6-24-2015

Order-Constrained Reference Priors with Implications for Bayesian Isotonic Regression, Analysis of Covariance and Spatial Models

Maozhen Gong

Maozhen Gong

Candidate

Mathematics and Statistics

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Gabriel Huerta , Chairperson

Erik Erhardt

Guoyi Zhang

Shuang Luan

Order-Constrained Reference Priors with Implications for Bayesian Isotonic Regression, Analysis of Covariance and Spatial Models

by

Maozhen Gong

B.S., Lanzhou University, 2003
M.S., University of New Mexico, 2008

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Statistics

The University of New Mexico

Albuquerque, New Mexico

May, 2015

Dedication

I dedicate this dissertation to Yuzhi Li.

Acknowledgments

There are a lot of people I would like to thank for their help in completing this dissertation. First, I would like to thank my wife, Yuzhi Li, for always taking care of me. Dr. Michael Sonksen, thanks for being so nice to me and teaching me so much. Professor Gabriel Huerta, thanks for your kindness and helping me so much in the last stage of my graduate study. There is no way I can finish this dissertation without your help. Professor Erik Erhardt, Professor Guoyi Zhang and Professor Shuang Luan, thanks for being in my defense committee and giving a few valuable suggestions. Dr. Mark Burge, thanks for providing the data. It is such a lucky thing to know all of you in my life.

Order-Constrained Reference Priors with Implications for Bayesian Isotonic Regression, Analysis of Covariance and Spatial Models

by

Maozhen Gong

B.S., Lanzhou University, 2003
M.S., University of New Mexico, 2008

Ph.D., Statistics, University of New Mexico, 2015

Abstract

Selecting an appropriate prior distribution is a fundamental issue in Bayesian Statistics. In this dissertation, under the framework provided by Berger and Bernardo [1992], I derive the reference priors for several models which include: Analysis of Variance (ANOVA)/Analysis of Covariance (ANCOVA) models with a categorical variable under common ordering constraints, the conditionally autoregressive (CAR) models and the simultaneous autoregressive (SAR) models with a spatial autoregression parameter ρ considered. The performances of reference priors for ANOVA/ANCOVA models are evaluated by simulation studies with comparisons to Jeffreys' prior and Least Squares Estimation (LSE). The priors are then illustrated in a Bayesian model of the "Risk of Type 2 Diabetes in New Mexico" data, where the relationship between the type 2 diabetes risk (through Hemoglobin A1c) and different smoking levels is investigated. In both simulation studies and real data set

modeling, the reference priors that incorporate internal order information show good performances and can be used as default priors. The reference priors for the CAR and SAR models are also illustrated in the "1999 SAT State Average Verbal Scores" data with a comparison to a Uniform prior distribution. Due to the complexity of the reference priors for both CAR and SAR models, only a portion (12 states in the Midwest) of the original data set is considered. The reference priors can give a different marginal posterior distribution compared to a Uniform prior, which provides an alternative for prior specifications for areal data in Spatial statistics.

KEY WORDS: Bayesian Models, Non-informative priors, Markov chain Monte Carlo.

Contents

List of Figures

List of Figures

List of Tables

Chapter 1

Introduction

The primary topic of this dissertation considers order-constrained reference priors. Classical statistics treats parameters as fixed and relies on maximizing the likelihood function to make parameter estimation, while Bayesian statistics introduces a prior distribution for the parameters and commonly approximates the posterior distribution by some stochastic simulation algorithms. It may be difficult for a data analyst to specify an appropriate subjective prior for Bayesian analysis, either because sufficient knowledge is unavailable or difficult to incorporate into a prior distribution. However, researchers can still conduct Bayesian analysis using priors that limit subjective knowledge. These priors are known as non-informative, default or objective priors. Among non-informative priors, the reference prior of Bernardo [1979] stands out since: (1) It maximizes the limiting expected Kullback-Leibler (KL) divergence between posterior and prior densities with respect to the marginal distribution of the data. (2) Although "not well understood" in Berger [2006], the reference prior approach seems to guard successfully from posterior impropriety that may occur when adopting other improper non-informative prior distributions. (3) In the single parameter case the reference prior often defaults to Jeffreys' prior [Kass and Wasserman, 1996]. (4) Yang [1995] showed that for the most important types of

reparameterizations, the reference prior is invariant.

This dissertation proceeds as follows. In this chapter Bayesian statistics are reviewed, with a focus on noninformative prior distributions. In Chapters 2 and 3, the reference priors for population means and analysis of covariance (ANCOVA) models under common ordering constraints are derived. Their effectiveness is evaluated via Markov chain Monte Carlo (MCMC) methods and simulation studies while comparing it to Jeffreys' prior and Least Squares Estimation (LSE). Furthermore, the proposed prior for the ANCOVA model with an increasing internal ordering of the categorical variable of interest is applied in a collaboration with Dr. Mark Burge from the University of New Mexico (UNM) Health Sciences Center, where the relationship between type 2 diabetes risk (through HbA1c), smoking levels (categorical) and other variables is modeled by these reference priors. In Chapter 4, the reference priors for the spatial conditionally autoregressive (CAR) models and the simultaneous autoregressive (SAR) models with considering the auto-regression parameter are derived. A state level data set related to the 1999 SAT college entrance exam scores is considered. The analysis includes comparisons with a Uniform prior and Maximum Likelihood Estimation (MLE).

1.1 Bayesian Statistics

Bayesian statistics revolves around Bayes Theorem

$$\begin{aligned} p(\boldsymbol{\theta}|x) &= \frac{f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{m(x)} \\ &\propto f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta}), \end{aligned} \tag{1.1}$$

which can be traced back to Bayes [1763] and Laplace [1774]. The fundamental idea for Bayesian inference is that we have observed data X, with an assumed probability

distribution $f(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are the unobserved parameters following a prior distribution $\pi(\boldsymbol{\theta})$. Relying on Bayes Theorem, we calculate and summarize the posterior distribution, $p(\boldsymbol{\theta}|x)$, the conditional distribution of the unobserved parameters given the observed data, which updates prior beliefs with the observed data.

To employ Bayesian methodology, one must specify a prior distribution and a likelihood function. It may not be hard to specify a likelihood in many problems due to the long-time developments of likelihood inference. However, challenges arise when it comes to selection of an appropriate prior distribution. A researcher's actual beliefs on the unobserved parameters should be incorporated and default options are also needed when subjective knowledge regarding the parameters is unavailable. This has evolved into two major schools of thought about specifying the prior distributions, "subjective" and "objective" approaches. In this dissertation, I will focus on objective priors.

1.1.1 Objective Priors

The main goal in the study of objective priors is to find a prior distribution that does not favor any particular parameter value and thus, Bayesian inference can be impacted minimally by the selection of the prior. This means that the observed data will dominate the posterior distribution for a given likelihood function. Without actual knowledge or information for the unknown parameters, researchers can use objective priors as default options. Common objective priors include the Uniform prior, Jeffreys' prior and the reference prior as in Ghosh [2011]. Other less common ones are maximum entropy priors as in Jaynes [1982], the probability matching priors as in Tibshirani [1989], the maximum chi-squared distance priors as in Clarke and Sun [1997] and many others.

Uniform and Jeffreys' Prior. The use of Uniform priors can be traced back to Bayes and Laplace. For a Binomial likelihood, Bayes [1763] assumed that the probability of success was uniformly distributed on the interval (0, 1), while Laplace [1774] assumed that the mean $\boldsymbol{\theta}$ of a normal likelihood was proportional to one. In this case, $\pi(\boldsymbol{\theta}) \propto 1$, is not a proper distribution, which means the integral of the prior distribution over the parameter space does not equal one. Objective Bayesian analysis allows the use of an improper prior distribution if the corresponding posterior distribution is proper, which is axiomatically permissible under finite additivity as in Sun and Berger [2006].

A Uniform prior seems adequate since it does not favor any particular value in the parameter space. However, criticism has arisen to Uniform priors since these priors are not invariant to transformations of the parameters. For example, assume that $\pi(\mu) \propto 1$ and that $\tau = \exp(\mu)$. The implied prior for τ is

$$\pi(\tau) = \pi(\mu = \log(\tau)) \times \left| \frac{d}{d\tau} \log(\tau) \right| = \frac{1}{\tau}. \tag{1.2}$$

Note that the prior for τ is no longer uniform, which turns to be an awkward situation if uniformity is decided to reflect ignorance of any parameter.

This concern was first raised by Jeffreys [1961], among others. Later he proposed a prior distribution based on the determinant of the Fisher information matrix. That is,

$$\pi_J(\boldsymbol{\theta}) \propto \sqrt{\det(I(\boldsymbol{\theta}))}, \tag{1.3}$$

where $I(\boldsymbol{\theta})$ denotes the Fisher information matrix. $\pi_J(\boldsymbol{\theta})$ is known as Jeffreys' prior which can be shown to be invariant to one-to-one transformations of $\boldsymbol{\theta}$. However, Jeffreys' prior is often improper and gives no guarantee to produce a proper posterior distribution. At the same time, Jeffreys' prior does not work well in multi-parameter

problems, which includes marginalization paradoxes as in Stone and Dawid [1972] and strong inconsistency as in Stein [1959] and Stone [1976]. In addition, Jeffreys suggested to find the Jeffreys' prior for each individual parameter independently and use the product of the individual priors as a prior for all parameters. This is a common approach for finding objective priors, where modification of a general approach is introduced to make the prior work well for a specific situation. Regardless of these issues, Jeffreys' prior is very popular and widely used as a default/objective prior in many situations.

Reference Prior. The reference prior approach of Bernardo [1979], further refined in Berger and Bernardo [1992] and Berger et al. [2009], is a popular choice for objective priors in many situations. The idea is good in that the prior is the function that maximizes the expected KL divergence between the posterior and the prior densities with regard to the marginal distribution of the data. This makes the data have the maximum effect on the posterior estimates. Let us consider an inference scenario where we have data X with a distribution $p(x|\boldsymbol{\theta})$. The KL divergence between the posterior and prior is

$$KL(p(\boldsymbol{\theta}|x), \pi(\boldsymbol{\theta})) = \int p(\boldsymbol{\theta}|x) \log \frac{p(\boldsymbol{\theta}|x)}{\pi(\boldsymbol{\theta})} d\boldsymbol{\theta} \tag{1.4}$$

and the expected KL divergence over the marginal distribution of the data is

$$\begin{aligned} I_{KL}(\Theta, x) &= E\left[\int p(\boldsymbol{\theta}|x) \log \frac{p(\boldsymbol{\theta}|x)}{\pi(\boldsymbol{\theta})} d\boldsymbol{\theta}\right] \\ &= \int p(x) \int p(\boldsymbol{\theta}|x) \log \frac{p(\boldsymbol{\theta}|x)}{\pi(\boldsymbol{\theta})} d\boldsymbol{\theta} dx \\ &= \int\int p(\boldsymbol{\theta}, x) \log \frac{p(\boldsymbol{\theta}|x)}{\pi(\boldsymbol{\theta})} d\boldsymbol{\theta} dx. \end{aligned} \tag{1.5}$$

Now the reference prior, $\pi^*(\boldsymbol{\theta})$, is the one that maximizes this mutual information.

This means

$$\pi^*(\boldsymbol{\theta}) \;=\; arg\max_{\pi(\boldsymbol{\theta})} I_{KL}(\Theta, x). \tag{1.6}$$

This idea seems straightforward, however, direct maximization is often not analytically tractable because the prior is part of the posterior. Furthermore, it is quite common to find that the prior that maximizes the expected KL is discrete, which is not adequate if the parameter space is continuous. Instead Bernardo [1979] suggests to maximize

$$\begin{aligned} E[KL]_t &= E^{Z_t}\left[\int p(\boldsymbol{\theta}|x)\log\frac{p(\boldsymbol{\theta}|x)}{\pi(\boldsymbol{\theta})}d\boldsymbol{\theta}\right] \\ &= \int p(z_t)\int p(\boldsymbol{\theta}|z_t)\log\frac{p(\boldsymbol{\theta}|z_t)}{\pi(\boldsymbol{\theta})}d\boldsymbol{\theta}dz_t, \end{aligned} \tag{1.7}$$

where $Z_t = \{X_1, X_2, \ldots, X_t\}$ are t conditionally independent replicates of the original experiment and $p(z_t|\boldsymbol{\theta}) = \prod_{i=1}^{t} p(x_i|\boldsymbol{\theta})$. Berger and Bernardo [1992] point out that for Equation 1.7, when t $\rightarrow \infty$, Z_t will have all information about $\boldsymbol{\theta}$. In addition $E[KL]_\infty = \lim_{t\rightarrow\infty} E[KL]_t$ could be interpreted as the missing information about the parameters $\boldsymbol{\theta}$ relative to the prior $\pi(\boldsymbol{\theta})$. Hence $\pi(\boldsymbol{\theta})$ is a noninformative prior because it maximizes the missing information. These approaches still have several problems. First, $E[KL]_\infty$ is often infinite. The fix is to find a $\pi_t(\boldsymbol{\theta})$ at some t from maximizing $E[KL]_t$ and then use $\lim_{t\rightarrow\infty}\pi_t(\boldsymbol{\theta})$ to determine the final reference prior. Secondly, if the parameter space, Θ, is non-compact, $E[KL]_t$ is often infinite. The recommendation then is to define a sequence of compact subsets, Θ^l, such that $\lim_{l\rightarrow\infty}\Theta^l = \Theta$. A sequence of reference priors $\pi^l(\boldsymbol{\theta})$ is then found and the final reference prior is calculated by taking the limit, i.e., $\pi(\boldsymbol{\theta}) = \lim_{l\rightarrow\infty}\pi^l(\boldsymbol{\theta})$.

If the model is regular, which means $p(z_t|\boldsymbol{\theta})$ satisfies the conditions for asymptotic normality, Berger and Bernardo [1992] show that the procedure of deriving reference priors can be done in an explicit way. First, let us separate the elements of

$\boldsymbol{\theta}$ into m groups and order them as $\boldsymbol{\theta} = (\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}, \ldots, \boldsymbol{\theta}_{(m)})$. The specific grouping and ordering usually depend on the relative inferential importance. Suppose that group j includes m_j elements, that is, $\boldsymbol{\theta}_{(j)} = (\theta_{j_1}, \theta_{j_2}, \ldots, \theta_{j_{m_j}})$ and define $\boldsymbol{\theta}_{(1:j)} = (\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}, \ldots, \boldsymbol{\theta}_{(j)})$. We first calculate the Fisher information matrix $I(\boldsymbol{\theta})$ with $I(\boldsymbol{\theta})_{i,j} = E\left[(\frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i})(\frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j})\right]$ and then find the inverse of the Fisher information matrix,

$$\begin{aligned} S(\boldsymbol{\theta}) &= (I(\boldsymbol{\theta}))^{-1} \\ &= \begin{pmatrix} A_{11} & A_{21}^t & \ldots & A_{m1}^t \\ A_{21} & A_{22} & \ldots & A_{m2}^t \\ \vdots & \vdots & \ldots & \vdots \\ A_{m1} & A_{m2} & \ldots & A_{mm} \end{pmatrix}, \end{aligned} \tag{1.8}$$

where A_{ij} is an $m_i \times m_j$ matrix. Now let us define $N_j = \sum_{i=1}^{j} m_i$, S_j be the upper left $N_j \times N_j$ corner of $S(\boldsymbol{\theta})$, $H_j = S_j^{-1}$ and h_j be the lower right $m_j \times m_j$ corner of H_j. If the determinant of $h_j(\boldsymbol{\theta})$, $|h_j(\boldsymbol{\theta})|$, depends only on $\boldsymbol{\theta}_{(1:j)}$, for $j = 1, \ldots, m$, calculation of this m-group reference prior can be vastly simplified and gives,

$$\pi(\boldsymbol{\theta}) \propto \frac{\Pi_{j=1}^{m} |h_j(\boldsymbol{\theta})|^{1/2}}{\Pi_{j=1}^{m} \int_{\Theta_j|\Theta_{(1:(j-1))}} |h_j(\boldsymbol{\theta})|^{1/2} d\boldsymbol{\theta}_{(j)}} I_{\Theta}(\boldsymbol{\theta}). \tag{1.9}$$

Here $\Theta_j|\Theta_{(1:(j-1))}$ is the parameter space for group j given previous groups. For a single parameter case, the reference prior defaults to Jeffreys' prior under regularity condition as in Kass and Wasserman [1996]. For the multi-parameter case, the reference prior usually does not have a unique form, since the algorithm by Berger and Bernardo [1992] includes grouping and ordering parameters by inferential importance. This causes the concern that one may get different reference prior distributions for different groupings and orderings. The common way to handle this is to try several intuitive groupings or orderings and then conduct a sensitivity study

based on different resulting reference priors. Further development in reference priors can be seen in Sun and Berger [1998], Ghosal [1999] and Berger and Sun [2008]. In the following chapters of this dissertation I will rely on this algorithm to derive reference priors for order-constrained models and CAR/SAR models.

1.1.2 Bayesian Model Selection and Diagnostics

There exist various techniques used for Bayesian model selection and diagnostics, which include Bayes factors, predictive P-values and the Deviance Information Criterion (DIC). Generally the Bayes factors and DIC are used for model selection and predictive P-values are used for detecting lack-of-fit, although there is an overlap between these two goals. In this dissertation, I compare the DICs of Bayesian models under different priors because Bayes factors are undefined when the prior distribution is improper (which is the case for most of our models).

DIC. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), described by Akaike [1974] and Schwarz [1978] respectively, are widely used in classical model selection. They usually include a model fit term and a penalty term for model complexity. However, they ignore the prior distribution and their applications in Bayesian goodness-of-fit are limited. The DIC, a Bayesian alternative to AIC/BIC and proposed by Spiegelhalter et al. [2002], is used by many practitioners, although difficulties with DIC have been noted as in Celeux et al. [2006] and Plummer [2008]. For a likelihood $f(x|\boldsymbol{\theta})$, the deviance can be defined as

$$D(\boldsymbol{\theta}) \;=\; -2\log(f(x|\boldsymbol{\theta})) \;+\; 2\log[m(x)], \tag{1.10}$$

where the logarithm of the marginal likelihood, $\log[m(x)]$, is a constant that cancels out in the calculation. The posterior expectation of the deviance is $\bar{D} \;=\; E_{\boldsymbol{\theta}|x}[D(\boldsymbol{\theta})]$,

which has been suggested as a measure of how well the model fits the data. The smaller it is, the better the fit. Since more complex models (for example, model with larger effective number of parameters) fit the data better and hence give smaller $\bar{D}$, a measure of model complexity, p_D, is also needed to penalize $\bar{D}$. This measure is,

$$\begin{aligned} p_D &= E_{\boldsymbol{\theta}|x}[D(\boldsymbol{\theta})] - D(E_{\boldsymbol{\theta}|x}[\boldsymbol{\theta}]) \\ &= \bar{D} - D(\bar{\boldsymbol{\theta}}), \end{aligned} \tag{1.11}$$

which is actually the posterior mean deviance minus the deviance evaluated at the posterior mean of the parameters. The DIC is then defined as

$$DIC = \bar{D} + p_D. \tag{1.12}$$

Generally, models with smaller DIC are better supported by the data. DIC can be directly calculated from the posterior samples generated by MCMC methods. To compute DIC, simply calculate $\bar{D}$ from the average of $D(\boldsymbol{\theta})$, over the posterior samples of $\boldsymbol{\theta}$ and $D(\bar{\boldsymbol{\theta}})$ as the value of D evaluated at the posterior mean of $\boldsymbol{\theta}$. Claeskens and Hjort [2008] show that the DIC is equivalent to the natural model-robust version of the AIC for large samples.

1.2 Computation

The approximation of posterior distributions in modern Bayesian statistics heavily relies on stochastic simulation algorithms, among which MCMC is the most popular one. This is a powerful tool that I use throughout this thesis in my studies about reference priors. On the other hand, the Expectation-Maximization (EM) algorithm, or one of its variants, the Expectation Conditional Maximization (ECM) algorithm, can be used to find the maximum *a posteriori* probability (MAP) estimator. The MAP estimator is also known as the posterior mode.

1.2.1 Markov Chain Monte Carlo

Suppose that we have observed data, X, with a likelihood $f(x|\boldsymbol{\theta})$ and a prior distribution $\pi(\boldsymbol{\theta})$. MCMC methods attempt to construct a stationary Markov chain where its stationary distribution is approximately the posterior distribution $p(\boldsymbol{\theta}|x)$. This means that once the chain is stable, each MCMC iteration becomes an approximate realization from the posterior distribution and Bayesian inference can be easily drawn by summarizing these iterations through histograms, box plots, sample means and quantiles, which implicitly uses the strong law of large numbers. Although these realizations are not independent samples from the posterior distribution, ergodic theorems described in Karlin and Taylor [1975] and Tierney [1994] guarantee its convergence. The most popular MCMC methods include Gibbs Sampling and the Metropolis-Hastings algorithm.

Gibbs Sampling. This method was first proposed by Geman and Geman [1984] and takes advantage of the conditional conjugacy structure that many Bayesian models have. To employ Gibbs sampling, the initial values of the parameters, $\boldsymbol{\theta}^{(1)}$, are arbitrarily set at the beginning of the algorithm. At the t-th iteration, each element of $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_k^{(t)})$ is updated by drawing from the full conditional distributions in sequence and substitution,

$$\begin{aligned}
\theta_1^{(t)} &\sim \pi(\theta_1|\theta_2^{(t-1)},\ \theta_3^{(t-1)}, \ldots,\ \theta_k^{(t-1)},\ X),\\
\theta_2^{(t)} &\sim \pi(\theta_2|\theta_1^{(t)},\ \theta_3^{(t-1)}, \ldots,\ \theta_k^{(t-1)},\ X),\\
\theta_3^{(t)} &\sim \pi(\theta_3|\theta_1^{(t)},\ \theta_2^{(t)},\ \theta_4^{(t-1)}, \ldots,\ \theta_k^{(t-1)},\ X),\\
&\vdots\\
\theta_k^{(t)} &\sim \pi(\theta_1|\theta_1^{(t)},\ \theta_2^{(t)}, \ldots,\ \theta_{k-1}^{(t)},\ X).
\end{aligned}$$

As t approaches to ∞, the Markov chain reaches equilibrium and each $\boldsymbol{\theta}^{(t)}$ is

approximately a draw from the posterior distribution. In practice, the convergence is commonly monitored by trace plots or history of the sampled parameter values. Theoretically, the Gibbs sampler is not sensitive to the choice of initial values, however, crude estimates (least squares) can be adopted as initial values to achieve fast convergence. When the posterior distribution is multimodal, the Gibbs sampler may be trapped at one of the modes. This is a problem that may be solved by blocking the parameters or considering alternative algorithms such as slice sampling as in Neal [2003].

Metropolis-Hastings. Although the Gibbs sampling is very general, the algorithm relies on the availability to sample from the full conditional of the target distribution. If full conditionals are not available, a powerful alternative is the Metropolis-Hastings algorithm by Hastings [1970], which is a general version of the Metropolis algorithm by Metropolis et al. [1953]. Similar to the Gibbs sampling, the Metropolis-Hastings algorithm needs to be initialized at a point $\boldsymbol{\theta}^{(1)}$. For an inference scenario where we have a likelihood function $f(x|\boldsymbol{\theta})$, the updating at iteration t relies on a proposal distribution $q(\cdot|\boldsymbol{\theta}^{(t-1)})$. At iteration t, $\boldsymbol{\theta}^*$ is sampled from $q(\cdot|\boldsymbol{\theta}^{(t-1)})$ and accepted with probability $\alpha = \min\{1, r\}$ where

$$r = \frac{f(x|\boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^*)}{f(x|\boldsymbol{\theta}^{(t-1)})\pi(\boldsymbol{\theta}^{(t-1)})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})} \tag{1.13}$$

and $\pi(\boldsymbol{\theta})$ is the prior. If the proposal distribution is symmetric, $q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^*) = q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})$, the right side of Equation 1.13 can be simplified and the algorithm becomes the Metropolis algorithm. This algorithm can be iterated until the Markov chain attains stationary status, which is usually monitored by trace plots. Convergence diagnostics, such as the Gelman and Rubin [1992] diagnostic, can also be calculated for detecting convergence of the algorithm.

In theory, the algorithm converges regardless the choice of the reasonable proposal

density q. However, in practice some proposals may converge faster or slower. In this dissertation, independent normal (or truncated) proposals are mainly adopted with variances specified to produce good mixing of the MCMC. Studies about the acceptance rates of the Metropolis-Hastings by Roberts et al. [1997] showed the optimal rate is around 30% under certain conditions, but this result varies with the dimension of the model parameters $\boldsymbol{\theta}$. Typically researchers target for a 40-60% acceptance rate as in Gelman et al. [2013].

More advanced techniques for Bayesian computation are also available, such as hybrid Monte Carlo by Duane et al. [1987], slice sampling by Neal [2003], simulated tempering by Geyer and Thompson [1993] and reversible jump MCMC by Green [1995] among others.

1.2.2 Expectation-Maximization Algorithm

The Expectation-Maximization (EM) algorithm was introduced in Dempster et al. [1977] and has since become a very popular tool for calculating Maximum Likelihood Estimators (MLE) and posterior modes. Suppose that for some probability distribution with parameter $\boldsymbol{\theta}$, there is data X which is incomplete. If the complete data Z = (X, Y) where Y is the missing data, then the likelihood function can be written as

$$\begin{aligned} p(z|\boldsymbol{\theta}) &= p(x, y|\boldsymbol{\theta}) \\ &= p(y|x, \boldsymbol{\theta}) \times p(x|\boldsymbol{\theta}). \end{aligned} \tag{1.14}$$

Y may be either actual missing data or random variables used to facilitate easy computation. The EM algorithm allows one to find the MLE of $p(x|\boldsymbol{\theta})$ by working with $p(z|\boldsymbol{\theta})$. It involves two steps. First, in the E-step, one finds the expected value of the log likelihood, $\log[p(z|\boldsymbol{\theta})]$. The expectation is taken with regard to the

conditional distribution of Y given the observed data X and the current parameter estimates $\boldsymbol{\theta}^{(t)}$. That is:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) \quad = \quad E_{p(y|x,\boldsymbol{\theta}^{(t)})} \left[\log p(x, y|\boldsymbol{\theta})\right], \tag{1.15}$$

where $p(y|x, \boldsymbol{\theta}^{(t)})$ is the conditional density of Y give the actual data X and $\boldsymbol{\theta}^{(t)}$. Next, in the M-step, $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ is maximized, which gives a new estimate of $\boldsymbol{\theta}$, $\boldsymbol{\theta}^{(t+1)}$. That is

$$\boldsymbol{\theta}^{(t+1)} \quad = \quad arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}). \tag{1.16}$$

These two steps are repeated until convergence and then an MLE solution is obtained. The algorithm is guaranteed to converge to a local maximum of the likelihood function as in Wu [1983], Redner and Walker [1984], Jordan and Xu [1996] and Xu and Jordan [1996].

