Date 19.02.09

UC
UNIVERSITY OF
CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Exact Markov Chain Monte Carlo and Bayesian Linear Regression

A thesis submitted in partial fulfillment of the requirements
for the Degree of

Masters of Science in Statistics

in the University of Canterbury

by

Jason Bentley

ABSTRACT

In this work we investigate the use of perfect sampling methods within the context of Bayesian linear regression. We focus on inference problems related to the marginal posterior model probabilities. Model averaged inference for the response and Bayesian variable selection are considered. Perfect sampling is an alternate form of Markov chain Monte Carlo that generates exact sample points from the posterior of interest. This approach removes the need for burn-in assessment faced by traditional MCMC methods. For model averaged inference, we find the monotone Gibbs coupling from the past (CFTP) algorithm is the preferred choice. This requires the predictor matrix be orthogonal, preventing variable selection, but allowing model averaging for prediction of the response. Exploring choices of priors for the parameters in the Bayesian linear model, we investigate sufficiency for monotonicity assuming Gaussian errors. We discover that a number of other sufficient conditions exist, besides an orthogonal predictor matrix, for the construction of a monotone Gibbs Markov chain. Requiring an orthogonal predictor matrix, we investigate new methods of orthogonalizing the original predictor matrix. We find that a new method using the modified Gram-Schmidt orthogonalization procedure performs comparably with existing transformation methods, such as generalized principal components. Accounting for the effect of using an orthogonal predictor matrix, we discover that inference using model averaging for in-sample prediction of the response is comparable between the original and orthogonal predictor matrix. The Gibbs sampler is then investigated for sampling when using the original predictor matrix and the orthogonal predictor matrix. We find that a hybrid method, using a standard Gibbs sampler on the orthogonal space in conjunction with the monotone CFTP Gibbs sampler, provides the fastest computation and convergence to the posterior distribution. We conclude the hybrid approach should be used when the monotone Gibbs CFTP sampler becomes impractical, due to large backwards coupling times. We demonstrate large backwards coupling times occur when the sample size is close to the number of predictors, or when hyper-parameter choices increase model competition. The monotone Gibbs CFTP sampler should be taken advantage of when the backwards coupling time is small. For the problem of variable selection we turn to the exact version of the independent Metropolis-Hastings (IMH) algorithm. We reiterate the notion that the exact IMH sampler is redundant, being a needlessly complicated rejection sampler. We then determine a rejection sampler is feasible for variable selection when the sample size is close to the number of predictors and using Zellner's prior with a small value for the hyper-parameter *c*. Finally, we use the example of simulating from the posterior of *c* conditional on a model to demonstrate how the use of an exact IMH view-point clarifies how the rejection sampler can be adapted to improve efficiency.

This work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, this thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Jason Phillip Bentley

(19th February 2009)

ACKNOWLEDGEMENTS

It is a great honor and pleasure to be able to, for the first time in my studies, put down on paper those to whom I owe my deepest gratitude for my education and improvement. To those who have provided moral support and advice in times, good or bad: my family, my friends and my Grandmother (whom I miss greatly), this thesis is dedicated to you. To my supervisor, Dominic Lee, no words could express the thanks I wish to convey. You took a chance on someone who through your very own courses obtained a semester's experience in Computational statistics, Bayesian analysis, and MCMC. Now, having completed a thesis on such topics after my introduction to them only four months prior to beginning this work, you are without a doubt a dedicated and patient supervisor. To Christian Robert, my associate supervisor, thank you for your teaching in Bayesian analysis, now the Bayesian Core, when you taught at the University of Canterbury in 2006. Your comments, direction and subtle suggestions have no doubt improved the quality of this thesis, especially from our discussions at the ISBA 2008 conference. From discussions at the ISBA 2008 conference, I would also like to thank Gareth Roberts and Chris Holmes for comments that greatly improved chapters 2 and 3. Finally, to the support team including lecturers, administrative and technical support staff within the Mathematics and Statistics department here at the University of Canterbury, and those whom I had the pleasure of meeting and talking with at the conferences I had attended while completing this thesis, thank you for all you had done to aid me in this work and this thesis both directly and indirectly.

Jason Bentley

CONTENTS

CHAPTER 5

CHAPTER 6

NOTATION

f - density function.

X - random variable(s).

X_n - Markov chain.

$\mathbf{y}$ - response vector (data).

$\mathbf{X}$ - design matrix (data).

$\mathbf{H}$ - hat matrix for $\mathbf{X}$.

c - hyper-parameter for Zellner's prior.

$\boldsymbol{\gamma}$ - variable selection parameter (binary vector).

$\boldsymbol{\Gamma}$ - state space for $\boldsymbol{\gamma}$.

$\boldsymbol{\beta}$ - vector of regression coefficients.

e - random perturbations.

σ^2 - variance.

$\Gamma(x)$ - gamma function.

$|\mathbf{A}|$ - determinant of the matrix $\mathbf{A}$.

$\mathbf{W}$ - orthogonal design matrix.

$\mathbb{E}$ - expectation.

$\mathbb{V}$ - variance.

$\| a \|$ is the norm of a.

ABBREVIATIONS

AIC: Aikake Information Criterion.

BF: Bayes Factor.

BIC: Bayesian Information Criterion.

BMA: Bayesian Model Averaging.

BVS: Bayesian Variable Selection.

CFTP: Coupling From the Past.

DIC: Deviance Information Criterion.

EB: Empirical Bayes.

i.i.d.: independently and identically distributed.

IMH: Independent Metropolis Hastings.

LSE: Least Squares Estimator.

MAP: Maximum A Posteriori.

MC: Markov Chain(s).

MCMC: Markov Chain Monte Carlo.

MIP: Marginal Inclusion Probability.

MLE: Maximum Likelihood Estimate.

MPM: Median Probability Model.

MSE: Mean Square Error.

PPD: Posterior Predictive Density.

RIC: Risk Inflation Criterion.

CHAPTER 1

INTRODUCTION

"*In every phenomenon the beginning remains always the most notable moment*."

