\beta)$, "only γ": $(\gamma\gamma)$, "$\boldsymbol{z}$ and β": $(\boldsymbol{z}\beta)$, "$\boldsymbol{z}$ and γ": $(\boldsymbol{z}\gamma)$ and "β and γ": $(\beta\gamma)$. The case "$\boldsymbol{z}$ and β and γ" for P is ignored because it has derivatives equal to 0. Table 7.2 shows the order of each component in (7.8) for the different cases. Keep in mind that an index that ranges over the values of $\boldsymbol{z}$ contributes an extra factor of order k.

	Φ_{p_i}	r_{p_i}	ψ_{p_i}	$\psi_{\xi\xi}^{q_j}$
$\boldsymbol{zz}$	1 or 0^a	0	n or 1^b	n^{-1} or $k^{-1}n^{-1b}$
$\beta\beta$	0	1	$k\,n$	$k^{-1}n^{-1}$
$\gamma\gamma$	1	1	k	k^{-1}
$\boldsymbol{z}\beta$	0	0	n	$k^{-1}n^{-1}$
$\boldsymbol{z}\gamma$	1 or 0^a	0	1	$k^{-1}n^{-1}$
$\beta\gamma$	0	1	0	$k^{-1}n^{-1}$

Table 7.2: Order of magnitude of the derivatives for different cases. ^{a}1 for at most second derivative, 0 for higher order derivatives. ^{b}Corresponding to if the differentiation is with respect to the same component of $\boldsymbol{z}$ or not.

Based on the expansion in (7.8) and the orders shown in Table 7.2, an approximation to the Bayesian predictive density is given by

$$\tilde{F}(z_0|\boldsymbol{y}) = \hat{\Phi} + \frac{1}{2}\hat{\Phi}_{i_1 i_2}\hat{\psi}^{i_1 i_2} - \frac{1}{2}\hat{\Phi}_{i_1}\hat{\psi}_{i_2 i_2 i_2}\hat{\psi}^{i_1 i_2}\hat{\psi}^{i_2 i_2}$$

$$+\frac{1}{2}\hat{\Phi}_{j_1j_2}\hat{\psi}^{j_1j_2}+\hat{\Phi}_{j_1}\hat{r}_{j_2}\hat{\psi}^{j_1j_2}-\frac{1}{2}\hat{\Phi}_{j_1}\hat{\psi}_{j_2j_3j_4}\hat{\psi}^{j_1j_2}\hat{\psi}^{j_3j_4}+O(k^{-2}\vee k^2n^{-2}) \quad (7.14)$$

where here and subsequently we will use the index i to refer to the components of $\boldsymbol{z}$, and j to refer to the components of γ. The order of the approximation is $O(k^{-2}\vee k^2n^{-2})$ because for example terms such as $\hat{\Phi}_{j_1j_2}\hat{\psi}_{j_3j_4j_5}\hat{\psi}_{j_6j_7j_8}\hat{\psi}^{j_1j_3}\hat{\psi}^{j_2j_6}\hat{\psi}^{j_4j_5}\hat{\psi}^{j_7j_8}$ have order $O(k^{-2})$, while terms such as $\hat{\Phi}_{i_1i_2}\hat{\psi}_{i_3i_3i_3}\hat{\psi}_{i_4i_4i_4}\hat{\psi}^{i_1i_3}\hat{\psi}^{i_2i_4}\hat{\psi}^{i_3i_3}\hat{\psi}^{i_4i_4}$ have order $O(k^2n^{-2})$.

In (7.14),

$$\begin{aligned}
\Phi &= \Phi(z_0|\boldsymbol{z};\gamma)\\
\Phi_i &= -\frac{\mu_i}{\tau}\phi\left(\frac{z_0-\mu}{\tau}\right)\\
\Phi_{i_1i_2} &= -\frac{\mu_{i_1}\mu_{i_2}}{\tau^2}\frac{z_0-\mu}{\tau}\phi\left(\frac{z-\mu}{\tau}\right)\\
\Phi_j &= -\left(\frac{\mu_j}{\tau}+\frac{\tau_j}{\tau}\frac{z_0-\mu}{\tau}\right)\phi\left(\frac{z_0-\mu}{\tau}\right)\\
\Phi_{j_1j_2} &= -\left(\frac{\mu_{j_1j_2}}{\tau}-\frac{\mu_{j_1}\tau_{j_2}}{\tau^2}-\frac{\mu_{j_2}\tau_{j_1}}{\tau^2}+\frac{\tau_{j_1j_2}}{\tau}\frac{z_0-\mu}{\tau}-\frac{2\tau_{j_1}\tau_{j_2}}{\tau^2}\frac{z_0-\mu}{\tau}\right)\phi\left(\frac{z_0-\mu}{\tau}\right)\\
&\quad -\left(\frac{\mu_{j_1}}{\tau}+\frac{\tau_{j_1}}{\tau}\frac{z_0-\mu}{\tau}\right)\left(\frac{\mu_{j_2}}{\tau}+\frac{\tau_{j_2}}{\tau}\frac{z_0-\mu}{\tau}\right)\frac{z_0-\mu}{\tau}\phi\left(\frac{z_0-\mu}{\tau}\right)
\end{aligned} \quad (7.15)$$

with

$$\begin{aligned}
\mu_{\boldsymbol{z}} &= \mathbf{c}^\intercal\Sigma^{-1}\\
\mu_j &= \mathbf{c}_j^\intercal\Sigma^{-1}\boldsymbol{Z}-\mathbf{c}^\intercal\Sigma^{-1}\Sigma_j\Sigma^{-1}\boldsymbol{Z}\\
\mu_{j_1j_2} &= \mathbf{c}_{j_1j_2}^\intercal\Sigma^{-1}\boldsymbol{Z}-\mathbf{c}_{j_1}^\intercal\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\boldsymbol{Z}-\mathbf{c}_{j_2}^\intercal\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\boldsymbol{Z}-\mathbf{c}^\intercal\Sigma^{-1}\Sigma_{j_1j_2}\Sigma^{-1}\boldsymbol{Z}\\
&\quad +\mathbf{c}^\intercal\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\boldsymbol{Z}+\mathbf{c}^\intercal\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\boldsymbol{Z}\\
\mu_{j_1j_2j_3} &= \mathbf{c}_{j_1j_2j_3}^\intercal\Sigma^{-1}\boldsymbol{Z}-[3]\mathbf{c}_{j_1j_2}^\intercal\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\boldsymbol{Z}-[3]\mathbf{c}_{j_1}^\intercal\Sigma^{-1}\Sigma_{j_2j_3}\Sigma^{-1}\boldsymbol{Z}\\
&\quad +[6]\mathbf{c}_{j_1}^\intercal\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\boldsymbol{Z}-\mathbf{c}^\intercal\Sigma^{-1}\Sigma_{j_1j_2j_3}\Sigma^{-1}\boldsymbol{Z}+[3]\mathbf{c}^\intercal\Sigma^{-1}\Sigma_{j_1j_2}\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\boldsymbol{Z}\\
&\quad +[3]\mathbf{c}^\intercal\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\Sigma_{j_2j_3}\Sigma^{-1}\boldsymbol{Z}-[6]\mathbf{c}^\intercal\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\boldsymbol{Z}\\
\tau_j &= \left(\sigma_{0j}^2-2\mathbf{c}_j^\intercal\Sigma^{-1}\mathbf{c}+\mathbf{c}^\intercal\Sigma^{-1}\Sigma_j\Sigma^{-1}\mathbf{c}\right)/(2\tau)\\
\tau_{j_1j_2} &= \left(\sigma_{0j_1j_2}^2-2\mathbf{c}_{j_1j_2}^\intercal\Sigma^{-1}\mathbf{c}+2\mathbf{c}_{j_1}^\intercal\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\mathbf{c}+2\mathbf{c}_{j_2}^\intercal\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\mathbf{c}\right.\\
&\quad \left.-2\mathbf{c}_{j_1}^\intercal\Sigma^{-1}\mathbf{c}_{j_2}-2\mathbf{c}^\intercal\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\mathbf{c}+\mathbf{c}^\intercal\Sigma^{-1}\Sigma_{j_1j_2}\Sigma^{-1}\mathbf{c}\right)/(2\tau)
\end{aligned}$$

$$- \left(\sigma^2_{0j_1} - 2\mathbf{c}^\top_{j_1}\Sigma^{-1}\mathbf{c} + \mathbf{c}^\top\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\mathbf{c}\right)\left(\sigma^2_{0j_2} - 2\mathbf{c}^\top_{j_2}\Sigma^{-1}\mathbf{c} + \mathbf{c}^\top\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\mathbf{c}\right)/(4\tau^3)$$

$$\begin{aligned}\tau_{j_1j_2j_3} = &\left(\sigma^2_{0j_1j_2j_3} - 2\mathbf{c}^\top_{j_1j_2j_3}\Sigma^{-1}\mathbf{c} - [3]2\mathbf{c}^\top_{j_1j_2}\Sigma^{-1}\mathbf{c}_{j_3} + [3]2\mathbf{c}^\top_{j_1j_2}\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\mathbf{c}\right.\\
&+ [3]2\mathbf{c}^\top_{j_1}\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\mathbf{c}_{j_3} + [3]2\mathbf{c}^\top_{j_1}\Sigma^{-1}\Sigma_{j_2j_3}\Sigma^{-1}\mathbf{c} - [3]2\mathbf{c}^\top_{j_1}\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\mathbf{c}\\
&+ \mathbf{c}^\top\Sigma^{-1}\Sigma_{j_1j_2j_3}\Sigma^{-1}\mathbf{c} - [3]2\Sigma^{-1}\Sigma_{j_1j_2}\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\mathbf{c}\\
&\left. + [3]2\mathbf{c}^\top\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\mathbf{c}\right)/(2\tau) - [3]\tau_{j_1j_2}\tau_{j_3}/(2\tau)\end{aligned}$$

Bayesian prediction intervals

Write (7.14) in the form

$$\tilde{F}(z_0|\boldsymbol{y}) = \Phi(z_0|\hat{\boldsymbol{z}};\hat{\gamma}) + Q(z_0|\boldsymbol{y}) + O(k^{-2}\vee k^2n^{-2}) \tag{7.16}$$

with

$$\begin{aligned}Q(z_0|\boldsymbol{y}) = &\frac{1}{2}\hat{\Phi}_{i_1i_2}\hat{\psi}^{i_1i_2} - \frac{1}{2}\hat{\Phi}_{i_1}\hat{\psi}_{i_2i_2i_2}\hat{\psi}^{i_1i_2}\hat{\psi}^{i_2i_2}\\
&+ \frac{1}{2}\hat{\Phi}_{j_1j_2}\hat{\psi}^{j_1j_2} + \hat{\Phi}_{j_1}\hat{r}_{j_2}\hat{\psi}^{j_1j_2} - \frac{1}{2}\hat{\Phi}_{j_1}\hat{\psi}_{j_2j_3j_4}\hat{\psi}^{j_1j_2}\hat{\psi}^{j_3j_4}\end{aligned} \tag{7.17}$$

which has order $O(k^{-1}\vee k\,n^{-1})$. Next, we suggest two methods for constructing prediction intervals.

Define $\hat{\zeta}_\alpha$ to be the α-quantile of $\Phi(z_0|\hat{\boldsymbol{z}};\hat{\gamma})$. That is

$$\Phi(\hat{\zeta}_\alpha|\hat{\boldsymbol{z}};\hat{\gamma}) = \alpha \tag{7.18}$$

Also let $\tilde{z}_\alpha$ to be the α-quantile of $\tilde{F}(z_0|\boldsymbol{y})$, i.e.

$$\tilde{F}(\tilde{z}_\alpha|\boldsymbol{y}) = \alpha.$$

$\tilde{z}_\alpha$ can only be calculated via MCMC but $\hat{\zeta}_\alpha$ can be obtained analytically.

By (7.14), $\tilde{F}(z_0|\boldsymbol{y}) = \hat{\Phi}(z_0|\boldsymbol{y}) + O(k^{-1}\vee k\,n^{-1})$, so by Taylor expansion

$$0 = \tilde{F}(\tilde{z}_\alpha|\boldsymbol{y}) - \hat{\Phi}(\hat{\zeta}_\alpha|\boldsymbol{y})$$

$$\approx \hat{\Phi}(\tilde{z}_\alpha|\boldsymbol{y}) - \hat{\Phi}(\hat{\zeta}_\alpha|\boldsymbol{y})$$
$$\approx (\tilde{z}_\alpha - \hat{\zeta}_\alpha)\hat{\Phi}'(\hat{\zeta}_\alpha|\boldsymbol{y})$$

hence, $\tilde{z}_\alpha - \hat{\zeta}_\alpha = O(k^{-1} \vee k\, n^{-1})$ and $F(\hat{\zeta}_\alpha|\boldsymbol{y}) = \alpha + O(k^{-1} \vee k\, n^{-1})$

In view of (7.16), our first suggestion for the approximation of $\tilde{z}_\alpha$ is to use $\hat{\zeta}_{\alpha_1}$ where

$$\alpha_1 = \alpha - Q(\hat{\zeta}_\alpha|\boldsymbol{y}).$$

Then, by Taylor expansion, $\hat{\zeta}_{\alpha_1} - \zeta_\alpha = O(k^{-1} \vee k\, n^{-1})$ and

$$\begin{aligned}
\tilde{F}(\hat{\zeta}_{\alpha_1}|\boldsymbol{y}) &= \Phi(\hat{\zeta}_{\alpha_1}|\hat{\boldsymbol{z}};\hat{\gamma}) + Q(\hat{\zeta}_{\alpha_1}|\boldsymbol{y}) + O(k^{-2} \vee k^2 n^{-2}) \\
&= \alpha - Q(\hat{\zeta}_\alpha|\boldsymbol{y}) + Q(\hat{\zeta}_{\alpha_1}|\boldsymbol{y}) + O(k^{-2} \vee k^2 n^{-2}) \\
&= \alpha - Q(\hat{\zeta}_\alpha|\boldsymbol{y}) + Q(\hat{\zeta}_\alpha|\boldsymbol{y}) + O(k^{-2} \vee k^2 n^{-2}) \\
&= \alpha + O(k^{-2} \vee k^2 n^{-2})
\end{aligned}$$

The second approximation is to use

$$\tilde{z}^*_\alpha = \hat{\zeta}_\alpha - Q(\hat{\zeta}_\alpha|\boldsymbol{y})/\Phi'(\hat{\zeta}_\alpha|\hat{\boldsymbol{z}};\hat{\gamma}) \tag{7.19}$$

In this case, by Taylor expansion,

$$\begin{aligned}
\tilde{F}(\tilde{z}^*_\alpha|\boldsymbol{y}) &= \Phi(\tilde{z}^*_\alpha|\hat{\boldsymbol{z}};\hat{\gamma}) + Q(\tilde{z}^*_\alpha|\boldsymbol{y}) + O(k^{-2} \vee k^2 n^{-2}) \\
&= \Phi(\hat{\zeta}_\alpha|\hat{\boldsymbol{z}};\hat{\gamma}) - Q(\hat{\zeta}_\alpha|\boldsymbol{y}) + Q(\tilde{z}^*_\alpha|\boldsymbol{y}) + O(k^{-2} \vee k^2 n^{-2}) \\
&= \alpha + O(k^{-2} \vee k^2 n^{-2})
\end{aligned}$$

It turns out that

$$\begin{aligned}
\tilde{z}^*_\alpha = \hat{\mu} + \hat{\tau}\Phi^{-1}(\alpha) + \frac{1}{2}\hat{\tau}^{-1}\Phi^{-1}(\alpha)\hat{\mu}_{i_1}\hat{\mu}_{i_2}\hat{\psi}^{i_1 i_2} - \frac{1}{2}\hat{\mu}_{i_1}\hat{\psi}_{i_2 i_2 i_2}\hat{\psi}^{i_1 i_2}\hat{\psi}^{i_2 i_2} + \frac{1}{2}\hat{\mu}_{j_1 j_2}\hat{\psi}^{j_1 j_2} \\
- \hat{\tau}^{-1}(1 - \Phi^{-1}(\alpha)^2)\hat{\mu}_{j_1}\hat{\tau}_{j_2}\hat{\psi}^{j_1 j_2} + \frac{1}{2}\Phi^{-1}(\alpha)\hat{\tau}_{j_1 j_2}\hat{\psi}^{j_1 j_2} - \frac{1}{2}\hat{\tau}^{-1}\Phi^{-1}(\alpha)(2 - \Phi^{-1}(\alpha)^2)\hat{\tau}_{j_1}\hat{\tau}_{j_2}\hat{\psi}^{j_1 j_2} \\
+ \frac{1}{2}\hat{\tau}^{-1}\Phi^{-1}(\alpha)\hat{\mu}_{j_1}\hat{\mu}_{j_2}\hat{\psi}^{j_1 j_2} - \hat{\mu}_{j_1}\hat{r}_{j_2}\hat{\psi}^{j_1 j_2} - \Phi^{-1}(\alpha)\hat{\tau}_{j_1}\hat{r}_{j_2}\hat{\psi}^{j_1 j_2} - \frac{1}{2}\hat{\mu}_{j_1}\hat{\psi}_{j_2 j_3 j_4}\hat{\psi}^{j_1 j_2}\hat{\psi}^{j_3 j_4}
\end{aligned}$$

$$-\frac{1}{2}\Phi^{-1}(\alpha)\hat{\tau}_{j_1}\hat{\psi}_{j_2j_3j_4}\hat{\psi}^{j_1j_2}\hat{\psi}^{j_3j_4} \quad (7.20)$$

7.2 Coverage Probability Bias

Consider the predictive density $F(z_0|\boldsymbol{y};\gamma)$ constructed if the parameter γ was known:

$$F(z_0|\boldsymbol{y};\gamma) = \frac{\iint \Phi(z_0|\boldsymbol{z};\gamma)\exp\{-\psi(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\beta}{\iint \exp\{-\psi(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\beta} \quad (7.21)$$

Suppose we estimate the α-quantile of (7.21) by $\tilde{z}_\alpha$. The coverage probability bias is defined as

$$\mathbb{E}F(\tilde{z}_\alpha|\boldsymbol{Y};\gamma) - \alpha \quad (7.22)$$

Consider the log-likelihood for γ given the sample $\boldsymbol{y}$

$$\ell(\gamma|\boldsymbol{y}) = \iint \exp\{-\psi(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)\}\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\beta \quad (7.23)$$

Then

$$\tilde{F}(z_0|\boldsymbol{y}) = \frac{\int F(z_0|\boldsymbol{y};\gamma)\exp\{r(\gamma)\}\exp\{\ell(\gamma|\boldsymbol{y})\}\,\mathrm{d}\gamma}{\int \exp\{r(\gamma)\}\exp\{\ell(\gamma|\boldsymbol{y})\}\,\mathrm{d}\gamma} \quad (7.24)$$

We express the derivatives of the log-likelihood in (7.23) with respect to the components of γ in the form

$$U_{j_1}(\gamma|\boldsymbol{Y}) = k^{1/2}W_{j_1}(\gamma|\boldsymbol{Y}) \quad (7.25)$$

$$U_{j_1j_2}(\gamma|\boldsymbol{Y}) = k\kappa_{j_1j_2} + k^{1/2}W_{j_1j_2}(\gamma|\boldsymbol{Y}) \quad (7.26)$$

$$U_{j_1j_2j_3}(\gamma|\boldsymbol{Y}) = k\kappa_{j_1j_2j_3} + k^{1/2}W_{j_1j_2j_3}(\gamma|\boldsymbol{Y}) \quad (7.27)$$

and so on for higher order derivatives. This is justified because the leading terms in the expansions of the above derivatives are the same as the ones that would result if the distribution was Gaussian (see Section 4.2). Note that we treat the likelihood and its derivatives as being functions of $\boldsymbol{Y}$, and therefore, random variables.