Under the Bayesian framework, MAP estimator can be found at the mode of the marginal posterior distribution. Similar to finding the MLE, the EM algorithm can help to find the MAP without the need to explicitly manipulate the marginal posterior $p(\boldsymbol{\theta}|x)$. If the likelihood in Equation 1.15 is replaced by the posterior distribution, then

$$R(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) \quad = \quad E_{p(y|x,\boldsymbol{\theta}^{(t)})} \left[\log p(x, y|\boldsymbol{\theta})\right] + \log \pi(\boldsymbol{\theta}), \tag{1.17}$$

where $\pi(\boldsymbol{\theta})$ is the prior distribution. Applying the EM algorithm with $R(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ can give the MAP estimate.

The EM algorithm is typically applied under two situations. The first one is when the data indeed has missing values. The second one occurs when there are no missing values, however, direct optimization of the likelihood function/marginal

posterior distribution is not easy but this process can be simplified by introducing missing or latent variables. In this dissertation we utilize the latter of the two approaches.

Marschner [2010] described an extension of the EM algorithm, in which the observed data is treated as a summation of some latent/unobserved variables. He did not find the expectation of the log likelihood directly. Instead, he found the expectations of the latent variables and substituted these expectations into the log likelihood corresponding to the unobserved variables. These expectations are taken with regard to the conditional distribution of the unobserved variables given the observed data and the current parameter estimates. In Chapter 3, I will follow this extension to find MAP estimates.

Chapter 2

Reference Priors for Means with Common Order Restrictions

The prior distribution plays a central role in Bayesian analysis and statisticians spend a considerable amount of time looking for a prior that suits their needs (subjective, objective, or other). In data analyses, a common situation is that the analyst has some known *a priori* information about the parameters. For example, in many applications inequality constraints among population means θ_i, $i = 1, 2, \ldots, k$, may be adopted. Some common order restrictions of interest are,

Simple order: $\theta_1 < \theta_2 < \ldots < \theta_k$ (increasing or decreasing).

Simple tree order: $\theta_1 < \theta_i,\ i = 2, \ldots, k$ (no constraint among $\theta_i's$).

Umbrella order(with peak at i): $\theta_1 < \theta_2 < \ldots < \theta_i > \theta_{i+1} > \ldots > \theta_k$.

One example, explored by Morrissette and McDermott [2013], concerns patient outcomes and drug dosages. It may be known that the effect of the placebo is lower than any effects corresponding to dosage amounts of a drug (simple tree order). Another reasonable assumption is that higher dosages correspond to a larger

effect (simple order). Obviously, incorporating this additional information into a prior distribution is extremely attractive as it can produce better inference for the parameters, especially when the sample size is small and the variability of the data is large.

When presented with this information, the statistician must somehow incorporate it into a functional form for a prior. One option is to select a subjective prior and simply add the ordering restrictions. However, unless care is taken in the subjective prior elicitation, the resulting prior may be much more influential than originally envisioned. A similar problem can occur when the constraints are naively applied to a standard non-informative prior.

In this work, I utilize the reference prior framework of Berger and Bernardo [1992] to construct reference priors conditional on these common order restrictions. The derivation of the reference priors involves the typical sequential maximization of the Kullback-Leibler divergence between the posterior and the prior, which utilizes an iterative algorithm and requires model parameters to be grouped and ordered by inferential importance. A reference prior is then derived for the given likelihood function, conditional on the specified grouping and ordering. Sonksen and Peruggia [2012] constructed prior distributions on the occurrence rates for count data which accommodate a monotonic relationship between the rates and a single covariate. In Sonksen and Peruggia [2014], this idea was extended to multiple covariates. Following a similar path, I developed the reference priors for models with different ordered group means.

The rest of this chapter is organized as follows: In Section 1, an example of the reference prior with increasing normal means is described. In Section 2, the general expressions for means under common ordering constraints are derived and discussed. In Section 3, the performance of the reference priors is evaluated in simulation studies, with comparison to Jeffreys' prior and Least Squares Estimation (LSE).

2.1 Reference Prior Derivation: A Simple Example

In this section, I will show how to derive a reference prior in detail with a predetermined grouping and ordering. Suppose $X_{ij} \overset{ind}{\sim} N(\mu_i, 1)$, $i = 1, 2, 3$, $j = 1, \ldots, n$ and $\mu_1 < \mu_2 < \mu_3$. We define $\boldsymbol{\theta} = \{\mu_1,\ \mu_2,\ \mu_3\}$, with $\boldsymbol{\theta} \in \Theta_{Incr} = \{\boldsymbol{\theta}\colon -\infty < \mu_1 < \mu_2 < \mu_3 < \infty\}$. Since Θ_{Incr} is noncompact, from Chapter 1 we know it is useful to define a compact subset $\Theta^l = \{\boldsymbol{\theta}\colon -l < \mu_1 < \mu_2 < \mu_3 < l\}$. The log-likelihood function is

$$\log[L(\boldsymbol{\theta})] = \sum_{i=1}^{3}\sum_{j=1}^{n}\left[-\frac{1}{2}(X_{ij}-\mu_i)^2\right] + c. \tag{2.1}$$

The diagonal elements of the Fisher information matrix, $I(\boldsymbol{\theta})$, are

$$\begin{aligned} I_{kk}(\boldsymbol{\theta}) &= -E_{\boldsymbol{\theta}}\left(\frac{\partial^2}{\partial {\mu_k}^2}\log[L(\boldsymbol{\theta})]\right) \\ &= n. \end{aligned}$$

The non-diagonal elements of $I(\theta)$ are

$$\begin{aligned} I_{kl}(\boldsymbol{\theta}) &= -E_{\boldsymbol{\theta}}\left(\frac{\partial^2}{\partial\mu_k\partial\mu_l}\log[L(\boldsymbol{\theta})]\right) \\ &= 0. \end{aligned}$$

So,

$$I(\boldsymbol{\theta}) = \begin{pmatrix} n & 0 & 0 \\ 0 & n & 0 \\ 0 & 0 & n \end{pmatrix}. \tag{2.2}$$

I will show that with different ordering and grouping for the parameters, one can end up with different reference priors. If the parameters are grouped into 2 groups and ordered as $\boldsymbol{\theta} = (\boldsymbol{\theta}_{(1)},\ \boldsymbol{\theta}_{(2)}) = (\{\mu_1,\ \mu_3\},\ \{\mu_2\})$, then as defined in Chapter 1,

$$\begin{aligned} S(\boldsymbol{\theta}) &= [I(\boldsymbol{\theta})]^{-1} \\ &= \begin{pmatrix} 1/n & 0 & 0 \\ 0 & 1/n & 0 \\ 0 & 0 & 1/n \end{pmatrix}. \end{aligned} \tag{2.3}$$

$$h_1(\boldsymbol{\theta}) = \begin{pmatrix} n & 0 \\ 0 & n \end{pmatrix} \textit{ and } \ h_2(\boldsymbol{\theta}) = n.$$

The condition that $|h_j(\boldsymbol{\theta})|$ depends only on $\boldsymbol{\theta}_{(1:j)}$ is not violated here, for $j = 1, 2$. So we have

$$\pi^l(\boldsymbol{\theta}) = \frac{\Pi_{j=1}^2 |h_j(\boldsymbol{\theta})|^{1/2}}{\Pi_{j=1}^2 \int_{\Theta^l \cap [\Theta_j | \Theta_{(1:(j-1))}]} |h_j(\boldsymbol{\theta})|^{1/2} d\boldsymbol{\theta_{(j)}}} I_{\Theta^l}(\boldsymbol{\theta}). \tag{2.4}$$

For j=1,

$$\Theta^l \cap [\Theta_j | \Theta_{(1:(j-1))}] = \{(\mu_1, \mu_3) : -l < \mu_1 < l \textit{ and } \mu_1 < \mu_3 < l\}.$$

For j=2,

$$\Theta^l \cap [\Theta_j | \Theta_{(1:(j-1))}] = \{\mu_2 : \mu_1 < \mu_2 < \mu_3\}.$$

So,

$$\begin{aligned} \pi^l(\boldsymbol{\theta}) &= \frac{n}{\int_{\mu_1}^{l} \int_{-l}^{l} n d\mu_1 d\mu_3} \frac{n^{1/2}}{\int_{\mu_1}^{\mu_3} n^{1/2} d\mu_2} \\ &= \frac{1}{2l(l-\mu_1)} \frac{1}{(\mu_3 - \mu_1)}. \end{aligned}$$

Finally,

$$\begin{aligned} \pi_{\text{ref}}(\boldsymbol{\theta}) &= \lim_{l\to\infty} \frac{\pi^l(\boldsymbol{\theta})}{\pi^l(\boldsymbol{\theta}^*)} \\ &\propto \frac{1}{(\mu_3-\mu_1)} \times I_{\Theta_{Incr}}(\boldsymbol{\theta}), \end{aligned} \tag{2.5}$$

where $\boldsymbol{\theta}^*$ is any fixed point in Θ with positive density for all π^l, which is a constant with regard to $\boldsymbol{\theta}$. On the other hand, if the parameters are grouped into 3 groups and ordered as $\boldsymbol{\theta} = (\boldsymbol{\theta}_{(1)},\ \boldsymbol{\theta}_{(2)},\ \boldsymbol{\theta}_{(3)}) = (\{\mu_1\},\ \{\mu_2\},\ \{\mu_3\})$, then $h_1(\boldsymbol{\theta}) = h_1(\boldsymbol{\theta}) = h_1(\boldsymbol{\theta}) = n$. Similarly,

$$\begin{aligned} \pi^l(\boldsymbol{\theta}) &= \frac{n^{1/2}}{\int_{-l}^{l} n^{1/2} d\mu_1} \frac{n^{1/2}}{\int_{\mu_1}^{l} n^{1/2} d\mu_2} \frac{n^{1/2}}{\int_{\mu_2}^{l} n^{1/2} d\mu_3} \\ &= \frac{1}{(2l)} \frac{1}{(l-\mu_1)} \frac{1}{(l-\mu_2)} \end{aligned}$$

and

$$\begin{aligned} \pi_{\text{ref}}(\boldsymbol{\theta}) &= \lim_{l\to\infty} \frac{\pi^l(\boldsymbol{\theta})}{\pi^l(\boldsymbol{\theta}^*)} \\ &\propto I_{\Theta_{Incr}}(\boldsymbol{\theta}). \end{aligned} \tag{2.6}$$

It is clear that with different ordering and grouping for the parameters, one can end up with different reference priors. Suggestions for common grouping and ordering can be found in Berger and Bernardo [1992] and Sonksen and Peruggia [2012] among others.

2.2 Reference Prior Derivation: General Formula for Different Constraints

As described in Section 2.1, given the likelihood function, reference priors for more than one parameter can be derived under different grouping and ordering, where the grouping and ordering are closely related to the inferential importance for each parameter. This usually means that there is no uniquely defined expression for reference priors. Under the common order restrictions, I derive closed-form general expressions of the reference priors under the normal likelihood functions, which have not been seen in any literature. With these reference priors, the resulting models are a compromise between using the subjective information and letting the data drive the inferences.

Let us assume $X_{ij} \overset{ind}{\sim} N(\theta_i, 1)$, with $i = 1, 2, \ldots, k$ and $j = 1, \ldots, n$. Let $\boldsymbol{\theta} = \{\theta_1,\ \theta_2, \ldots,\ \theta_k\}$ with $\boldsymbol{\theta} \in \Theta$ and the θ's follow a certain order. Setting the variance equal to 1 does not lose generality for this problem. Since Θ is noncompact, a compact subset Θ^l is needed, where l is any real number that denotes the boundary of the compact subset. As described in Section 1.1.1 in Chapter 1, the elements of $\boldsymbol{\theta}$ are first partitioned into m groups and ordered by relative inferential importance, which gives $\boldsymbol{\theta} = (\boldsymbol{\theta}_{(1)},\ \boldsymbol{\theta}_{(2)}, \ldots,\ \boldsymbol{\theta}_{(m)})$. Suppose that group j contains m_j elements, that is, $\boldsymbol{\theta}_{(j)} = (\theta_{j_1},\ \theta_{j_2}, \ldots,\ \theta_{j_{m_j}})$. Actually, the user is totally in control of the specific ordering and grouping, which may have a noticeable influence on the resulting prior distribution.

Paralleling the grouping and ordering that we have above, if we follow the same definition of $h_j(\boldsymbol{\theta})$ as in Chapter 1, the Fisher information matrix for this Gaussian likelihood can be written as

$$I(\boldsymbol{\theta}) = \text{diag}[h_1(\boldsymbol{\theta}),\ h_2(\boldsymbol{\theta}), \ldots,\ h_m(\boldsymbol{\theta})].$$

For this specific example, $h_j(\boldsymbol{\theta}) = \text{diag}[n,\ n, \ldots,\ n]_{m_j \times m_j}$. Let us define $\boldsymbol{\theta}_{(1:j)} = (\boldsymbol{\theta}_{(1)},\ \boldsymbol{\theta}_{(2)}, \ldots,\ \boldsymbol{\theta}_{(j)})$. Because our model is regular and the determinant of $h_j(\boldsymbol{\theta})$, $|h_j(\boldsymbol{\theta})| = n^{m_j}$, we can use the simplified expression for the reference prior that is given in Lemma 1 of Berger and Bernardo [1992] and obtain

$$\pi^l(\boldsymbol{\theta}) \ = \ \frac{\Pi_{j=1}^m |h_j(\boldsymbol{\theta})|^{1/2}}{\Pi_{j=1}^m \int_{\Theta^l \cap [\Theta_j | \Theta_{(1:(j-1))}]} |h_j(\boldsymbol{\theta})|^{1/2} d\boldsymbol{\theta_{(j)}}} I_{\Theta^l}(\boldsymbol{\theta}), \tag{2.7}$$

where $[\Theta_j | \Theta_{(1:(j-1))}]$ is the parameter space of $\boldsymbol{\theta}_{(j)}$ given $\boldsymbol{\theta}_{(1:(j-1))}$.

To derive a general expression for the reference prior, we need to determine the integrals in the denominator of Equation 2.7. We define $\boldsymbol{\theta}_{(1:j),\,k}$ to be the k^{th} element of the vector $\boldsymbol{\theta}_{(1:j)}$. The term $|h_j(\boldsymbol{\theta})|^{1/2}$ can be canceled out from Equation 2.7 because it is only a function of n. Under regularity conditions, if the Fisher information matrix of the model satisfies Lemma 1 in Berger and Bernardo [1992], careful calculation can prove the following innovative theorems:

Theorem 1. *For a simple order, $\theta_1 < \theta_2 < \ldots < \theta_k$,*

$$\pi^l(\boldsymbol{\theta}) \ \propto \ \frac{1}{\Pi_{j=2}^m (\gamma_j - \eta_j)^{m_j}} I_{\Theta^l}(\boldsymbol{\theta})$$

with

$$\gamma_{j+1} = \begin{cases} \min\limits_k \{\boldsymbol{\theta}_{(1:j),\,k} : \boldsymbol{\theta}_{(1:j),\,k} > \max[\boldsymbol{\theta}_{(j+1)}]\} & , \ if \ \max[\boldsymbol{\theta}_{(1:j)}] > \max[\boldsymbol{\theta}_{(j+1)}] \\ l & , \ if \ \max[\boldsymbol{\theta}_{(1:j)}] < \max[\boldsymbol{\theta}_{(j+1)}] \end{cases}$$

and

$$\eta_{j+1} = \begin{cases} \max\limits_k \{\boldsymbol{\theta}_{(1:j),\,k} : \boldsymbol{\theta}_{(1:j),\,k} < \min[\boldsymbol{\theta}_{(j+1)}]\} & , \ if \ \min[\boldsymbol{\theta}_{(1:j)}] < \min[\boldsymbol{\theta}_{(j+1)}] \\ -l & , \ if \ \min[\boldsymbol{\theta}_{(1:j)}] > \min[\boldsymbol{\theta}_{(j+1)}]. \end{cases}$$

Theorem 2. *For a simple tree order, $\theta_1 < \theta_i,\ i = 2, \ldots, k$,*

$$\pi(\boldsymbol{\theta}) \ \propto \ I_{(\theta_1 < \theta_i)}.$$

Theorem 3. *For an umbrella order,* $\theta_1 < \theta_2 < \ldots < \theta_i > \theta_{i+1} > \ldots > \theta_k$, *parameters in group* j *can be separately treated as (1) with increasing order and (2) with decreasing order, then,*

$$\pi^l(\boldsymbol{\theta}) \propto \frac{1}{\Pi_{j=2}^{m}(\gamma_{j1}-\eta_{j1})^{m_{j1}}(\gamma_{j2}-\eta_{j2})^{m_{j2}}} I_{\Theta^l}(\boldsymbol{\theta}).$$

γ_{j1}, η_{j1}, γ_{j2} and η_{j2} can be determined by the definitions in Theorem 1 with $m_{j1}+m_{j2}=m_j$. The final reference priors in the true parameter space in Theorems 1 and 3 can be obtained by making $l \to \infty$.

These three theorems are generalizations of the results in Sonksen and Peruggia [2012]. They provide the general expressions that can be used to determine the reference priors of any grouping and ordering. These innovative results turn to be important contributions of this dissertation. In fact, these results can be generalized to other likelihood with the same kernels, such as Poisson and Binomial, etc. as long as the regularity conditions are satisfied. In addition, if the variance σ^2 is introduced in the model, it can be grouped by itself and considered as the first grouping. Then the theorems derived above can be adopted without any adjustment.

2.3 Simulation Study with Specific Orderings and Groupings

In this section, I will apply reference priors to a balanced one-way analysis of variance (ANOVA) model with a simple order, which is

$$\boldsymbol{y} = X\boldsymbol{\beta}+\boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon}\sim N(\mathbf{0},\sigma^2 I),$$

where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \; X = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}_{n \times k}, \; \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix}$$

and $\mu_1 < \mu_2 < \ldots < \mu_k$ and $k \geq 3$.

To derive the reference priors for this model, the user has to specify the ordering and grouping of the parameters, which may have a considerable impact on the resulting prior distribution and further inference. Berger and Bernardo [1992] suggest to completely separate the parameters with groups of one element each. On the other hand, Sonksen and Peruggia [2012] follow the Nicholls and Jones [2001] approach and suggest that the primary attention should be given to the extreme parameters, i.e., μ_1 and μ_k. Based on the general expression derived in Theorem 1, I consider these two ways of grouping and ordering and label them as $\pi(\boldsymbol{\theta})_{uni}$ and $\pi(\boldsymbol{\theta})_{u1k}$, respectively. The resulting prior distributions are listed in Table 2.1.

Table 2.1: Reference priors for one-way ANOVA model with a simple order

Label	Parameter Grouping and Ordering	Reference Prior
$\pi(\boldsymbol{\theta})_{uni}$	$(\{\sigma^2\}, \{\mu_1\}, \ldots, \{\mu_k\})$	$\frac{1}{\sigma^2} \times I_\Theta(\boldsymbol{\theta})$
$\pi(\boldsymbol{\theta})_{u1k}$	$(\{\sigma^2\}, \{\mu_1,\ \mu_k\}, \{\mu_2,\ \ldots,\ \mu_{k-1}\})$	$\frac{1}{\sigma^2} \times \frac{1}{(\mu_k - \mu_1)^{k-2}} \times I_\Theta(\boldsymbol{\theta})$

Note: $\boldsymbol{\theta} = \{\boldsymbol{\beta},\ \sigma^2\} = \{\mu_1, \ldots, \mu_k,\ \sigma^2\}$. $\Theta = \{(\boldsymbol{\beta}, \sigma^2) : -\infty < \mu_1 < \ldots < \mu_k < +\infty \text{ and } 0 < \sigma^2 < +\infty\}$.

2.3.1 Estimation of the Models

In the following simulation studies, different n, k and σ values are considered for the ANOVA model. The parameters can be estimated by Least Squares Estimation (LSE), or by posterior means/medians under Jeffreys' prior, or the reference priors $\pi(\boldsymbol{\theta})_{uni}$ and $\pi(\boldsymbol{\theta})_{u1k}$. LSEs can be calculated from closed-form solutions with $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}$ and $\hat{\sigma}^2 = \frac{SSE}{n-k}$, where SSE $= \boldsymbol{y}'(I-M)\boldsymbol{y}$ and $M = X(X'X)^{-1}X'$ as in Christensen [2011].

The Jeffreys' prior for this model is $\pi(\boldsymbol{\theta})_J \propto \frac{1}{\sigma^2}$ without any order restriction. With this prior

$$\begin{aligned}
p(\boldsymbol{\beta},\sigma^2|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\times\pi(\boldsymbol{\theta})_J \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \times \exp\left[-\frac{(\boldsymbol{y}-X\boldsymbol{\beta})'(\boldsymbol{y}-X\boldsymbol{\beta})}{2\sigma^2}\right]\times\frac{1}{\sigma^2} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \times \exp\left[-\frac{(X\boldsymbol{\beta}-M\boldsymbol{y})'(X\boldsymbol{\beta}-M\boldsymbol{y})}{2\sigma^2}\right] \\
&\quad \times \exp\left[-\frac{(\boldsymbol{y}-M\boldsymbol{y})'(\boldsymbol{y}-M\boldsymbol{y})}{2\sigma^2}\right]\times\frac{1}{\sigma^2} \\
&\propto \left(\frac{1}{\sigma^2}\right)^{k/2} \times \exp\left[-\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'X'X(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{2\sigma^2}\right] \\
&\quad \times \exp\left[-\frac{(\boldsymbol{y}-M\boldsymbol{y})'(\boldsymbol{y}-M\boldsymbol{y})}{2\sigma^2}\right]\times\left(\frac{1}{\sigma^2}\right)^{\frac{n-k}{2}+1} \\
&= N\left(\hat{\boldsymbol{\beta}},\sigma^2(X'X)^{-1}\right)\times IG\left(\frac{n-k}{2},\frac{SSE}{2}\right) \\
&= p(\boldsymbol{\beta}|\sigma^2,\boldsymbol{y})\times p(\sigma^2|\boldsymbol{y}). \qquad (2.8)
\end{aligned}$$

This implies that with $\pi(\boldsymbol{\theta})_J \propto \frac{1}{\sigma^2}$ the marginal posterior of σ^2 has an exact form, which is an inverse gamma distribution, $IG\left(\frac{n-k}{2},\frac{SSE}{2}\right)$. The posterior mean for σ^2 can be calculated from this inverse gamma distribution, which is $\frac{\boldsymbol{y}'(I-M)\boldsymbol{y}}{n-k-2}$. The conditional distribution of $\boldsymbol{\beta}$ is a normal distribution with mean $\hat{\boldsymbol{\beta}}$ and variance $\sigma^2(X'X)^{-1}$. Further derivation can show the marginal posterior of $\boldsymbol{\beta}$ is a multivariate

t distribution, where

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{y}) &= \int_0^\infty p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) d\sigma^2 \\ &\propto \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \times \exp\left[-\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})' X' X (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + SSE}{2\sigma^2}\right] d\sigma^2 \\ &\propto \frac{\Gamma\left(\frac{n}{2}\right)}{\left(\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})' X' X (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + SSE}{2}\right)^{n/2}} \\ &\propto \frac{\Gamma\left(\frac{n-k+k}{2}\right)}{\left[1 + \frac{1}{n-k} \times \frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})' X' X (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{SSE/(n-k)}\right]^{(n-k+k)/2}}. \end{aligned} \tag{2.9}$$

Each individual parameter of $\boldsymbol{\beta}$ follows a non-central univariate t distribution, so

$$\frac{\boldsymbol{\beta}_i - \hat{\boldsymbol{\beta}}_i}{\sqrt{\frac{SSE}{n-k}(X'X)^{-1}_{ii}}} \sim t_{n-k}, \tag{2.10}$$

where t_{n-k} represents a central t distribution with $n-k$ degrees of freedom. With these exact marginal posterior distributions, Gibbs sampling is not computationally necessary (unless for calculating DIC). This implies the posterior means and credible intervals can be easily obtained from the t and inverse gamma distributions. It is worth pointing out that Jeffreys' prior and LSE do not agree when estimating σ^2 although they do match at $\hat{\boldsymbol{\beta}}$. For Jeffreys' prior the estimator of σ^2 is $\hat{\sigma}^2 = \frac{SSE}{n-k-2}$, while the LSE for σ^2 corresponds to a Bayesian estimate with a prior of the form $\pi(\boldsymbol{\theta}) \propto (\frac{1}{\sigma^2})^2$. This implies the estimates of σ^2 from Jeffreys' prior are systematically larger than the ones from LSE.

For our reference priors, since there is no easy way to derive the full conditionals for $\boldsymbol{\beta}$ under these two priors, I will rely on the Metropolis-Hastings algorithm and utilize independent truncated normal proposals centered at the previous iteration with a variance adjustment to achieve good acceptance rates and stable Markov chains. At the t-th iteration, the parameter μ_1^* is sampled from a truncated normal

density

$$TN\left(\mu_1^{(t-1)},\xi\right) \propto \frac{1}{\sqrt{2\pi\xi}}\exp\left[-\frac{1}{2\xi}(\mu_1^* - \mu_1^{(t-1)})\right] I_{\{-\infty,\,\mu_2^{(t-1)}\}}(\mu_1^*) \tag{2.11}$$

and the parameter μ_k^* is sampled from a truncated normal density

$$TN\left(\mu_k^{(t-1)},\xi\right) \propto \frac{1}{\sqrt{2\pi\xi}}\exp\left[-\frac{1}{2\xi}(\mu_k^* - \mu_k^{(t-1)})\right] I_{\{\mu_{k-1}^{(t-1)},\,\infty\}}(\mu_k^*). \tag{2.12}$$

μ_1^* and μ_k^* are accepted as $\mu_1^{(t)}$ and $\mu_k^{(t)}$ with a probability of $\alpha = \min\{1, r\}$ where r is calculated by

$$r \;=\; \frac{f(\boldsymbol{y}|\boldsymbol{\beta}^*, {\sigma^{(t-1)}}^2) \times \pi(\boldsymbol{\theta}^*) \times TN\left(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^{(*)},\xi\right)}{f(\boldsymbol{y}|\boldsymbol{\beta}^{(t-1)}, {\sigma^{(t-1)}}^2) \times \pi(\boldsymbol{\theta}^{(t-1)}) \times TN\left(\boldsymbol{\theta}^{(*)}|\boldsymbol{\theta}^{(t-1)},\xi\right)}. \tag{2.13}$$

$\mu_2,\ \mu_3,\ \ldots, \mu_{(k-1)}$ can be sampled and updated respectively by similar procedures. Notice that once a μ_i has been updated, it will be adopted as the new truncation limit for the next adjacent $\mu_{(i+1)}$. For σ^2, the full condition distribution can be derived as follows,

$$\begin{aligned} f(\sigma^2|\boldsymbol{\beta},\boldsymbol{y}) &\propto f(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2) \times \pi(\boldsymbol{\theta}) \\ &\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \times \exp\left[-\frac{(\boldsymbol{y}-X\boldsymbol{\beta})'(\boldsymbol{y}-X\boldsymbol{\beta})}{2\sigma^2}\right] \times \frac{1}{\sigma^2}, \end{aligned} \tag{2.14}$$

which is an inverse gamma distribution, $IG\left(\frac{n}{2}, \frac{(y-X\boldsymbol{\beta})'(y-X\boldsymbol{\beta})}{2}\right)$. It can be updated by a Gibbs sampling step after $\boldsymbol{\beta}$ is updated.