- Thomas Carlyle

The linear model is a common and widely applied statistical model that has received much attention in the Bayesian literature. It is often a starting place for methods of analysis that are extended or adapted to more general classes of model. Of particular interest in this setting is the use of a random variable representing model configuration (Smith and Kohn, 1996).

Using the Bayesian paradigm a posterior distribution can be derived providing a probability mass function for the model space. In creating such a posterior distribution, high probability models can be selected, variable selection may be performed, and the probabilities can be used in model averaging for either inference or prediction.

The posterior distribution of model probabilities must be sampled from when the number of predictors is large. The posterior is typically proportional only, so Markov chain Monte Carlo methods are employed. The standard approach is to use a Gibbs sampler that samples from the conditional distribution of a binary vector representing model configuration. The state space is discrete and of size 2^k where k is the number of explanatory variables. For large k it may take a long time to explore the state space and diagnosing convergence can be difficult.

Advances in MCMC, known as perfect sampling, have yielded methods that eliminate the burn-in problem. The most famous of the perfect sampling methods is coupling from the past (CFTP) (Propp and Wilson, 1996). CFTP samples exactly from the posterior distribution and provides independently, and identically distributed (i.i.d.) sample points. CFTP requires a monotone structure in the update function to avoid having to start a Markov chain (MC) from every state. CFTP can be applied to the Bayes linear model to sample from the posterior model probabilities. However, when using a Gibbs MC the predictor matrix must be orthogonal for the update structure to be monotone. Monotonicity is a useful property for greatly reducing the computational burden of CFTP. This restriction has prevented perfect sampling from being applied routinely to linear regression, as an orthogonal predictor matrix does not allow variable selection. Thus, not much work exists on tackling the problem from a linear regression point of view.

Monotone CFTP has been of great use in signal reconstruction using orthogonal wavelets. The choices of priors and hyper-parameters for such applications are well defined, so little has been done to explore how robust the construction of a monotone Gibbs MC is to hyper-parameters and priors. However, model averaging can be used to great effect for modeling the response. Orthogonalization may have other consequences such as shrinking the model space and reducing computation time. Most work using wavelets and perfect sampling do not address the additional comparison between perfect sampling and the use of a standard Gibbs sampler on the orthogonal space. The specific aims of this research are:

1. Assuming an orthogonal predictor matrix, check the robustness in the construction of monotone Gibbs MC to choices of priors and hyper-parameters.

2. Determine the effect of using an orthogonal predictor matrix on inference using model averaging and the linear regression model.

3. From three versions of the Gibbs sampler; standard with the original predictor matrix, standard with an orthogonal predictor matrix and perfect with an orthogonal predictor

matrix, determine which is the best choice according to computational efficiency and rate of convergence to the stationary distribution.

4. Provide further exploration of the application of the perfect sampling version of the independence Metropolis-Hastings algorithm for Bayesian variable selection.

With these aims in mind the thesis is outlined as follows:

In the remainder of this Chapter we review the use of a binary vector ($\boldsymbol{\gamma}$) for model selection in the Bayes linear model. This provides a posterior distribution for model averaging and variable selection. We introduce and review the posterior mass function for $\boldsymbol{\gamma}$. We cover issues relating to the use of the posterior model probabilities namely, the difficulties faced when the number of predictors becomes large and sampling is required. We review the fundamentals of simulating random variables from a posterior distribution using MCMC. The use of MCMC is then expanded upon with a discussion of perfect sampling variants of MCMC. Under the property of uniform ergodicity, the construction of couplings with Markov chains, CFTP, and perfect forwards simulation are covered. Finally, monotonicity properties and applications of perfect sampling in Bayesian statistics are discussed.

In Chapter 2 we review the sufficient conditions for a monotone Gibbs MC. We consider common priors for the regression coefficients ($\boldsymbol{\beta}$) and error variance (σ^2). This includes the conjugate formulation, which includes Zellner's prior as a special case, and an adjusted form of Jeffreys prior. We also consider common choices for the model space prior such as the Bernoulli and truncated Poisson distributions, and general priors for the model size. We consider both a fully Bayes and special cases of an empirical Bayes (EB) approach. We also consider integration over hyper-parameters, such as c in Zellner's prior. Finally, we consider three examples during the chapter, two for Zellner's prior where one is a special informative case and the other an adjustment for outlier detection, and an example of a conjugate prior for the regression variance designed to provide a posterior with the mode equal to the classical unbiased estimate of the variance.

In Chapter 3 we indulge in a numerical demonstration of the monotonicity of the Gibbs sampler for an orthogonal design matrix. We then show the relation between the partial ordering required for monotonicity, and the nested model structure in linear regression. We then move to a discussion of orthogonalization methods and in particular, introduce the Lowdin transformation and two variants of the modified Gram-Schmidt orthogonalization procedure, yet to be explored in the literature. We briefly discuss the impact an orthogonal design matrix, and the required partial order has on posterior estimates. The second half of Chapter 3 is devoted to an exact exploration of the effect of using an orthogonal design matrix with four real datasets. We compare the four orthogonalization methods by looking at the expected model size and model competition. We then assess the use of **W** compared to **X** for in-sample prediction, using the deviance information criterion (DIC) extended to include integration over the model space. We also discuss the problems of using an orthogonal design matrix for out-of-sample prediction, and cross-validation methods for outlier detection.

In Chapter 4 we look at the efficiency of sampling with variants of the Gibbs sampler. We review the standard and monotone Gibbs CFTP algorithms. We also discuss using a standard Gibbs sampler in tandem with an initial run of the monotone Gibbs CFTP sampler, to remove the need for burn-in assessment. Using the four real datasets from the previous Chapter 3, we investigate the convergence of the three methods to the posterior distribution of γ. We also record the convergence of quantities such as the DIC, expected model size, and the model averaged fitted values. These results are summarized using the computational time to provide a comparison of convergence in cpu time. A simulation study is conducted to compare the computational time of the three methods. Returning to the real datasets, information and the backwards coupling times are investigated.