Let $\hat{\gamma} = \text{argmax}_\gamma \, \ell(\gamma|\boldsymbol{Y})$ and write the difference $\hat{\gamma} - \gamma$ as

$$\hat{\gamma}^j - \gamma^j = k^{-1/2}\varepsilon_1^j + k^{-1}\varepsilon_2^j + \dots \tag{7.28}$$

Then,

$$\varepsilon_1^j = W_{j_1}\kappa^{j_1,j} \tag{7.29}$$

$$\begin{aligned}\varepsilon_{m+1}^j = {} & W_{j_1 j_2}\varepsilon_m^{j_1}\kappa^{j_2,j} \\ & + \frac{1}{2!}\kappa_{j_1 j_2 j_3}\left(\sum_{i_1+i_2=m+1}\varepsilon_{i_1}^{j_1}\varepsilon_{i_2}^{j_2}\right)\kappa^{j_3,j} + \frac{1}{2!}W_{j_1 j_2 j_3}\left(\sum_{i_1+i_2=m}\varepsilon_{i_1}^{j_1}\varepsilon_{i_2}^{j_2}\right)\kappa^{j_3,j} \\ & + \dots + \frac{1}{m!}\kappa_{j_1\dots j_{m+1}}\left(\sum_{i_1+\dots+i_m=m+1}\varepsilon_{i_1}^{j_1}\dots\varepsilon_{i_m}^{j_m}\right)\kappa^{j_{m+1},j} \\ & + \frac{1}{m!}W_{j_1,\dots,j_{m+1}}\left(\sum_{i_1+\dots+i_m=m}\varepsilon_{i_1}^{j_1}\dots\varepsilon_{i_m}^{j_m}\right)\kappa^{j_{m+1},j} + \frac{1}{(m+1)!}\kappa_{j_1,\dots,j_{m+2}}\varepsilon_1^{j_1}\dots\varepsilon_1^{j_{m+1}}\kappa^{j_{m+2},j}\end{aligned} \tag{7.30}$$

Then the following result holds

Theorem 1. *The difference between the Bayesian and the true predictive distribution function is given by*

$$\tilde{F}(z_0|\boldsymbol{y}) - F(z_0|\boldsymbol{y};\gamma) = k^{-1/2}R(z_0|\boldsymbol{y};\gamma) + k^{-1}S(z_0|\boldsymbol{y};\gamma) + k^{-3/2}T(z_0|\boldsymbol{y};\gamma) + O(k^{-2}) \tag{7.31}$$

where

$$R = F_{j_1}\varepsilon_1^{j_1} \tag{7.32}$$

$$S = F_{j_1}\varepsilon_2^{j_1} + \frac{1}{2}F_{j_1 j_2}\varepsilon_1^{j_1}\varepsilon_1^{j_2} + \frac{1}{2}F_{j_1}\kappa_{j_2 j_3 j_4}\kappa^{j_1,j_2}\kappa^{j_3,j_4} + \frac{1}{2}F_{j_1 j_2}\kappa^{j_1,j_2} + F_{j_1}r_{j_2}\kappa^{j_1,j_2} \tag{7.33}$$

$$\begin{aligned}T = {} & F_{j_1}\varepsilon_3^{j_1} + \frac{1}{2}F_{j_1 j_2}(\varepsilon_1^{j_1}\varepsilon_2^{j_2} + \varepsilon_2^{j_1}\varepsilon_1^{j_2}) + \frac{1}{6}F_{j_1 j_2 j_3}\varepsilon_1^{j_1}\varepsilon_1^{j_2}\varepsilon_1^{j_3} \\ & + \frac{1}{2}F_{j_1}\{(W_{j_2 j_3 j_4} + \kappa_{j_2 j_3 j_4 j_5}\varepsilon_1^{j_5})\kappa^{j_1,j_2}\kappa^{j_3,j_4} + \kappa_{j_2 j_3 j_4}(W_{j_5 j_6} + \kappa_{j_5 j_6 j_7}\varepsilon_1^{j_7})\kappa^{j_1,j_5}\kappa^{j_2,j_6}\kappa^{j_3,j_4} \\ & \qquad + \kappa_{j_2 j_3 j_4}(W_{j_5 j_6} + \kappa_{j_5 j_6 j_7}\varepsilon_1^{j_7})\kappa^{j_1,j_2}\kappa^{j_3,j_5}\kappa^{j_4,j_6}\} \\ & + \frac{1}{2}F_{j_1 j_5}\kappa_{j_2 j_3 j_4}\varepsilon_1^{j_5}\kappa^{j_1,j_2}\kappa^{j_3,j_4} + \frac{1}{2}(F_{j_1,j_2} + 2F_{j_1}r_{j_2})(W_{j_3 j_4} + \kappa_{j_3 j_4 j_5}\varepsilon_1^{j_5})\kappa^{j_1,j_3}\kappa^{j_2,j_4}\end{aligned} \tag{7.34}$$

$$+\frac{1}{2}F_{j_1j_2j_3}\varepsilon_1^{j_3}\kappa^{j_1,j_2}+(F_{j_1}r_{j_2j_3}+F_{j_1j_3}r_{j_2})\varepsilon_1^{j_3}\kappa^{j_1,j_2} \tag{7.34}$$

and

$$F_{j_1\ldots j_s}=\frac{\partial^s}{\partial\gamma_{j_1}\ldots\partial\gamma_{j_s}}F(z_0|\boldsymbol{y};\gamma) \tag{7.35}$$

The proof of this result is given in section A.2 of the Appendix.

Evaluating (7.31) at $z_0=\tilde{z}_\alpha$,

$$F(\tilde{z}_\alpha|\boldsymbol{y};\gamma)=\alpha-k^{-1/2}R(\tilde{z}_\alpha|\boldsymbol{y};\gamma)-k^{-1}S(\tilde{z}_\alpha|\boldsymbol{y};\gamma)-k^{-3/2}T(\tilde{z}_\alpha|\boldsymbol{y};\gamma)+O(k^{-2}\vee k^2n^{-2}) \tag{7.36}$$

we obtain an estimate for the bias of the Bayesian predictive density. In section 7.4 we describe how the above expression can be computed. Knowing the bias is important in assessing the performance of the prior when used for the construction of the predictive density.

7.3 Kullback-Leibler Divergence

Let $\tilde{f}(z_0|\boldsymbol{y})$ be the Bayesian predictive density evaluated with prior $\gamma\sim\exp\{r(\gamma)\}$ and let $f(z_0|\boldsymbol{y};\gamma)$ be the predictive density if γ was known. We would like to see how close $\tilde{f}$ is to f by approximating the Kullback-Leibler divergence of $\tilde{f}$ from f.

7.3.1 Approximation to the Bayesian predictive density

Here we follow section 6 of Barndorff-Nielsen and Cox (1996).

Define $\tilde{z}_\alpha$ to be the α-quantile of the Bayesian predictive distribution using the prior $\gamma\sim\exp\{r(\gamma)\}$, i.e.

$$\tilde{F}(\tilde{z}_\alpha|\boldsymbol{y})=\alpha \tag{7.37}$$

Differentiating (7.37) with respect to α, we have

$$\left\{\frac{\mathrm{d}}{\mathrm{d}\alpha}\tilde{z}_\alpha\right\}\tilde{f}(\tilde{z}_\alpha|\boldsymbol{y})=1$$

hence

$$\tilde{f}(\tilde{z}_\alpha|\boldsymbol{y}) = \left\{ \frac{\mathrm{d}}{\mathrm{d}\alpha}\tilde{z}_\alpha \right\}^{-1} \tag{7.38}$$

Note that by (7.16), if $\hat{\zeta}_\alpha$ is given by (7.18), then

$$\begin{aligned} 0 &= \tilde{F}(\tilde{z}_\alpha|\boldsymbol{y}) - \Phi(\hat{\zeta}_\alpha|\hat{\boldsymbol{z}};\hat{\gamma}) \\ &= Q(\tilde{z}_\alpha|\boldsymbol{y}) + \Phi(\tilde{z}_\alpha|\hat{\boldsymbol{z}};\hat{\gamma}) - \Phi(\hat{\zeta}_\alpha|\hat{\boldsymbol{z}};\hat{\gamma}) \\ &= Q(\tilde{z}_\alpha|\boldsymbol{y}) + (\tilde{z}_\alpha - \hat{\zeta}_\alpha)\,\Phi'(\tilde{z}_\alpha|\hat{\boldsymbol{z}};\hat{\gamma}) \end{aligned}$$

Hence,

$$\tilde{z}_\alpha = \hat{\zeta}_\alpha - Q(\tilde{z}_\alpha|\boldsymbol{y})/\Phi'(\tilde{z}_\alpha|\hat{\boldsymbol{z}};\hat{\gamma}) + O(k^{-2} \vee k^2 n^{-2}) \tag{7.39}$$

Differentiating (7.39) with respect to $\tilde{z}_\alpha$,

$$1 = \frac{\mathrm{d}\hat{\zeta}_\alpha}{\mathrm{d}\tilde{z}_\alpha} - \frac{\mathrm{d}}{\mathrm{d}\tilde{z}_\alpha}\frac{Q(\tilde{z}_\alpha|\boldsymbol{y})}{\Phi'(\tilde{z}_\alpha|\hat{\boldsymbol{z}};\hat{\gamma})} + O(k^{-2} \vee k^2 n^{-2}) \tag{7.40}$$

so

$$\begin{aligned} \frac{\mathrm{d}\hat{\zeta}_\alpha}{\mathrm{d}\tilde{z}_\alpha} &= 1 + \frac{\mathrm{d}}{\mathrm{d}\tilde{z}_\alpha}\frac{Q(\tilde{z}_\alpha|\boldsymbol{y})}{\Phi'(\tilde{z}_\alpha|\hat{\boldsymbol{z}};\hat{\gamma})} + O(k^{-2} \vee k^2 n^{-2}) \\ \Rightarrow \frac{\mathrm{d}\tilde{z}_\alpha}{\mathrm{d}\hat{\zeta}_\alpha} &= \left(1 + \frac{\mathrm{d}}{\mathrm{d}\tilde{z}_\alpha}\frac{Q(\tilde{z}_\alpha|\boldsymbol{y})}{\Phi'(\tilde{z}_\alpha|\hat{\boldsymbol{z}};\hat{\gamma})} + O(k^{-2} \vee k^2 n^{-2}) \right)^{-1} \\ &= 1 - \frac{\mathrm{d}}{\mathrm{d}\tilde{z}_\alpha}\frac{Q(\tilde{z}_\alpha|\boldsymbol{y})}{\Phi'(\tilde{z}_\alpha|\hat{\boldsymbol{z}};\hat{\gamma})} + O(k^{-2} \vee k^2 n^{-2}) \end{aligned}$$

On the other hand, differentiating $\Phi(\hat{\zeta}_\alpha|\hat{\boldsymbol{z}};\hat{\gamma}) = \alpha$ with respect to α,

$$\frac{\mathrm{d}\hat{\zeta}_\alpha}{\mathrm{d}\alpha} = \left\{ \Phi'(\hat{\zeta}_\alpha|\hat{\boldsymbol{z}};\hat{\gamma}) \right\}^{-1} \tag{7.41}$$

Hence,

$$\begin{aligned} \frac{\mathrm{d}\tilde{z}_\alpha}{\mathrm{d}\alpha} &= \frac{\mathrm{d}\tilde{z}_\alpha}{\mathrm{d}\hat{\zeta}_\alpha}\frac{\mathrm{d}\hat{\zeta}_\alpha}{\mathrm{d}\alpha} \\ &= \left(1 - \frac{\mathrm{d}}{\mathrm{d}\tilde{z}_\alpha}\frac{Q(\tilde{z}_\alpha|\boldsymbol{y})}{\Phi'(\tilde{z}_\alpha|\hat{\boldsymbol{z}};\hat{\gamma})} + O(k^{-2} \vee k^2 n^{-2}) \right) \left\{ \Phi'(\hat{\zeta}_\alpha|\hat{\boldsymbol{z}};\hat{\gamma}) \right\}^{-1} \end{aligned}$$

$$= \left(1 - \frac{\mathrm{d}}{\mathrm{d}\tilde{z}_\alpha}\frac{Q(\tilde{z}_\alpha|\boldsymbol{y})}{\Phi'(\tilde{z}_\alpha|\hat{\boldsymbol{z}};\hat{\gamma})} + O(k^{-2}\vee k^2 n^{-2})\right)$$
$$\times \left\{\Phi'(\tilde{z}_\alpha + Q(\tilde{z}_\alpha|\boldsymbol{y})/\Phi'(\tilde{z}_\alpha|\hat{\boldsymbol{z}};\hat{\gamma}) + O(k^{-2}\vee k^2 n^{-2})|\hat{\boldsymbol{z}};\hat{\gamma})\right\}^{-1} \tag{7.42}$$

Using (7.42) into (7.38), we obtain the following approximation

$$\tilde{f}(z_0|\boldsymbol{y}) = \left(1 + \frac{\mathrm{d}}{\mathrm{d}z_0}\frac{Q(z_0|\boldsymbol{y})}{\Phi'(z_0|\hat{\boldsymbol{z}};\hat{\gamma})}\right)\Phi'\left(z_0 + \frac{Q(z_0|\boldsymbol{y})}{\Phi'(z_0|\hat{\boldsymbol{z}};\hat{\gamma})}\middle|\,\hat{\boldsymbol{z}};\hat{\gamma}\right)\left(1 + O(k^{-2}\vee k^2 n^{-2})\right) \tag{7.43}$$

Note that $\Phi(z_0|\hat{\boldsymbol{z}};\hat{\gamma}) = \Phi\left((z_0-\hat{\mu})/\hat{\tau}\right)$ and $Q(z_0|\boldsymbol{y})$ is given by (7.17), therefore

$$\begin{aligned}\frac{Q(z_0|\boldsymbol{y})}{\Phi'(z_0|\hat{\boldsymbol{z}};\hat{\gamma})} = &-\frac{1}{2}\hat{\tau}^{-1}\hat{\mu}_{i_1}\hat{\mu}_{i_2}\frac{z_0-\hat{\mu}}{\hat{\tau}}\hat{\psi}^{i_1 i_2} + \frac{1}{2}\hat{\mu}_{i_1}\hat{\psi}_{i_2 i_2 i_2}\hat{\psi}^{i_1 i_2}\hat{\psi}^{i_2 i_2} - \frac{1}{2}\hat{\mu}_{j_1 j_2}\hat{\psi}^{j_1 j_2} + \hat{\tau}^{-1}\hat{\mu}_{j_1}\hat{\tau}_{j_2}\hat{\psi}^{j_1 j_2}\\
&- \frac{1}{2}\hat{\tau}_{j_1 j_2}\frac{z_0-\hat{\mu}}{\hat{\tau}}\hat{\psi}^{j_1 j_2} + \hat{\tau}^{-1}\hat{\tau}_{j_1}\hat{\tau}_{j_2}\frac{z_0-\hat{\mu}}{\hat{\tau}}\hat{\psi}^{j_1 j_2} - \frac{1}{2}\hat{\tau}^{-1}\hat{\mu}_{j_1}\hat{\mu}_{j_2}\frac{z_0-\hat{\mu}}{\hat{\tau}}\hat{\psi}^{j_1 j_2}\\
&- \hat{\tau}^{-1}\hat{\mu}_{j_1}\hat{\tau}_{j_2}\left(\frac{z_0-\hat{\mu}}{\hat{\tau}}\right)^2\hat{\psi}^{j_1 j_2} - \frac{1}{2}\hat{\tau}^{-1}\hat{\tau}_{j_1}\hat{\tau}_{j_2}\left(\frac{z_0-\hat{\mu}}{\hat{\tau}}\right)^3\hat{\psi}^{j_1 j_2} - \hat{\mu}_{j_1}\hat{r}_{j_2}\hat{\psi}^{j_1 j_2}\\
&- \hat{\tau}_{j_1}\hat{r}_{j_2}\frac{z_0-\hat{\mu}}{\hat{\tau}}\hat{\psi}^{j_1 j_2} + \frac{1}{2}\hat{\mu}_{j_1}\hat{\psi}_{j_2 j_3 j_4}\hat{\psi}^{j_1 j_2}\hat{\psi}^{j_3 j_4} + \frac{1}{2}\hat{\tau}_{j_1}\frac{z_0-\hat{\mu}}{\hat{\tau}}\hat{\psi}_{j_2 j_3 j_4}\hat{\psi}^{j_1 j_2}\hat{\psi}^{j_3 j_4}\end{aligned} \tag{7.44}$$

and

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}z_0}\frac{Q(z_0|\boldsymbol{y})}{\Phi'(z_0|\hat{\boldsymbol{z}};\hat{\gamma})} = &-\frac{1}{2}\frac{\hat{\mu}_{i_1}\hat{\mu}_{i_2}}{\hat{\tau}^2}\hat{\psi}^{i_1 i_2} - \frac{1}{2}\frac{\hat{\tau}_{j_1 j_2}}{\hat{\tau}}\hat{\psi}^{j_1 j_2} + \frac{\hat{\tau}_{j_1}\hat{\tau}_{j_2}}{\hat{\tau}^2}\hat{\psi}^{j_1 j_2} - \frac{1}{2}\frac{\hat{\mu}_{j_1}\hat{\mu}_{j_2}}{\hat{\tau}^2}\hat{\psi}^{j_1 j_2}\\
&- 2\frac{\hat{\mu}_{j_1}\hat{\tau}_{j_2}}{\hat{\tau}^2}\frac{z_0-\hat{\mu}}{\hat{\tau}}\hat{\psi}^{j_1 j_2} - \frac{3}{2}\frac{\hat{\tau}_{j_1}\hat{\tau}_{j_2}}{\hat{\tau}^2}\left(\frac{z_0-\hat{\mu}}{\hat{\tau}}\right)^2\hat{\psi}^{j_1 j_2} - \frac{\hat{\tau}_{j_1}\hat{r}_{j_2}}{\hat{\tau}}\hat{\psi}^{j_1 j_2}\\
&+ \frac{1}{2}\frac{\hat{\tau}_{j_1}}{\hat{\tau}}\hat{\psi}_{j_2 j_3 j_4}\hat{\psi}^{j_1 j_2}\hat{\psi}^{j_3 j_4}\end{aligned} \tag{7.45}$$

7.3.2 Approximation to the Kullback-Leibler divergence

A similar expression exists for the predictive density when γ is known

$$f(z_0|\boldsymbol{y};\gamma) = \left(1 + \frac{\mathrm{d}}{\mathrm{d}z_0}\frac{Q(z_0|\boldsymbol{y};\gamma)}{\Phi'(z_0|\dot{\boldsymbol{z}};\gamma)}\right)\Phi'\left(z_0 + \frac{Q(z_0|\boldsymbol{y};\gamma)}{\Phi'(z_0|\dot{\boldsymbol{z}};\gamma)}\middle|\,\dot{\boldsymbol{z}};\gamma\right)\left(1 + O(k^{-2}\vee k^2 n^{-2})\right) \tag{7.46}$$

where $(\dot{\boldsymbol{z}},\dot{\beta}) = \mathrm{argmin}_{(\boldsymbol{z},\beta)}\,\psi(\boldsymbol{y},\boldsymbol{z};\beta,\gamma)$ and

$$\frac{Q(z_0|\boldsymbol{y};\gamma)}{\Phi'(z_0|\dot{\boldsymbol{z}};\gamma)} = -\frac{1}{2}\hat{\tau}^{-1}\mu_{i_1}\mu_{i_2}\frac{z_0-\dot{\mu}}{\tau}\dot{\psi}^{i_1 i_2} + \frac{1}{2}\mu_{i_1}\dot{\psi}_{i_2 i_2 i_2}\dot{\psi}^{i_1 i_2}\dot{\psi}^{i_2 i_2}$$

$$\frac{\mathrm{d}}{\mathrm{d}z_0}\frac{Q(z_0|\boldsymbol{y};\gamma)}{\Phi'(z_0|\dot{\boldsymbol{z}};\gamma)} = -\frac{1}{2}\frac{\mu_{i_1}\mu_{i_2}}{\tau^2}\dot{\psi}^{i_1 i_2}$$

hence Normal with mean

$$\mathring{\mu} = \dot{\mu} - \frac{1}{2}\tau^{-1}\mu_{i_1}\dot{\psi}_{i_2 i_2 i_2}\dot{\psi}^{i_1 i_2}\dot{\psi}^{i_2 i_2}\left(1 - \frac{1}{2}\tau^{-2}\mu_{i_1}\mu_{i_2}\dot{\psi}^{i_1 i_2}\right)^{-1}$$

and standard deviation

$$\mathring{\tau} = \tau\left(1 - \frac{1}{2}\tau^{-2}\mu_{i_1}\mu_{i_2}\dot{\psi}^{i_1 i_2}\right)^{-1}$$