2.3.2 Simulation Studies

The total number of simulated studies is 1000. Each study is analyzed by obtaining 11000 MCMC iterations and the first 1000 iterations are treated as burn-in. With an acceptance rate around 0.4, the posterior medians are used as the Bayesian estimates.

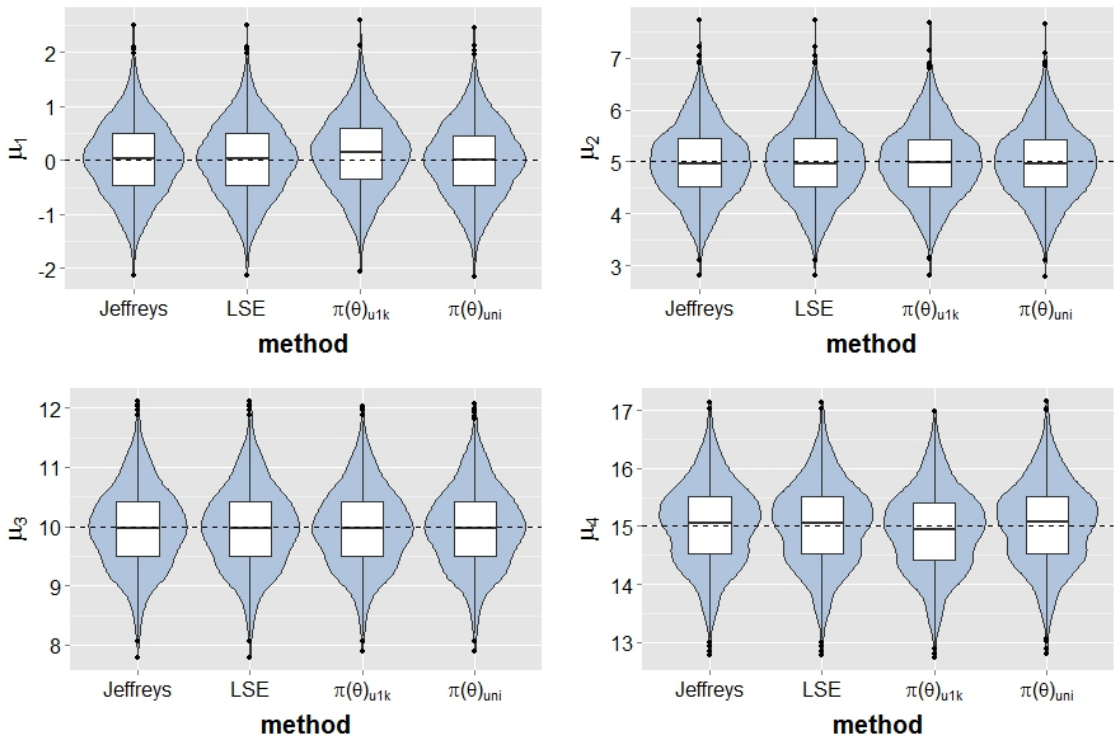

Figure 2.1: Point estimation of $\boldsymbol{\beta}$ by different methods: $n = 8$, $\sigma = 1$, $\boldsymbol{\beta} = (0,\ 5,\ 10,\ 15)'$.

For each study, the 95% credible or confidence intervals are determined and the true parameter values are checked to see if covered by the 95% intervals for each method. The empirical coverages of the intervals are then computed based on these 1000 simulations. The root mean square error (RMSE) for each parameter between estimates and real parameter value is also calculated. The average DIC from simulations is determined for each prior as an important tool for Bayesian model comparison and selection. The detailed settings and results are shown in the following figures and later discussed in detail.

Figure 2.1 shows the box plots of the μ estimates from 1000 simulations with $n = 8$, $\boldsymbol{\beta} = (0,\ 5,\ 10,\ 15)'$ and $\sigma = 1$. Under balanced design, there are two observations in each group. For this setting, the estimates from LSE and the Bayesian models

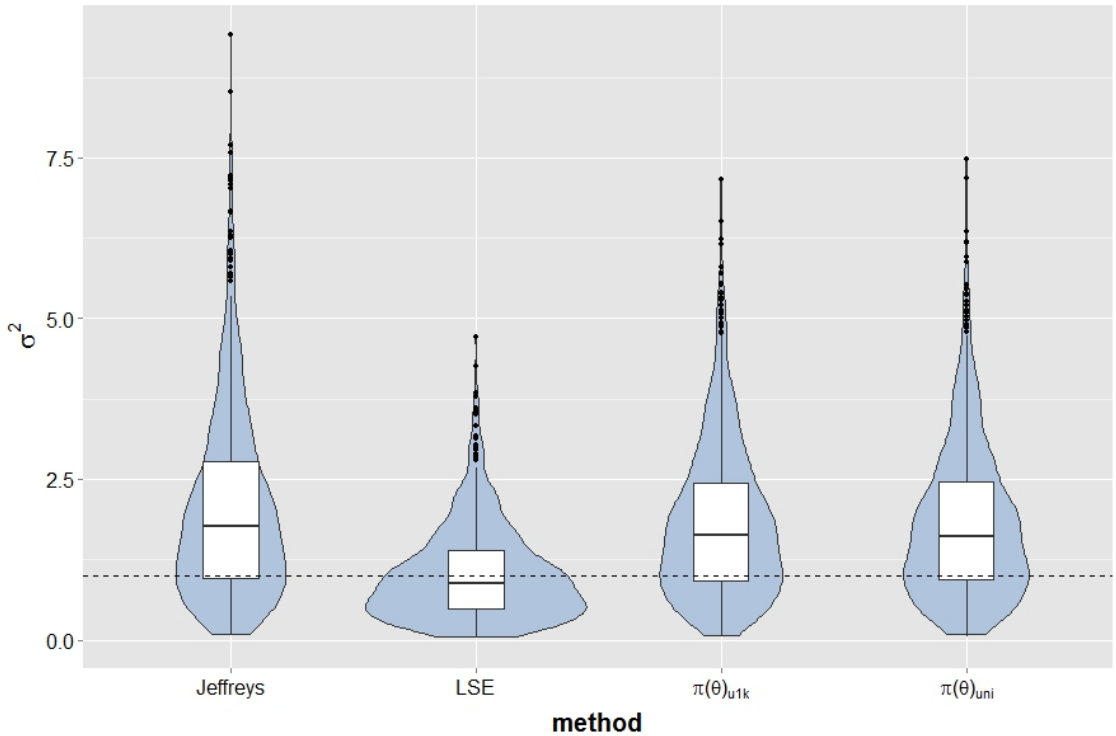

Figure 2.2: Point estimation of σ^2 by different methods: $n = 8$, $\sigma = 1$, $\boldsymbol{\beta} = (0,\ 5,\ 10,\ 15)'$.

are similar based on the box plots. The box plots of estimates for σ^2 are shown in Figure 2.2. For this parameter, it appears that the LSE gives the best result, while the results from Bayesian methods seem to have heavy tails regardless the choice of the priors. Jeffreys' prior is the worst and the two reference priors show a similar pattern. The results are not surprising in that Least Squares Estimation always performs well in an ANOVA model with equal variance and balanced design. Although our reference priors can take advantage of incorporating the internal order of the parameters, this effect may be ignorable as there are only four μ's in this model. The reference priors do give better estimations than Jeffreys' prior, however, it is worse when comparing with LSE.

Hence I decide to increase the number of parameters in the model. With more

parameters considered, the ordering information may become very important to the model fitting and the reference priors with this information may be able to show overwhelming dominance compared to Jeffreys' prior and LSE. I set $k = 10$ with $\boldsymbol{\beta} = (0, 1, 6, 7, 12, 13, 18, 19, 24, 25)'$ and $\sigma = 1$ or 5. Both balanced and unbalanced design are considered with the same sample size $n = 40$.

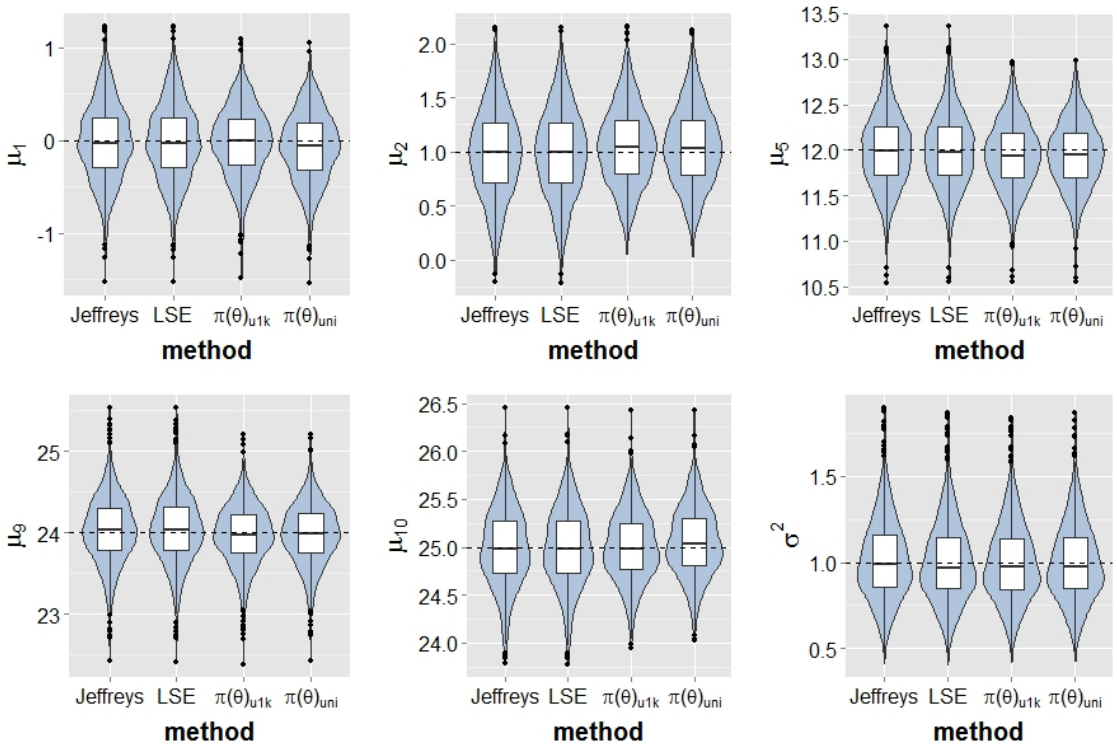

Figure 2.3: Point estimation of selected parameters by different methods: $n = 40$, $\sigma = 1$ and $k = 10$ with balanced design.

Figure 2.3 shows the simulation results with $\sigma = 1$ under balanced design. With the number of μ's increased to 10, the estimates from reference priors are more concentrated around the true values and show less variability, which is true for every μ in the figure. In addition, the reference priors perform at least as good as the LSE and Jeffreys' prior in estimating σ^2 as shown in the box plot. In order to clearly show the better performance of the reference priors in estimating the parameters,

Figure 2.4: RMSE comparison of different methods: $n = 40$, $\sigma = 1$ and $k = 10$ with balanced design.

I calculate the Root Mean Square Error (RMSE) between the estimates and true parameter values as shown in Figure 2.4. Smaller RMSE implies better estimation. It is clear from the figure that the reference priors consistently give smaller RMSE when estimating μ's, while the four methods provide similar RMSE for σ^2. This is an inspiring result as it confirms that when the number of parameters is large, incorporating the internal ordering information is important and helpful for model fitting. When the design is unbalanced or σ is increased to 5, a similar conclusion is drawn.

To further investigate the performance of the reference priors, I increase k to 20 with $\boldsymbol{\beta} = (0,\ 1,\ 6,\ 7,\ 12,\ 13,\ 18,\ 19,\ 24,\ 25,\ 30,\ 31,\ 36,\ 37,\ 42,\ 43,\ 48,\ 49,\ 54,\ 55)'$ and $\sigma = 1$ or 5. The sample size n is set to be 80 and both balanced and

unbalanced design are considered. A similar conclusion can be drawn, where the reference priors consistently give better estimates for the μ's in term of providing smaller RMSE and less variability. At the same time, they performs at least as well as LSE or even better when estimating σ^2. The figures are not shown here but they are similar as Figures 2.3 and 2.4. It is obvious that in this one way ANOVA model, when the number of parameters is large, the internal ordering information becomes important. Incorporating this ordering information into a prior distribution is helpful in model fitting and hence the reference priors can give good estimates with smaller uncertainty and RMSEs for all parameters.

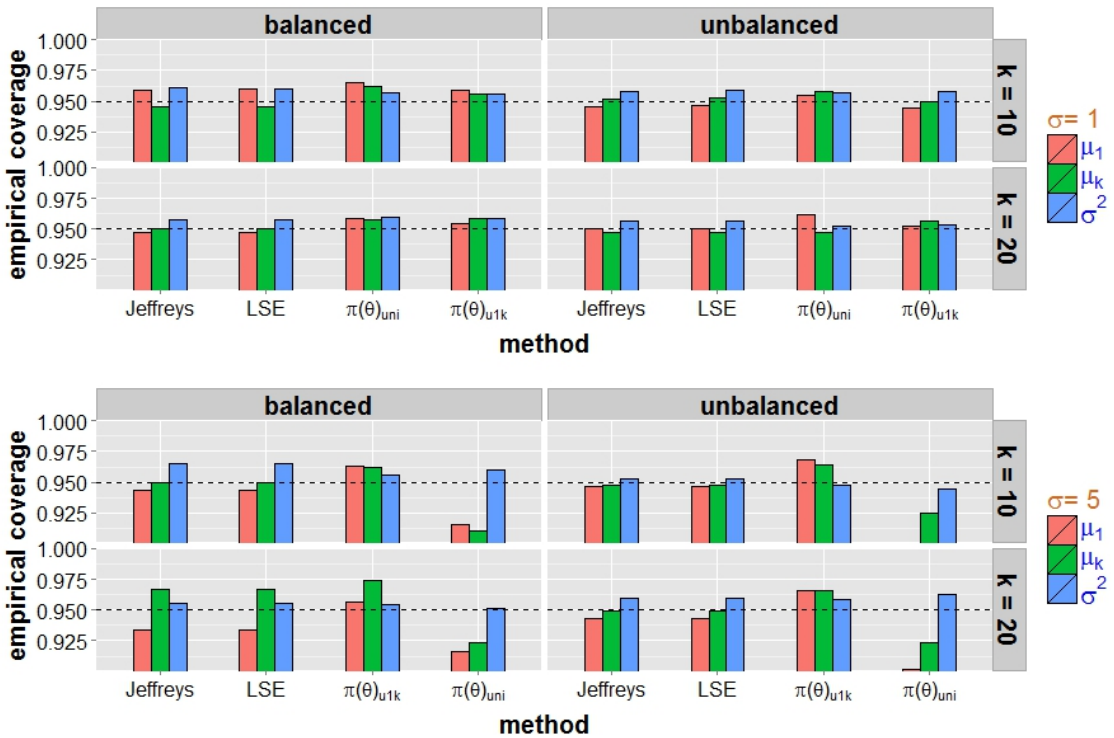

Figure 2.5: Empirical coverage of 95% CI under different settings and methods.

At the end, all the simulation results from 8 different settings are summarized in Figures 2.5-2.7 for $k = 10$ or 20, $\sigma = 1$, 5 and $n = 40$, 80. Figure 2.5 shows the

empirical coverages of 95% confidence or credible intervals. All these coverages look close to 0.95 except when $\sigma = 5$, the coverage of $\pi(\boldsymbol{\theta})_{uni}$ is relatively low especially for unbalanced design. This is in accordance with the fact that when the variance is large, the estimates from $\pi(\boldsymbol{\theta})_{uni}$ are somewhat off the true values for μ_1 and μ_k.

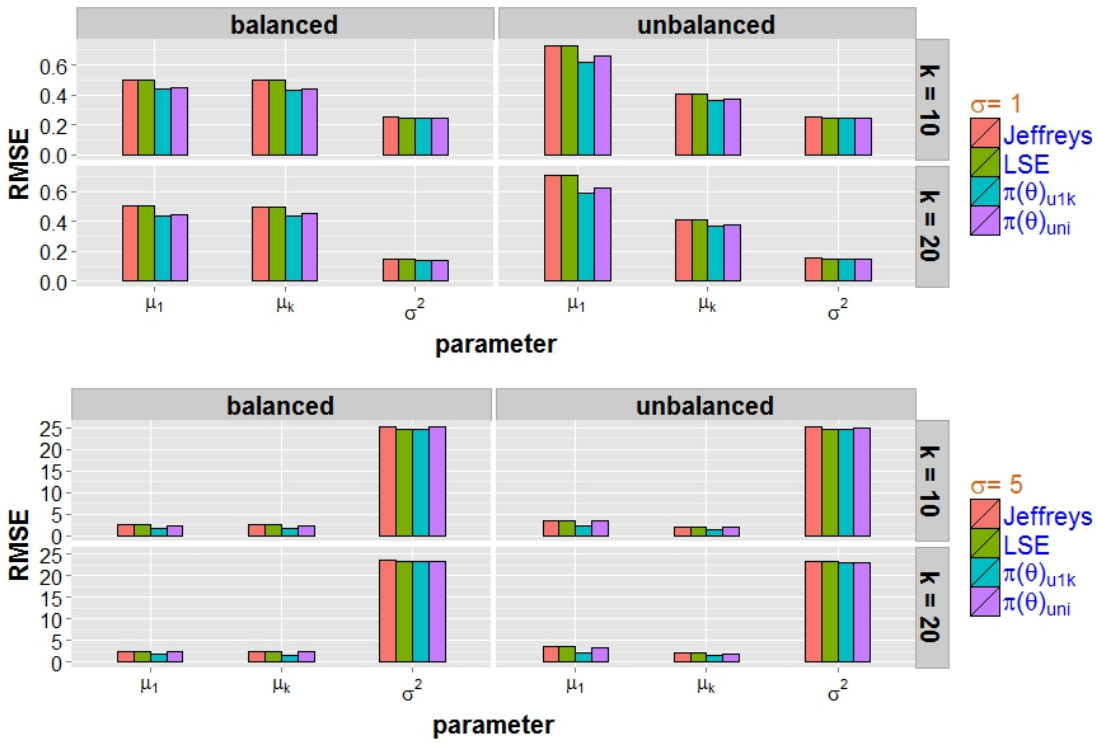

Figure 2.6: RMSE comparisons of μ_1, μ_k and σ^2 under different settings and methods.

Figure 2.6 shows the RMSEs of μ_1, μ_k and σ^2 under different settings. The reference prior $\pi(\boldsymbol{\theta})_{u1k}$ always gives the smallest RMSE when estimating σ^2, although the RMSEs from these four methods are actually close. When estimating μ_1 and μ_k, the reference prior $\pi(\boldsymbol{\theta})_{u1k}$ tends to give the smallest RMSE under all settings. The reference prior $\pi(\boldsymbol{\theta})_{uni}$ also gives pretty small RMSEs when $\sigma = 1$, however, this is not obvious when $\sigma = 5$.

Figure 2.7 shows simulation averages for DIC under the three Bayesian priors.

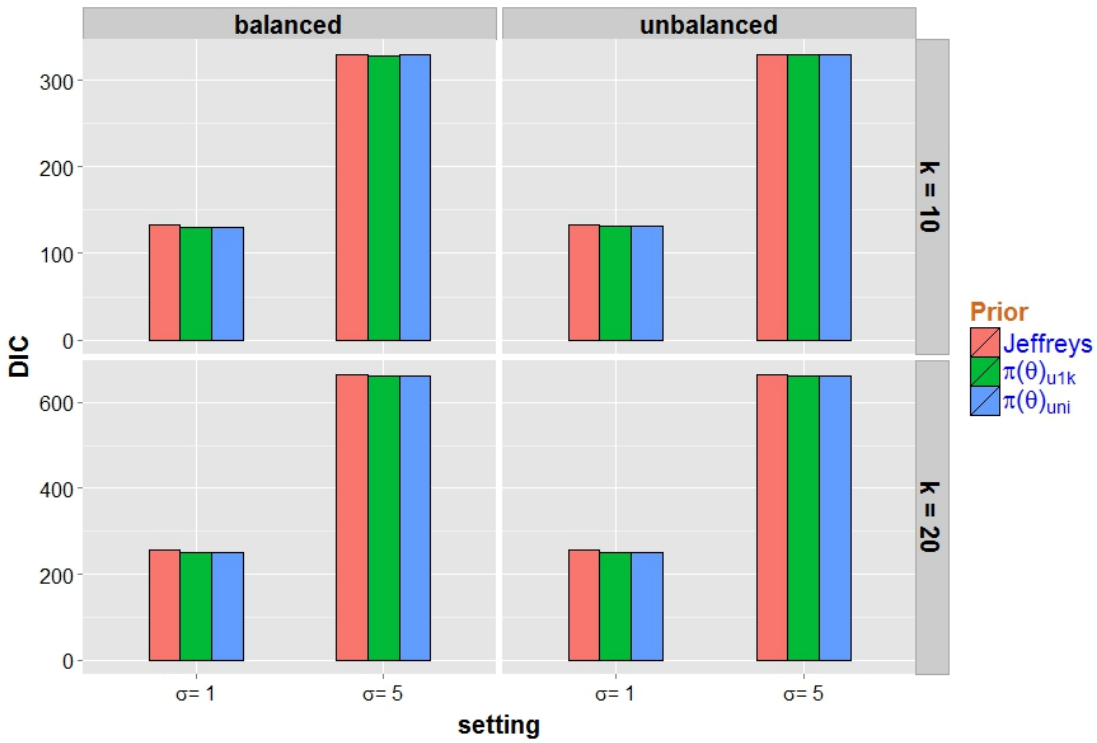

Figure 2.7: Average DIC comparisons of three Bayesian methods.

The calculation of DIC follows the method introduced in Chapter 1, where $\bar{D}$ is calculated from the average of $D(\boldsymbol{\theta})$, over the samples of $\boldsymbol{\theta}$ and $D(\bar{\boldsymbol{\theta}})$ is calculated as the value of D evaluated at the average of the samples of $\boldsymbol{\theta}$. Then the DIC can be determined. The resulting values are close and the reference prior $\pi(\boldsymbol{\theta})_{u1k}$ always gives the smallest average DIC value while Jeffreys' prior seems the worst.

Based on all our simulation results, we can conclude the reference priors that consider the internal order information are good choices when dealing with isotonic models, especially when the number of parameters is large. Under this situation the internal ordering information is important and the reference priors that incorporates this information stand out and work really well when looking for default priors.

Chapter 3

Constrained Reference Priors for Analysis of Covariance Model

Our reference priors presented in Chapter 2 can be easily extended to more complex models. For example, researchers can be conducting a regression analysis with several predictors and know that one of their categorical variables has a common ordering relationship with the response. Their inferences may be enhanced by formally incorporating this information. This model can be expressed as

$$\boldsymbol{y} \;=\; X_1\boldsymbol{\mu} + X_2\boldsymbol{\gamma} + X_3\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon}\sim N(\mathbf{0}, \sigma^2 I), \tag{3.1}$$

where $\boldsymbol{y} = (y_1,\ y_2, \ldots, y_n)'$ and $X_1 \;=\; (\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1k})$ is an $n \times k$ matrix of indicator variables designating the k levels of the grouping variable of interest. This means that the entry is 1 in $\boldsymbol{x}_{1i}$ if the observation belongs to group i, 0 otherwise and so forth. X_2 represents the $n \times r$ design matrix for the collection of r continuous and categorical covariates that do not interact with the grouping variable of interest. The final t continuous or categorical covariates that interact with the grouping variable of interest can be represented by an $n \times t$ matrix Z. Then $X_3 \;=\; (\boldsymbol{x}_{11} \times Z, \ldots, \boldsymbol{x}_{1k} \times Z)$,

where the j-th column of $\boldsymbol{x}_{1i} \times Z$ is the Hadamard product of the indicator vector $\boldsymbol{x}_{1i}$ with the jth column of Z, for $i = 1, \ldots, k$. $\boldsymbol{\mu}$ may follow one of the three order restrictions presented in Chapter 2, while there is no restriction for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

The model defined in Equation 3.1 can be expressed in a vector notation as

$$\boldsymbol{y} \;=\; X\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \tag{3.2}$$

where $X = (X_1,\ X_2,\ X_3)$ is an $n \times (k + r + kt)$ matrix, $\boldsymbol{\alpha} = (\boldsymbol{\mu}',\ \boldsymbol{\gamma}',\ \boldsymbol{\beta}')'$ is $(k + r + kt) \times 1$ vector and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$.

When the predictors in this model contains both categorical and quantitative variables, the model is known as an Analysis of Covariance (ANCOVA) model. The rest of this chapter is organized as follows: Section 3.1 shows the derivation for the general expression of the reference priors in ANCOVA models with a simple order restriction. Section 3.2 provides the results from simulation studies under different priors. Finally, in Section 3.3, a real data set application is considered for inference via LSE and Bayesian approaches.