In Chapter 5 we review the particulars of the exact IMH algorithm, and then discuss and explore the relation to rejection sampling. This is followed by an investigation of the difficulties in finding an efficient bound for the marginal posterior of γ. We do however, find a way to obtain the optimal value when reducing the posterior to a function of the residual sum of squares only. Under these circumstances, we explore how efficient the rejection sampler is for various choices of hyper-parameters. The second part of this chapter moves to the posterior for the hyper-parameter c in Zellner's prior conditional on γ. We review the use of rejection sampling and a second approach that allows a refinement of the proposal distribution, reducing the expected waiting time for exact i.i.d.

sample points. In the final chapter, Chapter 6, we summarize and discuss the findings of this work and provide topics of future research.

1.1 Linear Regression and Bayesian Variable Selection

The likelihood function for an independent and identically distributed (i.i.d.) sample **D** = (x_1, ….., x_n), with unknown parameter(s) θ, for a given density function f_θ is

$$f(\mathbf{D} \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) . \tag{1.1}$$

Standard likelihood methods maximize (1.1) to obtain estimates of θ, $\hat{\theta}$ known as the maximum likelihood estimate (MLE). In Bayesian statistics, inference about θ involves the product of the likelihood function and a prior on θ, $f(\theta)$, to obtain a posterior distribution, $f(\theta \mid \mathbf{D})$, using Bayes' theorem for distributions:

$$f(\theta \mid \mathbf{D}) = \frac{f(\mathbf{D} \mid \theta) f(\theta)}{\int f(\mathbf{D} \mid \theta) f(\theta) d\theta} \propto f(\mathbf{D} \mid \theta) f(\theta) . \tag{1.2}$$

When the denominator is not available in closed form, we can specify the posterior up to a normalizing constant.

1.1.1 The Model

Let **y** be an n x 1 vector of measured responses and **X** be an n x (k + 1) matrix where the first column is a constant and the remaining k columns are the recorded predictors. We assume **y** may be modeled as a linear combination of the k + 1 columns of the predictor matrix **X**, plus a random perturbation (**e**), having a normal distribution with mean zero and constant variance σ^2, i.e.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \text{ where } \mathbf{e} \sim \mathbf{N}(0, \sigma^2 \mathbf{I}_n) . \tag{1.3}$$

$\boldsymbol{\beta}$ = (β_0,β_1,...,β_k) is a vector of regression coefficients measuring the effect of each column in **X** where β_0 corresponds to the intercept. Including a column of ones in the predictor matrix to fit an intercept is common practice.

Heavy tail distributions such as the *t* distribution or Cauchy distribution may be used for **e**, which is common in econometrics (Draper and Smith, 1998). More recently, work on using symmetric exponentials and epsilon skewed distributions has been investigated (Elsalloukh *et al*, 2005). The aim of such choices is to improve the robustness of (1.3), reducing the sensitivity of estimation to extreme values of **y**.

Typically we seek to find a subset of predictors that adequately models **y**; this is variable selection. In the Bayesian sense, we extend the standard linear regression model to treat variable selection by introducing a binary parameter vector $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, ..., \gamma_k)$ that represents the configuration of a specific model (Smith and Kohn, 1996; Kuo and Mallick, 1998), so that

$$\boldsymbol{\beta}_\gamma = \beta_i \mathbb{I}_{\{1\}}(\gamma_i), \text{ for } i = 0,...,k, \tag{1.4}$$

where $\mathbb{I}_A$ denotes the indicator function for the set A. Hence $\boldsymbol{\gamma}$ acts as a subset indicator ($\gamma_i = 0$ removes while $\gamma_i = 1$ includes the *i*-th predictor) on **X**, denoted $\mathbf{X}_\gamma$. All models are assumed to contain the intercept term so that $\gamma_0 = 1$ and $\boldsymbol{\gamma} \in \boldsymbol{\Gamma} = \{1\} \text{ x } \{0,1\}^k$. Thus the model space contains 2^k models. The linear regression model (1.3) conditional on $\boldsymbol{\gamma}$, becomes

$$\mathbf{y} = \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma + \mathbf{e}, \tag{1.5}$$

and the likelihood function for (1.5) is

$$f(\mathbf{y} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}_\gamma, \sigma^2, \mathbf{X}) \sim \mathbf{N}_n(\mathbf{X}_\gamma \boldsymbol{\beta}_\gamma, \sigma^2 \mathbf{I}_n). \tag{1.6}$$

Assigning priors to the unknown parameters $\boldsymbol{\beta}_\gamma$, σ^2, and $\boldsymbol{\gamma}$, the joint posterior is

$$f(\boldsymbol{\beta}_\gamma, \sigma^2, \boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) \propto f(\mathbf{y} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}_\gamma, \sigma^2, \mathbf{X}) f(\boldsymbol{\beta}_\gamma, \sigma^2, \boldsymbol{\gamma} \mid \mathbf{X}), \tag{1.7}$$

which can be factored as

$$f(\boldsymbol{\beta}_\gamma, \sigma^2, \boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) = f(\boldsymbol{\beta}_\gamma \mid \sigma^2, \boldsymbol{\gamma}, \mathbf{y}, \mathbf{X}) f(\sigma^2 \mid \boldsymbol{\gamma}, \mathbf{y}, \mathbf{X}) f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}), \tag{1.8}$$

in terms of the conditional posterior distributions for $\boldsymbol{\beta}_\gamma$, and σ^2, and the marginal posterior distribution for $\boldsymbol{\gamma}$. Traditionally, some form of dependence structure is assumed for the priors

$$f(\boldsymbol{\beta}_\gamma, \sigma^2, \boldsymbol{\gamma} \mid \mathbf{X}) = f(\boldsymbol{\beta}_\gamma \mid \sigma^2, \boldsymbol{\gamma}, \mathbf{X})\, f(\sigma^2 \mid \mathbf{X})\, f(\boldsymbol{\gamma}), \tag{1.9}$$

although alternate forms of dependence are possible. The most common alternative is

$$f(\boldsymbol{\beta}_\gamma, \sigma^2, \boldsymbol{\gamma} \mid \mathbf{X}) = f(\boldsymbol{\beta}_\gamma \mid \sigma^2, \boldsymbol{\gamma}, \mathbf{X})\, f(\sigma^2 \mid \boldsymbol{\gamma}, \mathbf{X})\, f(\boldsymbol{\gamma}), \tag{1.10}$$

which has been explored in the literature, for example George and McCulloch (1997).