The log-difference of (7.43) from (7.46) is approximately

$$\begin{aligned}\log\tilde{f}(z_0|\boldsymbol{y}) - \log f(z_0|\boldsymbol{y};\gamma) = {}& \frac{\mathrm{d}}{\mathrm{d}z_0}\frac{Q(z_0|\boldsymbol{y})}{\Phi'(z_0|\hat{\boldsymbol{z}};\hat{\gamma})} + \log\frac{\mathring{\tau}}{\hat{\tau}} - \frac{1}{2\hat{\tau}^2}(z_0-\hat{\mu})^2 \\ & + \frac{1}{\hat{\tau}^2}(z_0-\hat{\mu})\frac{Q(z_0|\boldsymbol{y})}{\Phi'(z_0|\hat{\boldsymbol{z}};\hat{\gamma})} + \frac{1}{2\mathring{\tau}^2}(z_0-\mathring{\mu})^2 + O(k^{-2}\vee k^2 n^{-2})\end{aligned}$$

The Kullback-Leibler divergence of $\tilde{f}$ from f given $\boldsymbol{y}$, $\mathrm{KL}(\tilde{f}, f|\boldsymbol{y})$, is

$$\begin{aligned}\mathrm{KL}(\tilde{f}, f|\boldsymbol{y}) = {}& -\int\{\log\tilde{f}(z_0|\boldsymbol{y}) - \log f(z_0|\boldsymbol{y};\gamma)\}f(z_0|\boldsymbol{y};\gamma)\,\mathrm{d}z_0 \\ = {}& \frac{1}{2}\frac{\hat{\mu}_{i_1}\hat{\mu}_{i_2}}{\hat{\tau}^2}\hat{\psi}^{i_1 i_2} + \frac{1}{2}\frac{\hat{\tau}_{j_1 j_2}}{\hat{\tau}}\hat{\psi}^{j_1 j_2} - \frac{\hat{\tau}_{j_1}\hat{\tau}_{j_2}}{\hat{\tau}^2}\hat{\psi}^{j_1 j_2} + \frac{1}{2}\frac{\hat{\mu}_{j_1}\hat{\mu}_{j_2}}{\hat{\tau}^2}\hat{\psi}^{j_1 j_2} \\ & + 2\frac{\mathring{\mu}-\hat{\mu}}{\hat{\tau}}\frac{\hat{\mu}_{j_1}\hat{\tau}_{j_2}}{\hat{\tau}^2}\hat{\psi}^{j_1 j_2} + \frac{1}{2}\frac{\mathring{\tau}^2}{\hat{\tau}^2}\frac{\hat{\tau}_{j_1}\hat{\tau}_{j_2}}{\hat{\tau}^2}\hat{\psi}^{j_1 j_2} + \frac{1}{2}\frac{(\mathring{\mu}-\hat{\mu})^2}{\hat{\tau}^2}\frac{\hat{\tau}_{j_1}\hat{\tau}_{j_2}}{\hat{\tau}^2}\hat{\psi}^{j_1 j_2} + \frac{\hat{\tau}_{j_1}\hat{r}_{j_2}}{\hat{\tau}}\hat{\psi}^{j_1 j_2} \\ & - \frac{1}{2}\frac{\hat{\tau}_{j_1}}{\hat{\tau}}\hat{\psi}_{j_2 j_3 j_4}\hat{\psi}^{j_1 j_2}\hat{\psi}^{j_3 j_4} - \log\frac{\mathring{\tau}}{\hat{\tau}} + \frac{\mathring{\tau}^2}{2\hat{\tau}^2} + \frac{(\mathring{\mu}-\hat{\mu})^2}{2\hat{\tau}^2} \\ & + \frac{1}{2}\frac{(\mathring{\mu}-\hat{\mu})^2}{\hat{\tau}^2}\frac{\hat{\mu}_{i_1}\hat{\mu}_{i_2}}{\hat{\tau}^2}\hat{\psi}^{i_1 i_2} + \frac{1}{2}\frac{\mathring{\tau}^2}{\hat{\tau}^2}\frac{\hat{\mu}_{i_1}\hat{\mu}_{i_2}}{\hat{\tau}^2}\hat{\psi}^{i_1 i_2} - \frac{1}{2}\frac{\mathring{\mu}-\hat{\mu}}{\hat{\tau}}\frac{\hat{\mu}_{i_1}}{\hat{\tau}}\hat{\psi}_{i_2 i_2 i_2}\hat{\psi}^{i_1 i_2}\hat{\psi}^{i_2 i_2} \\ & + \frac{1}{2}\frac{\mathring{\mu}-\hat{\mu}}{\hat{\tau}}\frac{\hat{\mu}_{j_1 j_2}}{\hat{\tau}}\hat{\psi}^{j_1 j_2} - \frac{\mathring{\mu}-\hat{\mu}}{\hat{\tau}}\frac{\hat{\mu}_{j_1}\hat{\tau}_{j_2}}{\hat{\tau}^2}\hat{\psi}^{j_1 j_2} + \frac{1}{2}\frac{(\mathring{\mu}-\hat{\mu})^2}{\hat{\tau}^2}\frac{\hat{\tau}_{j_1 j_2}}{\hat{\tau}}\hat{\psi}^{j_1 j_2} \\ & + \frac{1}{2}\frac{\mathring{\tau}^2}{\hat{\tau}^2}\frac{\hat{\tau}_{j_1 j_2}}{\hat{\tau}}\hat{\psi}^{j_1 j_2} + \frac{1}{2}\frac{(\mathring{\mu}-\hat{\mu})^2}{\hat{\tau}^2}\frac{\hat{\mu}_{j_1}\hat{\mu}_{j_2}}{\hat{\tau}^2}\hat{\psi}^{j_1 j_2} \\ & + \frac{1}{2}\frac{\mathring{\tau}^2}{\hat{\tau}^2}\frac{\hat{\mu}_{j_1}\hat{\mu}_{j_2}}{\hat{\tau}^2}\hat{\psi}^{j_1 j_2} + \frac{(\mathring{\mu}-\hat{\mu})^3}{\hat{\tau}^3}\frac{\hat{\mu}_{j_1}\hat{\tau}_{j_2}}{\hat{\tau}^2}\hat{\psi}^{j_1 j_2} + 3\frac{(\mathring{\mu}-\hat{\mu})\mathring{\tau}^2}{\hat{\tau}^3}\frac{\hat{\mu}_{j_1}\hat{\tau}_{j_2}}{\hat{\tau}^2}\hat{\psi}^{j_1 j_2} \\ & + \frac{1}{2}\frac{(\mathring{\mu}-\hat{\mu})^4}{\hat{\tau}^4}\frac{\hat{\tau}_{j_1}\hat{\tau}_{j_2}}{\hat{\tau}^2}\hat{\psi}^{j_1 j_2} + 3\frac{(\mathring{\mu}-\hat{\mu})^2\mathring{\tau}^2}{\hat{\tau}^4}\frac{\hat{\tau}_{j_1}\hat{\tau}_{j_2}}{\hat{\tau}^2}\hat{\psi}^{j_1 j_2} + \frac{3}{2}\frac{\mathring{\tau}^4}{\hat{\tau}^4}\frac{\hat{\tau}_{j_1}\hat{\tau}_{j_2}}{\hat{\tau}^2}\hat{\psi}^{j_1 j_2} \\ & + \frac{\mathring{\mu}-\hat{\mu}}{\hat{\tau}}\frac{\hat{\mu}_{j_1}\hat{r}_{j_2}}{\hat{\tau}}\hat{\psi}^{j_1 j_2} + \frac{(\mathring{\mu}-\hat{\mu})^2}{\hat{\tau}^2}\frac{\hat{\tau}_{j_1}\hat{r}_{j_2}}{\hat{\tau}}\hat{\psi}^{j_1 j_2} + \frac{\mathring{\tau}^2}{\hat{\tau}^2}\frac{\hat{\tau}_{j_1}\hat{r}_{j_2}}{\hat{\tau}}\hat{\psi}^{j_1 j_2} \\ & - \frac{1}{2}\frac{\mathring{\mu}-\hat{\mu}}{\hat{\tau}}\frac{\hat{\mu}_{j_1}}{\hat{\tau}}\hat{\psi}_{j_2 j_3 j_4}\hat{\psi}^{j_1 j_2}\hat{\psi}^{j_3 j_4} - \frac{1}{2}\frac{(\mathring{\mu}-\hat{\mu})^2}{\hat{\tau}^2}\frac{\hat{\tau}_{j_1}}{\hat{\tau}}\hat{\psi}_{j_2 j_3 j_4}\hat{\psi}^{j_1 j_2}\hat{\psi}^{j_3 j_4} \\ & - \frac{1}{2}\frac{\mathring{\tau}^2}{\hat{\tau}^2}\frac{\hat{\tau}_{j_1}}{\hat{\tau}}\hat{\psi}_{j_2 j_3 j_4}\hat{\psi}^{j_1 j_2}\hat{\psi}^{j_3 j_4} - \frac{1}{2} + O(k^{-2}\vee k^2 n^{-2}) \qquad (7.47)\end{aligned}$$

Ideally we would like to be able to obtain the Kullback-Leibler divergence unconditionally, but the expression in (7.47) is too complicated to allow us to have a closed form expression for its expectation with respect to $\boldsymbol{y}$.

7.4 Computations

The expressions in (7.36) and (7.47) involve quantities that cannot be computed explicitly. Here we show how they can be approximated given a random sample from the distribution $f(\boldsymbol{z}|\boldsymbol{y};\beta,\gamma)$ which is obtained for predicting $Z_0|\boldsymbol{Y}$.

7.4.1 Log-likelihood derivatives and Cumulants

The derivatives of the log-likelihood as defined in (7.26) and (7.27) are needed. Here we propose a method to obtain them approximately using the random sample from the distribution of $\boldsymbol{Z}|\boldsymbol{Y}$.

Write the likelihood:

$$L(\beta,\gamma|\boldsymbol{y}) = \int f(\boldsymbol{y}|\boldsymbol{z};\beta)\, f(\boldsymbol{z};\gamma)\,\mathrm{d}\boldsymbol{z} \tag{7.48}$$

and the log-likelihood:

$$\ell(\beta,\gamma|\boldsymbol{y}) = \log L(\beta,\gamma|\boldsymbol{y}) \tag{7.49}$$

To simplify the notation, we will use ∂_{j_1} for $\partial/\partial\gamma_{j_1}$, $\partial^2_{j_1j_2}$ for $\partial^2/\partial\gamma_{j_1}\partial\gamma_{j_2}$ and so on. Then (see section A.3)

$$\frac{\partial_{j_1} L(\beta,\gamma|\boldsymbol{y})}{L(\beta,\gamma|\boldsymbol{y})} = U_{j_1}(\boldsymbol{Y};\gamma) \tag{7.50}$$

$$\frac{\partial^2_{j_1j_2} L(\beta,\gamma|\boldsymbol{y})}{L(\beta,\gamma|\boldsymbol{y})} = U_{j_1j_2}(\boldsymbol{Y};\gamma) + U_{j_1}(\boldsymbol{Y};\gamma)U_{j_2}(\boldsymbol{Y};\gamma) \tag{7.51}$$

$$\frac{\partial^3_{j_1j_2j_3} L(\beta,\gamma|\boldsymbol{y})}{L(\beta,\gamma|\boldsymbol{y})} = U_{j_1j_2j_3}(\boldsymbol{Y};\gamma) + [3]U_{j_1j_2}(\boldsymbol{Y};\gamma)U_{j_3}(\boldsymbol{Y};\gamma) + U_{j_1}(\boldsymbol{Y};\gamma)U_{j_2}(\boldsymbol{Y};\gamma)U_{j_3}(\boldsymbol{Y};\gamma) \tag{7.52}$$

$$\begin{aligned}\frac{\partial^4_{j_1j_2j_3j_4} L(\beta,\gamma|\boldsymbol{y})}{L(\beta,\gamma|\boldsymbol{y})} &= U_{j_1j_2j_3j_4}(\boldsymbol{Y};\gamma) + [4]U_{j_1j_2j_3}(\boldsymbol{Y};\gamma)U_{j_4}(\boldsymbol{Y};\gamma) + [3]U_{j_1j_2}(\boldsymbol{Y};\gamma)U_{j_3j_4}(\boldsymbol{Y};\gamma) \\ &\quad + [6]U_{j_1j_2}(\boldsymbol{Y};\gamma)U_{j_3}(\boldsymbol{Y};\gamma)U_{j_4}(\boldsymbol{Y};\gamma)\end{aligned}$$

$$+ U_{j_1}(\boldsymbol{Y};\gamma)U_{j_2}(\boldsymbol{Y};\gamma)U_{j_3}(\boldsymbol{Y};\gamma)U_{j_4}(\boldsymbol{Y};\gamma) \tag{7.53}$$

from where the log-likelihood derivatives, $U_{j_1}(\boldsymbol{Y};\gamma)$, $U_{j_1 j_2}(\boldsymbol{Y};\gamma)$, etc, can be obtained given the quantities in the left hand side of equations (7.50) – (7.53).

Using (7.48),

$$\begin{aligned}\frac{\partial_{j_1} L(\beta,\gamma|\boldsymbol{y})}{L(\beta,\gamma|\boldsymbol{y})} &= \frac{\int f(\boldsymbol{y}|\boldsymbol{z};\beta) f_{j_1}(\boldsymbol{z};\gamma)\,\mathrm{d}\boldsymbol{z}}{\int f(\boldsymbol{y}|\boldsymbol{z};\beta) f(\boldsymbol{z};\gamma)\,\mathrm{d}\boldsymbol{z}} \\ &= \frac{\int f(\boldsymbol{y}|\boldsymbol{z};\beta) f(\boldsymbol{z};\gamma)\{\partial_{j_1}\log f(\boldsymbol{z};\gamma)\}\,\mathrm{d}\boldsymbol{z}}{\int f(\boldsymbol{y}|\boldsymbol{z};\beta) f(\boldsymbol{z};\gamma)\,\mathrm{d}\boldsymbol{z}} \\ &= \int f(\boldsymbol{z}|\boldsymbol{y};\beta,\gamma)\{\partial_{j_1}\log f(\boldsymbol{z};\gamma)\}\,\mathrm{d}\boldsymbol{z} \end{aligned} \tag{7.54}$$

and similarly,

$$\begin{aligned}\frac{\partial^2_{j_1 j_2} L(\beta,\gamma|\boldsymbol{y})}{L(\beta,\gamma|\boldsymbol{y})} &= \int f(\boldsymbol{z}|\boldsymbol{y};\beta,\gamma)\{\partial^2_{j_1 j_2}\log f(\boldsymbol{z};\gamma)\}\,\mathrm{d}\boldsymbol{z} \\ &\quad + \int f(\boldsymbol{z}|\boldsymbol{y};\beta,\gamma)\{\partial_{j_1}\log f(\boldsymbol{z};\gamma)\}\{\partial_{j_2}\log f(\boldsymbol{z};\gamma)\}\,\mathrm{d}\boldsymbol{z} \end{aligned} \tag{7.55}$$

$$\begin{aligned}\frac{\partial^3_{j_1 j_2 j_3} L(\beta,\gamma|\boldsymbol{y})}{L(\beta,\gamma|\boldsymbol{y})} &= \int f(\boldsymbol{z}|\boldsymbol{y};\beta,\gamma)\{\partial^3_{j_1 j_2 j_3}\log f(\boldsymbol{z};\gamma)\}\,\mathrm{d}\boldsymbol{z} \\ &\quad + [3]\int f(\boldsymbol{z}|\boldsymbol{y};\beta,\gamma)\{\partial^2_{j_1 j_2}\log f(\boldsymbol{z};\gamma)\}\{\partial_{j_3}\log f(\boldsymbol{z};\gamma)\}\,\mathrm{d}\boldsymbol{z} \\ &\quad + \int f(\boldsymbol{z}|\boldsymbol{y};\beta,\gamma)\{\partial_{j_1}\log f(\boldsymbol{z};\gamma)\}\{\partial_{j_2}\log f(\boldsymbol{z};\gamma)\}\{\partial_{j_3}\log f(\boldsymbol{z};\gamma)\}\,\mathrm{d}\boldsymbol{z} \end{aligned} \tag{7.56}$$

$$\begin{aligned}\frac{\partial^4_{j_1 j_2 j_3 j_4} L(\beta,\gamma|\boldsymbol{y})}{L(\beta,\gamma|\boldsymbol{y})} &= \int f(\boldsymbol{z}|\boldsymbol{y};\beta,\gamma)\{\partial^4_{j_1 j_2 j_3 j_4}\log f(\boldsymbol{z};\gamma)\}\,\mathrm{d}\boldsymbol{z} \\ &\quad + [4]\int f(\boldsymbol{z}|\boldsymbol{y};\beta,\gamma)\{\partial^3_{j_1 j_2 j_3}\log f(\boldsymbol{z};\gamma)\}\{\partial_{j_4}\log f(\boldsymbol{z};\gamma)\}\,\mathrm{d}\boldsymbol{z} \\ &\quad + [3]\int f(\boldsymbol{z}|\boldsymbol{y};\beta,\gamma)\{\partial^2_{j_1 j_2}\log f(\boldsymbol{z};\gamma)\}\{\partial^2_{j_3 j_4}\log f(\boldsymbol{z};\gamma)\}\,\mathrm{d}\boldsymbol{z} \\ &\quad + [6]\int f(\boldsymbol{z}|\boldsymbol{y};\beta,\gamma)\{\partial^2_{j_1 j_2}\log f(\boldsymbol{z};\gamma)\}\{\partial_{j_3}\log f(\boldsymbol{z};\gamma)\}\{\partial_{j_4}\log f(\boldsymbol{z};\gamma)\}\,\mathrm{d}\boldsymbol{z} \\ &\quad + \int f(\boldsymbol{z}|\boldsymbol{y};\beta,\gamma)\{\partial_{j_1}\log f(\boldsymbol{z};\gamma)\}\{\partial_{j_2}\log f(\boldsymbol{z};\gamma)\} \\ &\qquad\qquad \{\partial_{j_3}\log f(\boldsymbol{z};\gamma)\}\{\partial_{j_4}\log f(\boldsymbol{z};\gamma)\}\,\mathrm{d}\boldsymbol{z} \end{aligned} \tag{7.57}$$

where

$$\partial_{j_1} \log f(\boldsymbol{z};\gamma) = \frac{1}{2}\boldsymbol{z}^\intercal\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\boldsymbol{z} - \frac{1}{2}\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1}) \tag{7.58}$$

$$\begin{aligned}\partial^2_{j_1j_2} \log f(\boldsymbol{z};\gamma) = &-\boldsymbol{z}^\intercal\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\boldsymbol{z} + \frac{1}{2}\boldsymbol{z}^\intercal\Sigma^{-1}\Sigma_{j_1j_2}\Sigma^{-1}\boldsymbol{z} \\ &+ \frac{1}{2}\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\Sigma_{j_2}) - \frac{1}{2}\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1j_2})\end{aligned} \tag{7.59}$$

$$\begin{aligned}\partial^3_{j_1j_2j_3} \log f(\boldsymbol{z};\gamma) = &-[3]\boldsymbol{z}^\intercal\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\boldsymbol{z} + [3]\boldsymbol{z}^\intercal\Sigma^{-1}\Sigma_{j_1j_2}\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\boldsymbol{z} \\ &+ \frac{1}{2}\boldsymbol{z}^\intercal\Sigma^{-1}\Sigma_{j_1j_2j_3}\Sigma^{-1}\boldsymbol{z} - [2]\frac{1}{2}\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\Sigma_{j_3}) \\ &+ [3]\frac{1}{2}\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1j_2}\Sigma^{-1}\Sigma_{j_3}) - \frac{1}{2}\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1j_2j_3})\end{aligned} \tag{7.60}$$

$$\begin{aligned}\partial^4_{j_1j_2j_3j_4} \log f(\boldsymbol{z};\gamma) = &[12]\boldsymbol{z}^\intercal\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\Sigma_{j_4}\Sigma^{-1}\boldsymbol{z} \\ &- [12]\boldsymbol{z}^\intercal\Sigma^{-1}\Sigma_{j_1j_2}\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\Sigma_{j_4}\Sigma^{-1}\boldsymbol{z} \\ &- [6]\boldsymbol{z}^\intercal\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\Sigma_{j_2j_3}\Sigma^{-1}\Sigma_{j_4}\Sigma^{-1}\boldsymbol{z} + [4]\boldsymbol{z}^\intercal\Sigma^{-1}\Sigma_{j_1j_2j_3}\Sigma^{-1}\Sigma_{j_4}\Sigma^{-1} \\ &+ [3]\boldsymbol{z}^\intercal\Sigma^{-1}\Sigma_{j_1j_2}\Sigma^{-1}\Sigma_{j_3j_4}\Sigma^{-1} + \frac{1}{2}\boldsymbol{z}^\intercal\Sigma^{-1}\Sigma_{j_1j_2j_3j_4}\Sigma^{-1}\boldsymbol{z} \\ &+ [6]\frac{1}{2}\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\Sigma_{j_4}) \\ &- [12]\frac{1}{2}\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1j_2}\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\Sigma_{j_4}) + [3]\frac{1}{2}\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1j_2}\Sigma^{-1}\Sigma_{j_3j_4}) \\ &+ [4]\frac{1}{2}\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1j_2j_3}\Sigma^{-1}\Sigma_{j_4}) - \frac{1}{2}\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1j_2j_3j_4})\end{aligned} \tag{7.61}$$

Equations (7.54) – (7.57) involve the expectations over the distribution of $\boldsymbol{Z}|\boldsymbol{Y}$. Using the random sample from $f(\boldsymbol{z}|\boldsymbol{y};\beta,\gamma)$, (7.58) – (7.61) are computed for every simulation and then taking the average we obtain an approximation to (7.54) – (7.57).