3.1 Reference Prior Derivation: General Formula for an Increasing Constraint

For the model of Equation 3.2, we assume there is an increasing constraint (simple order) with regard to the parameter $\boldsymbol{\mu}$, which is $\mu_1 < \mu_2 < \ldots < \mu_k$. There are no restrictions imposed for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

Since the reference priors depend on different groupings and orderings, I will derive the general formula for the increasing constraint on $\boldsymbol{\mu}$. Based on the model

described by Equation 3.2, its log-likelihood function is

$$l(\boldsymbol{\alpha},\ \sigma^2) \quad = \quad -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{(\boldsymbol{y}-X\boldsymbol{\alpha})'(\boldsymbol{y}-X\boldsymbol{\alpha})}{2\sigma^2}. \tag{3.3}$$

To derive the Fisher information matrix, I first find the partial derivatives of the log-likelihood with respect to the model parameters and then find the expected values of these second derivatives. Therefore

$$\frac{\partial l(\boldsymbol{\alpha},\ \sigma^2)}{\partial \boldsymbol{\alpha}} \quad = \quad \frac{1}{\sigma^2}X'(\boldsymbol{y}-X\boldsymbol{\alpha}),$$

$$\frac{\partial^2 l(\boldsymbol{\alpha},\ \sigma^2)}{\partial \boldsymbol{\alpha}^2} = -\frac{1}{\sigma^2}X'X \Rightarrow -E\left[\frac{\partial^2 l(\boldsymbol{\alpha},\ \sigma^2)}{\partial \boldsymbol{\alpha}^2}\right] = \frac{1}{\sigma^2}X'X,$$

$$\frac{\partial^2 l(\boldsymbol{\alpha},\ \sigma^2)}{\partial \boldsymbol{\alpha}\partial\sigma^2} = -\frac{1}{\sigma^4}X'(\boldsymbol{y}-X\boldsymbol{\alpha}) \Rightarrow -E\left[\frac{\partial^2 l(\boldsymbol{\alpha},\ \sigma^2)}{\partial \boldsymbol{\alpha}\partial\sigma^2}\right] = 0,$$

$$\frac{\partial^2 l(\boldsymbol{\alpha},\ \sigma^2)}{\partial {\sigma^2}^2} = \frac{n}{2\sigma^4} - \frac{(\boldsymbol{y}-X\boldsymbol{\alpha})'(\boldsymbol{y}-X\boldsymbol{\alpha})}{\sigma^6} \Rightarrow -E\left[\frac{\partial^2 l(\boldsymbol{\alpha},\ \sigma^2)}{\partial {\sigma^2}^2}\right] = \frac{n}{2\sigma^4}.$$

So the Fisher information matrix is

$$I(\boldsymbol{\alpha},\sigma^2) \quad = \quad \begin{pmatrix} \frac{n}{2\sigma^4} & \mathbf{0} \\ \mathbf{0} & \frac{X'X}{\sigma^2} \end{pmatrix}. \tag{3.4}$$

The regularity conditions for asymptotic normality are still satisfied by this model. Note the Fisher information matrix does not depend on any elements of $\boldsymbol{\alpha}$. If the parameters, $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \sigma^2\}$, are grouped into m groups and ordered as: ($\{\sigma^2\}$, ...), then $|h_j(\boldsymbol{\theta})|$ depends only on $\boldsymbol{\theta}_{(1:j)}$, for $j = 1, \ldots, m$. This means we can use the simplified expression that was given in Berger and Bernardo [1992] again for the reference prior as in Chapter 2, which is

$$\pi^l(\boldsymbol{\theta}) \quad = \quad \frac{\Pi_{j=1}^m |h_j(\boldsymbol{\theta})|^{1/2}}{\Pi_{j=1}^m \int_{\Theta^l \cap [\Theta_j | \Theta_{(1:(j-1))}]} |h_j(\boldsymbol{\theta})|^{1/2} d\boldsymbol{\theta}_{(\boldsymbol{j})}} I_{\Theta^l}(\boldsymbol{\theta}), \tag{3.5}$$

where $h_j(\boldsymbol{\theta})$ and Θ^l follow the similar definitions as those introduced in Chapter 2.

Obviously $|h_1(\boldsymbol{\theta})| = \frac{n}{2\sigma^4}$ and other $|h_j(\boldsymbol{\theta})|$ terms can be canceled out from the numerator and denominator in Equation 3.5, because they are constants with regard to the integrals as the Fisher information matrix does not depend on $\boldsymbol{\alpha}$. Since there is no restriction for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, the integrals related to these two parameters are functions of the boundary l for the compact subset, Θ^l. They will be dropped when taking limit in Equation 3.7. The integrals related to $\boldsymbol{\mu}$ only depend on the grouping and ordering of $\boldsymbol{\mu}$ and the general expression for this kernel will be exactly the same as the one derived in Chapter 2, so

$$\begin{aligned} \pi^l(\boldsymbol{\theta}) &= \frac{\sqrt{\frac{n}{2\sigma^4}}}{\int_{1/l}^{l}\sqrt{\frac{n}{2\sigma^4}}d\sigma^2} \times \frac{1}{\Pi_{j=2}^{m}\int_{\Theta^l\cap[\Theta_j|\Theta_{(1:(j-1))}]} d\boldsymbol{\theta}_{(\boldsymbol{j})}} \times I_{\Theta^l}(\boldsymbol{\theta}) \\ &\propto \frac{1}{\sigma^2} \times \frac{1}{\Pi_{j=2}^{m_\mu}(\gamma_j-\eta_j)^{m_j}} \times I_{\Theta^l}(\boldsymbol{\theta}), \end{aligned} \tag{3.6}$$

where γ_j and η_j follow similar definitions as in Theorem 1. The actual reference priors can be found by

$$\pi_{\text{ref}}(\boldsymbol{\theta}) = \lim_{l\to\infty}\frac{\pi^l(\boldsymbol{\theta})}{\pi^l(\boldsymbol{\theta^*})}, \tag{3.7}$$

where $\boldsymbol{\theta}^*$ is any fixed point in Θ with positive density for $\pi^l(\boldsymbol{\theta})$, which is a constant with regard to $\boldsymbol{\theta}$.

3.2 Simulation Study with Specific Orderings and Groupings

In this section we consider a simulation study to get a better understanding of the properties of the posterior distribution under different reference priors for the proposed ANCOVA model. With the general expression, Equation 3.6, a specific prior

can be obtained under a given grouping and ordering of the parameters. Following the same rational as in Chapter 2, I consider two ways of grouping and ordering, which are shown in the following table along with the resulting reference priors. With these two ways of grouping and ordering, the resulting prior distributions have

Table 3.1: Reference priors for ANCOVA model with a simple order

Label	Parameter Grouping and Ordering	Reference Prior
$\pi(\boldsymbol{\theta})_{uni}$	$(\{\sigma^2\}, \{\mu_1\}, \ldots, \{\mu_k\}, \{\boldsymbol{\gamma}, \boldsymbol{\beta}\})$	$\frac{1}{\sigma^2} \times I_\Theta(\boldsymbol{\theta})$
$\pi(\boldsymbol{\theta})_{u1k}$	$(\{\sigma^2\}, \{\mu_1, \mu_k\}, \{\mu_2, \ldots, \mu_{k-1}\}, \{\boldsymbol{\gamma}, \boldsymbol{\beta}\})$	$\frac{1}{\sigma^2} \times \frac{1}{(\mu_k - \mu_1)^{k-2}} \times I_\Theta(\boldsymbol{\theta})$

Note: $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma^2\}$. Θ is the whole parameter space with $\mu_1 < \mu_2 < \ldots < \mu_k$.

a similar form to the priors for the ANOVA model of Chapter 2.

In my simulation study, I consider a categorical variable with a simple order and a continuous covariate which interacts with this categorical variable. The $\boldsymbol{\gamma}$ parameters are ignored since they typically are not the focus of inference. The parameters can be estimated by Least Squares Estimation (LSE), or by posterior means/medians under different priors or by MAP estimates.

Least Squares Estimators can be calculated directly by $\hat{\boldsymbol{\alpha}} = (X'X)^{-1}X'\boldsymbol{y}$ and $\hat{\sigma}^2 = \frac{SSE}{n-p}$ with SSE $= \boldsymbol{y}'(I - M)\boldsymbol{y}$ and $M = X(X'X)^{-1}X'$ as in Christensen [2011].

The Jeffreys' prior for this model is $\pi(\boldsymbol{\theta})_J \propto \frac{1}{\sigma^2}$ without any order restriction. As shown in Chapter 2, with this prior, the joint posterior, $p(\boldsymbol{\alpha}, \sigma^2|\boldsymbol{y})$, can be expressed as the product of $p(\boldsymbol{\alpha}|\sigma^2, \boldsymbol{y})$ and $p(\sigma^2|\boldsymbol{y})$, where $\boldsymbol{\alpha} = (\boldsymbol{\mu}', \boldsymbol{\gamma}', \boldsymbol{\beta}')'$. Further derivation can show that the marginal posterior of σ^2 is an inverse gamma distribution, $IG\left(\frac{n-p}{2}, \frac{\boldsymbol{y}'(I-M)\boldsymbol{y}}{2}\right)$. The marginal posterior of $\boldsymbol{\alpha}$ is a multivariate t distribution. Each individual parameter $\boldsymbol{\alpha}_i$ follows a non-central univariate t distribution, so

$$\frac{\boldsymbol{\alpha}_i - \hat{\boldsymbol{\alpha}}_i}{\sqrt{\frac{SSE}{n-p}\,(X'X)^{-1}_{ii}}} \quad \sim \quad t_{n-p}, \tag{3.8}$$

where t_{n-p} represents a central t-distribution with $n-p$ degrees of freedom. The posterior means and credible intervals can be easily obtained from the t and inverse gamma distributions.

Another alternative is to utilize the EM algorithm to find the MAP estimates as discussed in Chapter 1. For this model, it is convenient to define two latent independent variables, $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, such that $\boldsymbol{y} = \boldsymbol{y}_1 + \boldsymbol{y}_2$ with $\boldsymbol{y}_1 \sim N(X_1\boldsymbol{\mu}, \sigma^2 I/2)$ and $\boldsymbol{y}_2 \sim N(X_3\boldsymbol{\beta}, \sigma^2 I/2)$. It is obvious that $\boldsymbol{y}|\boldsymbol{y}_1, \boldsymbol{\theta} \sim N(\boldsymbol{y}_1 + X_3\boldsymbol{\beta}, \sigma^2 I/2)$ and $\boldsymbol{y}|\boldsymbol{y}_2, \boldsymbol{\theta} \sim N(\boldsymbol{y}_2 + X_1\boldsymbol{\mu}, \sigma^2 I/2)$. From Bayes theorem,

$$\begin{aligned} f(\boldsymbol{y}_1|\boldsymbol{y}, \boldsymbol{\theta}) &\propto f(\boldsymbol{y}_1, \boldsymbol{y}|\boldsymbol{\theta}) \\ &\propto f(\boldsymbol{y}|\boldsymbol{y}_1, \boldsymbol{\theta}) \times f(\boldsymbol{y}_1|\boldsymbol{\theta}) \\ &\propto \exp\left[-\frac{(\boldsymbol{y}_1 + X_3\boldsymbol{\beta} - \boldsymbol{y})'(\boldsymbol{y}_1 + X_3\boldsymbol{\beta} - \boldsymbol{y})}{\sigma^2}\right] \\ &\quad \times \exp\left[-\frac{(\boldsymbol{y}_1 - X_1\boldsymbol{\mu})'(\boldsymbol{y}_1 - X_1\boldsymbol{\mu})}{\sigma^2}\right], \end{aligned} \tag{3.9}$$

which clearly is a $N\left(\frac{\boldsymbol{y}+X_1\boldsymbol{\mu}-X_3\boldsymbol{\beta}}{2}, \frac{\sigma^2}{4}I\right)$ distribution. Similarly we have $\boldsymbol{y}_2|\boldsymbol{y}, \boldsymbol{\theta} \sim N\left(\frac{\boldsymbol{y}-X_1\boldsymbol{\mu}+X_3\boldsymbol{\beta}}{2}, \frac{\sigma^2}{4}I\right)$. As in Marschner [2010], in the E-step, I calculate the expected values for these two latent variables at the $(t+1)$-th iteration with regard to the estimates of the parameters at the t-th iteration. These expected values are

$$\begin{aligned} \hat{\boldsymbol{y}_1}^{(t+1)} &= E(\boldsymbol{y}_1|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(t)}) \\ &= \frac{\boldsymbol{y} + X_1\hat{\boldsymbol{\mu}}^{(t)} - X_3\hat{\boldsymbol{\beta}}^{(t)}}{2} \end{aligned} \tag{3.10}$$

and

$$\begin{aligned} \hat{\boldsymbol{y}_2}^{(t+1)} &= E(\boldsymbol{y}_2|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(t)}) \\ &= \frac{\boldsymbol{y} - X_1\hat{\boldsymbol{\mu}}^{(t)} + X_3\hat{\boldsymbol{\beta}}^{(t)}}{2}. \end{aligned} \tag{3.11}$$

The log posterior distribution for this model can be expressed as

$$\begin{aligned}\log f(\boldsymbol{\theta}|\boldsymbol{y}_1,\boldsymbol{y}_2) &\propto \log f(\boldsymbol{y}_1,\boldsymbol{y}_2|\boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta}) \\ &\propto -2n\log\sigma - \frac{1}{\sigma^2}\left[(\boldsymbol{y}_1 - X_1\boldsymbol{\mu})'(\boldsymbol{y}_1 - X_1\boldsymbol{\mu}) + (\boldsymbol{y}_2 - X_3\boldsymbol{\beta})'(\boldsymbol{y}_2 - X_3\boldsymbol{\beta})\right] \\ &\quad + \log \pi(\boldsymbol{\theta}). \end{aligned} \tag{3.12}$$

If $\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})_{uni}$, then $\log \pi(\boldsymbol{\theta})$ is a constant and can be dropped from Equation 3.12. The maximizer for this log posterior has a solution close to the MLE. In addition, $\boldsymbol{\mu}$ has to follow the simple ordering restriction, which can be controlled along the iterations of the EM algorithm. On the other hand, if $\pi(\boldsymbol{\theta})_{u1k}$ is used, the prior part cannot be dropped from the log posterior because it is not a constant. When estimating $\boldsymbol{\mu}$, only the estimation for μ_1 and μ_k will be affected by this change because the other μ's are not part of the prior. Hence, here I focus on the estimation of μ_1 and μ_k under the prior $\pi(\boldsymbol{\theta})_{u1k}$. For the following M-step, the expectations from Equations 3.10 and 3.11 are substituted into the log posterior distribution of Equation 3.12 to give

$$\begin{aligned}\log f(\boldsymbol{\theta}|\boldsymbol{y}_1,\boldsymbol{y}_2) &\propto \log f(\boldsymbol{y}_1,\boldsymbol{y}_2|\boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta}) \\ &= -2n\log\sigma - \frac{1}{\sigma^2}(\hat{\boldsymbol{y}}_1^{(t+1)} - X_1\boldsymbol{\mu})'(\hat{\boldsymbol{y}}_1^{(t+1)} - X_1\boldsymbol{\mu}) \\ &\quad -\frac{1}{\sigma^2}(\hat{\boldsymbol{y}}_2^{(t+1)} - X_3\boldsymbol{\beta})'(\hat{\boldsymbol{y}}_2^{(t+1)} - X_3\boldsymbol{\beta}) - (k-2)\log(\mu_k - \mu_1) - 2\log\sigma. \end{aligned} \tag{3.13}$$

Maximizing this expression gives the new estimates of the parameters at $t+1$. The maximization process include taking partial derivative for each parameter and then making the derivative equal to zero. It should be pointed out when dealing with μ_1 and μ_k, there are no close form solutions. Instead one remedy could invoke the ECM to find the conditional MAP estimate for μ_1 given $\mu_k = \hat{\mu}_k^{(t)}$ and then find the conditional MAP estimate for μ_k given $\mu_1 = \hat{\mu}_k^{(t+1)}$.

Compared to MCMC methods, the EM algorithm converges really fast. On the other hand, it does not produce standard errors for parameters automatically and requires extra work to obtain these. I did not compute standard errors for the EM estimates for this work. However, a parametric bootstrap could be easily employed to obtain these in the case of the ANCOVA model.

For a full Bayesian analysis with our reference priors, since there is no easy way to derive the full conditionals for $\boldsymbol{\mu}$, I rely on the Metropolis-Hastings algorithm and adopt independent truncated normal proposals centered at the previous iteration with a variance selected through trial and error to achieve good acceptance rate and a stable Markov chain. Similar to Chapter 2, at the t-th iteration, μ_i^* is sampled from a truncated normal density $TN(\mu_i^{(t-1)}, \xi_i)$ and accepted as $\mu_i^{(t)}$ with a probability of $\alpha = \min\{1, r_i\}$ where r_i can be calculated by

$$r_i = \frac{f(\boldsymbol{y}|\boldsymbol{\mu}^*, \boldsymbol{\beta}^{(t-1)}, {\sigma^{(t-1)}}^2) \times \pi(\boldsymbol{\theta}^*) \times TN(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^{(*)}, \xi)}{f(\boldsymbol{y}|\boldsymbol{\mu}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, {\sigma^{(t-1)}}^2) \times \pi(\boldsymbol{\theta}^{(t-1)}) \times TN(\boldsymbol{\theta}^{(*)}|\boldsymbol{\theta}^{(t-1)}, \xi)}. \tag{3.14}$$

The truncation limits are set to be between the two adjacent $\mu's$ from last iteration to make sure the resulting μ_i follow the simple order. Once μ_i has been updated, it will be adopted as the new truncation limit for the adjacent $\mu_{(i+1)}$. All other parameters can be updated via direct sampling of full conditionals, which are shown in Appendix A.

For the simulation studies, k = 10, 20 and σ =1, 5 with n = 60 or 120 are considered. The detailed settings are similar to the studies in Chapter 2 in that both balanced and unbalanced designs (with respect to the categorical variable) are considered. The total number of simulated studies for each setting is 1000. Each study is analyzed by obtaining 11000 MCMC iterations and the first 1000 iterations are treated as burn-in. With an acceptance rate at around 0.4, the posterior medians under different priors are calculated as the Bayesian estimates.

At each study, the 95% credible or confidence intervals are determined and the true parameter values are checked to see if covered by the 95% intervals for each method. The empirical coverages of the intervals are then computed based on these 1000 simulations. The root mean square error (RMSE) for each parameter between estimates and real parameter value is also calculated. The average DIC for each prior is determined as an important tool for Bayesian model comparison and selection. The detailed results are listed in Figures 3.1-3.3.

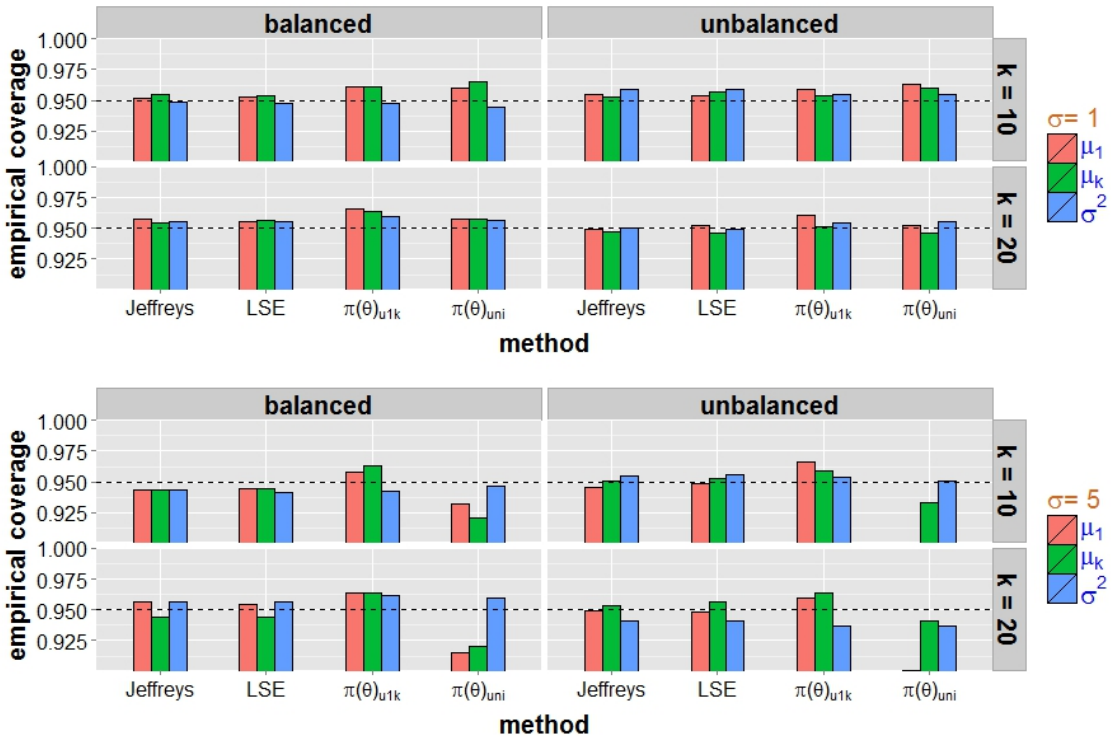

Figure 3.1: Empirical coverage of 95% CI under different settings and methods.

As discussed previously, $\boldsymbol{\mu}$, $\boldsymbol{\beta}$ and σ^2 are estimated and compared with the true values. Since there is no order restriction for $\boldsymbol{\beta}$, the four methods could give similar estimations for this parameter(s), which is confirmed by my simulation studies. Also, the simulation studies show that the performances for each method are reasonably

similar to what I have presented in Chapter 2 except the DIC comparisons.

Figure 3.1 shows the empirical coverages of 95% confidence or credible intervals. All these coverages look close to 0.95 except when $\sigma = 5$, the coverage of $\pi(\boldsymbol{\theta})_{uni}$ is relatively low especially for unbalanced design. This is in accordance with the fact that when the variance is large, the estimates from $\pi(\boldsymbol{\theta})_{uni}$ are somewhat off the true values for μ_1 and μ_k.

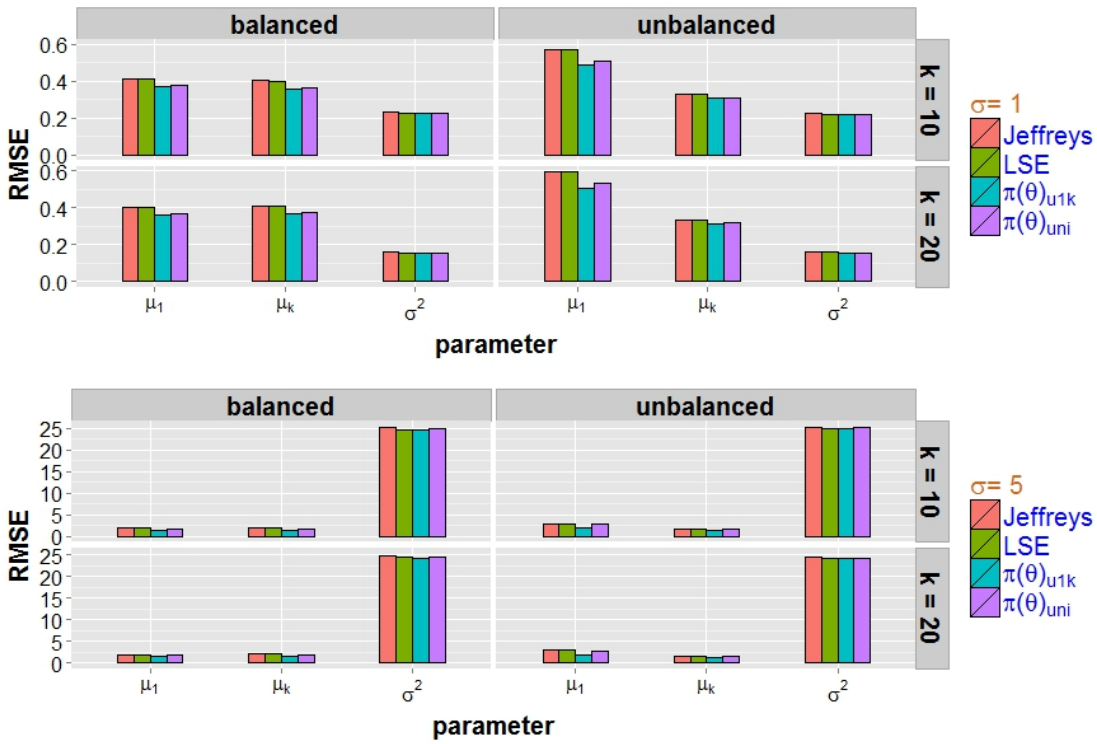

Figure 3.2: RMSE comparisons of μ_1, μ_k and σ^2 under different settings and methods.

Figure 3.2 shows the RMSEs of μ_1, μ_k and σ^2 under different settings. The reference prior $\pi(\boldsymbol{\theta})_{u1k}$ always gives the smallest RMSE when estimating σ^2, although the RMSEs from these four methods are actually close. When estimating μ_1 and μ_k, the reference prior $\pi(\boldsymbol{\theta})_{u1k}$ tends to give the smallest RMSE under all settings. The reference prior $\pi(\boldsymbol{\theta})_{uni}$ also gives pretty small RMSEs when $\sigma = 1$, however, this is

not obvious when $\sigma = 5$.

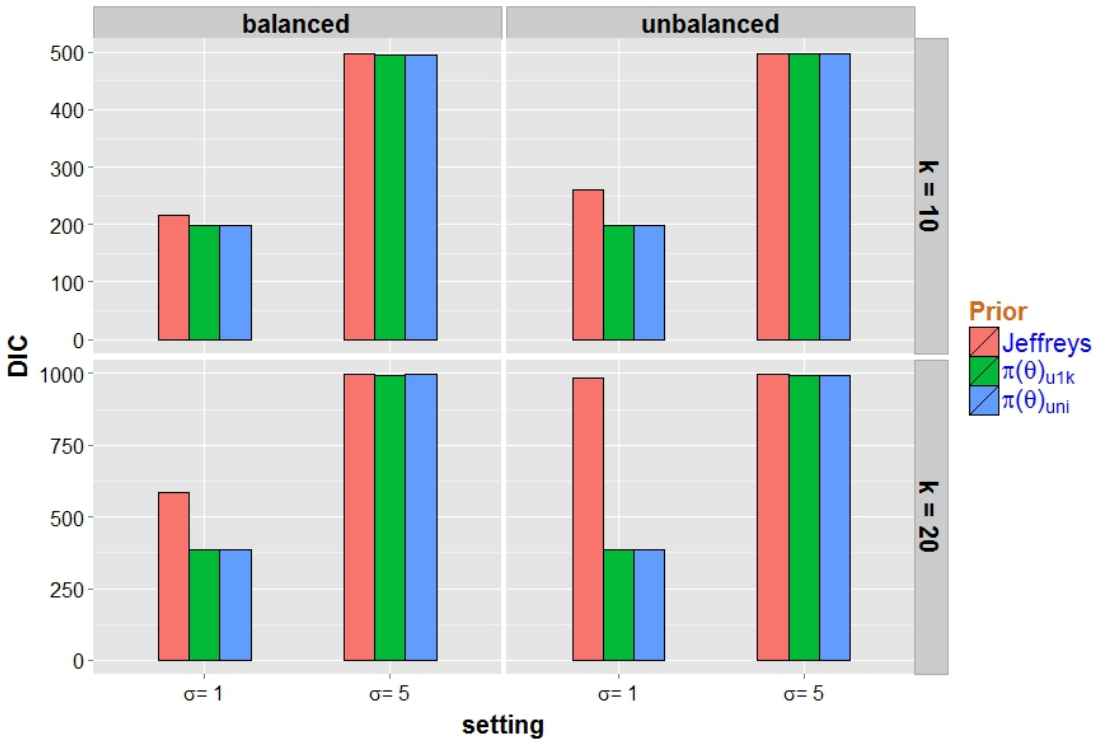

Figure 3.3: Average DIC comparisons of three Bayesian methods.

Figure 3.3 shows simulation averages for DIC under the three Bayesian priors. The reference priors always give small average DIC values while Jeffreys' prior seems the worst. This difference is obvious at σ =1, especially when $k = 20$.