1.1.2 Posterior Model Probabilities

The marginal posterior distribution, $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$ is

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) = f(\boldsymbol{\gamma}) \int\int f(\mathbf{y} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}_\gamma, \sigma^2, \mathbf{X}) f(\boldsymbol{\beta}_\gamma, \sigma^2 \mid \boldsymbol{\gamma}, \mathbf{X}) d\boldsymbol{\beta}_\gamma d\sigma^2, \tag{1.11}$$

and is a non-standard mass function on 2^k states. Using $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$ a common model choice is the mode of the posterior distribution, or the maximum aposteriori (MAP) estimate:

$$\hat{\boldsymbol{\gamma}}_{\mathrm{MAP}} = \underset{\boldsymbol{\gamma}}{\arg\max} \{ f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) \}. \tag{1.12}$$

This selects the model with the greatest posterior probability given the data. Maximization of $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$ is greatly simplified in the case where the predictor matrix is orthogonal (Chipman *et al* 2001). The marginal inclusion probability (MIP) of γ_i, is defined as

$$\mathrm{MIP}_i = \Pr(\gamma_i = 1) = \sum_{\{\boldsymbol{\gamma} \in \boldsymbol{\Gamma} : \gamma_i = 1\}} f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}). \tag{1.13}$$

This provides an intuitive measure of the relative importance of a predictor, and may be used to rank the predictors. Those predictors which appear in higher probability models frequently will have a high MIP and because we always include an intercept $\mathrm{MIP}_0 = 1$. The MIP can be used to define the median probability model (MPM) as

$$\text{MPM}_i = \mathbb{I}_{[0.5,1]}(\text{ MIP}_i \text{ }) \text{ for } i = 0,...,k. \tag{1.14}$$

This model has been shown to perform well for prediction and is optimal under squared error loss (Berger and Pericchi, 2001). To assess the model complexity of $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$, we can obtain the posterior distribution and expectation for the number of predictors, $q \in \{0,...,k\}$, as $q = \boldsymbol{\gamma}_{1:k}^T \boldsymbol{\gamma}_{1:k}$.

$$f(q \mid \mathbf{y}, \mathbf{X}) = \sum_{\{\boldsymbol{\gamma} \in \boldsymbol{\Gamma} : \sum_{i=1}^{k} \gamma_i = q\}} f\ (\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}), \tag{1.15}$$

and

$$\mathbb{E}[q] = \sum_{q=0}^{k} q f(q \mid \mathbf{y}, \mathbf{X}). \tag{1.16}$$

Model competition in $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$ can be visually assessed using a cumulative probability plot of sorted model probabilities: $p_{(1)},...,p_{(2^k)}$. $p_{(1)}$ is the maximum marginal posterior probability and $p_{(2^k)}$ is the minimum marginal posterior probability. The faster the cumulative sorted probability tends to 1, the indication of less model competition in the posterior. In particular for a given threshold $\alpha \in (0,1)$ we may define:

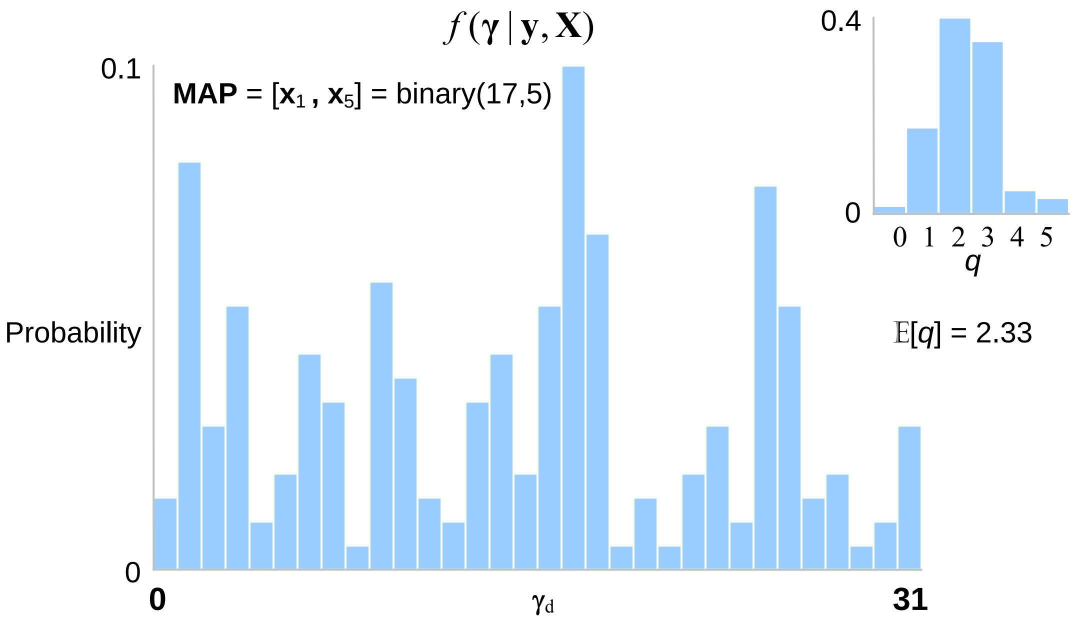
$$M_\alpha = \min\{j : \sum_{i=1}^{j} p_{(i)} \geq \alpha\}. \tag{1.17}$$

M_α represents the smallest number of highest probability models required to account for a probability of at least α in the posterior.