The cumulants are defined as the expectations of the derivatives of the log-likelihood:

$$\kappa_{j_1j_2} = \mathbb{E}\{U_{j_1j_2}(\boldsymbol{Y})\}$$

$$\kappa_{j_1j_2j_3} = \mathbb{E}\{U_{j_1j_2j_3}(\boldsymbol{Y})\}$$

and so on. The expectations at the right hand side are over non-linear functions of $\boldsymbol{Y}$ and it is not possible to calculate the cumulants explicitly. An approximation is obtained as follows. Note that the left hand sides of (7.50) – (7.53) have expectation 0 with respect to the distribution

of $\boldsymbol{Y}$. Taking expectations, after some simplifications (see Appendix section A.4), we obtain

$$\begin{aligned}
\kappa_{j_1 j_2} &= \mathbb{E}(U_{j_1 j_2}) \\
&= -\frac{1}{4}\mathbb{E}\left[\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top|\boldsymbol{Y})\}\ \operatorname{tr}\{\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top|\boldsymbol{Y})\}\right] \\
&\quad + \frac{1}{4}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_1}\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_2}\} \qquad (7.62)
\end{aligned}$$

$$\begin{aligned}
\kappa_{j_1 j_2 j_3} &= \mathbb{E}(U_{j_1 j_2 j_3}) \\
&= [3]\frac{1}{2}\mathbb{E}\left[\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top|\boldsymbol{Y}))\operatorname{tr}(\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top|\boldsymbol{Y}))\right] \\
&\quad - [3]\frac{1}{4}\mathbb{E}\left[\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1 j_2}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top|\boldsymbol{Y}))\operatorname{tr}(\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top|\boldsymbol{Y}))\right] \\
&\quad - [3]\frac{1}{8}\mathbb{E}\left[\mathbb{E}\left\{(\boldsymbol{Z}^\top\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\boldsymbol{Z})(\boldsymbol{Z}^\top\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\boldsymbol{Z})|\boldsymbol{Y}\right\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top|\boldsymbol{Y})\}\right] \\
&\quad + \frac{1}{4}\mathbb{E}\big[\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top|\boldsymbol{Y})\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top|\boldsymbol{Y})\} \\
&\qquad\qquad \operatorname{tr}\{\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top|\boldsymbol{Y})\}\big] \\
&\quad - [3]\frac{1}{4}\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\Sigma_{j_2})\operatorname{tr}(\Sigma^{-1}\Sigma_{j_3}) \\
&\quad + [3]\frac{1}{4}\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1 j_2})\operatorname{tr}(\Sigma^{-1}\Sigma_{j_3}) \\
&\quad + \frac{1}{2}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_1}\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_2}\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_3}\} \qquad (7.63)
\end{aligned}$$

There is no direct way of obtaining explicitly (7.62) or (7.63) but the expectation $\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top|\boldsymbol{Y})$ can be approximated using the random sample from $f(\boldsymbol{z}|\boldsymbol{y};\beta,\gamma)$. A second option would be to average over the log-likelihood derivatives computed for each simulation.

7.4.2 Derivatives of the distribution function

The derivatives of the distribution function (7.21) as defined at (7.35) are needed. Write

$$\begin{aligned}
F(z_0|\boldsymbol{y};\gamma) &= \frac{\iint \Phi(z_0|\boldsymbol{z};\gamma)f(\boldsymbol{z};\gamma)f(\boldsymbol{y}|\boldsymbol{z};\beta)e^{u(\beta)}\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\beta}{\iint f(\boldsymbol{z};\gamma)f(\boldsymbol{y}|\boldsymbol{z};\beta)e^{u(\beta)}\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\beta} \\
&= \iint \Phi(z_0|\boldsymbol{z};\gamma)f(\beta,\boldsymbol{z}|\boldsymbol{y};\gamma)\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\beta \\
&= \int \Phi(z_0|\boldsymbol{z};\gamma)f(\boldsymbol{z}|\boldsymbol{y};\gamma)\,\mathrm{d}\boldsymbol{z} \qquad (7.64)
\end{aligned}$$

Then

$$
\begin{aligned}
F_{j_1}(z_0|\boldsymbol{y};\gamma) &= \frac{\iint \Phi_{j_1}(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{z};\gamma) f(\boldsymbol{y}|\boldsymbol{z};\beta) e^{u(\beta)} \,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\beta}{\iint f(\boldsymbol{z};\gamma) f(\boldsymbol{y}|\boldsymbol{z};\beta) e^{u(\beta)} \,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\beta} \\
&\quad + \frac{\iint \Phi(z_0|\boldsymbol{z};\gamma) f_{j_1}(\boldsymbol{z};\gamma) f(\boldsymbol{y}|\boldsymbol{z};\beta) e^{u(\beta)} \,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\beta}{\iint f(\boldsymbol{z};\gamma) f(\boldsymbol{y}|\boldsymbol{z};\beta) e^{u(\beta)} \,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\beta} \\
&\quad - F(z_0|\boldsymbol{y};\gamma) \frac{\iint f_{j_1}(\boldsymbol{z};\gamma) f(\boldsymbol{y}|\boldsymbol{z};\beta) e^{u(\beta)} \,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\beta}{\iint f(\boldsymbol{z};\gamma) f(\boldsymbol{y}|\boldsymbol{z};\beta) e^{u(\beta)} \,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\beta} \\
&= \int \Phi_{j_1}(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{z}|\boldsymbol{y};\gamma) \,\mathrm{d}\boldsymbol{z} \\
&\quad + \int \Phi(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{z}|\boldsymbol{y};\gamma) \{\partial_{j_1} \log f(\boldsymbol{z};\gamma)\} \,\mathrm{d}\boldsymbol{z} \\
&\quad - F(z_0|\boldsymbol{y};\gamma) \int f(\boldsymbol{z}|\boldsymbol{y};\gamma) \{\partial_{j_1} \log f(\boldsymbol{z};\gamma)\} \,\mathrm{d}\boldsymbol{z} \qquad (7.65) \\
F_{j_1 j_2}(z_0|\boldsymbol{y};\gamma) &= \int \Phi_{j_1 j_2}(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{z}|\boldsymbol{y};\gamma) \,\mathrm{d}\boldsymbol{z} \\
&\quad + [2] \int \Phi_{j_1}(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{z}|\boldsymbol{y};\gamma) \{\partial_{j_2} \log f(\boldsymbol{z};\gamma)\} \,\mathrm{d}\boldsymbol{z} \\
&\quad + \int \Phi(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{z}|\boldsymbol{y};\gamma) \{\partial^2_{j_1 j_2} \log f(\boldsymbol{z};\gamma)\} \,\mathrm{d}\boldsymbol{z} \\
&\quad + \int \Phi(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{z}|\boldsymbol{y};\gamma) \{\partial_{j_1} \log f(\boldsymbol{z};\gamma)\} \{\partial_{j_2} \log f(\boldsymbol{z};\gamma)\} \,\mathrm{d}\boldsymbol{z} \\
&\quad - [2] F_{j_1}(z_0|\boldsymbol{y};\gamma) \int f(\boldsymbol{z}|\boldsymbol{y};\gamma) \{\partial_{j_2} \log f(\boldsymbol{z};\gamma)\} \,\mathrm{d}\boldsymbol{z} \\
&\quad - F(z_0|\boldsymbol{y};\gamma) \int f(\boldsymbol{z}|\boldsymbol{y};\gamma) \{\partial^2_{j_1 j_2} \log f(\boldsymbol{z};\gamma)\} \,\mathrm{d}\boldsymbol{z} \\
&\quad - F(z_0|\boldsymbol{y};\gamma) \int f(\boldsymbol{z}|\boldsymbol{y};\gamma) \{\partial_{j_1} \log f(\boldsymbol{z};\gamma)\} \{\partial_{j_2} \log f(\boldsymbol{z};\gamma)\} \,\mathrm{d}\boldsymbol{z} \qquad (7.66) \\
F_{j_1 j_2 j_3} &= \int \Phi_{j_1 j_2 j_3}(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{z}|\boldsymbol{y};\gamma) \,\mathrm{d}\boldsymbol{z} + [3] \int \Phi_{j_1 j_2}(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{z}|\boldsymbol{y};\gamma) \{\partial_{j_3} \log f(z;\gamma)\} \,\mathrm{d}\boldsymbol{z} \\
&\quad + [3] \int \Phi_{j_1}(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{z}|\boldsymbol{y};\gamma) \{\partial^2_{j_2 j_3} \log f(z;\gamma)\} \,\mathrm{d}\boldsymbol{z} \\
&\quad + [3] \int \Phi_{j_1}(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{z}|\boldsymbol{y};\gamma) \{\partial_{j_2} \log f(z;\gamma)\} \{\partial_{j_3} \log f(z;\gamma)\} \,\mathrm{d}\boldsymbol{z} \\
&\quad + \int \Phi(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{z}|\boldsymbol{y};\gamma) \{\partial^3_{j_1 j_2 j_3} \log f(z;\gamma)\} \,\mathrm{d}\boldsymbol{z} \\
&\quad + [3] \int \Phi(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{z}|\boldsymbol{y};\gamma) \{\partial^2_{j_1 j_2} \log f(z;\gamma)\} \{\partial_{j_3} \log f(z;\gamma)\} \,\mathrm{d}\boldsymbol{z} \\
&\quad + \int \Phi(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{z}|\boldsymbol{y};\gamma) \{\partial_{j_1} \log f(z;\gamma)\} \{\partial_{j_2} \log f(z;\gamma)\} \{\partial_{j_3} \log f(z;\gamma)\} \,\mathrm{d}\boldsymbol{z} \\
&\quad - F(z_0|\boldsymbol{y};\gamma) \int f(\boldsymbol{z}|\boldsymbol{y};\gamma) \{\partial^3_{j_1 j_2 j_3} \log f(z;\gamma)\} \,\mathrm{d}\boldsymbol{z} \\
&\quad - [3] F(z_0|\boldsymbol{y};\gamma) \int f(\boldsymbol{z}|\boldsymbol{y};\gamma) \{\partial^2_{j_1 j_2} \log f(z;\gamma)\} \{\partial_{j_3} \log f(z;\gamma)\} \,\mathrm{d}\boldsymbol{z}
\end{aligned}
$$

$$
\begin{aligned}
&- F(z_0|\boldsymbol{y};\gamma) \int f(\boldsymbol{z}|\boldsymbol{y};\gamma)\{\partial_{j_1} \log f(z;\gamma)\}\{\partial_{j_2} \log f(z;\gamma)\}\{\partial_{j_3} \log f(z;\gamma)\}\, \mathrm{d}\boldsymbol{z} \\
&- [3]F_{j_1}(z_0|\boldsymbol{y};\gamma) \int f(\boldsymbol{z}|\boldsymbol{y};\gamma)\{\partial^2_{j_2 j_3} \log f(z;\gamma)\}\, \mathrm{d}\boldsymbol{z} \\
&- [3]F_{j_1}(z_0|\boldsymbol{y};\gamma) \int f(\boldsymbol{z}|\boldsymbol{y};\gamma)\{\partial_{j_2} \log f(z;\gamma)\}\{\partial_{j_3} \log f(z;\gamma)\}\, \mathrm{d}\boldsymbol{z} \\
&- [3]F_{j_1 j_2}(z_0|\boldsymbol{y};\gamma) \int f(\boldsymbol{z}|\boldsymbol{y};\gamma)\{\partial_{j_3} \log f(z;\gamma)\}\, \mathrm{d}\boldsymbol{z} \qquad (7.67)
\end{aligned}
$$

where the derivatives of $\log f(\boldsymbol{z};\gamma)$ are given in (7.58) – (7.60) and the derivatives of Φ in (7.15). These expressions can also be computed by simulation.

7.4.3 Simulations

Nine predictive densities were considered. Eight of them were Bayesian predictive densities under different priors for the three parameters of the covariance matrix (see Table 7.3). The literature didn't give much focus in the nugget parameter and the only proposal we found was the use of a uniform prior. The first set of priors uses uniform bounded priors for all three parameters as in Diggle et al. (1998). Christensen et al. (2000) propose the use of improper inverse gamma for the partial sill and uniform or exponential for the range parameter. Following their suggestion, the second set of priors assigns uniform priors to the nugget and range parameters but improper inverse gamma for the partial sill. The third, fourth, and fifth set of priors assigns uniform prior to the nugget, improper inverse gamma to the partial sill and exponential to the range parameter with corresponding means 2, 4, and 6. For the Gaussian model Berger et al. (2001) suggested three other priors for the partial sill and range parameters, the reference prior, the Jeffreys independent prior and the Jeffreys rule prior (see section A.5 in the Appendix). These three sets of priors were also considered. We also considered the coverage probability bias for the plug-in predictive density as proposed by Zhang (2002) and also in Chapter 5 of this thesis.

Coverage probability bias

We perform simulations to compute the coverage probability bias under the setting of section 4.3.3. Using the locations in Figure 4.1, we simulate 1000 realizations of a Gaussian random field as described in section 4.3.3. Then we repeatedly draw $n = 60$ observations from

prior	nugget	partial sill	range
1	uniform	uniform	uniform
2	uniform	inverse gamma	uniform
3	uniform	inverse gamma	exponential w/ mean 2
4	uniform	inverse gamma	exponential w/ mean 4
5	uniform	inverse gamma	exponential w/ mean 6
6	uniform	Gaussian reference prior	
7	uniform	Gaussian Jeffreys indep. prior	
8	uniform	Gaussian Jeffreys rule prior	

Table 7.3: Priors used for the simulations.

the Bernoulli distribution. We consider prediction at location $(5, 5)$, the center of Figure 4.1. We computed the bias at $\gamma = (0.2, 2.0, 4.0)$ for the 2.5%, 5%, 50%, 95% and 97.5% quantiles.

For each simulation, using the `R` package `geoRglm` we drew an MCMC sample of size 1000 from the distribution of $Z_0|Y$ where Z_0 corresponds to the random effect at location $(5, 5)$. Equations (7.54) – (7.57) were computed by averaging (7.58) – (7.61) over the MCMC sample which in turn were used to obtain the log-likelihood derivatives using (7.50) – (7.53). In addition (7.64) – (7.67) were also computed by averaging.

The cumulants were calculated by averaging the log-likelihood derivatives over each simulation. All the quantities needed to obtain the coverage probability bias given in (7.36) have been computed. The mean, median and standard deviation over the 1000 simulations for the nine different predictive densities are shown in Table 7.4.

The plug-in method seems to have the smallest bias. We see that the second set of priors (inverse gamma for the partial sill and uniform for the other two) tends to have lower bias at the tails of the distribution which is the most significant when a prediction interval is constructed. Also the first set of priors (all uniform) has a small bias. On the other hand, the standard deviation is too large to allow for a clear answer as to which prior should be preferred. It's also interesting to see in how many cases over the 1000 simulations there are such that the absolute bias of the predictive density under one prior is less than the one under a different prior. This comparison is presented in Table 7.5. According to this table and in combination with Table 7.4, the second set of priors results to a predictive distribution with smaller coverage probability bias among the Bayesian predictive densities. The plug-in method also results to low coverage probability bias.

A chi-square goodness-of-fit test was performed testing the hypothesis that each of the

quantile		1	2	3	4	5
2.50%	Mean	1.137E-03	1.077E-03	1.393E-03	1.235E-03	1.182E-03
	Median	1.254E-03	1.211E-03	1.433E-03	1.319E-03	1.290E-03
	S.D.	5.074E-05	5.068E-05	5.164E-05	5.099E-05	5.085E-05
5%	Mean	1.759E-03	1.675E-03	2.121E-03	1.898E-03	1.824E-03
	Median	1.867E-03	1.784E-03	2.069E-03	1.946E-03	1.881E-03
	S.D.	7.454E-05	7.429E-05	7.696E-05	7.531E-05	7.490E-05
50%	Mean	-2.342E-06	-3.648E-08	-2.226E-05	-1.115E-05	-7.445E-06
	Median	8.091E-05	6.797E-05	3.234E-05	2.081E-05	1.502E-05
	S.D.	1.449E-04	1.415E-04	1.683E-04	1.537E-04	1.493E-04
95%	Mean	-1.749E-03	-1.664E-03	-2.123E-03	-1.893E-03	-1.817E-03
	Median	-1.855E-03	-1.813E-03	-2.104E-03	-1.946E-03	-1.912E-03
	S.D.	7.996E-05	7.967E-05	8.255E-05	8.082E-05	8.037E-05
97.50%	Mean	-1.127E-03	-1.067E-03	-1.391E-03	-1.229E-03	-1.175E-03
	Median	-1.264E-03	-1.193E-03	-1.432E-03	-1.315E-03	-1.290E-03
	S.D.	5.408E-05	5.399E-05	5.514E-05	5.441E-05	5.423E-05

quantile		6	7	8	plug-in	
2.50%	Mean	1.379E-03	1.306E-03	1.313E-03	-1.819E-05	
	Median	1.425E-03	1.380E-03	1.386E-03	1.338E-04	
	S.D.	5.157E-05	5.124E-05	5.131E-05	5.066E-05	
5%	Mean	2.101E-03	1.998E-03	2.007E-03	-2.190E-05	
	Median	2.047E-03	1.989E-03	1.990E-03	1.710E-04	
	S.D.	7.679E-05	7.598E-05	7.616E-05	7.410E-05	
50%	Mean	-2.130E-05	-1.616E-05	-1.849E-05	2.028E-05	
	Median	6.810E-06	4.291E-05	5.263E-05	1.633E-05	
	S.D.	1.669E-04	1.600E-04	1.617E-04	1.285E-04	
95%	Mean	-2.103E-03	-1.997E-03	-2.007E-03	3.968E-05	
	Median	-2.098E-03	-2.024E-03	-2.028E-03	-1.652E-04	
	S.D.	8.238E-05	8.153E-05	8.172E-05	7.891E-05	
97.50%	Mean	-1.377E-03	-1.302E-03	-1.310E-03	3.035E-05	
	Median	-1.424E-03	-1.378E-03	-1.388E-03	-1.193E-04	
	S.D.	5.506E-05	5.470E-05	5.478E-05	5.373E-05	

Table 7.4: Approximate coverage probability bias calculated for the eight different sets of priors and the plug-in method.

proportions in Table 7.5 is different from 0.5. If $\hat{p}$ is the observed proportion from Table 7.5 and N is the number of simulations, the test statistic for testing $p \neq 0.5$ is given by

$$T = 2\,N\,(\hat{p} - 0.5)^2 + 2\,N\,(1 - \hat{p} - 0.5)^2 = 4\,N\,(\hat{p} - 0.5)^2$$

which has, approximately, the chi square distribution with 1 degree of freedom. The null hypothesis is rejected for every pair of Table 7.5 which suggests the use of inverse gamma prior for the partial sill and uniform priors for the other two parameters.