Based on all our simulation results, we can conclude the reference priors that consider the internal order information are good choices when dealing with isotonic models, especially when the number of parameters is large. Under this situation the internal ordering information is important and the reference priors that incorporates this information stand out and work really well when looking for default priors. Based on all our simulation results, we can conclude the reference priors that consider the internal order information are good choices when dealing with this model, especially

when the number of parameters is large. Under this situation the internal ordering information is important. It seems a reference prior incorporating the internal order information is a good choice in a sense of giving good estimates around the true values with smaller uncertainty and RMSE, which makes it a possible default prior in ANCOVA model.

3.3 Application of Reference Priors: Smoking and Type 2 Diabetes

3.3.1 Introduction

Smoking tends to induce high risk of having type 2 diabetes [Xie et al., 2009]. At the same time, people with diabetes who smoke are more likely than nonsmokers to have trouble with insulin dosing. However, this relationship sometimes is masked by the variability in observational data. That is, smoking may show a non-significant effect, which will certainly hamper the interpretation of the model. In this section, I assume internal order information and that there is a simple order for the mean responses of smoking levels. I construct an ANCOVA model for a real data set to investigate the relationship of type 2 diabetes and smoking along with several other covariates. I adopt the reference priors derived in previous sections and perform an analysis under a Bayesian framework. For comparisons, I also consider Jeffreys' prior and LSE approaches.

3.3.2 Risk of Type 2 Diabetes in New Mexico

Diabetes is a major public health problem in the state of New Mexico. It is one of the ten leading causes of death in the state. During the last 15 years, the number

of persons having diabetes has increased dramatically, which now affects the health of about 10 percent of the adult population in New Mexico as in Centers for Disease Control and Prevention [2012]. Furthermore, this also can place a tremendous financial burden on the state. The total cost of diabetes in New Mexico has already exceeded $ 1.25 billion per year as described by Juvenile Diabetes Research Foundation. Deep studies in diabetes are needed to benefit the whole state economically and medically. In 2013 Dr. Mark Burge at UNM Health Science Center started a study with regard to type 2 diabetes. In this study, 218 adults in New Mexico at risk for type 2 diabetes were screened to determine their glucose homeostasis status. Hemoglobin A1c (HbA1c), a common variable used to measure diabetes status, was measured for each patient along with other predictors. Generally, a person with HbA1c $\leq 5.4\%$ can be recognized as normal or with no diabetes. A person with HbA1c $\geq 6.5\%$ can be diagnosed as a diabetic. Patients with a HbA1c in between are recognized as pre-diabetic. Other relevant information like the participant's high-density lipoprotein (HDL), body mass index (BMI) and age were also collected along with the participant's smoking levels, which were originally recorded as "number of cigarettes per day" and "how many years as a smoker?". The variable "Pack years", defined as the number of packs (20 cigarettes) times the number of years spent smoking, is a common measure of smoking intensity. I used three classifications for smoker status: High-level smokers (more than 10 pack-years), Low-level smokers (between 0 and 10 pack years) and non-smokers (0 pack years), or H, L, N. The whole data set was first cleaned to remove the entries with missing responses, which gave 207 remaining data entries. Covariate selection was done by classical regression and LDL, BMI and ages of the participants seem important in a sense that the p-values of their coefficients are less than 0.05. After centering these variables, I added the smoking effect, which contains three levels: High, Low and None.

3.3.3 Model Setup and Analysis

Figure 3.4 shows the scatter plots and box plot of HbA1c against each of the predictors. A data point shows different than others because the participant had a really

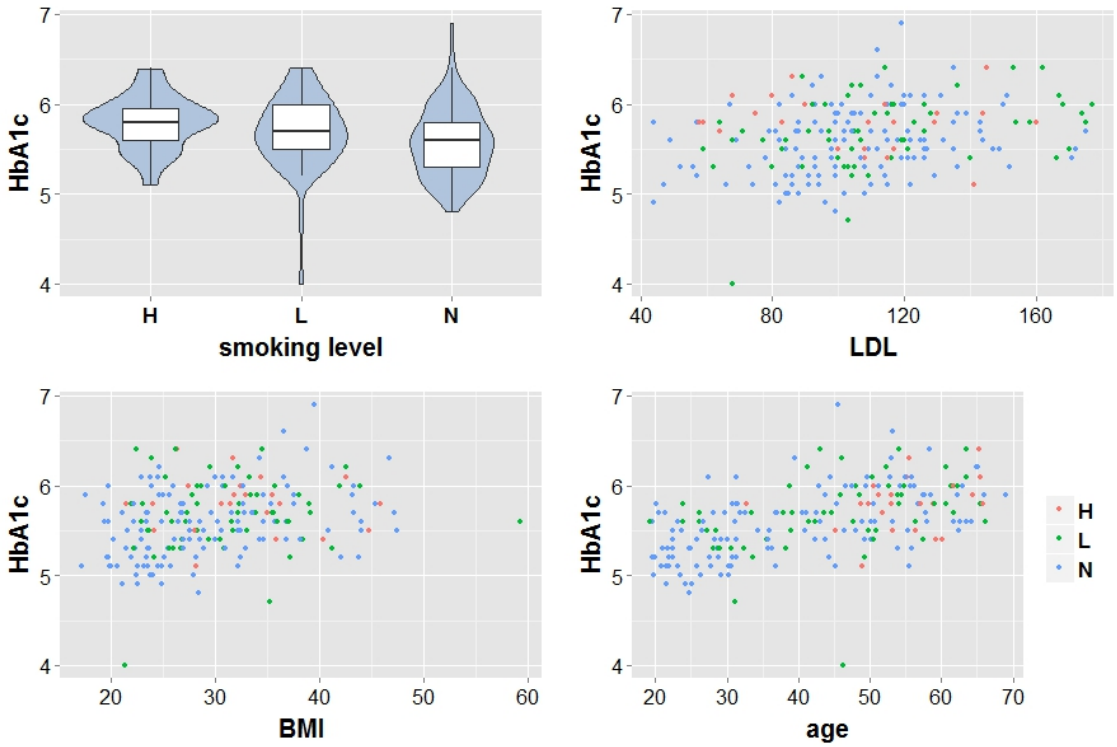

Figure 3.4: Plots of HbA1c vs covariates

low HbA1c value. An internal order of smoking levels corresponding to HbA1c may exist, however, this relationship is not clear due to the variability of the data. With this data, I consider an ANCOVA model with HbA1c as the response variable and other variables as predictors or independent variables. The fitted model by LSE is

shown below:

$$\begin{aligned} \hat{HbA1c} &= 5.6295 + 0.005688 \times I(\text{smoking level} = \text{"L"}) \\ &\quad -0.01524 \times I(\text{smoking level} = \text{"H"}) \\ &\quad +0.001736 \times LDL + 0.01009 \times BMI + 0.01321 \times age. \end{aligned} \tag{3.15}$$

Interaction among predictors are not considered. A simple regression analysis shows that the smoking effect is not significant and LDL, BMI and age are all positively associated with the response. Variance Inflation Factors (VIF) checks show no indication of multicollinearity. The summary of the regression model is shown in Table 3.2.

Table 3.2: Summary of regression model

	Estimate	Std. Error	P value
smoking level = "N"	5.6295	0.0298	0
smoking level = "L"	5.6352	0.0717	0
smoking level = "H"	5.6143	0.0428	0
LDL	1.7359E-3	8.428E-4	0.0407
BMI	1.0090E-2	3.2329E-3	0.0021
age	1.3215E-2	1.7694E-3	0

Note: Although the effects from three smoking levels are significant, they are not significantly different from one another.

On the other hand, a Bayesian analysis with a prior distribution that considers a simple order for the smoking effects can be performed, where heavier smokers induce higher risk of type 2 diabetes. It is a natural thing to adopt the two reference priors, $\pi(\boldsymbol{\theta})_{uni}$ and $\pi(\boldsymbol{\theta})_{u1k}$ derived and discussed in previous sections. Fitted models can be compared with Jeffreys' prior $\pi(\boldsymbol{\theta})_J \propto \frac{1}{\sigma^2}$ and LSE. Figures 3.5 and 3.6 show the marginal posterior distributions of the parameters under the different priors. Posterior distributions from two reference priors show similar patterns, which is not

surprising since the sample size is fairly large. The results from Jeffreys' prior are close to LSE, where the estimates for the three smoking levels tend to be mixed up. Compared with the reference priors, the marginal posterior from Jeffreys' prior seems to have heavier tail when estimating the mean responses for different smoking levels, while they behave similarly when estimating other parameters.

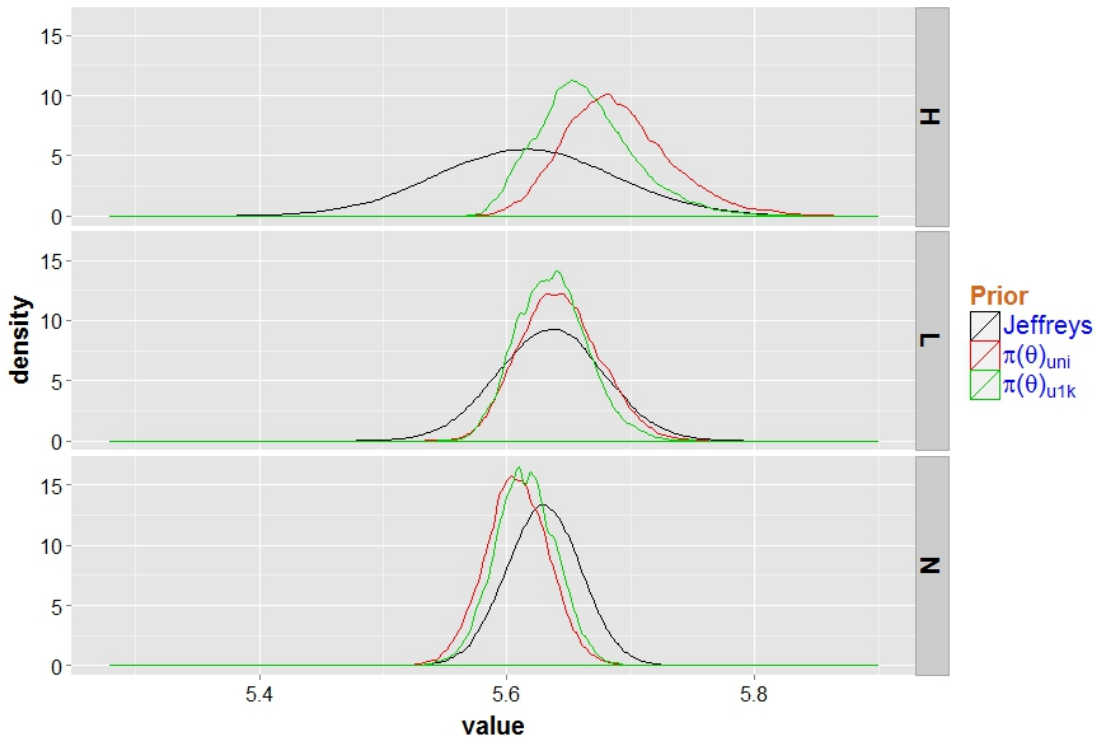

Figure 3.5: Marginal posterior distributions for different smoking levels under three priors

The MCMC results based on different priors is summarized in Table 3.3. Although Jeffreys' prior gives similar results as classical regression, its DIC is the largest. The one with the reference prior $\pi(\boldsymbol{\theta})_{u1k}$ gives the smallest DIC, which turns to be the evidence of better fitting and less complexity of the model. If we consider the differences between different smoking levels, the reference priors incorporating the simple order show there is a significant difference between high level smokers and

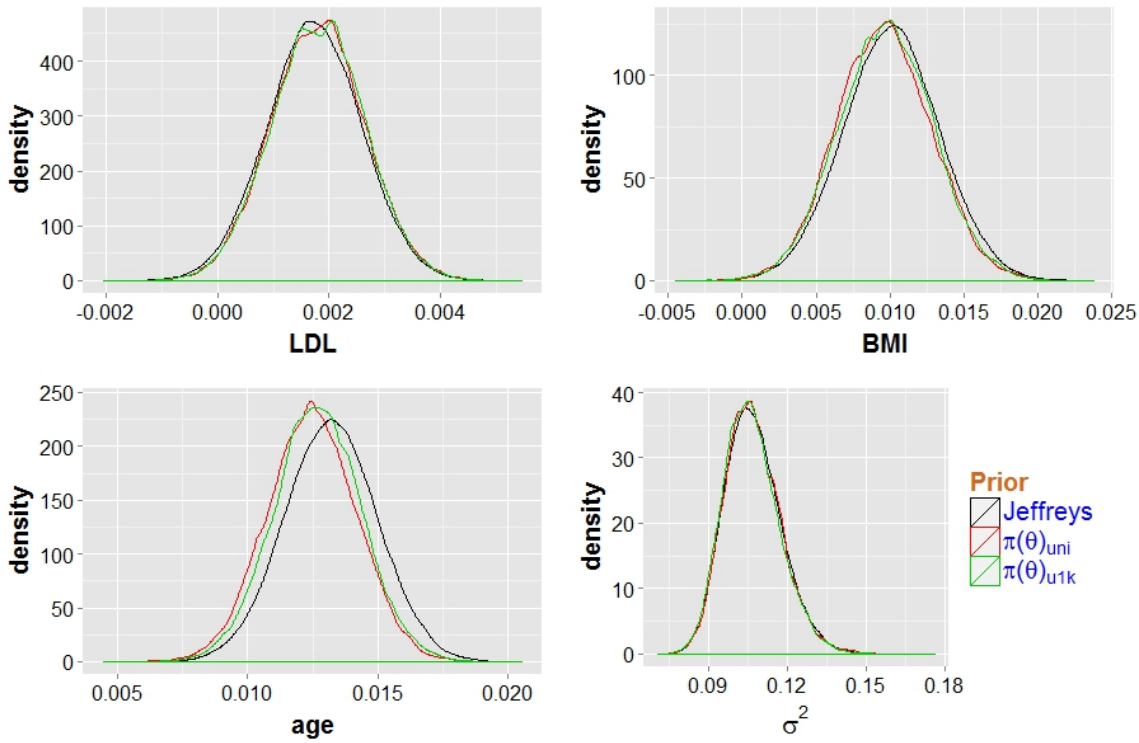

Figure 3.6: Marginal posterior distributions for LDL, BMI, age and σ^2 under three priors

non-smokers, while the results from Jeffreys' prior cannot show this relationship. Our conclusion is that once the internal ordering information is incorporated, it seems there is a progress for the model fitting. HbA1c seems positively associated with LDL, BMI and age and considering the order of smoking levels along with response seems more reasonable.

Table 3.3: MCMC results of Bayesian analysis with three priors

		Estimate	95% credible interval	DIC
Jeffreys	smoking level = "N"	5.6302	(5.5735, 5.6916)	1542.362
	smoking level = "L"	5.6359	(5.5524, 5.7175)	
	smoking level = "H"	5.6175	(5.4732, 5.7706)	
	LDL	1.7527E-3	(1.2256E-4, 3.4259E-3)	
	BMI	1.0058E-2	(3.7499E-3, 1.6283E-2)	
	age	1.3122E-2	(9.8467E-3, 1.6479E-2)	
	σ^2	0.1066	(0.0881, 0.1284)	
$\pi(\boldsymbol{\theta})_{uni}$	smoking level = "N"	5.6084	(5.5560, 5.6605)	1531.634
	smoking level = "L"	5.6434	(5.5838, 5.7100)	
	smoking level = "H"	5.6917	(5.6190, 5.7874)	
	LDL	1.8337E-3	(1.6109E-4, 3.4335E-3)	
	BMI	9.4074E-3	(3.2307E-3, 1.6214E-2)	
	age	1.2430E-2	(9.0153E-3, 1.5759E-2)	
	σ^2	0.1066	(0.0887, 0.1303)	
$\pi(\boldsymbol{\theta})_{u1k}$	smoking level = "N"	5.6088	(5.5570, 5.6575)	1513.375
	smoking level = "L"	5.6449	(5.5875, 5.7076)	
	smoking level = "H"	5.6924	(5.6192, 5.7854)	
	LDL	1.8225E-3	(7.9451E-5, 3.5334E-3)	
	BMI	9.2048E-3	(2.6986E-3, 1.5427E-2)	
	age	1.2385E-2	(8.7402E-3, 1.5909E-2)	
	σ^2	0.1073	(0.0881, 0.1300)	

Chapter 4

Reference Priors for Spatial CAR and SAR Models

Bayesian spatial models have been growing in popularity due to advances in computation and the presence of many spatial data sets. However, the number of default prior options is still very small for spatial modeling when the spatial autoregression parameter, ρ, is considered. Researchers tend to adopt the intrinsic conditionally autoregressive (CAR) model and simultaneous autoregressive (SAR) model specification, where the spatial effect is introduced as a random effect in the second stage of a hierarchical setting. On the other hand, although the introduction of ρ can guarantee the propriety of the likelihood function, it complicates the model, which makes it difficult in exploring priors. There is an urgent need for default priors under these models. Under the framework provided by Berger and Bernardo [1992], the reference priors for CAR and SAR models are worth deriving and studying for model fitting.

This chapter is organized as follows: Section 4.1 gives a brief introduction to spatial statistics with different data types. Section 4.2 talks about the spatial CAR

and SAR models for areal data sets. In Section 4.3, the derivation of the reference priors for these two models are discussed in detail. And finally, in Section 4.4, the reference priors are applied to a Bayesian analysis for the state level 1999 SAT scores in the United States, with comparison to a Uniform prior.

4.1 Spatial Statistics and Its Data Types

When statisticians deal with quantitative study of phenomena referenced in space, the regular independence assumption of observations is not valid, since the attributes of location i may have influence on the attributes of location j. Hence spatial statistics can be interpreted as a class of methods that consider the spatial correlation among observations. Spatial correlations are very common when analyzing data in epidemiology, criminology, agriculture, econometrics and geography, etc.

Spatial data sets are generally classified into three types: point-reference data, areal data and point pattern data. For point-reference data, the response, $\boldsymbol{y}(s)$, is a random vector or scalar at a location s while s varies continuously over a fixed subset of an r dimensional space, D. For areal data, the location information, D, is partitioned into a finite number of areal parts (either regular or irregular) with well-defined boundaries. For the point pattern data, D is random and $\boldsymbol{y}(s)$ simply shows the occurrence of the event of interest at some s $\in D$.

It is natural that for different types of data sets, people consider different models. For example, for point-reference data, the distances between locations are used to describe the strength of spatial association. General speaking, the longer the distance, the smaller or weaker the spatial association. The variance-covariance structure of the model can then be modeled by a function of the Matérn family. For areal data, a similar idea can be adopted. That is, for each region the response can be assumed to have been observed at the centroid location and distances among centroids are

used to develop the spatial variance-covariance structure of the data. However, this approach for areal data is problematic since it is not possible for the observations to occur continuously in space.

Another way to model areal data in spatial statistics considers a neighborhood or a contiguity structure based on the shape of the lattice where the data is observed. Figure 4.1 shows the ways of defining contiguity: Rook's case, Bishop's case and Queen's case. This figure is originally drawn by Sawada [2009].

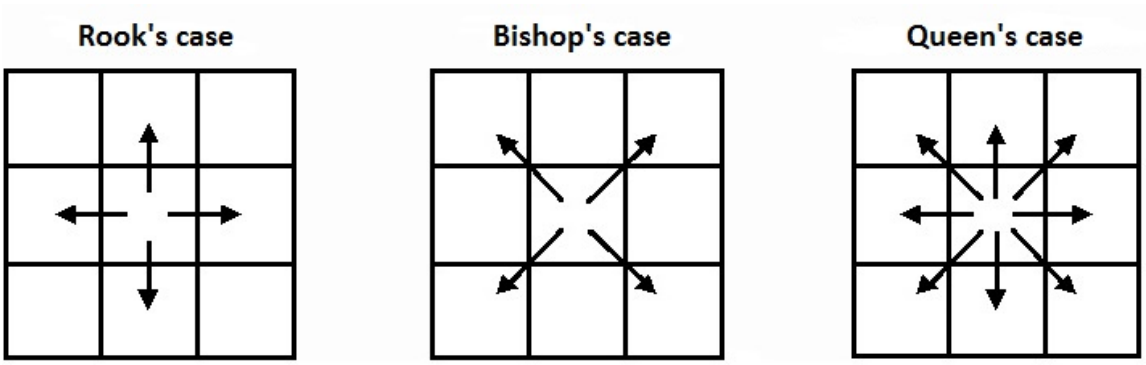

Figure 4.1: Three ways of defining contiguity for areal data.

Rook contiguity only uses common boundaries and Bishop uses only common vertexes, while Queen's case considers both to determine the neighbors. Once this contiguity structure is defined, the neighboring information can be stored into a symmetric proximity matrix W with entries w_{ij}, such that w_{ij}=1 if two regions are neighbors, w_{ij}=0 otherwise. Models resembling autoregressive models in time series are considered. Two very popular models that incorporate this discrete neighboring information are, as mentioned before, CAR and SAR models, which were originally introduced by Besag [1974] and Whittle [1954], respectively.

4.2 Introduction to CAR and SAR Models

In this section I will formally define the CAR and SAR models, with a spatial autocorrelation ρ being introduced. If ρ is considered, direct modeling of the data with CAR and SAR models is possible.

4.2.1 CAR Model

Given a vector $\boldsymbol{\phi} = (\phi_1,\ \phi_2,\ \ldots,\ \phi_n)'$ as a Gaussian process from a lattice D, the zero-centered CAR specification is given by a full conditional distribution

$$\phi_i|\phi_j,\ j \neq i \ \sim \ N\left(\sum_j b_{ij}\phi_j,\ \tau_i^2\right),\ \text{for } i\ = 1, \ldots, n. \tag{4.1}$$

where b_{ij} denotes the weight from neighbor j. By Brook's Lemma as in Banerjee et al. [2014] we obtain the joint distribution

$$p(\phi_1, \phi_2, \ldots, \phi_n) \ \propto \ \exp\left\{-\frac{1}{2}\boldsymbol{\phi}^T D_{\tau^2}^{-1}(I - B)\boldsymbol{\phi}\right\}. \tag{4.2}$$

Here B $= \{b_{ij}\}_{n\times n}$ with $b_{ii} = 0$ and D_{τ^2} is a diagonal matrix with $D_{ii} = \tau_i^2$. Since $D_{\tau^2}^{-1}(I - B)$ is the inverse of the variance-covariance matrix, the requirement that $D_{\tau^2}^{-1}(I - B)$ be symmetric yields

$$\frac{b_{ij}}{\tau_i^2} \ = \ \frac{b_{ji}}{\tau_j^2}, \quad \text{for all } i,\ j. \tag{4.3}$$

Recall the definition of the proximity matrix W in last section. Suppose we set $b_{ij} = w_{ij}/w_{i+}$ and $\tau_i^2 = \tau^2/w_{i+}$, where $w_{i+} = \sum_j w_{ij}$ is the sum of row i, then Equation 4.3 is satisfied. The way of setting B and D_{τ^2} here is actually called weighted (heterogeneous) CAR (WCAR) model. Others include homogeneous CAR (HCAR) model and autocorrelation CAR (ACAR) model. A good review of different

ways of setting CAR models can be found in Cressie and Kapat [2008]. If we define D_w to be diagonal with entries w_{i+}, Equation 4.1 becomes

$$\phi_i|\phi_j,\ j \neq i \ \ \sim \ \ N\left(\sum_j w_{ij}\phi_j/w_{i+},\ \tau^2/w_{i+}\right),\ \text{for } i = 1,\ldots,n. \tag{4.4}$$

and Equation 4.2 is

$$p(\phi_1,\phi_2,\ldots,\phi_n) \ \ \propto \ \ \exp\left\{-\frac{1}{2\tau^2}\boldsymbol{\phi}^T(D_w - W)\boldsymbol{\phi}\right\}. \tag{4.5}$$

If the variance-covariance matrix for the distribution in Equation 4.5 is denoted by $\sum_{\boldsymbol{\phi}}$, it can be shown that $\sum_{\boldsymbol{\phi}}^{-1} \times \mathbf{1} = \frac{1}{\tau^2}(D_w - W) \times \mathbf{1} = \mathbf{0}$, i.e., $\sum_{\boldsymbol{\phi}}^{-1}$ is singular and Equation 4.5 does not give a proper distribution. A suggested approach to have a proper model is to introduce a parameter ρ and redefine $\sum_{\boldsymbol{\phi}}^{-1} = D_{\tau^2}^{-1}(I - \rho B) = \frac{1}{\tau^2}(D_w - \rho W)$ as in Banerjee et al. [2014], where ρ is known as the "spatial autoregression parameter". $\sum_{\boldsymbol{\phi}}$ can be nonsingular if ρ is carefully chosen. Banerjee et al. [2014] show that this is guaranteed if $\rho \in (\lambda_{(1)}^{-1},\ \lambda_{(n)}^{-1})$, where $\lambda_{(1)} < \lambda_{(2)} < \ldots < \lambda_{(n)}$ are the ordered eigenvalues of $D_w^{-1/2} W D_w^{-1/2}$.

The ideas presented in this section can be easily extended to a Gaussian process of the form

$$y_i|y_j,\ j \neq i,\ \tau_i^2 \ \ \sim \ \ N\left(\mu_i + \sum_j b_{ij}(y_j - \mu_j),\ \tau_i^2\right),\ \text{for } i \ = \ 1,\ldots,n. \tag{4.6}$$

By adding linear regressors, the model can be re-written with a likelihood function of the form

$$f(\boldsymbol{y}|\boldsymbol{\beta},\tau^2,\rho) \ \ \propto \ \ [\det(V)]^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\beta})^T V^{-1}(\boldsymbol{y} - X\boldsymbol{\beta})\right\}, \tag{4.7}$$

where $\boldsymbol{y} = (y_1,\ y_2,\ldots,\ y_n)'$ and $V \ = \ \sum_{\boldsymbol{y}} \ = \ \tau^2[D_w - \rho W]^{-1}$.