For plotting $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$ it is useful to use the decimal representation of $\boldsymbol{\gamma}$:

$$\boldsymbol{\gamma}_\text{d} = \text{Decimal}(\boldsymbol{\gamma}) = \sum_{i=1}^{k} 2^{k-i} \mathbb{I}_{\{1\}}(\gamma_i). \tag{1.18}$$

Binary(γ_d,k) is the reverse operation, which recovers the binary sequence from the decimal representation with k bits. Figure 1.1 provides an illustration of the quantities defined above.



	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
MIP	0.502	0.431	0.336	0.492	0.568
Rank	2	4	5	3	1

MPM = [$\mathbf{x}_1$, $\mathbf{x}_5$]

Figure 1.1 Representation of the marginal posterior for γ and related quantities for a fictitious data set. The larger bar plot is the posterior for γ and the smaller plot is the posterior distribution for model size. Also shown are the MAP and MPM models along with the expected model size and ranks based on the MIP (Table: The green highlighted squares indicated those variables included in the MPM). The sorted cumulative probability plot has been omitted, and in this example M_α = 25, for α = 0.95.

1.1.3 Model Averaging and Inference

Being able to obtain $f(\gamma \mid \mathbf{y}, \mathbf{X})$ lends itself naturally to BMA which deals systematically with uncertainty in model selection (Brown *et al*, 2002; Hoeting, 2002; Hoeting *et al*, 1999; Liang *et al*, 2001; Raftery *et al*, 1997; Wasserman 2000). The benefit of this, is to avoid over-stating the precision of inference by avoiding conditioning on a single model. Further, under squared error loss, BMA is optimal when performing out-of-sample prediction. For each model in $\mathbf{\Gamma}$, we require inference about a quantity of interest (θ) such as parameters, or predicted response, conditional on the data. Using BMA we weight the posterior distribution of $\theta|\gamma$ by the posterior probability of γ, producing an average distribution for θ across $\mathbf{\Gamma}$:

$$f(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X}) = \sum_{\gamma \in \mathbf{\Gamma}} f(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X}, \gamma) f(\gamma \mid \mathbf{y}, \mathbf{X}) \ . \tag{1.19}$$

The practical implementation of BMA can be hindered by computation of $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$ for large $\mathbf{\Gamma}$ which may not be available in closed form, the choice of prior probabilities for each model *f*(γ), and the number of models to be averaged over. Weights for BMA may be constructed using the Aikake information criterion (AIC), the Bayesian information criterion (BIC), or even Bayes factors (BF) using hyper-G priors and Zellner-Siow priors (Montgomery and Nyhan, 2008). This differs from the approach we take explicitly using the marginal posterior for $\boldsymbol{\gamma}$ rather than a model selection criterion or BF. We consider this approach more natural as weights based on selection criteria is strictly speaking "model averaging", whereas using a posterior distribution to obtain the weights is decidedly BMA. Montgomery and Nyhan (2008) also recommend that multiple priors should be investigated for BMA to assess sensitivity. The $\boldsymbol{\gamma}$ formulation is an example of discrete model expansion. This is a special case of the more general continuous model expansion where components are assigned Dirichlet priors (Draper, 1995). An example of the continuous case is mixing over different forms of random effects in random effects models (Lawson and Clark, 2002). Attempts have even been made to account for uncertainty when selecting a link function in GLM's (Czado and Raftery, 2006).

When conducting linear regression analysis, there are three main inference problems. The first is $\boldsymbol{\beta}$ or $\mathbf{X}$, which answers questions about the effect of $\boldsymbol{\beta}$ in terms of magnitude and direction, the importance of predictors in $\mathbf{X}$, and in explaining the response $\mathbf{y}$. The second is in sample prediction and capturing features of $\mathbf{y}$. The third is out of sample prediction for future responses, given that we have already observed $\mathbf{y}$ and $\mathbf{X}$. Most studies of prediction in BMA use some form of cross validation approach, the data is partitioned into a training set used for model fitting and a test set used to assess predictive performance. The problem with such an approach is that the choice of size for the training and test sets involves a bias-variance trade-off that can be difficult to optimize in practice.

In model averaging the posterior distribution of $\boldsymbol{\beta}$ is not straight forward, unlike that for σ^2. Due to the $\boldsymbol{\gamma}$ formulation, the distribution and any point estimate of any component of $\boldsymbol{\beta}$ is conditional on the inclusion probability for that predictor. Model averaging avoids the difficult interpretation of the effect of a predictor which can vary depending on $\boldsymbol{\gamma}$ due to correlation with other predictors. The model-averaged posterior for $\boldsymbol{\beta}$ is

$$f(\beta_i \mid \mathbf{y},\mathbf{X},\sigma^2) \propto \sum_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}} f(\beta_i \mid \mathbf{y},\mathbf{X},\sigma^2,\boldsymbol{\gamma}) f(\boldsymbol{\gamma} \mid \mathbf{y},\mathbf{X}) \mathbb{I}_{\{1\}}(\gamma_i)\,. \tag{1.20}$$

While this representation is relatively simple it is possible for the posterior to be more complex i.e. appear multi-modal, due to the behavior of the estimated regression coefficients in the presence of strong correlations.

The model-averaged posterior for σ^2 is more straightforward being required for all models. The posterior is estimated as

$$f(\sigma^2 \mid \mathbf{y},\mathbf{X}) = \sum_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}} f(\sigma^2 \mid \boldsymbol{\gamma},\mathbf{y},\mathbf{X}) f(\boldsymbol{\gamma} \mid \mathbf{y},\mathbf{X})\,. \tag{1.21}$$

When k is small the un-normalized probability for every $\boldsymbol{\gamma}\in\boldsymbol{\Gamma}$ can be calculated and then normalized, providing $f(\boldsymbol{\gamma} \mid \mathbf{y},\mathbf{X})$ exactly without the need for sampling using the following steps:

1. Calculate: $[\tilde{f}(\boldsymbol{\gamma}^1 \mid \mathbf{y},\mathbf{X}),...,\tilde{f}(\boldsymbol{\gamma}^d \mid \mathbf{y},\mathbf{X})]$, where $d = 2^k$ and $\tilde{f}(\,)$ is the un-normalized probability mass function.