The same computations were performed for different values of the range parameter but for fixed nugget and partial sill parameter. The values of the range parameter were $2, 3, \ldots, 8$ while

the nugget was fixed at 0.2 and the partial sill at 2.0. We plotted the estimate of the coverage probability bias with respect to the value of the range parameter (Figures 7.1–7.2). The plug-in has clearly the smallest coverage probability bias. Among the Bayesian predictive densities, the one that seems to have the lowest bias is the second (inverse gamma for the partial sill and uniform for the other two parameters). The decreasing pattern can be explained by the fact that for higher values of the range, more locations can be used for the prediction at a certain location, hence the bias is smaller. Also notice, that the third set of priors (exponential with mean 2 for the range) had good performance when the true value is less than or equal to 4 but separates from the others when the true value is higher.

The coverage probability bias was also computed for different values of the partial sill parameter: $0.15, 0.2, 0.35$ and $0.5, 1, 1.5, \ldots, 5$, while the nugget was fixed at 0.2 and the range was fixed at 4.0. Figures 7.3 – 7.4 show how the coverage probability bias changes. There is an increase for the values of the sill that are smaller than 0.5, then the bias is decreasing towards 0 while the sill becomes larger but starts increasing again when the sill is larger than 3. We actually expect the bias to decrease as the sill increases, as is the case for the values between 0.5 and 3.0, because the signal-to-noise ratio is increased. Perhaps the false pattern for very small and very large values of the sill is due to instability of the simulations. A third set of simulations was performed with the nugget varying at values $\frac{0.4}{0.15}, \frac{0.4}{0.2}, \frac{0.4}{0.35}$ and $\frac{0.4}{0.5}, \frac{0.4}{1}, \frac{0.4}{1.5}, \ldots, \frac{0.4}{5}$ with range and partial sill being fixed at values 4 and 2 respectively. The plots are shown in Figures 7.5 – 7.6. Besides the low values for the nugget, the bias seems to increase as the value of the nugget increases which can be explained by the fact that when the nugget is high, the variability for each observation is higher introducing more uncertainty and measurement error.

Alternative computation of the coverage probability bias by simulation

Empirically, the coverage probability bias can be computed by simulating from the distributions of $Z_0|\boldsymbol{Y}$ and $Z_0|\boldsymbol{Y};\gamma$ by noting that

$$f(z_0|\boldsymbol{y}) = \iint f(z_0|\boldsymbol{z},\gamma) f(\boldsymbol{y}|\boldsymbol{z};\beta) f(\gamma)\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\gamma$$

and

$$f(z_0|\boldsymbol{y};\gamma) = \int f(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{y}|\boldsymbol{z};\beta)\,\mathrm{d}\boldsymbol{z}$$

and by using the following scheme:

For a given vector of observations $\boldsymbol{Y}$ the predictive quantiles of the Bayesian predictive density are obtained as follows

1. Replicate the following steps:
   (a) Simulate from the distribution of $\gamma|\boldsymbol{Z},\boldsymbol{Y}$
   (b) Simulate from the distribution of $\boldsymbol{Z}|\boldsymbol{Y},\gamma$
2. Estimate γ by $\hat{\gamma}$ and predict $\boldsymbol{Z}$ by $\hat{\boldsymbol{Z}}$ by averaging over the simulations.
3. Construct the distribution of $Z_0|\hat{\boldsymbol{Z}};\hat{\gamma}$ and compute its quantiles

For the predictive density for fixed covariance parameters, say $\gamma = \gamma_{\mathsf{T}}$, the coverage probability of the quantiles obtained above is estimated in the following way

1. Replicate the following:
   (a) Simulate from the distribution of $\boldsymbol{Z}|\boldsymbol{Y};\gamma_{\mathsf{T}}$
2. Predict $\boldsymbol{Z}$ by $\tilde{\boldsymbol{Z}}$ by averaging over the simulations.
3. Construct the distribution of $Z_0|\tilde{\boldsymbol{Z}};\gamma_{\mathsf{T}}$
4. Compute the coverage probability for the given quantiles.

The simulations at step 1. were replicated 1000 times and the whole procedure was repeated 1000 times. The median of the coverage probability bias was computed under the first five sets of priors of Table 7.3 and for the plug-in method (we chose the median because the distribution of the coverage probability bias computed this way was not symmetric and had high variability). According to the simulations the coverage probability bias is as shown in Table 7.6. The simulations seem to agree with our approximation (within simulation error) in terms of the value of the coverage probability bias. The comparison between the simulations and the approximation can be easily seen in Figures 7.7 and 7.8 where in most cases there is an agreement between

the two but the approximation has much smaller variability. A different comparison is shown in Figures 7.9 – 7.11 where the approximation and the simulation methods are plotted against different sill values. Because of the high variability of the simulations, the interquartile range is too wide but we can see that our approximation falls within the two quartiles.

We also counted how many times one predictive density has smaller coverage probability bias than another (in absolute value) and tested if there is any significant advantage using one prior over another. Table 7.7 shows the observed proportions. The simulations recommend the first and second set of priors in agreement with our approximation.

Kullback-Leibler divergence

For the computation of the Kullback-Leibler divergence, we used a larger sample size to reduce the error of the approximation given in (7.47). We used $k = 100$ randomly chosen locations from where we drew $n = 100$ observations at each location and calculated (7.47) for the given simulation under the priors in Table 7.3. We repeated this 2000 times and the average was taken. The distribution of the Kullback-Leibler divergence was not symmetric. Most of the values were less than 10 with first quartile around 0.3 and third quartile around 8. We excluded any values larger than 15 and were left with 1122 simulated values of the Kullback-Leibler divergence. The average Kullback-Leibler divergence over the simulations for each set of priors is shown in Table 7.8

It is not very clear from Table 7.8 if any predictive density has an advantage but the one constructed using the second set of priors seem to be the best. The proportions of how many times one predictive density had smaller Kullback-Leibler divergence than another density were computed and displayed in Table 7.9.

A similar chi-square goodness-of-fit test was performed for the proportions in Table 7.9 testing if they are different from 0.5. The null hypothesis is rejected for every pair of table Table 7.9. This is another evidence that the predictive density constructed using the second set of priors has the smallest Kullback-Leibler divergence.

7.5 Summary

The main results of this chapter are the approximations to the coverage probability bias and Kullback-Leibler divergence of the Bayesian predictive distribution. In order to obtain these approximations we first derive expressions for the approximation of the Bayesian predictive distribution and the Bayesian probability density function. Our result also allows us to approximate the quantiles of the Bayesian predictive distribution under a given prior.

Our approximation to the coverage probability bias was compared against the bias computed using simulations. Due to the high variability of the simulated coverage probability bias, it is hard to provide a golden standard to the exact value of the coverage probability bias, but out approximation seem to agree (within error) with the simulated one.

The computations were performed for different values of the parameters. We find that the bias is reduced as the range parameter increases which can be explained by the fact that when the range is larger, more locations can be used for the prediction at a given location, which reduces the bias. In addition, when the sill parameter is increased while the nugget remains constant, the signal-to-noise ratio is increased and the bias of the prediction should be reduced. This property is not completely captured by our approximation (except for the range 0.5 – 2.5) but neither by the simulated bias. A similar reasoning can be applied when the nugget parameter is increased while the sill and range parameters remain constant. In this case, the signal-to-noise ratio is reduced and the bias of the prediction should be increased. This feature is captured by our approximation except for very low values of the nugget parameter, i.e., those lower than 0.2.

Based on our approximation to the coverage probability bias in Tables 7.4 and 7.5, the exponential prior is not a good choice for the range parameter. This result is also verified by our approximation to the Kullback-Leibler divergence in Table 7.9. For the sill parameter, either uniform or inverse gamma is a good choice. In addition, priors proposed for the Gaussian Spatial model don't perform very well for the Generalized Linear Mixed Model.

	2	3	4	5	6	7	8	plug-in
1	0.256	0.737	0.711	0.683	0.736	0.728	0.717	0.322
	0.260	0.716	0.690	0.677	0.716	0.706	0.697	0.312
	0.307	0.764	0.738	0.730	0.762	0.751	0.754	0.347
	0.264	0.740	0.719	0.691	0.738	0.732	0.724	0.316
	0.254	0.742	0.721	0.704	0.741	0.731	0.729	0.328
2		0.742	0.728	0.722	0.741	0.736	0.728	0.327
		0.723	0.712	0.711	0.721	0.712	0.703	0.315
		0.753	0.723	0.716	0.750	0.738	0.744	0.363
		0.742	0.723	0.717	0.742	0.736	0.732	0.319
		0.746	0.728	0.719	0.745	0.736	0.730	0.343
3			0.250	0.253	0.238	0.245	0.228	0.300
			0.268	0.270	0.257	0.263	0.238	0.282
			0.227	0.235	0.206	0.222	0.235	0.298
			0.244	0.252	0.228	0.237	0.220	0.303
			0.243	0.247	0.229	0.232	0.214	0.314
4				0.263	0.748	0.747	0.720	0.313
				0.282	0.731	0.728	0.703	0.304
				0.254	0.772	0.762	0.766	0.335
				0.261	0.753	0.747	0.732	0.313
				0.260	0.756	0.748	0.735	0.326
5					0.747	0.744	0.727	0.317
					0.729	0.724	0.707	0.308
					0.762	0.758	0.762	0.340
					0.748	0.742	0.734	0.315
					0.751	0.746	0.737	0.327
6						0.246	0.223	0.300
						0.264	0.235	0.287
						0.222	0.241	0.299
						0.240	0.220	0.305
						0.233	0.205	0.314
7							0.595	0.310
							0.591	0.296
							0.766	0.321
							0.611	0.305
							0.611	0.319
8								0.309
								0.296
								0.319
								0.305
								0.320

Table 7.5: Comparison of the absolute coverage probability bias. The numbers in each cell show what proportion over 1000 simulations, the set of priors in the corresponding row had a smaller bias compared to the prior in the corresponding column for the different quantiles. Within each cell, the quantiles compared were, from top to bottom, 2.5%, 5%, 50%, 95%, and 97.5%

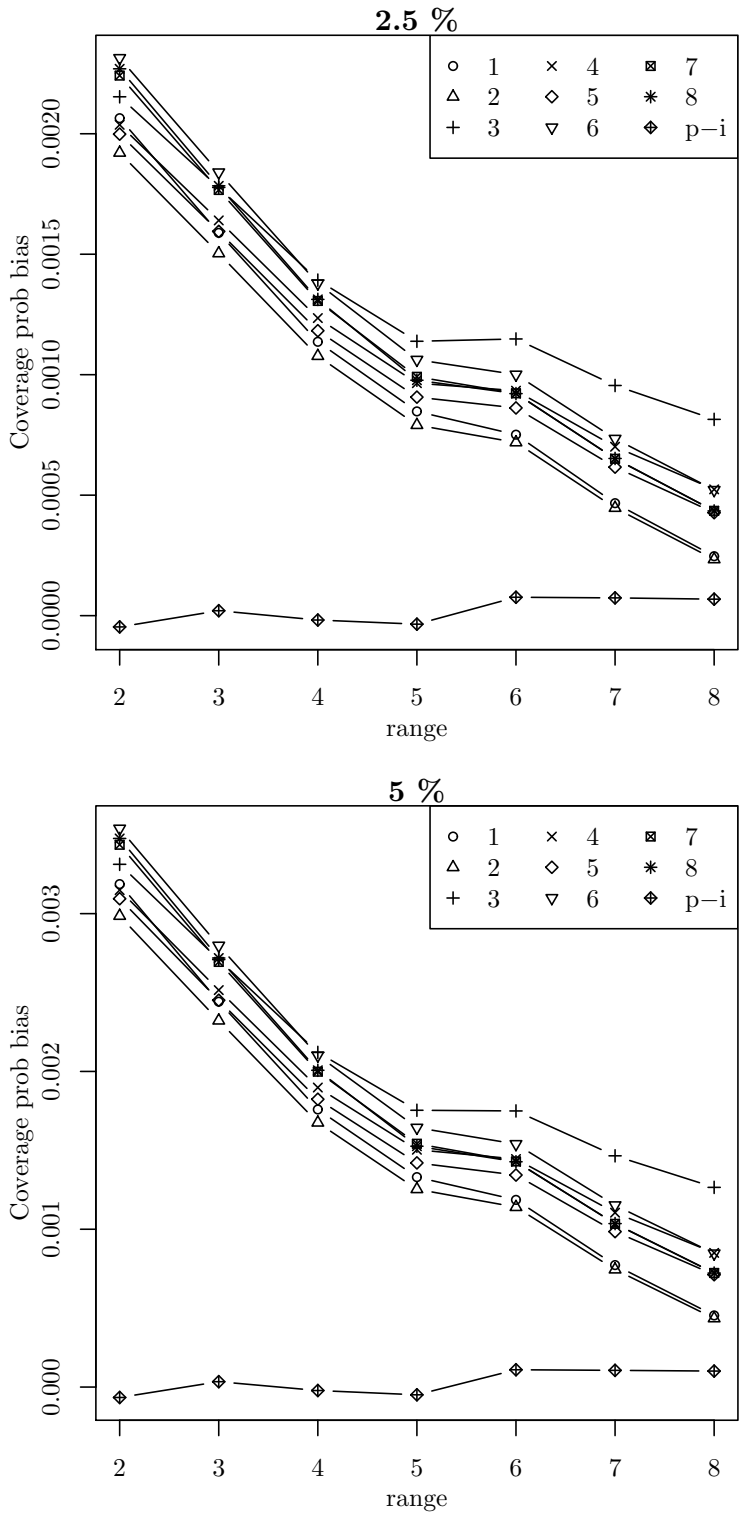

Figure 7.1: Coverage probability bias for eight different priors and plug-in at the 2.5% and 5% quantiles with range varying.

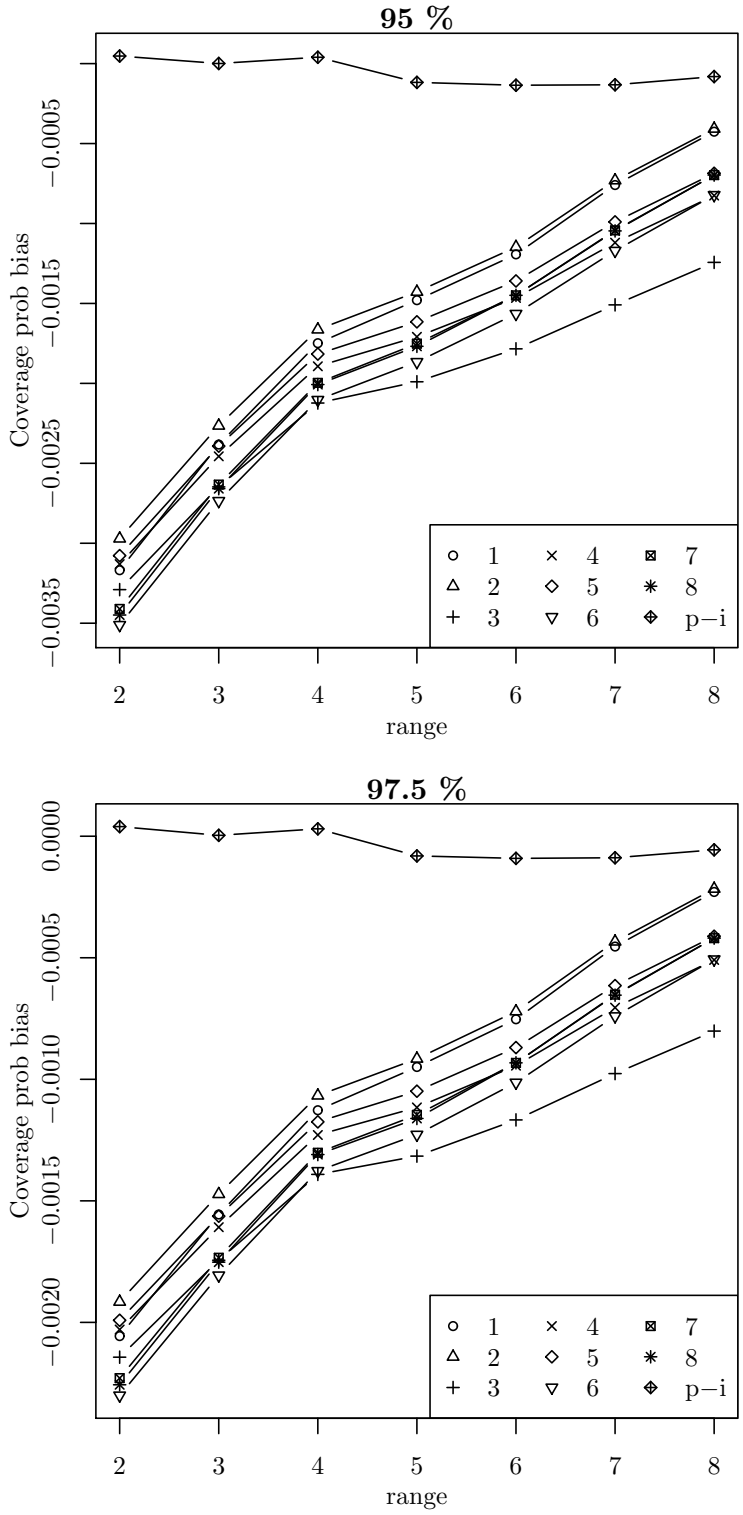

Figure 7.2: Coverage probability bias for eight different priors and plug-in at the 95% and 97.5% quantiles with range varying.

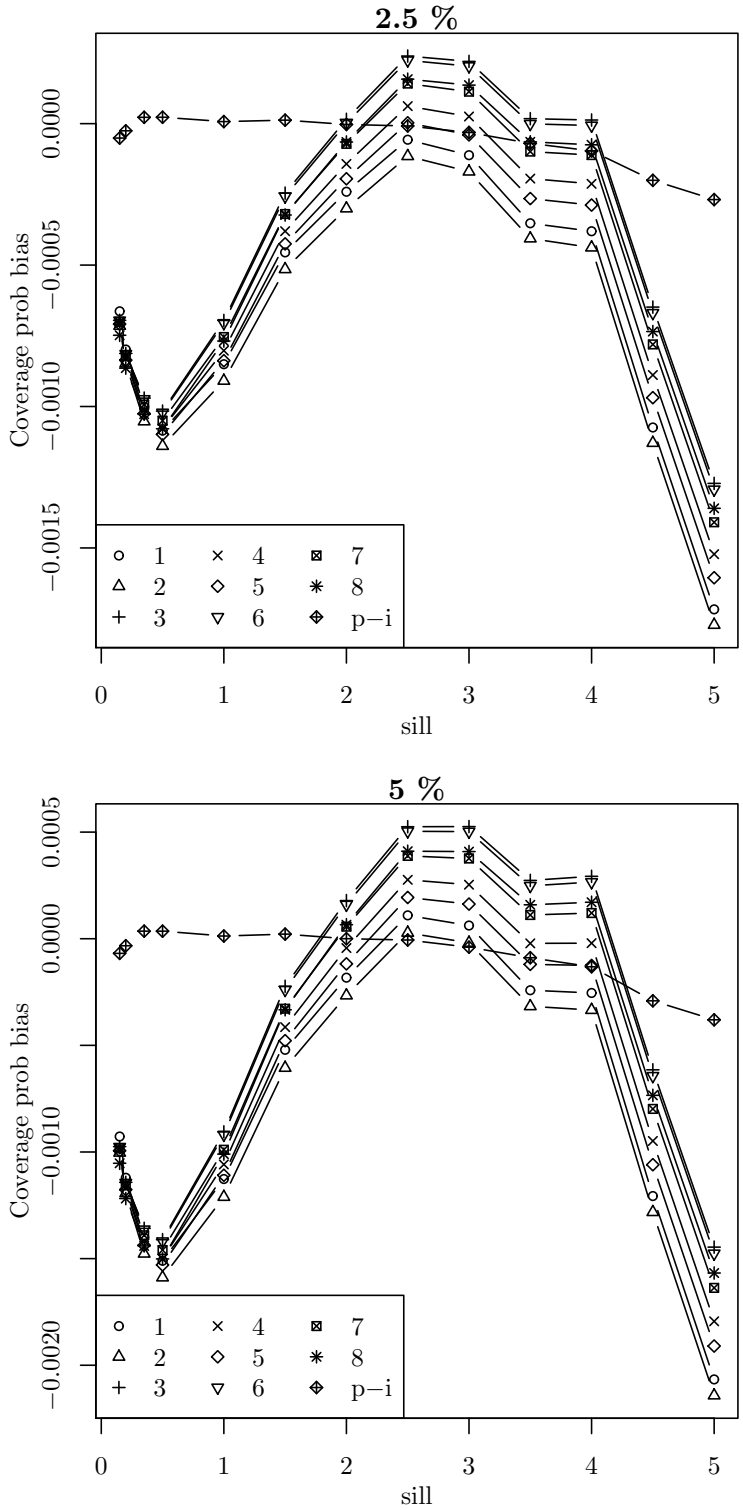

Figure 7.3: Coverage probability bias for eight different priors and plug-in at the 2.5% and 5% quantiles with sill varying.