4.2.2 SAR Model

The definition of the SAR model is similar to the CAR process and can be followed by Equation 4.6, where we may rewrite the model for observation y_i as,

$$y_i \;=\; \mu_i + \sum_j b_{ij}(y_j - \mu_j) + \varepsilon_i. \tag{4.8}$$

In matrix notation, we have

$$\boldsymbol{y} \;=\; \boldsymbol{\mu} + B(\boldsymbol{y} - \boldsymbol{\mu}) + \boldsymbol{\varepsilon}, \tag{4.9}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)'$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$. Equivalently,

$$(I - B)(\boldsymbol{y} - \boldsymbol{\mu}) \;=\; \boldsymbol{\varepsilon}. \tag{4.10}$$

This expression implies that a probability distribution for $\boldsymbol{\varepsilon}$ can induce a distribution for $\boldsymbol{y}$. The model is called *simultaneous* because generally the error terms ε_i's are correlated with $\boldsymbol{y}$. Now let's assume $\boldsymbol{\varepsilon} \sim N(\mathbf{0},\, \boldsymbol{\Lambda})$ with $\boldsymbol{\Lambda}$ a diagonal matrix. The distribution for $\boldsymbol{y}$ is then

$$\boldsymbol{y} \;\sim\; N\left(\boldsymbol{\mu},\, (I_n - B)^{-1}\boldsymbol{\Lambda}((I_n - B)^{-1})'\right). \tag{4.11}$$

There may be different choices for the $\boldsymbol{\Lambda}$ and B matrices. Wall [2004] suggests to set $\boldsymbol{\Lambda} = \sigma^2 D_w^{-1}$ where D_w and B follow the same definition as in the CAR model. In order for Equation 4.11 be proper distribution, an extra parameter(spatial autocorrelation) ρ is also introduced to the model. With the addition of linear regressors, the model can be re-written with a likelihood of the form

$$f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \rho) \;\propto\; [\det(V)]^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})'(I_n - \rho B)' D_w (I_n - \rho B)(\boldsymbol{y} - X\boldsymbol{\beta})\right\}, \tag{4.12}$$

where $V = \sum_{\boldsymbol{y}} = \sigma^2(I_n - \rho B)^{-1}D_w^{-1}((I_n - \rho B)^{-1})'$.

4.3 Derivation of the Reference Priors for CAR and SAR Models

It is always easy to adopt a Uniform prior on ρ as a non-informative choice. This can be seen in Bell and Broemeling [2000], Hepple [1995a] and Hepple [1995b], where $\pi(\rho) \propto 1_{(\lambda_{(1)}^{-1},\, \lambda_{(n)}^{-1})}(\rho)$ is used as a prior. Besag and Kooperberg [1995] point out that for datasets with a strong correlation between neighboring observations, this relationship is hard to be reproduced in CAR models unless the spatial autocorrelation is quite close to the boundaries, either $\lambda_{(1)}^{-1}$ or $\lambda_{(n)}^{-1}$. Since Uniform prior assigns equal probability in the whole parameter space, this common behavior is simply ignored by the prior. De Oliveira [2012] and De Oliveira and Song [2008] derived the independence Jeffreys' prior for CAR and SAR models respectively, where large prior mass is assigned to parameter values close to their boundaries. Ren and Sun [2014] also studied some objective priors including reference priors with nugget effects. However, these studies are limited to easy setting like HCAR models. The derivation of the reference priors for heterogeneous CAR and SAR models can be done similarly under the framework provided by Berger and Bernardo [1992]. Based on Equation 4.7, we can derive the Fisher information matrix for the CAR model, which is

$$I(\rho, \tau^2, \boldsymbol{\beta}) = \begin{pmatrix} \frac{tr[WVWV]}{2\tau^4} & \frac{tr[WV]}{2\tau^4} & 0 \\ \frac{tr[WV]}{2\tau^4} & \frac{1}{2\tau^4} & 0 \\ 0 & 0 & \frac{1}{\tau^2}X'(D_w - \rho W)X \end{pmatrix}.$$

Suppose $\boldsymbol{\theta} = \{\rho, \tau^2, \boldsymbol{\beta}\}$. If the parameters are grouped and ordered as $\{\rho\}$, $\{\tau^2\}$, $\{\boldsymbol{\beta}\}$, then the reference prior is

Theorem 4.

$$\pi(\boldsymbol{\theta})_{car} \propto \left|\frac{tr[WVWV]}{\tau^4} - \frac{tr[WV]tr[WV]}{\tau^4}\right|^{1/2} \times \frac{1}{\tau^2} \times 1_{(\lambda_{(1)}^{-1},\, \lambda_{(n)}^{-1})}(\rho), \quad (4.13)$$

where W is proximity matrix and V is the variance-covariance matrix in the CAR model that contains the parameter ρ. The detailed derivation steps for Theorem 4 are shown in Appendix B.

In the SAR model, suppose $\boldsymbol{\theta} = \{\rho, \sigma^2, \boldsymbol{\beta}\}$. If parameters are grouped and ordered as $\{\rho\}$, $\{\sigma^2\}$, $\{\boldsymbol{\beta}\}$, then a similar procedure can be adopted to derive the reference prior for this model as shown in Appendix C, which gives the reference prior

Theorem 5.

$$\pi(\boldsymbol{\theta})_{sar} \propto |h_1|^{1/2} \times \frac{1}{\sigma^2} \times 1_{(\lambda_{(1)}^{-1},\, \lambda_{(n)}^{-1})}(\rho) \quad (4.14)$$

with $h_1 = \frac{1}{\sigma^2} tr\left[B' D_w B V\right] - \frac{1}{2\sigma^2} tr\left[2B' D_w B V - \frac{AVAV}{\sigma^2}\right] - \frac{1}{2\sigma^4} tr\left[AV\right] tr\left[AV\right]$ and $A = -[(I_n - \rho B)' D_w B + B' D_w (I_n - \rho B)]$.

It should be pointed out that the reference priors for both CAR and SAR models require calculation of the trace from $n \times n$ matrices. Eventually the resulting priors are high order polynomials with a high dimensional proximity matrix. It could be a problem to find the exact form of the prior distribution when handling areal data with a large number of regions. However, it may be simplified to numerically evaluate the prior density for some specific ρ values. This is the main requirement needed for posterior inference with the Metropolis-Hastings algorithm. On the other hand, for the derivation of the reference priors in the CAR and SAR models, the spatial autoregression parameter, ρ, is always grouped and ordered in the first position, because this is the main parameter of interest for this study. Also, this grouping and ordering guarantees that the simplified expression for deriving the reference prior in Lemma 1 of Berger and Bernardo [1992] can be used.

4.4 Analysis of the 1999 SAT State Average Verbal Scores

In this section, a state level summary data set related to the 1999 SAT college entrance exam is considered. The data set contains the state average verbal SAT scores along with the percentage of eligible students taking the exam in the corresponding state for all US 48 contiguous states. A choropleth map of the data is presented in Figure 4.2. The black dots represent centroids of each state. A spatial pattern

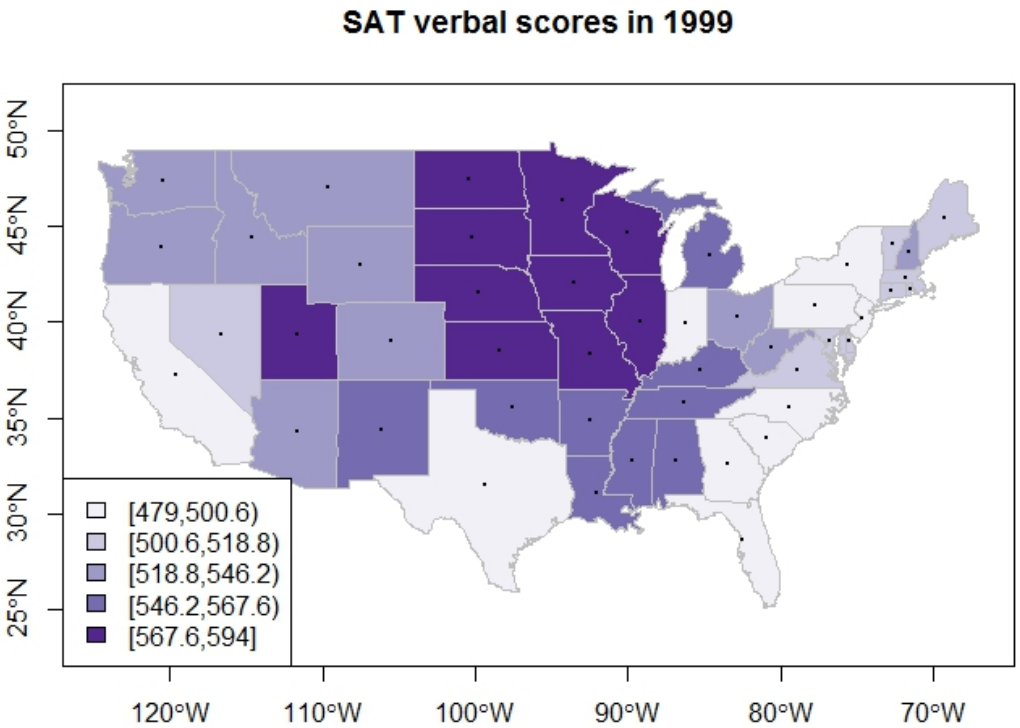

Figure 4.2: Choropleth map of 48 contiguous state average SAT verbal scores for 1999

may be existing since the states in the Midwest seem to have higher average SAT

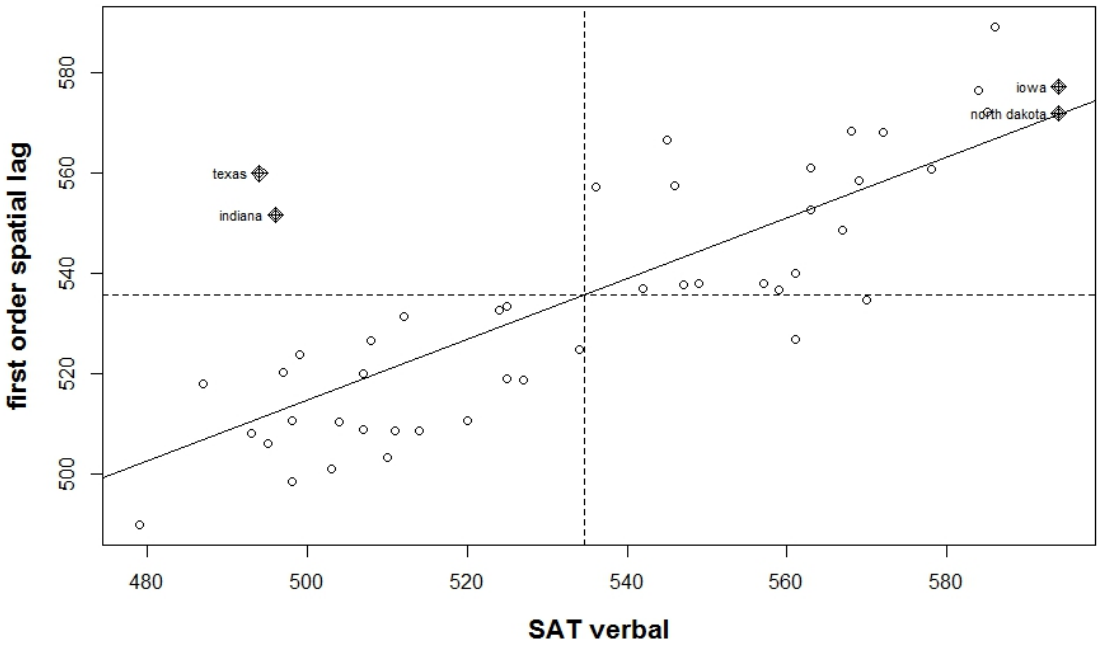

Figure 4.3: Plot of first order spatial lag vs state average SAT verbal scores

scores than those from the East or West portions of the U.S., although the scores from Texas and Indiana seem unusually low compared to their neighbors. The standard statistics used to measure spatial association for areal data are Moran's I and Geary's C as in Ripley [1981]. The Moran's I can be calculated by

$$I = \frac{n\sum_i\sum_j w_{ij}(Y_i-\bar{Y})(Y_j-\bar{Y})}{(\sum_{i\neq j})\sum_i(Y_i-\bar{Y})^2}. \tag{4.15}$$

Under the null model where the Y_i are iid, Moran's I is asymptotically normally distributed with mean $-\frac{1}{n-1}$ as in Sen [1976]. For this data, Moran's I gives 0.6055 with a p-value almost 0, which confirms there is a strong positive spatial association. Figure 4.3 shows this association graphically by presenting the scatter plot of the first order spatial lag against the state average SAT verbal scores, where the first order

spatial lag for each state is calculated by averaging the scores from its neighbors.

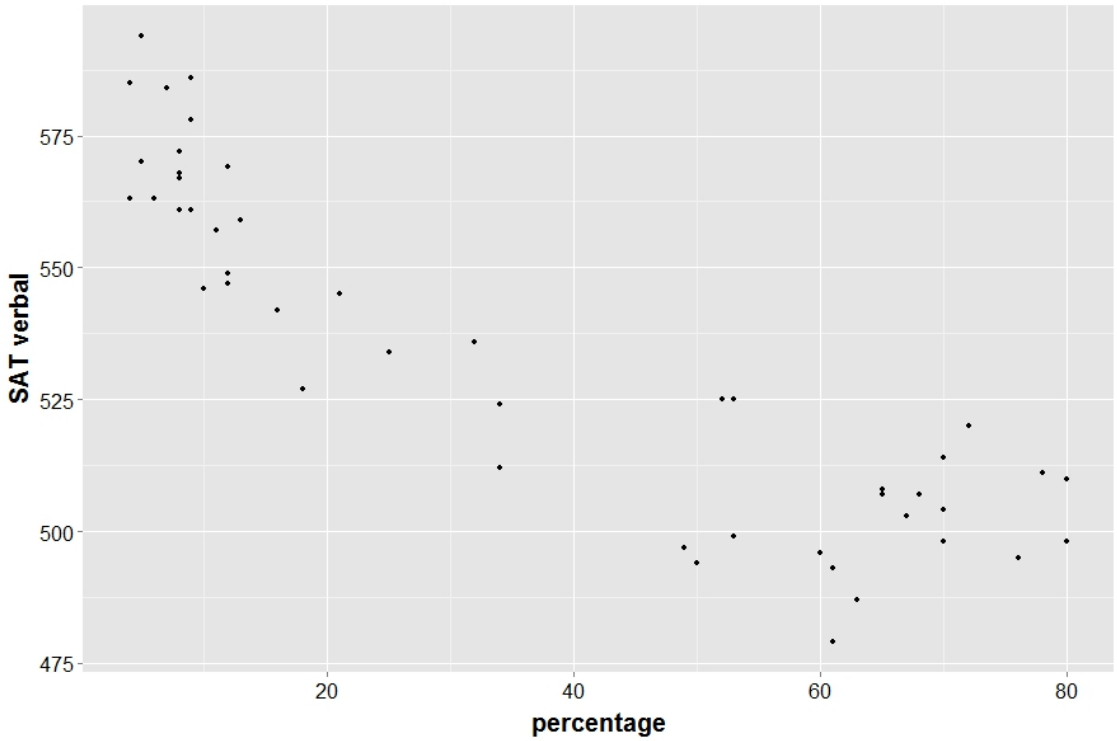

Figure 4.4: Scatter plot of state average 1999 SAT verbal scores vs percentage of eligible students taking the exam

Further analysis shows there is a strong negative association between the average SAT verbal scores and the percentages of eligible students taking the exam, as shown in Figure 4.4. Wall [2004] suggests to remove the trend of the data by using the following model:

$$Z(A_i) \quad = \quad \beta_0 + \beta_1 X(A_i) + \beta_2 (X(A_i))^2 + \mu(A_i), \tag{4.16}$$

where $Z(A_i)$ represents the average SAT verbal score in state A_i, $X(A_i)$ represents the percentage of eligible students taking the exam at state A_i and $\mu(A_i)$ is the error term with zero mean and a Normal distribution for $i = 1, \ldots, 48$.

If we define $\boldsymbol{\mu} = (\mu(A_1), \mu(A_2), \ldots, \mu(A_{48}))'$, then there are several ways to model the covariance structure for $\boldsymbol{\mu}$. A naive way is to assume $\boldsymbol{\mu} \sim N(\mathbf{0}, \sigma^2 I_{48})$, then the spatial associations are ignored and the data is modeled as *iid* observations with polynomial regression. As described at the beginning of this chapter, another possibility is to consider areal data as point-reference data by relying on the centroids of the 48 states. Then an isotropic variogram structure for the $\boldsymbol{\mu}$ can be considered where the centroids are used to define distances between states as mentioned in Wall [2004]. A Matérn covariance function can be adopted to describe the variance-covariance change along distances and the covariance could also contain a nugget effect.

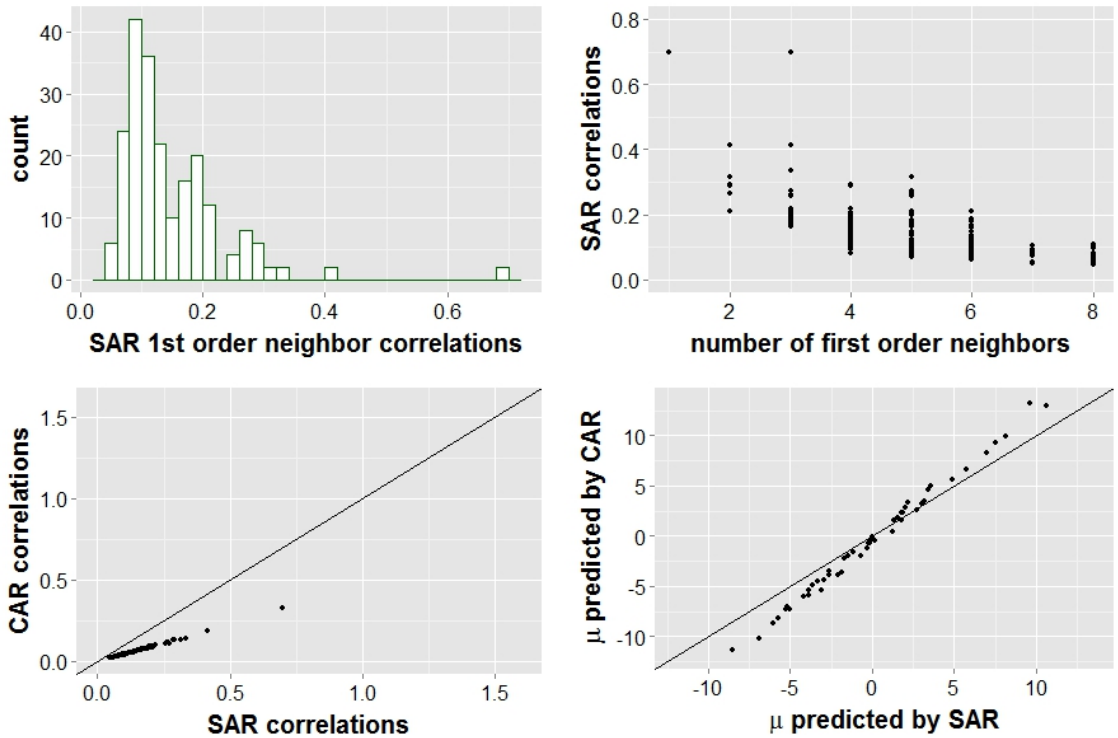

Figure 4.5: Top: Histogram of the 1st neighbor correlations (left) and stratified correlations from SAR model; Bottom: Comparison of SAR and CAR results

A more reasonable strategy is to adopt a SAR or CAR model which properly

take into account the spatial structure of the data. For these two models, Maximum Likelihood Estimation (MLE) can be obtained easily with the R package "spdep". Figure 4.5 shows some results based on MLE of the CAR and SAR models for the state average 1999 SAT verbal scores. The top left shows a histogram of all the first order neighbor correlations from the SAR model. Here first order neighbors are defined by neighboring states that share common boundaries. Notice that the largest correlation is around 0.6975. This corresponds to the correlation between New Hampshire and Maine, where Maine is the only state having just one neighbor (i.e., New Hampshire). The smallest correlation happens between Tennessee and Missouri, which is about 0.0476. These two states are the two states with largest number of neighbors (i.e., eight). The graph on top right also shows a general trend between the magnitude of correlations and the number of first order neighbors.

The bottom graphs provide a comparison between SAR and CAR models. The predicted values for $\boldsymbol{\mu}$ have a correlation around 0.92 and are scattered around the identity line. The correlations from two models are linearly related but the correlations from the CAR model are lower than for the SAR model.

The relation between ρ and the implied spatial correlation from the model was also studied as in Wall [2004]. It is easy to find that for this model $\rho \in (\lambda_{(1)}^{-1},\ \lambda_{(n)}^{-1}) = (-1.3924,\ 1)$. She notes the first order neighbor correlations increase as a function of ρ when $\rho > 0$, while this relation seems quite different when $\rho < 0$.

In terms of Bayesian analysis, due to the complexity of the reference priors for both CAR and SAR models, it is hard to find the analytical forms of the reference priors for all 48 states. Therefore, I just analyze the data for the 12 states in the Midwest instead as shown in Figure 4.6 via reference priors.

This portion of the data set still keeps similar properties as the full data, such as positive spatial associations with the first order spatial lag and a negative trend with

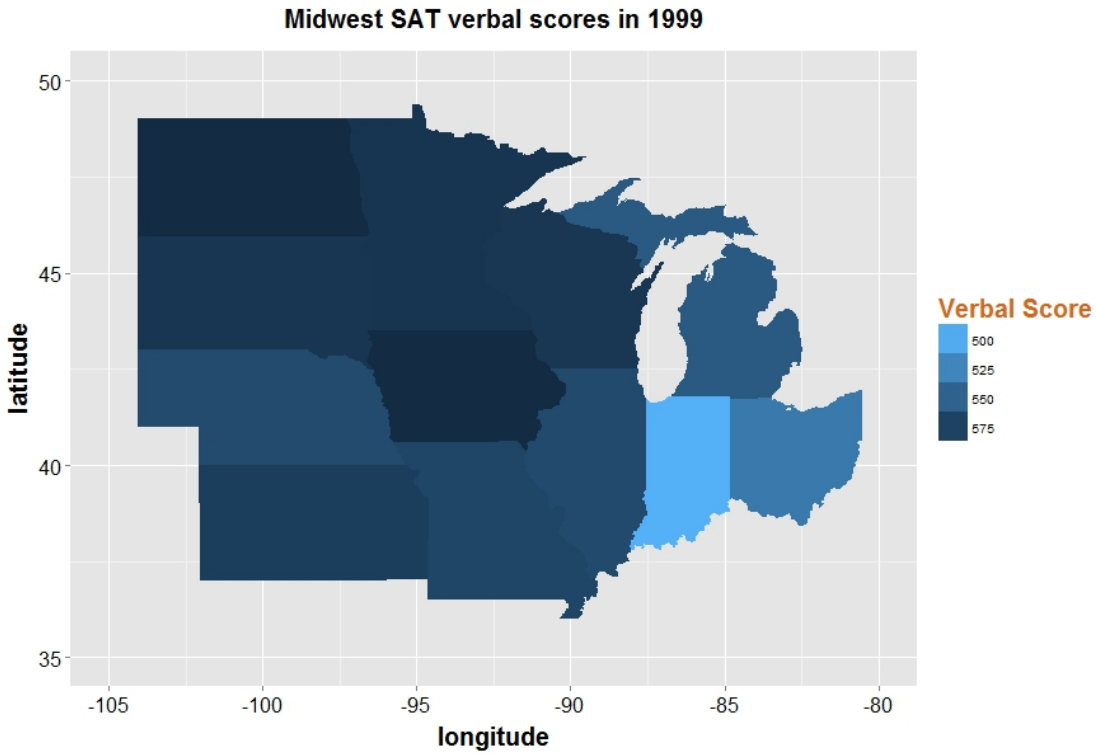

Figure 4.6: Choropleth map of state average SAT verbal scores in Midwest 12 states for 1999

regard to the percentage of eligible students taking the exam for these states. For this portion of the data, I use the percentage of eligible students taking the exam as a regressor to model the trend. The model is fitted as

$$Z(A_i) \quad = \quad \beta_0 + \beta_1 X(A_i) + \mu(A_i), \tag{4.17}$$

where $Z(A_i)$, $X(A_i)$ and $\mu(A_i)$ are defined as before. The variance-covariance structure of $\boldsymbol{\mu}$ can be modeled by either the CAR or SAR settings. The priors derived in Theorems 4 and 5 for these two models are considered for posterior inference. In particular, the Metropolis-Hastings algorithm is adopted for the spatial autoregression parameter ρ. At the t-th iteration, a proposed value ρ^* is sampled from a truncated normal density $TN(\rho^{(t-1)},\ \xi) \propto \frac{1}{\sqrt{2\pi\xi}} \exp[-\frac{1}{2\xi}(\rho^* - \rho^{(t-1)})] I_{(\lambda_{(1)}^{-1},\, \lambda_{(n)}^{-1})}(\rho^*)$, where $\lambda_{(1)}^{-1}$

= -1.5335 and $\lambda_{(n)}^{-1}$=1 for this portion of the data set. This ρ^* is accepted as $\rho^{(t)}$ with a probability of $\alpha = \min\{1,\ r\}$ where r is calculated by

$$r \quad = \quad \frac{f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \rho^*) \times \pi(\rho^*) \times TN(\rho^{(t-1)}|\rho^*, \xi)}{f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \rho^{(t-1)}) \times \pi(\rho^{(t-1)}) \times TN(\rho^*|\rho^{(t-1)}, \xi)}. \tag{4.18}$$

ξ can be adjusted to get an acceptance rate of 0.4. The other parameters can be

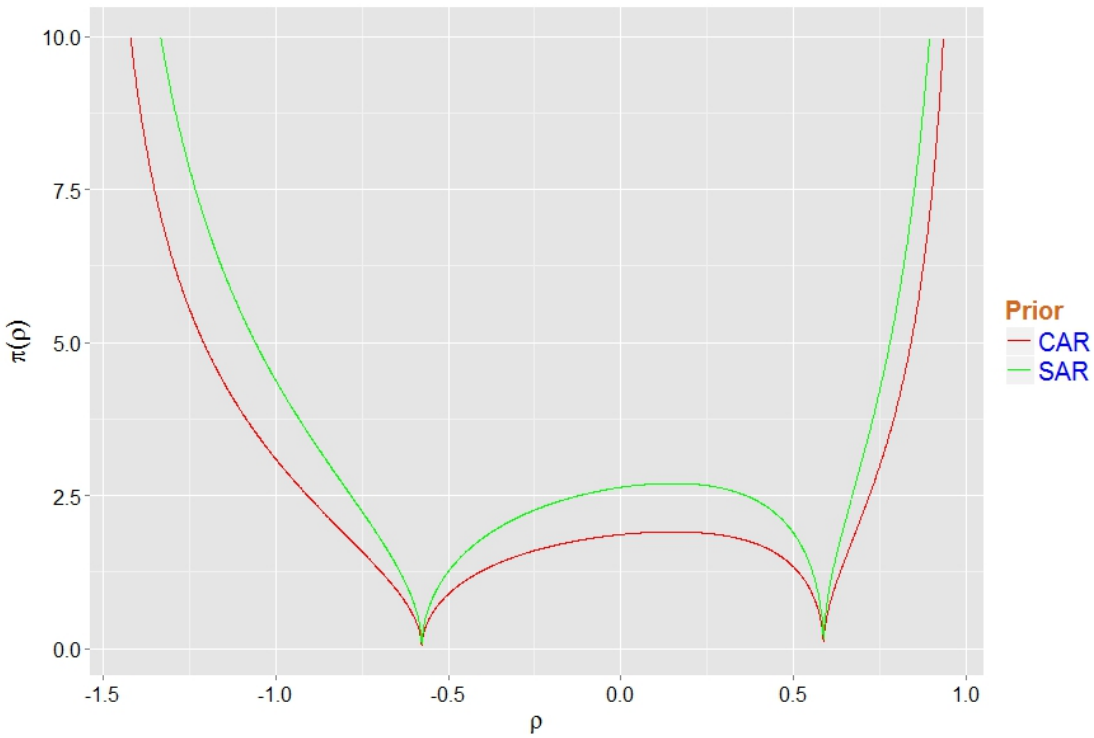

Figure 4.7: Prior densities on ρ of the two reference priors (area below each curve is not standardized).

updated by Gibbs sampling steps because it is easy to obtain their full conditional distributions. The full conditional distributions for these parameters in the CAR and SAR models are shown in Appendix B and C respectively. As a comparison, I also include the Bayesian analysis with a Uniform prior for ρ. The prior densities on ρ corresponding to the reference priors are shown in Figure 4.7. For our data

example, the marginal posterior distributions of all model parameters are shown in Figures 4.8 and 4.9.