2. Then for $i = 1,...,d$, calculate: $f(\boldsymbol{\gamma}^i \mid \mathbf{y},\mathbf{X}) = \tilde{f}(\boldsymbol{\gamma}^i \mid \mathbf{y},\mathbf{X}) / \sum_{j=1}^{d} \tilde{f}(\boldsymbol{\gamma}^j \mid \mathbf{y},\mathbf{X})$.

The Gray code (Savage, 1997) provides an efficient ordering of $\boldsymbol{\gamma}$ such that only one component is updated at a time while enumerating the model space. If k is large numerical underflow can also be problematic as the less probable models become negligible. Furthermore, it becomes time consuming and even infeasible (time-wise, memory-wise, or both) to proceed without some means of sampling from the posterior. It is necessary to be slightly vague about what constitutes a "large" *k*, as it depends on the amount of computing resources available. If the concern is not with model averaging, we may obtain the MAP estimate using some optimization method such as an annealed Gibbs sampler. Ultimately, a stochastic sampling method such as MCMC is typically required to estimate posterior model probabilities.

1.2 Markov chain Monte Carlo.

The posterior for $\boldsymbol{\gamma}$ is not a known standard parametric family mass function. Typically it is also known only up to a normalizing constant, and so MCMC methods are typically employed to generate sample points. If a MCMC approach is well implemented, inference from 30,000 sample points under model averaging may well be satisfactory even if we have 25 predictors (33,554,432 states!). Thus, at least for now and likely for some time yet, an MCMC sampling procedure is necessary. Notice that in problems with large *k*, the state space may become increasingly sparse so that the more probable models become lost in a sea of small probability models. MCMC methods are designed to find these high probability models by means of a stochastic search. Brute force calculation on the other hand has no such mechanism and so can be described as undirected.

1.2.1 Markov Chains and Simulation

Let the sequence of random variables (X_1, X_2, X_3, …), denoted $\{X_n\}$, be a stochastic process on a state space D with σ-algebra $\mathfrak{F}$ and let $X_n \in D$ and $E \in \mathfrak{F}$.

Definition 1.1: Markov Chain

The stochastic process $\{X_n\}$ *with the property:*

$$\Pr(X_{t+1} \in E \mid X_t, X_{t-1}, \ldots., X_1, X_0) = \Pr(X_{t+1} \in E \mid X_t), \tag{1.22}$$

is a Markov Chain (MC).

The new state is dependent only upon the previous, and not the entire history of the chain. This is the Markovian property. When D is discrete, the movement from time t to $t+1$ is defined by a matrix of transition probabilities $\boldsymbol{P}$. The probability of moving from state i at time t to state j at time t+1 is $p_{ij} = \Pr(x_{t+1} = j \mid x_t = i)$.For a finite state space with m states, the stationary distribution $f = (f_1,..,f_m)$ may be obtained by solving the equations:

$$f\mathbf{P} = f \text{ and } \sum_{i=1}^{m} f_i = 1. \tag{1.23}$$

When D is continuous, the transition rules are specified by a transition kernel K:

$$\Pr(X_{t+1} \in E \mid X_t) = \int_E K(x_t, dx). \tag{1.24}$$

In practice, when simulating X_n it is convenient to consider the update function ϕ which generates X_{t+1} as a function of X_t and a pseudo-random number U_{t+1}:

$$X_{t+1} = \phi(X_t, U_{t+1}) \text{ where } \phi : D \text{ x } \boldsymbol{U} \rightarrow D, \text{ and } U \in \boldsymbol{U}. \tag{1.25}$$

The update function represents the MC as a stochastic recursive sequence (SRS). In Bayesian applications, to obtain a sample from the posterior f the update function is constructed to have the limiting distribution f. The stationary distribution f is a limiting distribution with X_t converging in distribution to f. The simulation of samples by this method is known as MCMC. The necessary conditions of aperiodicity, irreducibility, reversibility, and recurrence ensure that the MC is ergodic so f is guaranteed to exist uniquely. For further details on these conditions, see Roberts and Casella (2004: Chapter 6). Under these regularity conditions the update function is a measure preserving transform. With an ergodic MC the time average and sample space average are the same. Consequently, a central limit theorem applies and so we may estimate expectations based on the sample generated:

$$\frac{1}{N}\sum_{n=1}^{N} h(X_n) \rightarrow E_f[h(X)], \tag{1.26}$$

where h is some measurable function. Traditional MCMC algorithms attempt to construct an update function for which these conditions are observed and generate states that are samples from f. The more common MCMC algorithms which are the random walk, independent proposal, and Gibbs samplers can be viewed as variants of the Metropolis Hastings (MH) algorithm (Metropolis *et al*, 1953; and Hastings, 1970).

1.2.2 Metropolis-Hastings

The Metropolis-Hastings (MH) algorithm forms the basis for most MCMC samplers and is presented in Algorithm I.

Algorithm I: Metropolis-Hastings.

```
Set: $x_1 \in D$
For $i = \{2,..., N\}$
    Propose: $y \sim q( y, x_{i-1} )$
    Generate: $u \sim U(0,1)$
    Calculate: $\alpha = \min\left\{1, \frac{f(y)q(x_{i-1}, y)}{f(x_{i-1})q(y, x_{i-1})}\right\}$
    If $u \leq \alpha$
        Set: $x_i = y$
    Else
        Set: $x_i = x_{i-1}$
```

q is a proposal density that is easy to simulate from, and generates a new candidate value conditional on the previous value in the MC. MCMC algorithms begin from an arbitrary initial state and run forward in time until it is believed that X_n has converged to *f*. The first *m* sample points are discarded as burn-in, so that

$$E_f[h(X)] \approx \frac{1}{N-m} \sum_{n=m+1}^{N} h(X_n) . \tag{1.27}$$

In practice, obtaining sample points that approximate *f* well requires knowledge of how large *m* must be, which can be difficult to determine beforehand. Finding a suitable *m* can be further complicated by the choice of starting state affecting the rate of convergence to *f* (e.g. geometrically ergodic X_n). Since the justification of MCMC is asymptotic, if it were possible to run the chain for an infinite time we would have no concerns. Practically however, there is a necessity to impose a finite burn-in that can bias results. Convergence diagnostics using the auto-correlation function are typically employed to estimate a suitable burn-in period. Inference conducted using transformations or functions of the original sample points must have the auto-correlation function calculated in each case to estimate the burn-in and variance.