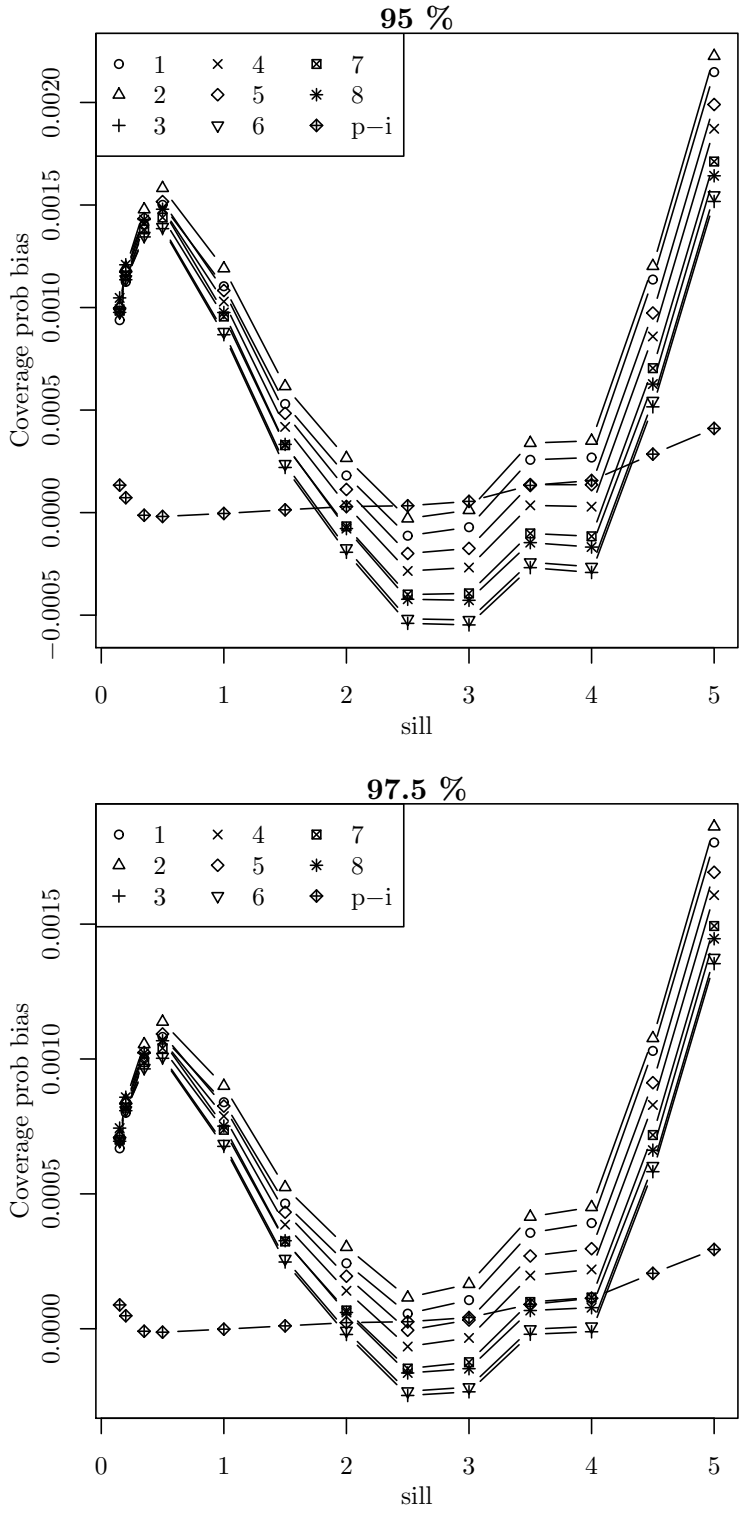

Figure 7.4: Coverage probability bias for eight different priors and plug-in at the 95% and 97.5% quantiles with sill varying.

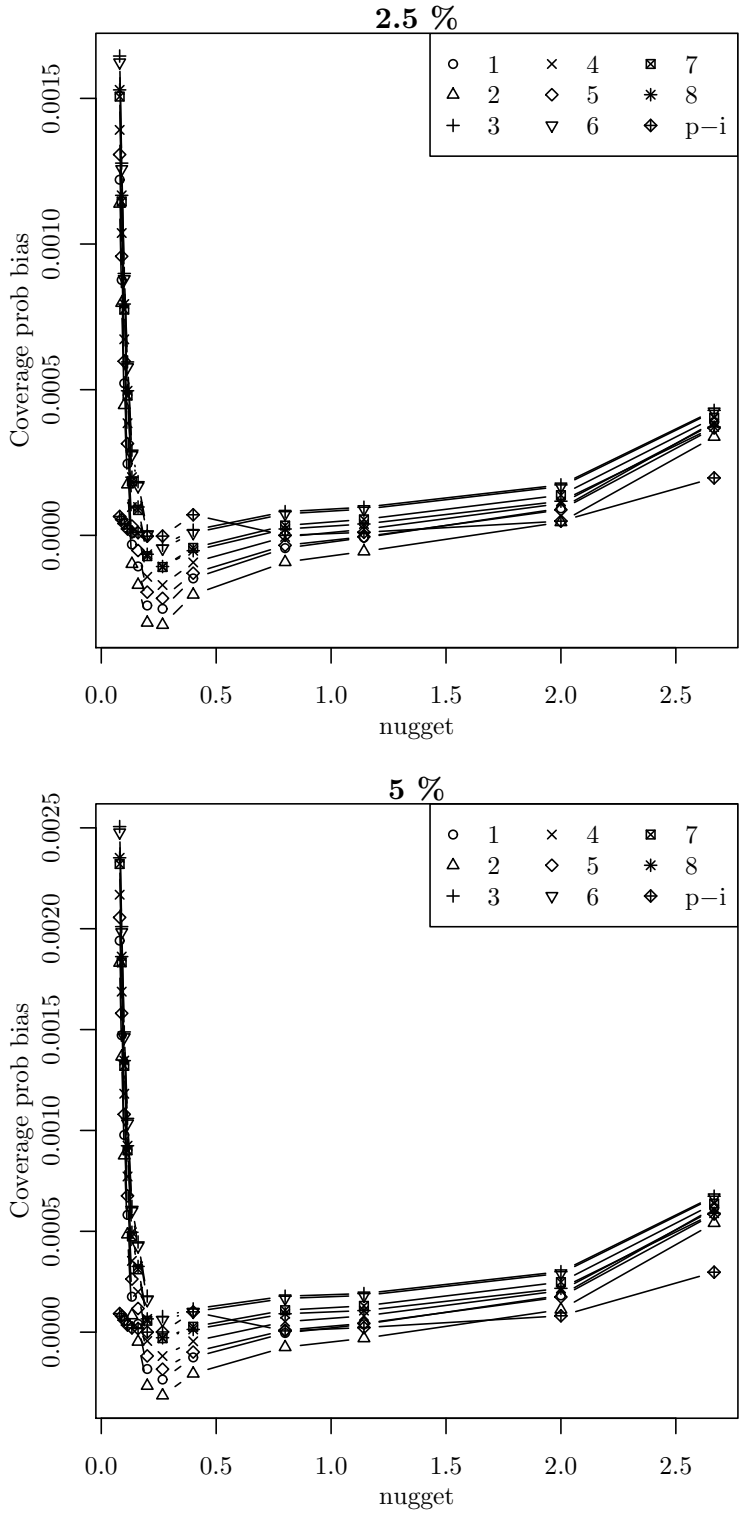

Figure 7.5: Coverage probability bias for eight different priors and plug-in at the 2.5% and 5% quantiles with nugget varying.

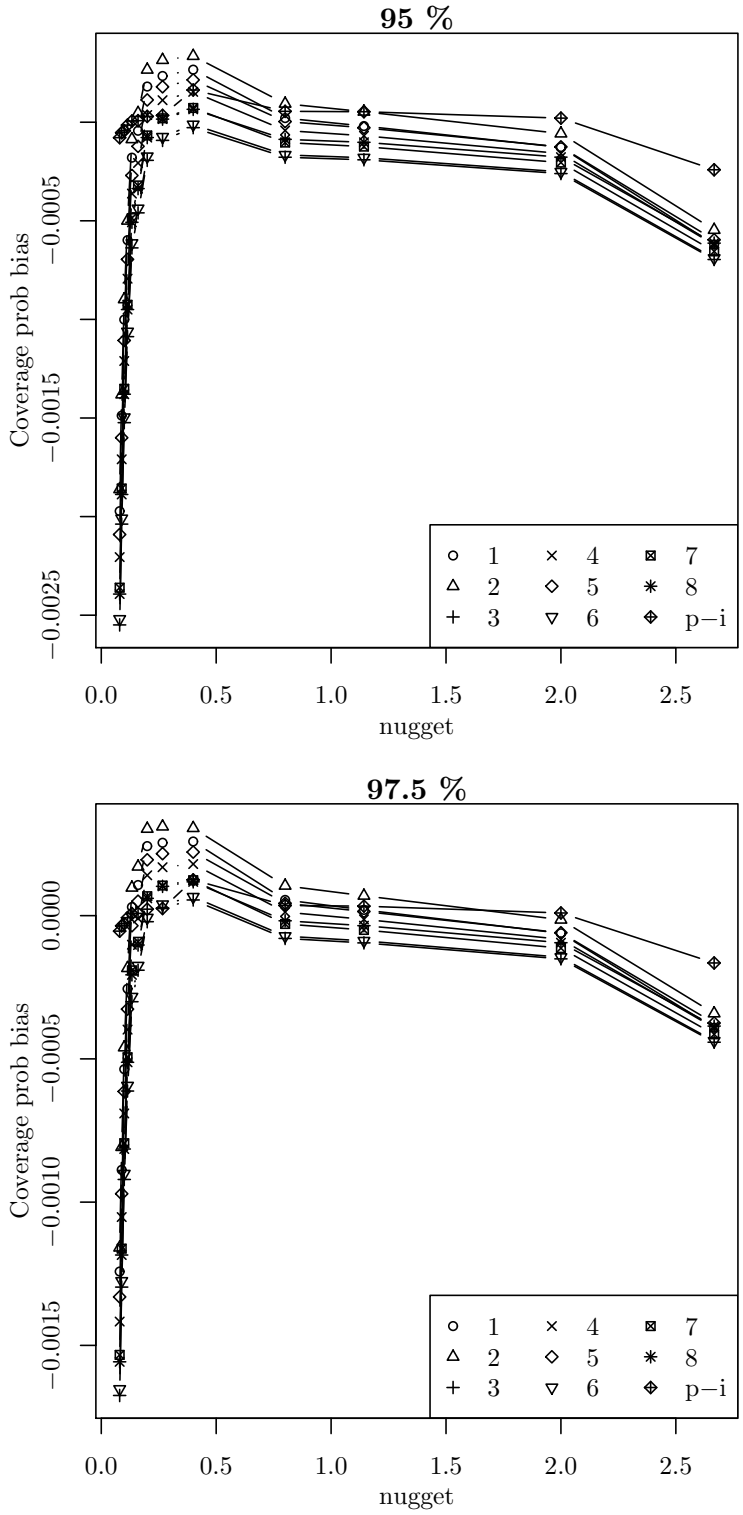

Figure 7.6: Coverage probability bias for eight different priors and plug-in at the 95% and 97.5% quantiles with nugget varying.

quantile		1	2	3	4	5	Plug-in
2.50%	Mean	4.630E-05	5.097E-04	-4.116E-03	2.782E-05	2.982E-03	2.593E-03
	Median	-1.049E-04	-1.056E-04	-4.030E-03	-1.252E-04	1.847E-03	3.687E-05
	S.D.	2.083E-04	2.113E-04	2.558E-04	2.146E-04	2.479E-04	1.420E-03
5%	Mean	1.773E-05	7.185E-04	-6.283E-03	2.567E-05	4.408E-03	2.717E-03
	Median	-1.847E-04	-2.857E-04	-5.795E-03	-2.347E-04	2.806E-03	9.803E-07
	S.D.	3.589E-04	3.570E-04	4.670E-04	3.656E-04	4.104E-04	1.424E-03
50%	Mean	-2.274E-04	3.124E-04	8.517E-04	8.460E-04	2.270E-03	4.385E-04
	Median	-1.398E-04	-1.726E-03	1.067E-03	1.455E-03	1.367E-03	2.026E-04
	S.D.	1.355E-03	1.239E-03	2.009E-03	1.325E-03	1.333E-03	1.188E-03
95%	Mean	-2.388E-04	-6.527E-04	6.615E-03	3.126E-04	-3.232E-03	-9.206E-04
	Median	4.124E-04	-5.394E-04	5.828E-03	7.826E-04	-1.985E-03	3.896E-05
	S.D.	3.970E-04	3.522E-04	4.804E-04	3.854E-04	3.932E-04	3.374E-04
97.50%	Mean	-1.955E-04	-4.960E-04	4.282E-03	1.410E-04	-2.317E-03	-6.737E-04
	Median	2.542E-04	-3.548E-04	3.908E-03	5.917E-04	-1.404E-03	3.382E-05
	S.D.	2.337E-04	2.087E-04	2.638E-04	2.281E-04	2.371E-04	2.173E-04

Table 7.6: Coverage probability bias computed by simulations.

	2	3	4	5	plugin
1	0.469	0.696	0.514	0.579	0.271
	0.475	0.683	0.516	0.567	0.277
	0.488	0.675	0.520	0.544	0.267
	0.483	0.695	0.513	0.566	0.274
	0.493	0.703	0.515	0.576	0.261
2		0.703	0.532	0.595	0.288
		0.690	0.528	0.592	0.299
		0.681	0.543	0.561	0.302
		0.690	0.521	0.558	0.279
		0.696	0.528	0.579	0.277
3			0.301	0.379	0.165
			0.307	0.395	0.167
			0.347	0.376	0.175
			0.308	0.375	0.168
			0.308	0.366	0.161
4				0.544	0.274
				0.552	0.269
				0.534	0.270
				0.566	0.273
				0.549	0.258
5					0.261
					0.266
					0.273
					0.256
					0.268

Table 7.7: Comparison of the absolute coverage probability bias by simulations.

	1	2	3	4	5	6	7	8
Mean	2.2659	2.2593	2.2845	2.2719	2.2677	4.0462	2.2911	2.2779
S.D.	0.0770	0.0770	0.0771	0.0771	0.0771	0.1152	0.0770	0.0772

Table 7.8: Mean and standard deviation of the Kullback-Leibler divergence for the simulations.

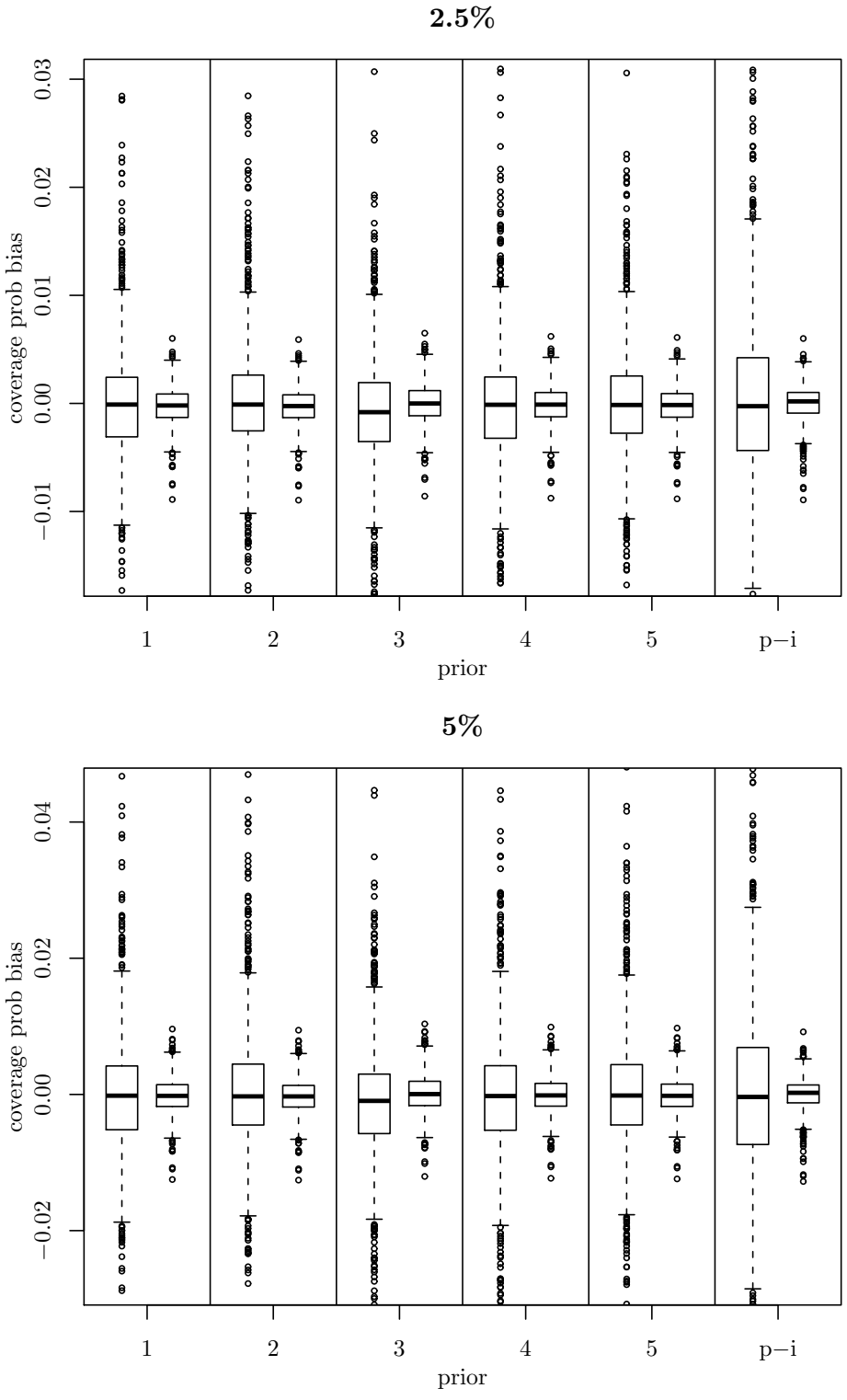

Figure 7.7: Boxplots to compare the coverage probability bias by simulation (left) and by approximation (right) for different priors and plug-in. The top corresponds to the 2.5% quantile and the bottom to the 5% quantile