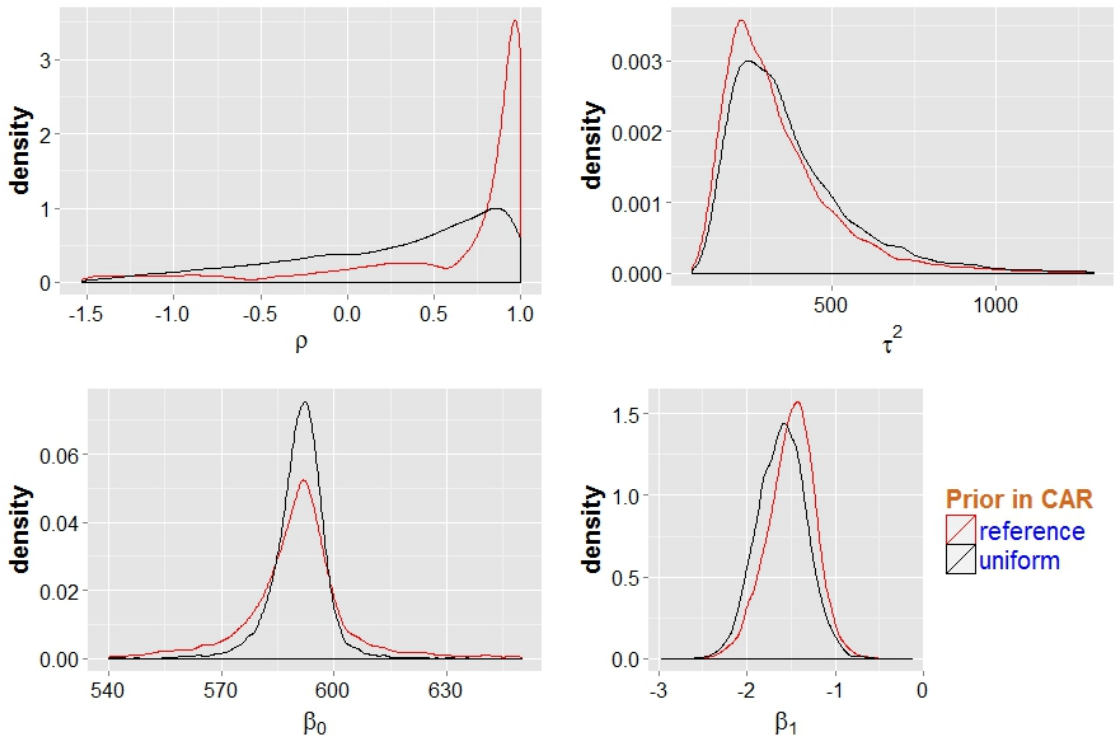

Figure 4.8: Marginal posterior distributions from CAR model for the Midwest SAT data.

For both CAR and SAR models, the reference priors give similar results as the Uniform prior when estimating $\boldsymbol{\beta}$, σ^2 or τ^2. However, the marginal posterior distributions for ρ show different patterns. In Figure 4.8, both posteriors have a mode close to 1. The posterior from the reference prior seems more peaked while the one from a Uniform prior seems flat. Similar comparisons can be noticed in Figure 4.9, except that there is a small spike at -0.6 for the reference prior. Table 4.1 gives the posterior means of the parameters for different priors and methods. It can be seen that after removing the trend, none of the analyses gives a significant estimate of ρ. A similar conclusion can be drawn by fitting the model with "spdep" in R.

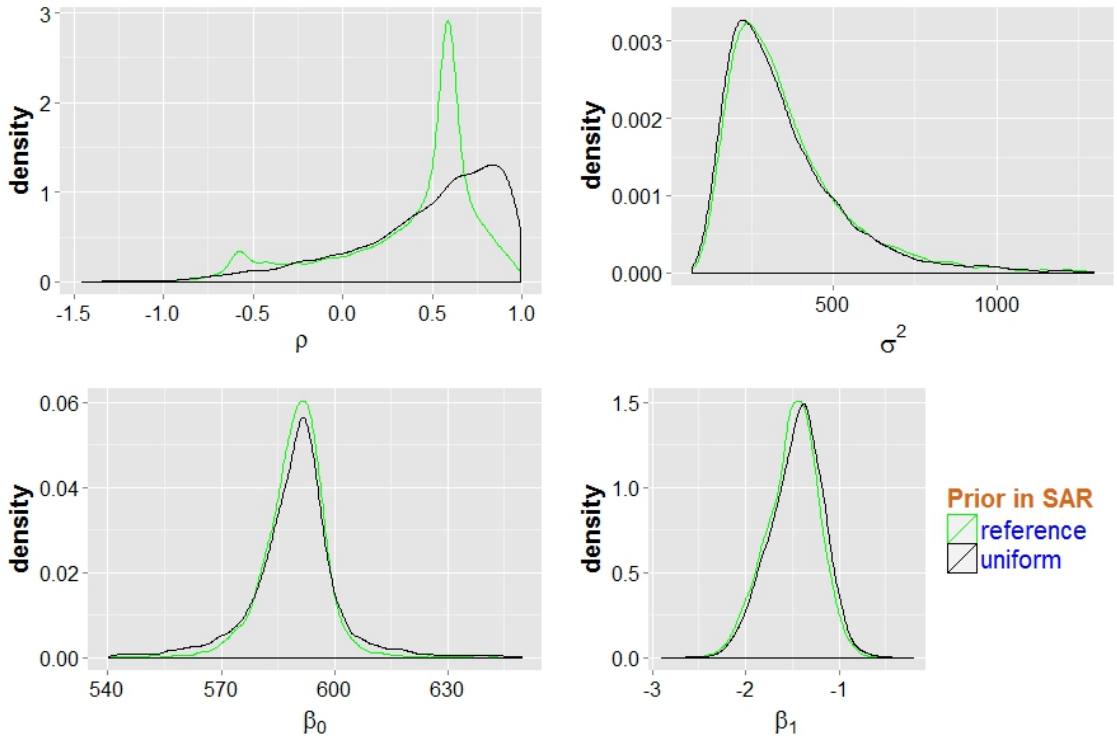

Figure 4.9: Marginal posterior distributions from SAR model for the Midwest SAT data.

Our data example shows that with a small number of areal units, reference priors can give a different marginal posterior compared to a Uniform prior under the CAR/SAR modeling framework.

Table 4.1: Summaries of the marginal posterior distributions

		Estimate	95% HPD interval	DIC	Model
Uniform prior	ρ	0.2680	(-0.9445, 0.9991)	71.6	CAR
	τ^2	377.4	(109.0, 784.4)		
	β_0	590.8	(575.4, 605.1)		
	β_1	-1.607	(-2.147,-1.073)		
reference prior	ρ	0.6158	(-0.9302, 0.9998)	71.8	
	τ^2	340.9	(95.7, 705.7)		
	β_0	588.9	(524.7, 657.8)		
	β_1	-1.501	(-2.069,-1.013)		
Uniform prior	ρ	0.4588	(-0.4228, 0.9928)	101.2	SAR
	σ^2	354.1	(97.0, 728.4)		
	β_0	589.3	(542.2, 635.8)		
	β_1	-1.474	(-2.060,-0.945)		
reference prior	ρ	0.3339	(-0.5879, 0.8696)	100.1	
	σ^2	363.1	(101.6, 750.6)		
	β_0	589.7	(572.2, 604.0)		
	β_1	-1.524	(-2.110,-0.993)		

Chapter 5

Discussion and Future Work

5.1 Discussion

Choosing an appropriate prior distribution is always important when fitting Bayesian models. In multi-parameter problem, people tend to rely on using Uniform priors. The reference prior of Bernardo [1979] and Berger and Bernardo [1992] is an alternative to the Uniform prior which considers the divergence (or distance) between the posterior and the prior distributions. This divergence is interpreted as the missing information about the parameters $\boldsymbol{\theta}$ relative to the prior $\pi(\boldsymbol{\theta})$. Prior distributions derived following this definition are naturally incorporating the principle that the data should dominate the posterior distribution. In addition to these theoretical properties, in practice reference priors have been shown to be useful default priors when little outside information is available. The largest obstacle for the wide adaptation of reference priors is that they can be difficult to calculate. The focus of this thesis is to derive the reference priors for several new settings.

In this dissertation, I adopt the reference prior framework and utilize the sequential maximization of the Kullback-Leibler divergence between the prior and the

posterior by Berger and Bernardo [1992]. In Chapters 2 and 3, I focus on deriving reference priors in Analysis of Variance (ANOVA)/Analysis of Covariance (ANCOVA) models with a categorical variable under common ordering constraints. This idea is a natural extension of Sonksen and Peruggia [2012], where a Poisson likelihood with ordered rate parameters is considered in a lung cancer data. I extended their results to many more likelihoods (focusing on the ANOVA/ANCOVA in particular) and different ordering constraints. Under different settings, the performances of reference priors in ANOVA/ANCOVA models can be evaluated by simulation studies, with comparisons to other methods, such as Jeffreys' priors and LSE. Part of the results from simulation studies shows the advantage of incorporating ordering information into the prior distribution. All common ordering constraints are considered and general expressions of the reference priors are derived. The priors for the simple order are then illustrated in a Bayesian model of the "Risk of Type 2 Diabetes in New Mexico" data, where the relationship between the type 2 diabetes risk (through Hemoglobin A1c) and different smoking levels is investigated. In both simulation studies and real data set modeling, the reference priors that incorporate internal order information show good performances and can be used as default priors.

In the second part of this dissertation (Chapter 4) I focus on deriving the reference priors for conditionally autoregressive (CAR) models and simultaneous autoregressive (SAR) models with a spatial autoregression parameter ρ. When handling these two models in Bayesian statistics, a common choice of a prior for ρ is to use a Uniform prior. Since ρ is bounded between $\lambda_{(1)}^{-1}$ and $\lambda_{(n)}^{-1}$, the resulting prior on ρ may be related to these two eigen values. The reference priors for the CAR and SAR models are illustrated in the "1999 SAT State Average Verbal Scores" data with a comparison to Uniform prior. In the process of calculating these reference priors, one has to deal with the traces of the proximity matrix and variance-covariance matrix. A practical problem arises when these matrices are large or non-sparse as the computation time can quickly increase. Due to the complexity of the reference priors for

both CAR and SAR models, I only consider a portion (12 states in the Midwest) of the original data. However, it should be not hard to evaluate the prior density for a specific ρ value and that is all we need for proceeding Metropolis-Hastings algorithm, which basically means that analyzing the 48 states on the whole map is not impossible. The reference priors can give a different marginal posterior distribution compared with Uniform prior, which provides another choice when facing areal data in spatial statistics.

In both of the major topics of this thesis I have come to two conclusions: first, constrained parameter spaces are a useful way to incorporate subjective information and to enforce regularity conditions. Second, the reference prior framework provides a way to build a prior taking the constraints into account. While finding the reference prior can be challenging, I have shown there is value in these models.

5.2 Future Work

After graduation, I plan to continue my research on reference priors and spatial statistics. I will describe several problems that I intend to study as follows.

Ordering constraints are popular among different likelihoods and incorporating this information into the prior distribution is attractive and helpful for inference. Sonksen and Peruggia [2012] and Sonksen and Peruggia [2014] talk about the simple order for several discrete likelihoods with an emphasis on Poisson. It may be worth digging further to find general expressions of the priors that could work for all common distributions.

The idea adopted in this dissertation can be easily extended to order-constrained variance-covariance structures. For example, if the data presents heteroscedasticity and common ordering constraints exist for the variances, it would be interesting

to derive and evaluate the reference priors that fit this situation. A related, and important problem, is how to test the existence of these constraints. Because the parameter space is non-compact and the priors could be improper, a testing procedure not based on Bayes factors will have to be developed.

Spatial statistics draw many people's interests and the importance of considering spatial association has been realized by researchers from different fields. After graduation, I will finish analyzing the whole U.S. 1999 state average SAT data set with the reference priors derived in this dissertation. Furthermore, I may move to other examples where CAR/SAR priors are used hierarchically. Recently, lots of attention has been paid on non-Gaussian spatial models, such as spatial or spatial-temporal model on extreme value data, for example, extreme weather events including droughts, downpours, heat waves, atmospheric rivers, tropical cyclones, and hurricanes. In these models, prediction should be emphasized since these events may severely affect human's daily life. Under the Bayesian framework, I would like to develop new methodology to model this type of data. As providing a universal prior for a given likelihood is always something that should be achieved at the beginning, I will start with a focus on the prior studies.

Appendices

Appendix A

Derivation of the Full Conditional Distributions for the ANCOVA Model

From Chapter 3 we know the ANCOVA model for the simulation studies is

$$\boldsymbol{y} \;=\; X\boldsymbol{\alpha}+\boldsymbol{\varepsilon} \;=\; X_1\boldsymbol{\mu}+X_3\boldsymbol{\beta}+\boldsymbol{\varepsilon},\; \boldsymbol{\varepsilon}\sim N(\mathbf{0},\sigma^2 I). \tag{A.1}$$

For the convenience of derivation, we can assume there are m observations in each group. In matrix form we can write

$$\begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1m} \\ y_{21} \\ \vdots \\ y_{km} \end{pmatrix}_{n\times 1} = \begin{pmatrix} \mu_1 + X_{11}\beta_1 \\ \mu_1 + X_{12}\beta_1 \\ \vdots \\ \mu_1 + X_{1m}\beta_1 \\ \mu_2 + X_{21}\beta_2 \\ \vdots \\ \mu_k + X_{km}\beta_k \end{pmatrix}_{n\times 1} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1m} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{km} \end{pmatrix}_{n\times 1}.$$

This tells us the likelihood function for this model can be written as

$$f(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\beta},\sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \times \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{m}(y_{ij}-\mu_i-X_{ij}\beta_i)^2\right]. \quad \text{(A.2)}$$

From the discussion in Chapter 3 we know the prior distribution, $\pi(\boldsymbol{\theta})$, does not contain $\boldsymbol{\beta}$, so

$$\begin{aligned} f(\beta_i|\boldsymbol{\mu},\beta_1,\ldots,\beta_k,\sigma^2,\boldsymbol{y}) &\propto f(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\beta},\sigma^2)\times\pi(\boldsymbol{\theta}) \\ &\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{m}(y_{ij}-\mu_i-X_{ij}\beta_i)^2\right] \\ &\propto \exp\left[-\frac{1}{2\sigma^2}\left(\beta_i^2\sum_{j=1}^{m}(X_{ij})^2-2\beta_i\sum_{j=1}^{m}[X_{ij}(y_{ij}-\mu_i)]\right)\right] \\ &\propto \exp\left[-\frac{\sum_{j=1}^{m}(X_{ij})^2}{2\sigma^2}\left(\beta_i-\frac{\sum_{j=1}^{m}[X_{ij}(y_{ij}-\mu_i)]}{\sum_{j=1}^{m}(X_{ij})^2}\right)^2\right] \\ &= N\left(\frac{\sum_{j=1}^{m}[X_{ij}(y_{ij}-\mu_i)]}{\sum_{j=1}^{m}(X_{ij})^2},\frac{\sigma^2}{\sum_{j=1}^{m}(X_{ij})^2}\right). \end{aligned} \quad \text{(A.3)}$$

Similarly,

$$\begin{aligned} f(\sigma^2|\boldsymbol{\mu},\boldsymbol{\beta},\boldsymbol{y}) &\propto f(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\beta},\sigma^2)\times\pi(\boldsymbol{\theta}) \\ &\propto \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \times \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-X\boldsymbol{\alpha})'(\boldsymbol{y}-X\boldsymbol{\alpha})\right\}\times\frac{1}{\sigma^2} \\ &\propto \exp\left\{-\frac{(\boldsymbol{y}-X\boldsymbol{\alpha})'(\boldsymbol{y}-X\boldsymbol{\alpha})}{2\sigma^2}\right\}\times\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \\ &= IG\left(\frac{n}{2},\frac{(\boldsymbol{y}-X\boldsymbol{\alpha})'(\boldsymbol{y}-X\boldsymbol{\alpha})}{2}\right). \end{aligned} \quad \text{(A.4)}$$

Appendix B

Derivation of the Reference Prior for the CAR Model

B.1 Derivation of the Reference Prior

From Chapter 4 we know that the likelihood function for CAR model with linear regressors is

$$f(\boldsymbol{y}|\boldsymbol{\beta},\tau^2,\rho) \;\propto\; [\det(V)]^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}(\boldsymbol{y}-X\boldsymbol{\beta})'V^{-1}(\boldsymbol{y}-X\boldsymbol{\beta})\right\}, \tag{B.1}$$

where $V = \sum_{\boldsymbol{y}} = \tau^2[D_w - \rho W]^{-1}$.

$$\begin{aligned} l &= \log f(\boldsymbol{y}|\boldsymbol{\beta},\tau^2,\rho) \\ &= C - \frac{1}{2}\log[\det(V)] - \frac{1}{2}\left\{(\boldsymbol{y}-X\boldsymbol{\beta})'V^{-1}(\boldsymbol{y}-X\boldsymbol{\beta})\right\}. \end{aligned} \tag{B.2}$$

$$\frac{\partial l}{\partial \beta} = \frac{1}{\tau^2}X'(D_w - \rho W)(\boldsymbol{y}-X\boldsymbol{\beta}).$$

$$\frac{\partial^2 l}{\partial \beta^2} = -\frac{1}{\tau^2} X'(D_w - \rho W)X \Rightarrow -E\left[\frac{\partial^2 l}{\partial \beta^2}\right] = \frac{1}{\tau^2} X'(D_w - \rho W)X. \tag{B.3}$$

$$\frac{\partial^2 l}{\partial \beta \partial \tau^2} = -\frac{1}{\tau^4} X'(D_w - \rho W)(\boldsymbol{y} - X\boldsymbol{\beta}) \Rightarrow -E\left[\frac{\partial^2 l}{\partial \beta \partial \tau^2}\right] = 0. \tag{B.4}$$

$$\frac{\partial^2 l}{\partial \beta \partial \rho} = -\frac{1}{\tau^2} X'W(\boldsymbol{y} - X\boldsymbol{\beta}) \Rightarrow -E\left[\frac{\partial^2 l}{\partial \beta \partial \rho}\right] = 0. \tag{B.5}$$

$$\begin{aligned}
\frac{\partial l}{\partial \tau^2} &= -\frac{1}{2} \times \frac{1}{\det(V)} \times \frac{\partial \det(V)}{\partial \tau^2} + \frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\beta})' \frac{1}{\tau^4}(D_w - \rho W)(\boldsymbol{y} - X\boldsymbol{\beta}) \\
&= -\frac{1}{2\det(V)} \times \det(V) \times tr\left[V^{-1} \times \frac{\partial V}{\partial \tau^2}\right] + \frac{1}{2\tau^4}(\boldsymbol{y} - X\boldsymbol{\beta})'(D_w - \rho W)(\boldsymbol{y} - X\boldsymbol{\beta}) \\
&= -\frac{1}{2} tr\left[\frac{1}{\tau^2}(D_w - \rho W)(D_w - \rho W)^{-1}\right] + \frac{1}{2\tau^4}(\boldsymbol{y} - X\boldsymbol{\beta})'(D_w - \rho W)(\boldsymbol{y} - X\boldsymbol{\beta}) \\
&= -\frac{1}{2\tau^2} + \frac{1}{2\tau^4}(\boldsymbol{y} - X\boldsymbol{\beta})'(D_w - \rho W)(\boldsymbol{y} - X\boldsymbol{\beta}).
\end{aligned}$$

$$\frac{\partial^2 l}{\partial (\tau^2)^2} = \frac{1}{2\tau^4} - \frac{1}{\tau^6}(\boldsymbol{y} - X\boldsymbol{\beta})'(D_w - \rho W)(\boldsymbol{y} - X\boldsymbol{\beta}).$$

$$\begin{aligned}
-E\left[\frac{\partial^2 l}{\partial (\tau^2)^2}\right] &= -\frac{1}{2\tau^4} + \frac{1}{\tau^6} E[(\boldsymbol{y} - X\boldsymbol{\beta})'(D_w - \rho W)(\boldsymbol{y} - X\boldsymbol{\beta})] \\
&= -\frac{1}{2\tau^4} + \frac{1}{\tau^6} tr[(D_w - \rho W)V] \\
&= \frac{1}{2\tau^4}.
\end{aligned} \tag{B.6}$$

$$\frac{\partial^2 l}{\partial \tau^2 \rho} = -\frac{1}{2\tau^4}(\boldsymbol{y} - X\boldsymbol{\beta})'W(\boldsymbol{y} - X\boldsymbol{\beta}).$$

$$\begin{aligned}
-E\left[\frac{\partial^2 l}{\partial \tau^2 \rho}\right] &= \frac{1}{2\tau^4} E[(\boldsymbol{y} - X\boldsymbol{\beta})'W(\boldsymbol{y} - X\boldsymbol{\beta})] \\
&= \frac{1}{2\tau^4} tr[WV].
\end{aligned} \tag{B.7}$$

Appendix B. Derivation of the Reference Prior for the CAR Model

$$\begin{aligned}
\frac{\partial l}{\partial \rho} &= -\frac{1}{2} \times \frac{1}{\det(V)} \times \frac{\partial \det(V)}{\partial \rho} + \frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\beta})' \frac{1}{\tau^2} W(\boldsymbol{y} - X\boldsymbol{\beta}) \\
&= -\frac{1}{2\det(V)} \times \det(V) \times tr\left[V^{-1} \times \frac{\partial V}{\partial \rho}\right] + \frac{1}{2\tau^2}(\boldsymbol{y} - X\boldsymbol{\beta})' W(\boldsymbol{y} - X\boldsymbol{\beta}) \\
&= -\frac{1}{2} tr\left[V^{-1} \times (-V) \times \frac{\partial V^{-1}}{\partial \rho} \times V\right] + \frac{1}{2\tau^2}(\boldsymbol{y} - X\boldsymbol{\beta})' W(\boldsymbol{y} - X\boldsymbol{\beta}) \\
&= -\frac{1}{2\tau^2} tr[WV] + \frac{1}{2\tau^2}(\boldsymbol{y} - X\boldsymbol{\beta})' W(\boldsymbol{y} - X\boldsymbol{\beta}).
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l}{\partial \rho^2} &= -\frac{1}{2\tau^2} tr\left[W \times \frac{\partial V}{\rho}\right] = -\frac{1}{2\tau^2} tr\left[W \times (-V) \times \frac{\partial V^{-1}}{\rho} \times V\right] \\
&= -\frac{1}{2\tau^4} tr[WVWV].
\end{aligned}$$

$$-E\left[\frac{\partial^2 l}{\partial \rho^2}\right] = \frac{1}{2\tau^4} tr[WVWV]. \tag{B.8}$$

Based on B.3, B.4, B.5, B.6, B.7 and B.8 we have the Fisher information matrix, which is

$$I(\rho, \tau^2, \boldsymbol{\beta}) = \begin{pmatrix} \frac{tr[WVWV]}{2\tau^4} & \frac{tr[WV]}{2\tau^4} & 0 \\ \frac{tr[WV]}{2\tau^4} & \frac{1}{2\tau^4} & 0 \\ 0 & 0 & \frac{1}{\tau^2} X'(D_w - \rho W)X \end{pmatrix}.$$

Let's group our parameters as $\{\rho\}$, $\{\tau^2\}$, $\{\boldsymbol{\beta}\}$, then

$$\begin{aligned}
S(\rho, \tau^2, \boldsymbol{\beta}) &= [I(\rho, \tau^2, \boldsymbol{\beta})]^{-1} \\
&= \begin{pmatrix} \frac{\frac{1}{2\tau^4}}{\frac{1}{2\tau^4}\frac{tr[WVWV]}{2\tau^4} - \left(\frac{tr[WV]}{2\tau^4}\right)^2} & \frac{-\frac{1}{2\tau^4} tr[WV]}{\frac{1}{2\tau^4}\frac{tr[WVWV]}{2\tau^4} - \left(\frac{tr[WV]}{2\tau^4}\right)^2} & 0 \\ \frac{-\frac{1}{2\tau^4} tr[WV]}{\frac{1}{2\tau^4}\frac{tr[WVWV]}{2\tau^4} - \left(\frac{tr[WV]}{2\tau^4}\right)^2} & \frac{\frac{1}{2\tau^4} tr[WVWV]}{\frac{1}{2\tau^4}\frac{tr[WVWV]}{2\tau^4} - \left(\frac{tr[WV]}{2\tau^4}\right)^2} & 0 \\ 0 & 0 & \frac{\tau^2}{X'(D_w - \rho W)X} \end{pmatrix}.
\end{aligned}$$

Appendix B. Derivation of the Reference Prior for the CAR Model

Suppose $\boldsymbol{\theta} = \{\rho, \tau^2, \boldsymbol{\beta}\} \in \Theta$, then

$$h_1(\boldsymbol{\theta}) = \frac{tr[WVWV]}{2\tau^4} - \frac{tr[WV]tr[WV]}{2\tau^4}. \tag{B.9}$$

$$h_2(\boldsymbol{\theta}) = \frac{1}{2\tau^4}. \tag{B.10}$$

$$h_3(\boldsymbol{\theta}) = \frac{1}{\tau^2} X'(D_w - \rho W)X. \tag{B.11}$$

$|h_j(\boldsymbol{\theta})|$ depends only on $\boldsymbol{\theta}_{(1:j)}$, for $j = 1, 2, 3$. We need to define a compact subset Θ^l. Now, for j=1,

$$\Theta^l \cap [\Theta_j | \Theta_{(1:(j-1))}] = \{\rho : \lambda_{(1)}^{-1} < \rho < \lambda_{(n)}^{-1}\}.$$

For j=2,

$$\Theta^l \cap [\Theta_j | \Theta_{(1:(j-1))}] = \{\tau^2 : l^{-1} < \tau^2 < l\}.$$