The first common implementation of the MH algorithm is to assume $q(y,x) = q(y)$ so the proposal density is independent of past values of x. The MH algorithm then simplifies to the independence Metropolis-Hastings (IMH) sampler in Algorithm II.

Algorithm II: Independence Metropolis-Hastings.

```
Set: $x_1 \in D$
For $i = \{2,..., N\}$
    Propose: $y \sim q( y )$
    Generate: $u \sim U(0,1)$
    Calculate: $\alpha = \min\left\{1, \frac{f(y)q(x_{i-1})}{f(x_{i-1})q(y)}\right\}$
    If $u \le \alpha$
        Set: $x_i = y$
    Else
        Set: $x_i = x_{i-1}$
```

Provided q is not too different than f and has heavier tails the IMH algorithm will generate well approximated samples from f. A second common implementation of the MH algorithm is the Gibbs sampler. Let $f(\mathbf{X})$ be a p dimensional density, assuming we can sample easily from the univariate conditional density $f(\mathbf{X}_i \mid \mathbf{X}_{-i})$ for all dimensions where $\mathbf{X}_{-i} = (\mathbf{X}_1,\ldots,\mathbf{X}_{i-1},\mathbf{X}_{i+1},\ldots,\mathbf{X}_p)$, the Gibbs sampler is Algorithm III.

The Gibbs sampler is a very adaptable algorithm and while the updated components must remain constant, they may be updated sequentially in random order or even in blocks. The Gibbs sampler will also permit the use of further MCMC algorithms such as the IMH sampler to generate candidate values from $f(\mathbf{X}_i \mid \mathbf{X}_{-i})$. Notice that $f(\mathbf{X}_{p-1} \mid (\mathbf{X}_1)_{i},,\ldots,(\mathbf{X}_{p-2})_i,$

$(\mathbf{X}_p)_{i\text{-}1}$) is both the proposal and target density in the standard Metropolis-Hastings algorithm.

Algorithm III: Gibbs Sampler.

```
Set: $\mathbf{X}_1 \in \mathrm{D}$
For $i = \{2,...,N\}$
    Generate: $(\mathbf{X}_1)_i \sim f(\mathbf{X}_1 \mid (\mathbf{X}_2)_{i\text{-}1},\ldots,(\mathbf{X}_p)_{i\text{-}1})$
    Generate: $(\mathbf{X}_2)_i \sim f(\mathbf{X}_2 \mid (\mathbf{X}_1)_i,(\mathbf{X}_3)_{i\text{-}1},\ldots, (\mathbf{X}_p)_{i\text{-}1})$
              □
    Generate: $(\mathbf{X}_{p\text{-}1})_i \sim f(\mathbf{X}_{p\text{-}1} \mid (\mathbf{X}_1)_i,,\ldots,(\mathbf{X}_{p\text{-}2})_i, (\mathbf{X}_p)_{i\text{-}1})$
    Generate: $(\mathbf{X}_p)_i \sim f(\mathbf{X}_p \mid (\mathbf{X}_1)_i,,\ldots,(\mathbf{X}_{p\text{-}1})_i)$
```

Thus, the Gibbs sampler is a Metropolis-Hastings algorithm with an acceptance probability always equal to 1.

Another common implementation is the random walk Metropolis-Hastings which we do not cover here. From these very simple and powerful approaches a number of tricks exist to improve mixing and convergence. For a great survey of such ideas and approaches along with some computational aspects see books by Givens and Hoeting (2005, Chapter 7), Gamerman and Lopes (2006). The Gibbs sampler will be the prominent focus of later chapters along with some attention to the IMH sampler. We now provide a brief history of some MCMC approaches that have been used for variable selection in linear regression.

1.2.3 MCMC for Variable Selection.

A vast number of MCMC methods exist for variable selection problems and we briefly mention a few here. The Gibbs sampler, stochastic search variable selection (SSVS), and the Swendsen-Wang algorithm use the conditional distribution for γ to either sequentially, randomly, or in clusters, update the binary model vector (Carlin and Chib, 1995; Dellaportas *et al*, 2002; George and McCulloch, 1993; Nott and Green, 2004). Trans-dimensional or reversible jump methods (Green, 1995), and birth and death or auxiliary variable methods (Stephens, 2000), generally allow the MCMC method to traverse the dimension of γ. In particular the trans-dimensional approach includes a transition between dimensions, while the birth and death process uses a stochastic mechanism for either introducing or removing a dimension. More recently, the adaptive IMH sampler (Nott, and Kohn 2005) uses adaptation of the proposal distribution while generating sample points to improve convergence and mixing.

1.3 Exact Sampling with Markov Chains.

Exact MCMC (or perfect/exact simulation/sampling/MCMC) methods are traditional MCMC without statistical error, the sample points are generated exactly according to the stationary distribution f. Consequently these methods remove the need for a burn-in period. Exact MCMC methods came about after Propp and Wilson (1996) laid the foundation with their coupling from the past (CFTP) algorithm. The mechanism required depends upon whether a backwards or forwards simulation method is chosen. In the backwards case we require a coupling construction of Markov chains, while in the forwards case a residual kernel is needed (coupling may also be required). In either case we are bounding the rate of convergence of the MC. If the true rate of convergence is poor or the bound is poor, then perfect sampling becomes impractical.