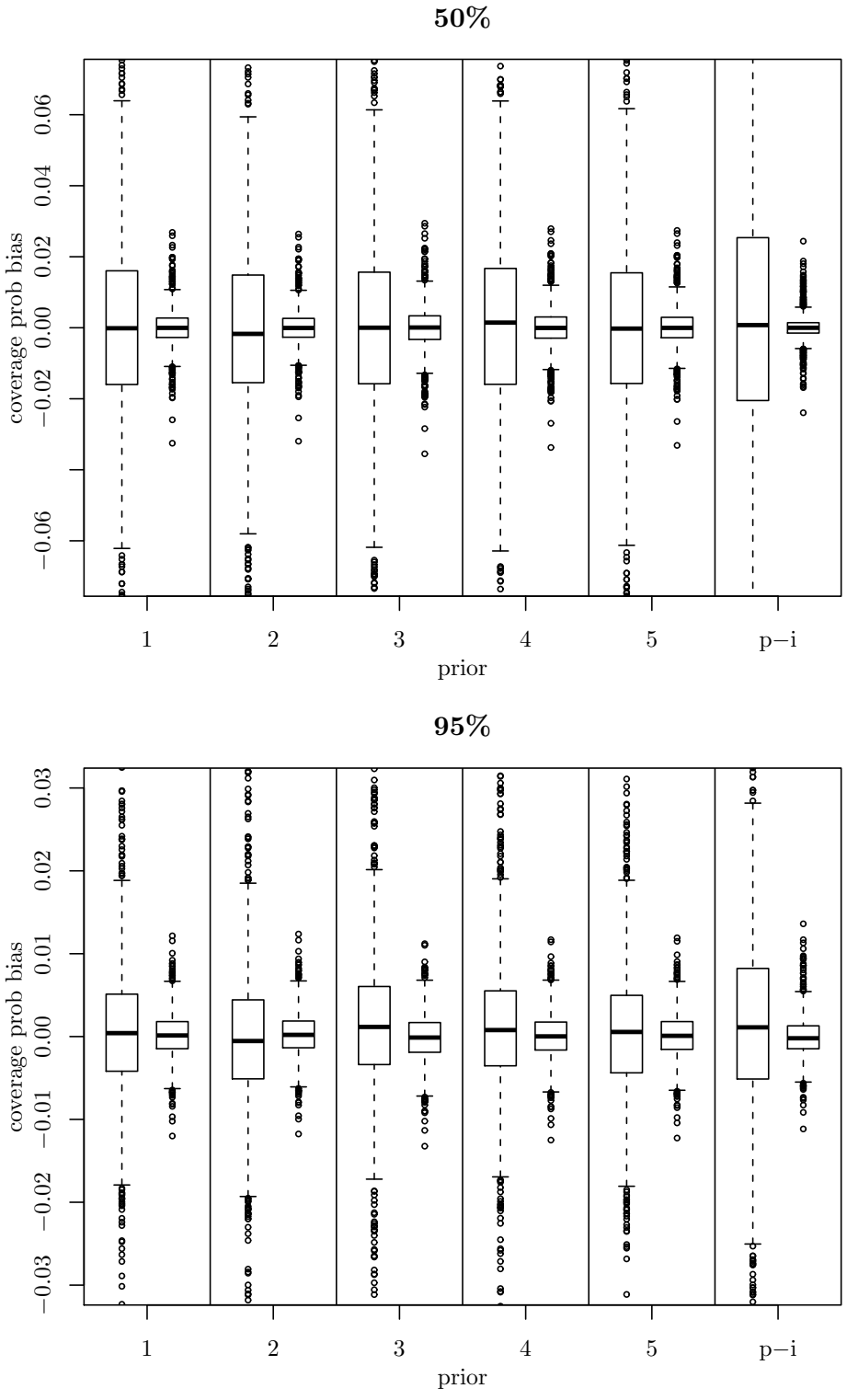

Figure 7.8: Boxplots to compare the coverage probability bias by simulation (left) and by approximation (right) for different priors and plug-in. The top corresponds to the 50% quantile and the bottom to the 95% quantile

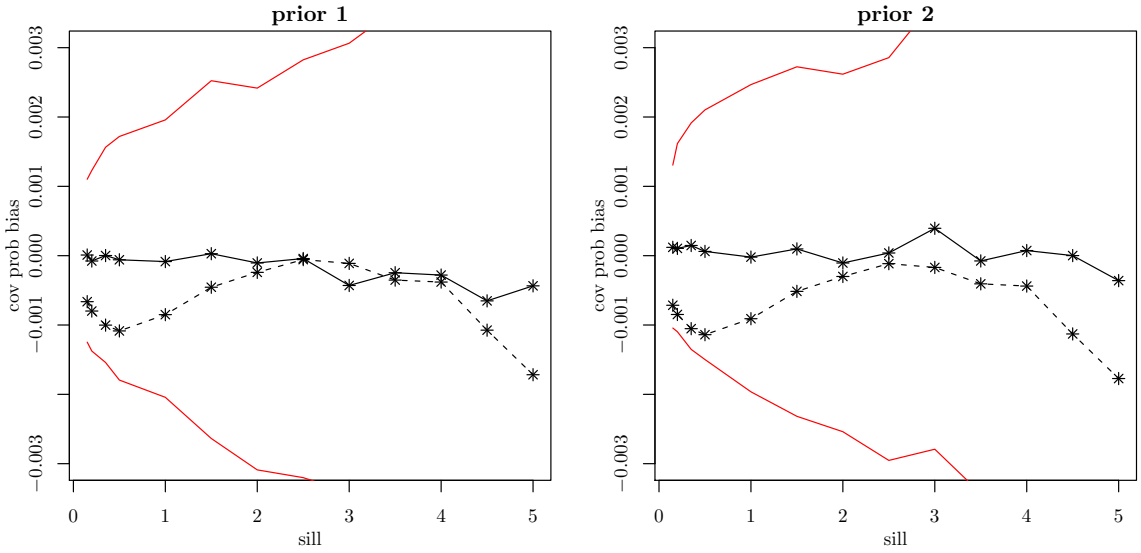

Figure 7.9: Approximated (dashed with $*$) and simulated (solid with $*$) coverage probability bias at the 2.5% quantile under priors 1 and 2 against sill with the interquartile range of the simulated coverage probability bias represented by a plain solid line.

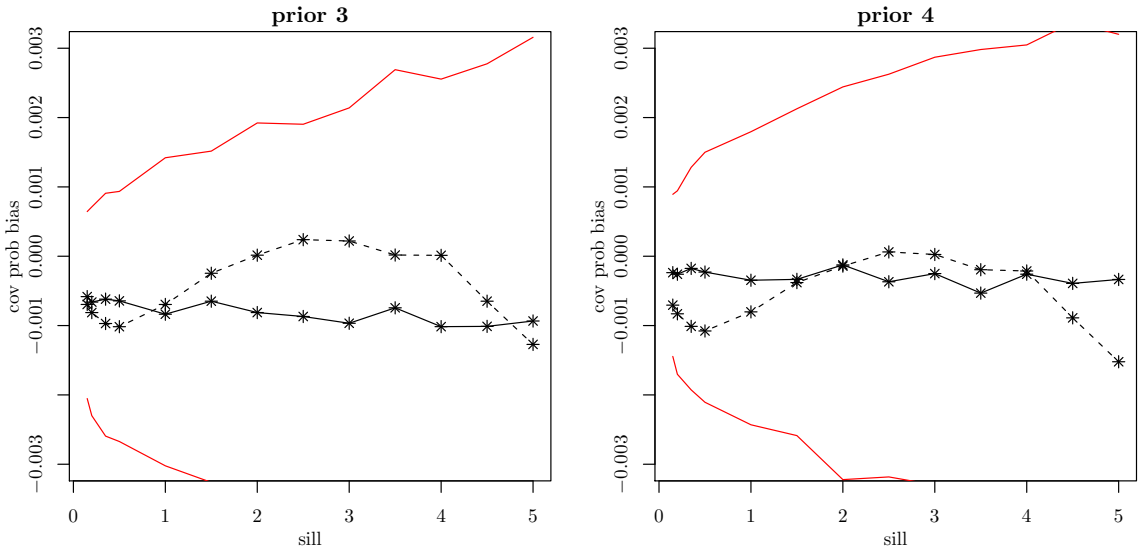

Figure 7.10: Approximated (dashed with $*$) and simulated (solid with $*$) coverage probability bias at the 2.5% quantile under priors 3 and 4 against sill with the interquartile range of the simulated coverage probability bias represented by a plain solid line.

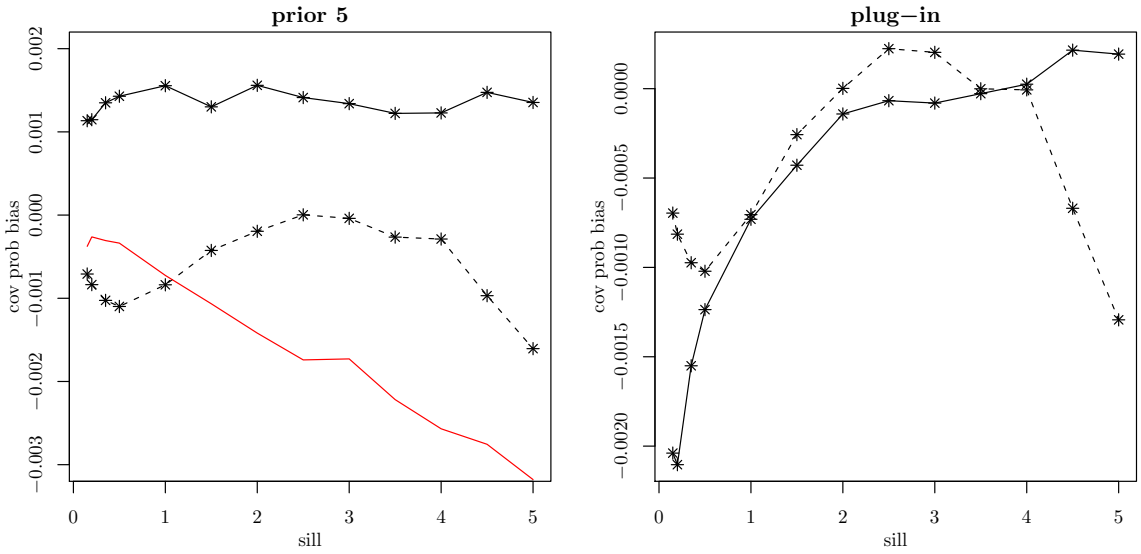

Figure 7.11: Approximated (dashed with $*$) and simulated (solid with $*$) coverage probability bias at the 2.5% quantile under priors 5 and plug-in against sill with the interquartile range of the simulated coverage probability bias represented by a plain solid line.

	2	3	4	5	6	7	8
1	0.3610	0.8396	0.7219	0.6301	0.9439	0.9189	0.7219
2		0.9537	0.9537	0.9537	0.9537	0.9537	0.8396
3			0.0463	0.0463	0.9537	0.7727	0.3610
4				0.0463	0.9537	0.9510	0.7219
5					0.9537	0.9537	0.7736
6						0.0463	0.0472
7							0.2718

Table 7.9: Proportions for comparing the Kullback-Leibler divergence. Each cell shows the proportion of how many times the predictive density constructed using the set of priors in the corresponding row had a smaller Kullback-Leibler divergence than the one constructed using the priors in the corresponding column.

CHAPTER 8

Summary and Future Work

In this thesis we provide methods that can be used for the analysis of data modeled under the Generalized Linear Mixed Model framework. Although the focus is toward spatial applications, these methods can easily be extended to other types of problems (e.g. longitudinal data) under the same model. Our methods are based on approximate techniques, mainly Taylor expansion and Laplace approximation to integrals described in section 3.2 and work well if the sample size is large.

We solve the problem of estimation by proposing an approximate likelihood method based on Laplace approximation. Although these types of approximations are known to be biased, we find that the mean square error is small. A further correction can be made to reduce bias by evaluating the error in the Taylor expansion of the likelihood as outlined at the end of section 4.3.

For the prediction of the random effects, we suggest using modified a plug-in method, derived by an application of Laplace approximation. Our proposed predictive distribution is Normal and prediction intervals can be easily computed. Our method is comparable to other existing methods in terms of prediction accuracy using three different scoring rules and it's much faster. Potential improvement can be made as we describe in section 5.1.1 although this has not been investigated yet.

Similar ideas are applied to the Bayesian predictive distribution where we suggest corrections to the predictive quantiles. The coverage probability bias and the Kullabck-Leibler divergence are also approximated. We use our approximation to compare the Bayesian predictive distributions constructed under different priors for the covariance parameters and concluded that the

best choice among the ones we compared is to use inverse gamma for the partial sill parameter and uniform for the nugget and range parameters.

Some ideas for future work

A potential improvement to the approximate likelihood can be made by computing the error of the Laplace approximation. An idea is presented at the bottom of section 4.3 which is based on an approximation to the conditional density of the random effects given the data as described in section 4.1.

An improvement to the plug-in prediction is outlined in section 5.1.1. The idea is to use numerical derivatives to obtain a higher order correction to the predictive quantiles. It will be interesting to apply the same idea to the quantiles of the Bayesian predictive distribution.

Another possibility would be to apply Laplace approximation to compute the posterior quantiles for the covariance parameters. If this idea works, it can be considered as an alternative to MCMC methods that are currently used for Bayesian estimation.

There is also the possibility of using Laplace approximation in the place of the E step of an EM algorithm as an alternative to Monte Carlo simulation. Alternatively, the approximate density in section 4.1 can be used to simulate i.i.d. samples from the conditional distribution of the random effects given the data and compute the expectation by averaging. The same density can be used for importance sampling when approximating the simulated likelihood.

Appendix

A.1 Convergence of $\hat{\boldsymbol{z}}(\boldsymbol{Y})$

We show that $\hat{\boldsymbol{z}}(\boldsymbol{Y})$ defined in (4.10) converges as $k \to \infty$ to $\boldsymbol{Z}$ almost surely. For the proof we make use of the following Lemma:

Lemma 4. *If $X_n|(Y=y) \xrightarrow{a.s} y$ for all y as $n \to \infty$, then $X_n \xrightarrow{a.s} Y$.*

Proof.

$$\Pr(X_n \to Y) = \int \Pr(X_n \to Y|Y=y)\,\mathrm{d}\Pr(y) = \int 1\,\mathrm{d}\Pr(y) = 1$$

□

Now note that $\hat{\boldsymbol{Z}} = \hat{\boldsymbol{z}}(\boldsymbol{Y})$ is chosen to maximize

$$\left(\sum_{i=1}^{k} Y_{i\cdot}\theta(Z_i) - \sum_{i=1}^{k} n_i b(\theta(Z_i))\right)/\omega - \frac{1}{2}\boldsymbol{Z}^{\intercal}\Sigma^{-1}\boldsymbol{Z} \tag{A.1}$$

where $\theta(z)$ is an increasing function relating the canonical parameter with the linear predictor.

Define $Z_i^* = (b' \circ \theta)^{-1}(n_i^{-1}Y_{i\cdot})$ where $(b' \circ \theta)^{-1}(\cdot)$ is the inverse function of $b'(\theta(\cdot))$.

By the Law of Large Numbers, $n_i^{-1}Y_{i\cdot}|(Z_i = z_i) \xrightarrow{a.s} b'(\theta(z_i))$ which implies that $Z_i^*|(Z_i = z_i) \xrightarrow{a.s} z_i$. Consequently, by Lemma 4, $Z_i^* \xrightarrow{a.s} Z_i$.

To show that $\hat{Z}_i \xrightarrow{a.s} Z_i$, define $E_i = \theta(\hat{Z}_i) - \theta(Z_i^*)$. By definition, $\hat{\boldsymbol{Z}}$ satisfies

$$\theta'(\hat{Z}_i)\left(Y_{i\cdot} - n_i b'(\theta(\hat{Z}_i))\right)/\omega - \sum_{j=1}^{k}\sigma^{ij}\hat{Z}_j = 0 \tag{A.2}$$

where σ^{ij} is the (i,j) element of Σ^{-1}. By the mean value theorem

$$b'(\theta(\hat{Z}_i)) = b'(\theta(Z_i^*) + E_i) = b'(\theta(Z_i^*)) + E_i b''(\theta(Z_i^*) + \Xi_i) \tag{A.3}$$

for some Ξ_i such that $|\Xi_i| \leq |E_i|$ a.s.. Note that the monotonicity of $b' \circ \theta$ implies that $E_i\, b''(\theta(Z_i^*) + \Xi_i) \xrightarrow{a.s} 0$ if and only if $E_i \xrightarrow{a.s} 0$.

Substituting (A.3) into (A.2),

$$-\theta'(\hat{Z}_i)\, b''(\theta(\hat{Z}_i) + \Xi_i)\, E_i = n_i^{-1} \omega \sigma^{i i_1} \hat{Z}_{i_1}$$

and letting $k \to \infty$, $E_i \to 0$, so $\hat{Z}_i \overset{a.s}{\to} Z_i$.

A.2 Proof of Theorem 1

By (7.29) and (7.30), for any four-times-differentiable function g, the bias of the estimate $g(\hat{\gamma}) - g(\gamma)$ is given by expanding $g(\hat{\gamma})$ around the true value γ

$$\begin{aligned}
g(\hat{\gamma}) - g(\gamma) &= g_{j_1}(\hat{\gamma}^{j_1} - \gamma^{j_1}) + \frac{1}{2} g_{j_1 j_2}(\hat{\gamma}^{j_1} - \gamma^{j_1})(\hat{\gamma}^{j_2} - \gamma^{j_2}) \\
&\quad + \frac{1}{6} g_{j_1 j_2 j_3}(\hat{\gamma}^{j_1} - \gamma^{j_1})(\hat{\gamma}^{j_2} - \gamma^{j_2})(\hat{\gamma}^{j_3} - \gamma^{j_3}) + O(k^{-2}) \\
&= g_{j_1}(k^{-1/2}\varepsilon_1^{j_1} + k^{-1}\varepsilon_2^{j_1} + k^{-3/2}\varepsilon_3^{j_1}) + \frac{1}{2} g_{j_1 j_2}(k^{-1/2}\varepsilon_1^{j_1} + k^{-1}\varepsilon_2^{j_1})(k^{-1/2}\varepsilon_1^{j_2} + k^{-1}\varepsilon_2^{j_2}) \\
&\quad + \frac{1}{6} g_{j_1 j_2 j_3}(k^{-1/2}\varepsilon_1^{j_1})(k^{-1/2}\varepsilon_1^{j_2})(k^{-1/2}\varepsilon_1^{j_3}) + O(k^{-2}) \\
&= k^{-1/2} g_{j_1}\varepsilon_1^{j_1} + k^{-1}\left\{ g_{j_1}\varepsilon_2^{j_1} + \frac{1}{2} g_{j_1 j_2}\varepsilon_1^{j_1}\varepsilon_1^{j_2} \right\} \\
&\quad + k^{-3/2}\left\{ g_{j_1}\varepsilon_3^{j_1} + \frac{1}{2} g_{j_1 j_2}(\varepsilon_1^{j_1}\varepsilon_2^{j_2} + \varepsilon_2^{j_1}\varepsilon_1^{j_2}) + \frac{1}{6} g_{j_1 j_2 j_3}\varepsilon_1^{j_1}\varepsilon_1^{j_2}\varepsilon_1^{j_3} \right\} + O(k^{-2})
\end{aligned} \tag{A.4}$$

Using a well known formula for matrix inversion,

$$U^{j,j'} = -k^{-1}\kappa^{j,j'} + k^{-3/2} W_{j_1 j_2}\kappa^{j,j_1}\kappa^{j',j_2} + O(k^{-2})$$

and applying (A.4) on the plug-in predictive distribution we have

$$\begin{aligned}
F(z_0|\boldsymbol{y};\hat{\gamma}) - F(z_0|\boldsymbol{y};\gamma) &= k^{-1/2} F_{j_1}\varepsilon_1^{j_1} + k^{-1}\left\{ F_{j_1}\varepsilon_2^{j_1} + \frac{1}{2} F_{j_1 j_2}\varepsilon_1^{j_1}\varepsilon_1^{j_2} \right\} \\
&\quad + k^{-3/2}\left\{ F_{j_1}\varepsilon_3^{j_1} + \frac{1}{2} F_{j_1 j_2}(\varepsilon_1^{j_1}\varepsilon_2^{j_2} + \varepsilon_2^{j_1}\varepsilon_1^{j_2}) + \frac{1}{6} F_{j_1 j_2 j_3}\varepsilon_1^{j_1}\varepsilon_1^{j_2}\varepsilon_1^{j_3} \right\} + O(k^{-2})
\end{aligned} \tag{A.5}$$

where $F_{j_1 \cdots j_s}$ is defined in (7.35).