For j=3,

$$\Theta^l \cap [\Theta_j | \Theta_{(1:(j-1))}] = \{\boldsymbol{\beta} : \boldsymbol{\beta} \in \boldsymbol{\beta}^l\}.$$

$$\begin{aligned}
\pi^l(\boldsymbol{\theta}) &= \prod_{i=1}^{m} \frac{|h_i(\boldsymbol{\theta})|^{1/2}}{\int_{\Theta^l \cap [\Theta_j | \Theta_{(1:(j-1))}]} |h_i(\boldsymbol{\theta})|^{1/2} d\boldsymbol{\theta}_{(i)}} I_{\Theta^l}(\boldsymbol{\theta}) \\
&= \frac{|h_1|^{1/2}}{\int_{\lambda_{(1)}^{-1}}^{\lambda_{(n)}^{-1}} |h_1|^{1/2} d\rho} \times \frac{1/\tau^2}{\int_{l^{-1}}^{l} 1/\tau^2 d\tau^2} \times \frac{1}{\int_{\beta^l} 1 d\boldsymbol{\beta}} \\
&= \frac{|h_1|^{1/2}}{\int_{\lambda_{(1)}^{-1}}^{\lambda_{(n)}^{-1}} |h_1|^{1/2} d\rho} \times \frac{1/\tau^2}{2 \log l} \times f(l).
\end{aligned}$$

Finally,

$$\begin{aligned}\pi(\boldsymbol{\theta}) &= \lim_{l\to\infty}\frac{\pi_1^l(\boldsymbol{\theta})}{\pi_1^l(\boldsymbol{\theta}^*)} \\ &\propto \left|\frac{tr[WVWV]}{\tau^4} - \frac{tr[WV]tr[WV]}{\tau^4}\right|^{1/2} \times \frac{1}{\tau^2} \times 1_{(\lambda_{(1)}^{-1},\, \lambda_{(n)}^{-1})}(\rho). \qquad \text{(B.12)}\end{aligned}$$

B.2 The Posterior and Full Conditional Distributions

$$\begin{aligned}p(\boldsymbol{\beta},\tau^2,\rho|\boldsymbol{y}) &\propto f(\boldsymbol{y}|\boldsymbol{\beta},\tau^2,\rho)\times\pi(\boldsymbol{\beta},\tau^2,\rho) \\ &\propto [\det(V)]^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}(\boldsymbol{y}-X\boldsymbol{\beta})'V^{-1}(\boldsymbol{y}-X\boldsymbol{\beta})\right\} \\ &\quad \times \left|\frac{tr[WVWV]}{\tau^4} - \frac{tr[WV]tr[WV]}{\tau^4}\right|^{1/2} \times \frac{1}{\tau^2} \times 1_{(\lambda_{(1)}^{-1},\, \lambda_{(n)}^{-1})}(\rho). \qquad \text{(B.13)}\end{aligned}$$

Hence,

$$\begin{aligned}p(\boldsymbol{\beta}|\tau^2,\rho,\boldsymbol{y}) &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{y}-X\boldsymbol{\beta})'V^{-1}(\boldsymbol{y}-X\boldsymbol{\beta})\right\} \\ &= \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-(X'V^{-1}X)^{-1}X'V^{-1}\boldsymbol{y})'X'V^{-1}X(\boldsymbol{\beta}-(X'V^{-1}X)^{-1}X'V^{-1}\boldsymbol{y})\right\} \\ &\sim MVN\left((X'V^{-1}X)^{-1}X'V^{-1}\boldsymbol{y}, (X'V^{-1}X)^{-1}\right). \qquad \text{(B.14)}\end{aligned}$$

$$\begin{aligned}p(\tau^2|\boldsymbol{\beta},\rho,\boldsymbol{y}) &\propto [\det(V)]^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}(\boldsymbol{y}-X\boldsymbol{\beta})'V^{-1}(\boldsymbol{y}-X\boldsymbol{\beta})\right\} \times \frac{1}{\tau^2} \\ &= \left\{(\tau^2)^n \det[(D_w-\rho W)^{-1}]\right\}^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}(\boldsymbol{y}-X\boldsymbol{\beta})'V^{-1}(\boldsymbol{y}-X\boldsymbol{\beta})\right\} \times \frac{1}{\tau^2} \\ &\propto (\tau^2)^{-\frac{n}{2}-1} \times \exp\left\{-\frac{(\boldsymbol{y}-X\boldsymbol{\beta})'(D_w-\rho W)(\boldsymbol{y}-X\boldsymbol{\beta})}{2\tau^2}\right\} \\ &\sim IG\left(\frac{n}{2}, \frac{(\boldsymbol{y}-X\boldsymbol{\beta})'(D_w-\rho W)(\boldsymbol{y}-X\boldsymbol{\beta})}{2}\right). \qquad \text{(B.15)}\end{aligned}$$

Appendix B. Derivation of the Reference Prior for the CAR Model

As I mentioned in Chapter 4, the reference prior on ρ in the CAR model is a high order polynomial. There is no easy way to derive the full conditional distribution for parameter ρ, which is also the reason why the Metropolis-Hastings algorithm is used for this parameter.

Appendix C

Derivation of the Reference Prior for the SAR Model

C.1 Derivation of the Reference Prior

From Chapter 4 we know that the likelihood function for SAR model with linear regressors is

$$f(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2,\rho) \;\propto\; [\det(V)]^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-X\boldsymbol{\beta})'(I_n-\rho B)'D_w(I_n-\rho B)(\boldsymbol{y}-X\boldsymbol{\beta})\right\}, \tag{C.1}$$

where $V = \sum_{\boldsymbol{y}} = \sigma^2(I_n-\rho B)^{-1}D_w^{-1}((I_n-\rho B)^{-1})'$.
First, it is useful to define

$$\begin{aligned} A &= \frac{\partial[(I_n-\rho B)'D_w(I_n-\rho B)]}{\partial\rho} \\ &= -[(I_n-\rho B)'D_wB + B'D_w(I_n-\rho B)]. \end{aligned}$$

So,

$$\frac{\partial A}{\partial \rho} = 2B' D_w B$$

$$\begin{aligned} l &= \log f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \rho) \\ &= c - \frac{1}{2}\log[\det(V)] - \frac{1}{2}\left\{(\boldsymbol{y} - X\boldsymbol{\beta})' V^{-1} (\boldsymbol{y} - X\boldsymbol{\beta})\right\}. \end{aligned} \tag{C.2}$$

$$\frac{\partial l}{\partial \beta} = \frac{1}{\sigma^2} X' (I_n - \rho B)' D_w (I_n - \rho B)(\boldsymbol{y} - X\boldsymbol{\beta}).$$

$$\frac{\partial^2 l}{\partial \beta^2} = -\frac{1}{\sigma^2} X' (I_n - \rho B)' D_w (I_n - \rho B) X.$$

$$-E\left[\frac{\partial^2 l}{\partial \beta^2}\right] = \frac{1}{\sigma^2} X' (I_n - \rho B)' D_w (I_n - \rho B) X. \tag{C.3}$$

$$\frac{\partial^2 l}{\partial \beta \partial \sigma^2} = -\frac{1}{\sigma^4} X' (I_n - \rho B)' D_w (I_n - \rho B)(\boldsymbol{y} - X\boldsymbol{\beta}).$$

$$-E\left[\frac{\partial^2 l}{\partial \beta \partial \sigma^2}\right] = 0. \tag{C.4}$$

$$\frac{\partial^2 l}{\partial \beta \partial \rho} = \frac{1}{\sigma^2} X' A (\boldsymbol{y} - X\boldsymbol{\beta}).$$

$$-E\left[\frac{\partial^2 l}{\partial \beta \partial \rho}\right] = 0. \tag{C.5}$$

$$\begin{aligned}
\frac{\partial l}{\partial \sigma^2} &= -\frac{1}{2} \times \frac{1}{\det(V)} \times \frac{\partial \det(V)}{\partial \sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{y} - X\boldsymbol{\beta})'(I_n - \rho B)' D_w (I_n - \rho B)(\boldsymbol{y} - X\boldsymbol{\beta}) \\
&= -\frac{1}{2\det(V)} \times \det(V) \times tr\left[V^{-1} \times \frac{\partial V}{\partial \tau^2}\right] \\
&\quad + \frac{1}{2\sigma^4}(\boldsymbol{y} - X\boldsymbol{\beta})'(I_n - \rho B)' D_w (I_n - \rho B)(\boldsymbol{y} - X\boldsymbol{\beta}) \\
&= -\frac{1}{2} tr\left[\frac{1}{\sigma^2}(I_n - \rho B)' D_w (I_n - \rho B)(I_n - \rho B)^{-1} D_w^{-1} ((I_n - \rho B)^{-1})'\right] \\
&\quad + \frac{1}{2\sigma^4}(\boldsymbol{y} - X\boldsymbol{\beta})'(I_n - \rho B)' D_w (I_n - \rho B)(\boldsymbol{y} - X\boldsymbol{\beta}) \\
&= -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{y} - X\boldsymbol{\beta})'(I_n - \rho B)' D_w (I_n - \rho B)(\boldsymbol{y} - X\boldsymbol{\beta}).
\end{aligned}$$

$$\frac{\partial^2 l}{\partial (\sigma^2)^2} = \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}(\boldsymbol{y} - X\boldsymbol{\beta})'(I_n - \rho B)' D_w (I_n - \rho B)(\boldsymbol{y} - X\boldsymbol{\beta}).$$

$$\begin{aligned}
-E\left[\frac{\partial^2 l}{\partial (\sigma^2)^2}\right] &= -\frac{1}{2\sigma^4} + \frac{1}{\sigma^6} E[(\boldsymbol{y} - X\boldsymbol{\beta})'(I_n - \rho B)' D_w (I_n - \rho B)(\boldsymbol{y} - X\boldsymbol{\beta})] \\
&= -\frac{1}{2\sigma^4} + \frac{1}{\sigma^6} tr[(I_n - \rho B)' D_w (I_n - \rho B) V] \\
&= \frac{1}{2\sigma^4}. \qquad \text{(C.6)}
\end{aligned}$$

$$\frac{\partial^2 l}{\partial \sigma^2 \rho} = \frac{1}{2\sigma^4}(\boldsymbol{y} - X\boldsymbol{\beta})' A (\boldsymbol{y} - X\boldsymbol{\beta}).$$

$$\begin{aligned}
-E\left[\frac{\partial^2 l}{\partial \sigma^2 \rho}\right] &= -\frac{1}{2\sigma^4} E[(\boldsymbol{y} - X\boldsymbol{\beta})' A (\boldsymbol{y} - X\boldsymbol{\beta})] \\
&= -\frac{1}{2\sigma^4} tr[AV]. \qquad \text{(C.7)}
\end{aligned}$$

Appendix C. Derivation of the Reference Prior for the SAR Model

$$\begin{aligned}
\frac{\partial l}{\partial \rho} &= -\frac{1}{2} \times \frac{1}{\det(V)} \times \frac{\partial \det(V)}{\partial \rho} - \frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})'A(\boldsymbol{y} - X\boldsymbol{\beta}) \\
&= -\frac{1}{2\det(V)} \times \det(V) \times tr\left[V^{-1} \times \frac{\partial V}{\partial \rho}\right] - \frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})'A(\boldsymbol{y} - X\boldsymbol{\beta}) \\
&= -\frac{1}{2}tr\left[V^{-1} \times (-V) \times \frac{\partial V^{-1}}{\partial \rho} \times V\right] - \frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})'A(\boldsymbol{y} - X\boldsymbol{\beta}) \\
&= \frac{1}{2\sigma^2}tr[AV] - \frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})'A(\boldsymbol{y} - X\boldsymbol{\beta}).
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l}{\partial \rho^2} &= \frac{1}{2\sigma^2}tr\left[(2B'D_wB)V + A\frac{\partial V}{\rho}\right] - \frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})'(2B'D_wB)(\boldsymbol{y} - X\boldsymbol{\beta}) \\
&= \frac{1}{2\sigma^2}tr\left[2B'D_wBV + A(-V) \times \frac{\partial V^{-1}}{\rho} \times V\right] - \frac{1}{\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})'B'D_wB(\boldsymbol{y} - X\boldsymbol{\beta}) \\
&= \frac{1}{2\sigma^2}tr\left[2B'D_wBV - \frac{AVAV}{\sigma^2}\right] - \frac{1}{\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})'B'D_wB(\boldsymbol{y} - X\boldsymbol{\beta}).
\end{aligned}$$

$$-E\left[\frac{\partial^2 l}{\partial \rho^2}\right] = -\frac{1}{2\sigma^2}tr\left[2B'D_wBV - \frac{AVAV}{\sigma^2}\right] + \frac{1}{\sigma^2}tr\left[B'D_wBV\right]. \qquad \text{(C.8)}$$

Based on C.3, C.4, C.5, C.6, C.7 and C.8 we have the information matrix, which is:

$$I(\rho, \sigma^2, \boldsymbol{\beta}) = \begin{pmatrix} C.8 & C.7 & 0 \\ C.7 & C.6 & 0 \\ 0 & 0 & C.3 \end{pmatrix}.$$

Let's group our parameters as $\{\rho\}$, $\{\tau^2\}$, $\{\boldsymbol{\beta}\}$, then

$$\begin{aligned}
S(\rho, \tau^2, \boldsymbol{\beta}) &= [I(\rho, \tau^2, \boldsymbol{\beta})]^{-1} \\
&= \begin{pmatrix} \cdot\frac{C.6}{C.6 \times C.8 - C.7 \times C.7} & \frac{-C.7}{C.6 \times C.8 - C.7 \times C.7} & 0 \\ \frac{-C.7}{C.6 \times C.8 - C.7 \times C.7} & \frac{C.8}{C.6 \times C.8 - C.7 \times C.7} & 0 \\ 0 & 0 & \frac{1}{C.3} \end{pmatrix}
\end{aligned}$$

Appendix C. Derivation of the Reference Prior for the SAR Model

Suppose $\boldsymbol{\theta} = \{\rho, \sigma^2, \boldsymbol{\beta}\} \in \Theta$, then

$$\begin{aligned} h_1(\boldsymbol{\theta}) &= \frac{C.6 \times C.8 - C.7 \times C.7}{C.6} \\ &= \frac{1}{\sigma^2} tr\left[B' D_w B V\right] - \frac{1}{2\sigma^2} tr\left[2B' D_w B V - \frac{AVAV}{\sigma^2}\right] - \frac{1}{2\sigma^4} tr[AV] tr[AV]. \end{aligned} \tag{C.9}$$

$$\begin{aligned} h_2(\boldsymbol{\theta}) &= C.6 \\ &= \frac{1}{2\tau^4}. \end{aligned} \tag{C.10}$$

$$\begin{aligned} h_3(\boldsymbol{\theta}) &= C.3 \\ &= \frac{1}{\sigma^2} X'(I_n - \rho B)' D_w (I_n - \rho B) X. \end{aligned} \tag{C.11}$$

$|h_j(\boldsymbol{\theta})|$ depends only on $\boldsymbol{\theta}_{(1:j)}$, for $j = 1, 2, 3$. We need to define a compact subset Θ^l. Now, for j=1,

$$\Theta^l \cap [\Theta_j | \Theta_{(1:(j-1))}] = \{\rho : \lambda_{(1)}^{-1} < \rho < \lambda_{(n)}^{-1}\}.$$

For j=2,

$$\Theta^l \cap [\Theta_j | \Theta_{(1:(j-1))}]. = \{\sigma^2 : l^{-1} < \tau^2 < l\}$$

For j=3,

$$\Theta^l \cap [\Theta_j | \Theta_{(1:(j-1))}] = \{\boldsymbol{\beta} : \boldsymbol{\beta} \in \boldsymbol{\beta}^l\}.$$

$$\begin{aligned}
\pi^l(\boldsymbol{\theta}) &= \prod_{i=1}^{m} \frac{|h_i(\boldsymbol{\theta})|^{1/2}}{\int_{\Theta^l \cap [\Theta_j | \Theta_{(1:(j-1))}]} |h_i(\boldsymbol{\theta})|^{1/2} d\boldsymbol{\theta}_{(i)}} I_{\Theta^l}(\boldsymbol{\theta}) \\
&= \frac{|h_1|^{1/2}}{\int_{\lambda_{(1)}^{-1}}^{\lambda_{(n)}^{-1}} |h_1|^{1/2} d\rho} \times \frac{1/\tau^2}{\int_{l^{-1}}^{l} 1/\tau^2 d\tau^2} \times \frac{1}{\int_{\beta^l} 1 d\boldsymbol{\beta}} \\
&= \frac{|h_1|^{1/2}}{\int_{\lambda_{(1)}^{-1}}^{\lambda_{(n)}^{-1}} |h_1|^{1/2} d\rho} \times \frac{1/\tau^2}{2 \log l} \times f(l).
\end{aligned}$$

Finally,

$$\begin{aligned}
\pi(\boldsymbol{\theta}) &= \lim_{l \to \infty} \frac{\pi_1^l(\boldsymbol{\theta})}{\pi_1^l(\boldsymbol{\theta}^*)} \\
&\propto |h_1|^{1/2} \times \frac{1}{\sigma^2} \times 1_{(\lambda_{(1)}^{-1},\, \lambda_{(n)}^{-1})}(\rho). \qquad \text{(C.12)}
\end{aligned}$$

C.2 The Posterior and Full Conditional Distributions

$$\begin{aligned}
p(\boldsymbol{\beta}, \sigma^2, \rho | \boldsymbol{y}) &\propto f(\boldsymbol{y} | \boldsymbol{\beta}, \sigma^2, \rho) \times \pi(\boldsymbol{\beta}, \sigma^2, \rho) \\
&\propto [\det(V)]^{-\frac{1}{2}} \times \exp\left\{ -\frac{1}{2} (\boldsymbol{y} - X\boldsymbol{\beta})' V^{-1} (\boldsymbol{y} - X\boldsymbol{\beta}) \right\} \\
&\quad \times |h_1|^{1/2} \times \frac{1}{\sigma^2} \times 1_{(\lambda_{(1)}^{-1},\, \lambda_{(n)}^{-1})}(\rho). \qquad \text{(C.13)}
\end{aligned}$$

Hence,

$$\begin{aligned}
p(\boldsymbol{\beta} | \sigma^2, \rho, \boldsymbol{y}) &\propto \exp\left\{ -\frac{1}{2} (\boldsymbol{y} - X\boldsymbol{\beta})' V^{-1} (\boldsymbol{y} - X\boldsymbol{\beta}) \right\} \\
&= \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - (X'V^{-1}X)^{-1} X'V^{-1}\boldsymbol{y})' X'V^{-1}X (\boldsymbol{\beta} - (X'V^{-1}X)^{-1} X'V^{-1}\boldsymbol{y}) \right\} \\
&\sim MVN\left((X'V^{-1}X)^{-1} X'V^{-1}\boldsymbol{y}, (X'V^{-1}X)^{-1} \right). \qquad \text{(C.14)}
\end{aligned}$$

$$\begin{aligned} p(\sigma^2|\boldsymbol{\beta},\rho,\boldsymbol{y}) &\propto [\det(V)]^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}(\boldsymbol{y}-X\boldsymbol{\beta})'V^{-1}(\boldsymbol{y}-X\boldsymbol{\beta})\right\} \times \frac{1}{\sigma^2} \\ &\propto [(\sigma^2)^n]^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}(\boldsymbol{y}-X\boldsymbol{\beta})'V^{-1}(\boldsymbol{y}-X\boldsymbol{\beta})\right\} \\ &= (\sigma^2)^{-\frac{n}{2}-1} \times \exp\left\{-\frac{(\boldsymbol{y}-X\boldsymbol{\beta})'(I_n-\rho B)'D_w(I_n-\rho B)(\boldsymbol{y}-X\boldsymbol{\beta})}{2\sigma^2}\right\} \\ &\sim IG\left(\frac{n}{2}, \frac{(\boldsymbol{y}-X\boldsymbol{\beta})'(I_n-\rho B)'D_w(I_n-\rho B)(\boldsymbol{y}-X\boldsymbol{\beta})}{2}\right). \end{aligned} \tag{C.15}$$

Similar as the CAR model, the reference prior on ρ in the SAR model is also a high order polynomial. There is no easy way to derive the full conditional distribution for parameter ρ and that is the reason why the Metropolis-Hastings algorithm is used for this parameter.

References

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall, 2nd edition, 2014. ISBN 1439819173.

T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 53:330–418, 1763.

B. S. Bell and L. D. Broemeling. A Bayesian analysis of spatial processes with application to disease mapping. *Statistics in Medicine*, 19:957–974, 2000.

J. O. Berger. The case for objective Bayesian analysis. *Bayesian Analysis*, 1:385–402, 2006.

J. O. Berger and J. M. Bernardo. On the development of reference priors. *Bayesian Statistics 4*, 4:35–60, 1992.

J. O. Berger and D. Sun. Objective priors for the bivariate normal model. *Annals of Statistics*, 36:963–982, 2008.

J. O. Berger, J. Bernardo, and D. Sun. The formal definition of reference priors. *Annals of Statistics*, 37:905–938, 2009.

REFERENCES

J. M. Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 41:113–147, 1979.

J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. Ser. B*, 36:192–236, 1974.

J. Besag and C. Kooperberg. On conditional and intrinsic autoregressions. *Biometrika*, 82:733–746, 1995.

G. Celeux, F. Forbes, C. Robert, and D. Titterington. Deviance information criteria for missing data models. *Bayesian Analysis*, 1:651–706, 2006.

Centers for Disease Control and Prevention. New Mexico surveillance data, 2012. URL `http://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html`.

R. Christensen. *Plane Answers to Complex Questions*. Springer, 4th edition, 2011. ISBN 1441998152.

G. Claeskens and N. L. Hjort. *Model Selection and Model Averaging*. Cambridge University Press, 1st edition, 2008. ISBN 0521852250.

B. Clarke and D. Sun. Reference priors under the chi-squared distance. *Sankhya: The Indian Journal of Statistics, Series A*, 59:215–231, 1997.

N. Cressie and P. Kapat. Some diagnostics for Markov random fields. *Journal of Computational and Graphical Statistics*, 17:726–749, 2008.

V. De Oliveira. Bayesian analysis of conditional autoregressive models. *Annals of the Institute of Statistical Mathematics*, 64:107–133, 2012.

V. De Oliveira and J. J. Song. Bayesian analysis of simultaneous autoregressive models. *Sankhya: The Indian Journal of Statistics, Series B*, 70:323–350, 2008.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39: 1–38, 1977.

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195:216–222, 1987.

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

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Dunson. *Bayesian Data Analysis*. CRC Press, 3rd edition, 2013. ISBN 9781439840955.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

C. J. Geyer and E. A. Thompson. Annealing Markov chain Monte Carlo with applications to pedigree analysis. *Technical report, School of Statistics, University of Minnesota*, 1993.

S. Ghosal. Probability matching priors for nonregular cases. *Biometrika*, 86:956–964, 1999.

M. Ghosh. Objective priors: An introduction for frequentists. *Statistical Science*, 26:187–202, 2011.

P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

L. W. Hepple. Bayesian techniques in spatial and network econometrics: 1. model comparison and posterior odds. *Environment and Planning A*, 27:447–469, 1995a.

L. W. Hepple. Bayesian techniques in spatial and network econometrics: 2. computational methods and algorithms. *Environment and Planning A*, 27:615–644, 1995b.

E. T. Jaynes. On the rationale of maximum entropy methods. *Proceedings of IEEE*, 70:939–952, 1982.

H. Jeffreys. *Theory of Probability*. London: Oxford University Press, 3rd edition, 1961. ISBN 158488388X.

M. Jordan and L. Xu. Convergence results for the EM approach to mixtures of experts architectures. *Neural Networks*, 8:1409–1431, 1996.

S. Karlin and H. M. Taylor. *A First Course in Stochastic Processes*. Academic Press, San Diego, CA, 1975.

R. E. Kass and L. Wasserman. The selection of prior distributions by formal rules. *The Annals of the American Statistical Association*, 91:1343–1370, 1996.

P. S. Laplace. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 41:113–147, 1774.

I. C. Marschner. Stable computation of maximum likelihood estimates in identity link Poisson regression. *Journal of Computational and Graphical Statistics*, 19: 666–683, 2010.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21: 1087–1091, 1953.

J. L. Morrissette and M. P. McDermott. Estimation and inference concerning ordered means in analysis of covariance models with interactions. *Journal of the American Statistical Association*, 108:832–839, 2013.

R. M. Neal. Slice sampling. *Annals of Statistics*, 31:705–767, 2003.

G. Nicholls and M. Jones. Radiocarbon dating with temporal order constraints. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50:503–521, 2001.

M. Plummer. Penalized loss functions for Bayesian model comparison. *Biostatistics*, 9:523–539, 2008.

R. Redner and H. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26, 1984.

C. Ren and D. Sun. Objective Bayesian analysis for autoregressive models with nugget effects. *Journal of Multivariate Analysis*, 124:260–280, 2014.

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

G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7:110–120, 1997.

M. Sawada. Global spatial autocorrelation indices - Moran's I, Geary's C and the general cross-product statistic., September 2009. URL `http://www.lpc.uottawa.ca/publications/moransi/moran.htm`.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 14:461–464, 1978.

A. Sen. Large sample-size distribution of statistics used in testing for spatial correlation. *Geographical Analysis*, 9:175–184, 1976.

M. D. Sonksen and M. Peruggia. Reference priors for constrained rate models of count data. *Journal of Statistical Planning and Inference*, 142:3023–3036, 2012.

M. D. Sonksen and M. Peruggia. Inferences on lung cancer mortality rates based on reference priors under partial ordering. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63:783–800, 2014.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64:583–639, 2002.

C. Stein. An example of wide discrepancy between fiducial and confidence intervals. *Annals of Mathematical Statistics*, 30:877–880, 1959.

M. Stone. Strong inconsistency from uniform priors. *Journal of the American Statistical Association*, 71:114–125, 1976.

M. Stone and A. P. Dawid. Un-Bayesian implications of improper Bayes inference in routine statistical problems. *Biometrika*, 59:369–375, 1972.

D. Sun and J. O. Berger. Reference priors with partial information. *Biometrika*, 85: 55–71, 1998.

D. Sun and J. O. Berger. Objective Bayesian analysis for the multivariate normal model. *ISBA 8th World Meeting on Bayesian Statistics*, 2006.

R. Tibshirani. Noninformative priors for one parameter of many. *Biometrica*, 76: 604–608, 1989.

L. Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22:1701–1728, 1994.

M. M. Wall. A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, 121:311–324, 2004.

P. Whittle. On stationary process in the plane. *Biometrika*, 41:434–449, 1954.

C. F. J. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11:95–103, 1983.

X. Xie, Q. Liu, J. Wu, and M Wakui. Impact of cigarette smoking in type 2 diabetes development. *Acta Pharmacologica Sinica*, 30:784–787, 2009.

L. Xu and M. I. Jordan. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, 8:129–151, 1996.

R. Yang. Invariance of the reference prior under reparametrization. *Test*, 4:83–94, 1995.