On the other hand, applying standard Laplace approximation for the ratio of two integrals

on $F(z_0|\boldsymbol{y})$

$$F(z_0|\boldsymbol{y}) - F(z_0|\boldsymbol{y};\hat{\gamma}) = \frac{1}{2}\hat{F}_{j_1}\hat{U}_{j_2j_3j_4}\hat{U}^{j_1j_2}\hat{U}^{j_3j_4} - \frac{1}{2}(\hat{F}_{j_1j_2} + \hat{F}_{j_1}\hat{r}_{j_2})\hat{U}^{j_1j_2} + O(k^{-2}) \tag{A.6}$$

Applying (A.4) on $\hat{U}_{j_1j_2}$ and $\hat{U}_{j_1j_2j_3}$ we have

$$\begin{aligned}
\hat{U}_{j_1j_2} &= U_{j_1j_2} + k^{-1/2}U_{j_1j_2j}\varepsilon_1^j + O(1) \\
&= k\kappa_{j_1j_2} + k^{1/2}\left\{W_{j_1j_2} + \kappa_{j_1j_2j}\varepsilon_1^j\right\} + O(1) \qquad \text{(A.7)} \\
\hat{U}_{j_1j_2j_3} &= U_{j_1j_2j_3} + k^{-1/2}U_{j_1j_2j_3j}\varepsilon_1^j + O(1) \\
&= k\kappa_{j_1j_2j_3} + k^{1/2}\left\{W_{j_1j_2j_3} + \kappa_{j_1j_2j_3j}\varepsilon_1^j\right\} + O(1) \qquad \text{(A.8)}
\end{aligned}$$

and

$$\hat{U}^{j_1j_2} = -k^{-1}\kappa^{j_1,j_2} - k^{-3/2}\left\{W_{j_3j_4} + \kappa_{j_3j_4j_5}\varepsilon_1^{j_5}\right\}\kappa^{j_1,j_3}\kappa^{j_2,j_4} + O(k^{-2}) \tag{A.9}$$

so

$$\begin{aligned}
\hat{U}_{j_1j_2j_3}\hat{U}^{j_1j_2}\hat{U}^{j_3j_4} &= k^{-1}\kappa_{j_1j_2j_3}\kappa^{j_1,j_2}\kappa^{j_3,j_4} + k^{-3/2}\left\{\left(W_{j_1j_2j_3} + \kappa_{j_1j_2j_3j}\varepsilon_1^j\right)\kappa^{j_1,j_2}\kappa^{j_3,j_4}\right. \\
&\left. + \kappa_{j_1j_2j_3}\left(W_{jj_*} + \kappa_{jj_*j_\dagger}\varepsilon_1^{j_\dagger}\right)\kappa^{j_1,j}\kappa^{j_2,j_*}\kappa^{j_3,j_4} + \kappa_{j_1j_2j_3}\left(W_{jj_*} + \kappa_{jj_*j_\dagger}\varepsilon_1^{j_\dagger}\right)\kappa^{j_1,j_2}\kappa^{j_3,j}\kappa^{j_4,j_*}\right\} \\
&\quad + O(k^{-2}) \qquad \text{(A.10)}
\end{aligned}$$

In addition,

$$\hat{F}_{j_1} = F_{j_1} + k^{-1/2}F_{j_1j}\varepsilon_1^j + O(k^{-1}) \tag{A.11}$$

$$\hat{F}_{j_1j_2} = F_{j_1j_2} + k^{-1/2}F_{j_1j_2j}\varepsilon_1^j + O(k^{-1}) \tag{A.12}$$

$$\hat{F}_{j_1}\hat{r}_{j_2} = F_{j_1}r_{j_2} + k^{-1/2}(F_{j_1}r_{j_2j}\varepsilon_1^j + F_{j_1j}r_{j_2}\varepsilon_1^j) + O(k^{-1}) \tag{A.13}$$

Using (A.9) – (A.13) into (A.6),

$$\begin{aligned}
2\{F(z_0|\boldsymbol{y}) - F(z_0|\boldsymbol{y};\hat{\gamma})\} &= k^{-1}(F_{j_1}\kappa_{j_2j_3j_4}\kappa^{j_1,j_2}\kappa^{j_3,j_4} + F_{j_1j_2}\kappa^{j_1,j_2} + 2F_{j_1}\pi_{j_2}\kappa^{j_1,j_2}) \\
&+ k^{-3/2}[F_{j_1}\{(W_{j_2j_3j_4} + \kappa_{j_2j_3j_4r}\varepsilon_1^r)\kappa^{j_1,j_2}\kappa^{j_3,j_4} + \kappa_{j_2j_3j_4}(W_{rs} + \kappa_{rst}\varepsilon_1^t)\kappa^{j_1,r}\kappa^{j_2,s}\kappa^{j_3,j_4}
\end{aligned}$$

$$+\kappa_{j_2j_3j_4}(W_{rs}+\kappa_{rst}\varepsilon_1^t)\kappa^{j_1,j_2}\kappa^{j_3,r}\kappa^{j_4,s}\}+F_{j_1r}\kappa_{j_2j_3j_4}\varepsilon_1^r\kappa^{j_1,j_2}\kappa^{j_3,j_4}$$

$$+(F_{j_1,j_2}+2F_{j_1}\pi_{j_2})(W_{rs}+\kappa_{rst}\varepsilon_1^t)\kappa^{j_1,r}\kappa^{j_2,s}+F_{j_1j_2r}\varepsilon_1^r\kappa^{j_1j_2}+2(F_{j_1}\pi_{j_2r}+F_{j_1r}\pi_{j_2})\varepsilon_1^r\kappa^{j_1,j_2}]+O(k^{-2}) \tag{A.14}$$

Combining the result of (A.14) with (A.5), we obtain the result.

A.3 Expressions for the log-likelihood derivatives

For the first derivative U_{j_1} we have

$$\begin{aligned} U_{j_1} &= \partial_{j_1}\log L(\beta,\gamma|\boldsymbol{y}) \\ &= \frac{\partial_{j_1}L(\beta,\gamma|\boldsymbol{y})}{L(\beta,\gamma|\boldsymbol{y})} \end{aligned} \tag{A.15}$$

Differentiating (A.15),

$$\frac{\partial^2_{j_1j_2}L(\beta,\gamma|\boldsymbol{y})}{L(\beta,\gamma|\boldsymbol{y})}-\frac{\partial_{j_1}L(\beta,\gamma|\boldsymbol{y})}{L(\beta,\gamma|\boldsymbol{y})}\frac{\partial_{j_2}L(\beta,\gamma|\boldsymbol{y})}{L(\beta,\gamma|\boldsymbol{y})}=U_{j_1j_2} \tag{A.16}$$

Hence, by substituting (A.15) into (A.16),

$$\frac{\partial^2_{j_1j_2}L(\beta,\gamma|\boldsymbol{y})}{L(\beta,\gamma|\boldsymbol{y})}=U_{j_1j_2}+U_{j_1}U_{j_2} \tag{A.17}$$

which is (7.51).Further differentiation of (A.17) results (7.52) and (7.53).

A.4 Approximating the cumulants

The second cumulant, $\kappa_{j_1j_2}$ is defined as $\kappa_{j_1j_2}=\mathbb{E}(U_{j_1j_2})$ where the expectation is taken with respect to the distribution of $\boldsymbol{Y}$. Using (7.50) and (7.51) and the fact that $\mathbb{E}\{\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y})\}=\Sigma$ we have

$$\begin{aligned} \kappa_{j_1j_2} &= \mathbb{E}(U_{j_1j_2}) \\ &= -\mathbb{E}\left[\frac{\{\partial/\partial\gamma_{j_1}\}L(\beta,\gamma|\boldsymbol{y})}{L(\beta,\gamma|\boldsymbol{y})}\frac{\{\partial/\partial\gamma_{j_2}\}L(\beta,\gamma|\boldsymbol{y})}{L(\beta,\gamma|\boldsymbol{y})}\right] \end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{4}\mathbb{E}\left[\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y})\}\ \operatorname{tr}\{\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y})\}\right] \\
&\quad + [2]\frac{1}{4}\mathbb{E}\left[\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y})\}\right]\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_2}\} - \frac{1}{4}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_1}\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_2}\} \\
&= -\frac{1}{4}\mathbb{E}\left[\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y})\}\ \operatorname{tr}\{\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y})\}\right] \\
&\quad + \frac{1}{4}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_1}\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_2}\} \qquad \text{(A.18)}
\end{aligned}
$$

Similarly, for $\kappa_{j_1j_2j_3}$,

$$
\begin{aligned}
\kappa_{j_1j_2j_3} &= \mathbb{E}(U_{j_1j_2j_3}) \\
&= -[3]\frac{1}{4}\mathbb{E}\left[\left\{-2\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y})) + \operatorname{tr}(\Sigma^{-1}\Sigma_{j_1j_2}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y}))\right.\right. \\
&\qquad \left.\left. + \operatorname{tr}(\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\Sigma_{j_2}) + \operatorname{tr}(\Sigma^{-1}\Sigma_{j_1j_2})\right\}\right. \\
&\qquad \left.\left\{\operatorname{tr}(\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y})) - \operatorname{tr}(\Sigma^{-1}\Sigma_{j_3})\right\}\right] \\
&\quad - [3]\frac{1}{8}\mathbb{E}\left[\mathbb{E}\left\{(\boldsymbol{Z}^\intercal\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\boldsymbol{Z} - \operatorname{tr}(\Sigma^{-1}\Sigma_{j_1}))(\boldsymbol{Z}^\intercal\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\boldsymbol{Z} - \operatorname{tr}(\Sigma^{-1}\Sigma_{j_2}))|\boldsymbol{Y}\right\}\right. \\
&\qquad \left.\left\{\operatorname{tr}(\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y})) - \operatorname{tr}(\Sigma^{-1}\Sigma_{j_3})\right\}\right] \\
&\quad + \frac{1}{4}\mathbb{E}\left[\left\{\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y})) - \operatorname{tr}(\Sigma^{-1}\Sigma_{j_1})\right\}\right. \\
&\qquad \left\{\operatorname{tr}(\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y})) - \operatorname{tr}(\Sigma^{-1}\Sigma_{j_2})\right\} \\
&\qquad \left.\left\{\operatorname{tr}(\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y})) - \operatorname{tr}(\Sigma^{-1}\Sigma_{j_3})\right\}\right] \\
&= [3]\frac{1}{2}\mathbb{E}\left[\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y}))\operatorname{tr}(\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y}))\right] \\
&\quad - [3]\frac{1}{2}\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\Sigma_{j_2})\operatorname{tr}(\Sigma^{-1}\Sigma_{j_3}) \\
&\quad - [3]\frac{1}{4}\mathbb{E}\left[\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1j_2}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y}))\operatorname{tr}(\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y}))\right] \\
&\quad + [3]\frac{1}{4}\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1j_2})\operatorname{tr}(\Sigma^{-1}\Sigma_{j_3}) \\
&\quad - [3]\frac{1}{8}\mathbb{E}\left[\mathbb{E}\left\{(\boldsymbol{Z}^\intercal\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\boldsymbol{Z})(\boldsymbol{Z}^\intercal\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\boldsymbol{Z})|\boldsymbol{Y}\right\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y})\}\right] \\
&\quad + [3]\frac{1}{4}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\Sigma_{j_2}\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_3}\} + \frac{3}{8}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_1}\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_2}\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_3}\} \\
&\quad + [3]\frac{1}{4}\mathbb{E}\left[\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y})\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y})\}\right]\operatorname{tr}(\Sigma^{-1}\Sigma_{j_3}) \\
&\quad - \frac{3}{8}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_1}\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_2}\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_3}\} \\
&\quad + \frac{1}{4}\mathbb{E}\big[\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y})\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y})\} \\
&\qquad\qquad \operatorname{tr}\{\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal|\boldsymbol{Y})\}\big]
\end{aligned}
$$

$$
\begin{aligned}
&- [3]\frac{1}{4}\mathbb{E}\left[\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top|\boldsymbol{Y})\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top|\boldsymbol{Y})\}\right]\operatorname{tr}(\Sigma^{-1}\Sigma_{j_3}) \\
&+ \frac{1}{2}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_1}\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_2}\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_3}\} \\
&= [3]\frac{1}{2}\mathbb{E}\left[\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top|\boldsymbol{Y}))\operatorname{tr}(\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top|\boldsymbol{Y}))\right] \\
&- [3]\frac{1}{4}\mathbb{E}\left[\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1j_2}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top|\boldsymbol{Y}))\operatorname{tr}(\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top|\boldsymbol{Y}))\right] \\
&- [3]\frac{1}{8}\mathbb{E}\left[\mathbb{E}\left\{(\boldsymbol{Z}^\top\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\boldsymbol{Z})(\boldsymbol{Z}^\top\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\boldsymbol{Z})|\boldsymbol{Y}\right\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top|\boldsymbol{Y})\}\right] \\
&+ \frac{1}{4}\mathbb{E}\big[\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top|\boldsymbol{Y})\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_2}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top|\boldsymbol{Y})\} \\
&\qquad\qquad \operatorname{tr}\{\Sigma^{-1}\Sigma_{j_3}\Sigma^{-1}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top|\boldsymbol{Y})\}\big] \\
&- [3]\frac{1}{4}\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1}\Sigma^{-1}\Sigma_{j_2})\operatorname{tr}(\Sigma^{-1}\Sigma_{j_3}) \\
&+ [3]\frac{1}{4}\operatorname{tr}(\Sigma^{-1}\Sigma_{j_1j_2})\operatorname{tr}(\Sigma^{-1}\Sigma_{j_3}) \\
&+ \frac{1}{2}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_1}\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_2}\}\operatorname{tr}\{\Sigma^{-1}\Sigma_{j_3}\}
\end{aligned}
\tag{A.19}
$$

A.5 Priors for the Gaussian geostatistical model

Here we describe the three priors for the covariance parameters proposed by Berger et al. (2001) for the analysis of spatial data under the Gaussian assumption. The predictive densities constructed under these priors were compared among others in section 7.4.3.

The model that they used assumes

$$\boldsymbol{Z} \sim \mathrm{N}_n(X\beta, \Sigma)$$

where n is the sample size, X is a matrix of covariates, and Σ is a spatial covariance matrix depending on the covariance parameters.

The general form of the prior suggested for the partial sill (γ_2) and range (γ_3) was

$$\pi(\gamma_2, \gamma_3) \propto \gamma_2^{-a}\,\pi(\gamma_3), a > 0 \tag{A.20}$$

These were

- Reference prior:

$a = 1$ and $\pi_R(\gamma_3) \propto \left\{\mathrm{tr}(W^2) - (n-p)^{-1}\,\mathrm{tr}(W)^2\right\}^{1/2}$

where p is the number of columns in X and

$W = ((\,\mathrm{d}/\,\mathrm{d}\gamma_3)\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}P$, P is the matrix orthogonal to the space of X.

- Jeffreys rule prior:

  $a = 1$ and $\pi_{J1}(\gamma_3) \propto \left\{\mathrm{tr}(U^2) - n^{-1}\,\mathrm{tr}(U)^2\right\}^{1/2}$

  where $U = ((\,\mathrm{d}/\,\mathrm{d}\gamma_3)\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}$

- Jeffreys independent prior:

  $a = 1$ and $\pi_{J2}(\gamma_3) \propto |X^\intercal\boldsymbol{\Sigma}^{-1}X|^{1/2}\pi_{J1}(\gamma_3)$

Bibliography

Abramowitz, M. and Stegun, I. A. (1964), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover.

Barndorff-Nielsen, O. E. and Cox, D. R. (1989), *Asymptotic Techniques for Use in Statistics*, Chapman & Hall Ltd.

— (1996), "Prediction and Asymptotics," *Bernoulli*, 2, 319–340.

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

Booth, J. G. and Hobert, J. P. (1999), "Maximizing Generalized Linear Mixed Model Likelihoods with an Automated Monte Carlo EM Algorithm," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 61, 265–285.

Breslow, N. E. and Clayton, D. G. (1993), "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, 88, 9–25.

Breslow, N. E. and Lin, X. (1995), "Bias Correction in Generalised Linear Mixed Models with a Single Component of Dispersion," *Biometrika*, 82, 81–91.

Christensen, O. F. (2004), "Monte Carlo Maximum Likelihood in Model-based Geostatistics," *Journal of Computational and Graphical Statistics*, 13, 702–718.

Christensen, O. F., Møller, J., and Waagepetersen, R. (2000), "Analysis of spatial data using generalized linear mixed models and Langevin-type Markov chain Monte Carlo," Tech. rep., Department of Mathematical Sciences, Aalborg University.

Christensen, O. F. and Ribeiro, P. J. (2002), "GeoRglm: A Package for Generalised Linear Spatial Models," *R News*, 2, 26–28.

Christensen, O. F. and Waagepetersen, R. (2002), "Bayesian Prediction of Spatial Count Data Using Generalized Linear Mixed Models," *Biometrics*, 58, 280–286.

Clayton, D. G. (1996), "Generalized linear mixed models," in *Markov chain Monte Carlo in practice*, London: Chapman & Hall, Interdiscip. Statist., pp. 275–301.

Cressie, N. A. (1993), *Statistics for Spatial Data*, John Wiley & Sons.

Davis, R. A. and Rodriguez-Yam, G. (2005), "Estimation for State-space Models Based on a Likelihood Approximation," *Statistica Sinica*, 15, 381–406.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998), "Model-based Geostatistics," *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 47, 299–326.

Efron, B. and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, Chapman & Hall Ltd.

Fan, Y., Leslie, D. S., and Wand, M. P. (2008), "Generalised linear mixed model analysis via sequential Monte Carlo sampling," *Electronic Journal of Statistics*, 2, 916–938.

Fisher, R. A. (1912), "On an absolute criterion for fitting frequency curves," *Messenger of Mathematics*, 41, 155–160.

Geisser, S. (1993), *Predictive Inference: an Introduction*, Chapman & Hall Ltd.

Gneiting, T. and Raftery, A. E. (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102, 359–378.

Henderson, C. R., Kempthorne, O., Searle, S. R., and von Krosigk, C. M. (1959), "The Estimation of Environmental and Genetic Trends from Records Subject to Culling," *Biometrics*, 15, 192–218.

Karim, M. R. and Zeger, S. L. (1992), "Generalized Linear Models with Random Effects; Salamander Mating Revisited," *Biometrics*, 48, 631–644.

Krige, D. G. (1951), "A statistical approach to some basic mine valuation problems on the Witwatersrand," *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 52, 119–139.

Lin, X. and Breslow, N. E. (1996), "Bias Correction in Generalized Linear Mixed Models with Multiple Components of Dispersion," *Journal of the American Statistical Association*, 91, 1007–1016.

Mardia, K. V. and Marshall, R. J. (1984), "Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression," *Biometrika*, 71, 135–146.

Matheron, G. (1962), "Traité de Géostatistique Appliquée, Tome I. Mémoires du Bureau de Recherches Géologiques et Minières," .

— (1963), "Principles of geostatistics," *Economic Geology*, 1246–1266.

McCullagh, P. (1987), *Tensor Methods in Statistics*, Chapman & Hall Ltd.

McCullagh, P. and Nelder, J. A. (1999), *Generalized Linear Models*, Chapman & Hall Ltd.

McCulloch, C. E. (1997), "Maximum Likelihood Algorithms for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 92, 162–170.

Natarajan, R. and Kass, R. E. (2000), "Reference Bayesian Methods for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 95, 227–237.

Nelder, J. A. and Wedderburn, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society, Series A: General*, 135, 370–384.

Noh, M. and Lee, Y. (2007), "REML Estimation for Binary Data in GLMMs," *Journal of Multivariate Analysis*, 98, 896–915.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Raudenbush, S. W., Yang, M.-L., and Yosef, M. (2000), "Maximum Likelihood for Generalized Linear Models with Nested Random Effects Via High-order, Multivariate Laplace Approximation," *Journal of Computational and Graphical Statistics*, 9, 141–157.

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations," *Journal of the Royal Statistical Society, Series B, Methodological*, 71, 1–35.

Shun, Z. (1997), "Another look at the Salamander mating data: A modified Laplace approximation approach," *Journal of the American Statistical Association*, 92, 341–349.

Shun, Z. and McCullagh, P. (1995), "Laplace approximation of high dimensional integrals," *Journal of the Royal Statistical Society, Series B, Methodological*, 57, 749–760.

Smith, R. L. (1999), "Bayesian and Frequentist Approaches to Parametric Predictive Inference," in *Bayesian Statistics 6 – Proceedings of the Sixth Valencia International Meeting*, eds. Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., Clarendon Press [Oxford University Press], pp. 589–612.

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

Tierney, L., Kass, R. E., and Kadane, J. B. (1989), "Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions," *Journal of the American Statistical Association*, 84, 710–716.

Vidoni, P. (2006), "Response prediction in mixed effects models," *Journal of Statistical Planning and Inference*, 136, 3948–3966.

Zeger, S. L. and Karim, M. R. (1991), "Generalized Linear Models with Random Effects: A Gibbs Sampling Approach," *Journal of the American Statistical Association*, 86, 79–86.

Zhang, H. (2002), "On Estimation and Prediction for Spatial Generalized Linear Mixed Models," *Biometrics*, 58, 129–136.