The definition and useful properties of a uniformly ergodic MC are introduced as for the work to follow we need only consider the simpler case (especially with regard to perfect sampling) of the uniformly ergodic case rather than the more general case of a geometrically ergodic MC. We now review some basic theory related to the construction and properties of coupling Markov chains.

1.3.1 Uniform Ergodicity

Let X_n be a Markov chain on a state space D with σ-algebra $\mathfrak{F}$ so that $x \in D$ and $E \in \mathfrak{F}$.

We may define the transition probabilities for m iterations as

$$P^m(x, E) = \Pr(X_{t+m} \in E \mid X_t = x)\,. \tag{1.28}$$

This definition of the transition probabilities will be used in the definitions to follow.

Definition 1.2: Uniformly ergodic MC

$X_n \to f$ at a geometric rate independent of the initial value X_0 *if, and only if there exist H > 0 and* $r \in (0,1)$ *such that:*

$$\| P^m(x,E) - f(E) \| \le Hr^m \quad , \tag{1.29}$$

for all m, $x \in D$ and $E \in \mathfrak{F}$, X_n is said to be uniformly ergodic.

H is necessarily bounded when *D* is finite, hence a bounded discrete state space will typically provide a uniformly ergodic MC. Uniform ergodicity is a stronger condition than geometric ergodicity which we mention later in this chapter. Uniform ergodicity is equivalent to the entire state space *D* being small in the sense of the minorization condition (or Doeblin's condition).

Definition 1.3: Minorization condition

The subset $S \subseteq D$ is (m, ε, q) small when

$$P^m(x,E) \ge \varepsilon\, q(E)\,, \tag{1.30}$$

for some probability measure q, ε > 0, and positive integer m, and for all $x \in S$ and $E \in \mathfrak{F}$. If S = D then the entire state space is small which is equivalent to uniform ergodicity.

If the minorization condition holds or the MC is uniformly ergodic, we may define a residual kernel as a mixture involving the original kernel and regeneration times.

Definition 1.4: The residual kernel

Where S satisfies the minorization condition the residual kernel is defined as:

$$R^m(x,E) = (1-\varepsilon)^{-1}[P^m(x,E) - \varepsilon\, q(E)]\,, \tag{1.31}$$

for some probability measure q( ), ε > 0, positive integer m and for all $x \in S$ and $E \in \mathfrak{F}$.

Definition 1.5: Regeneration times

If X_n is uniformly ergodic there is a set of times:

$$\{T_0, T_1, ...\} \sim \text{Geometric}(\varepsilon)\,, \tag{1.32}$$

such that when X_n is in the (m,ε,q) small set S at T_k it will begin again (*regenerate*) *from q at T_{k+1}.*

Hence, for each block $\{(X_{T_0}, X_{T_1-1}), (X_{T_1}, X_{T_2-1}), ...\}$ of X_n indexed by the regeneration times (referred to as a tour) the process essentially begins again independently of, but distributed identically to the last tour, i.e. the tours are i.i.d. If $T_0 \neq 0$ then we have a delayed renewal process. The expectation of $\{T_0, T_1, ...\}$ using the geometric distribution is $1/\varepsilon$.

1.3.2 Construction and Properties of Coupled Markov Chains.

Coupling of probability measures (**P**) and random elements (X) on a singly measurable space is useful for investigating the individual properties and similarities of **P** or X. Consider the probability measures $\mathbf{P}_1$ and $\mathbf{P}_2$ on the measurable space (E, ξ), for the set E with σ-algebra ξ.

Definition 1.6: Coupling of Probability Measures

The coupling $\{\mathbf{P}_1, \mathbf{P}_2\}$ *is a probability measure* $\hat{\mathbf{P}}$ *in* (E^2, ξ^2) *where* $\mathbf{P}_1$ *and* $\mathbf{P}_2$ *are the marginal distributions of* $\hat{\mathbf{P}}$.

Now consider the random variables X^1 and X^2 defined on their respective probability spaces $(\Omega_1, \mathfrak{F}_1, \mathbf{P}_1)$ and $(\Omega_2, \mathfrak{F}_2, \mathbf{P}_2)$ both in (E, ξ).

Definition 1.7: Coupling of Markov Chains

The coupling $\{\hat{X}^1, \hat{X}^2\}$ *is a new probability space* (Ω, $\mathfrak{F}$, $\mathbf{P}$) *in* (E^2, ξ^2) *where*:

$$X^1 \overset{D}{=} \hat{X}^1 \text{ and } X^2 \overset{D}{=} \hat{X}^2, \tag{1.33}$$

and $\overset{D}{=}$ *is "tends to in distribution", such that:*

$$\mathbf{P}(\hat{X}^1, \hat{X}^2)^{-1} \text{ is a coupling of } \mathbf{P}_1(X^1)^{-1} \text{ and } \mathbf{P}_2(X^2)^{-1}, \tag{1.34}$$

as in definition 1.6.

Definitions 1.6 and 1.7 are from Lindvall (2002). Perfect simulation deals with coupling constructions using MCs (random processes). Thus, definition 1.7 for coupling random elements is more useful. For good discussions on coupling and related ideas for MCs see the books by Haggstrom (2002), Lindvall (2002), and Thorisson (2000). It should also be noted that a number of definitions and conditions for coupling have been investigated and discussed before the advent of exact sampling methods. Coupling is not new to MCMC, or limited to use in perfect sampling. Convergence proofs and diagnostics have used coupling (Johnson, 1998), and have been used to improve MCMC samplers e.g. antithetic coupling for the Gibbs sampler (Frigessi *et al*, 2000). Furthermore, coupling measures have a much longer history than that of perfect sampling, evident by the book by Lindvall being first published in 1992.

Consider the two ergodic MCs X_n^1 and X_n^2 on the state space D.

Definition 1.8: Coupling with the Update Function

X_n^1*, and* X_n^2 *are coupled if:*

$$\phi(\{X_t^1, X_t^2\}, U_{t+1}) \rightarrow \{X_{t+1}^1, X_{t+1}^2\}. \tag{1.35}$$

That is, the pair-wise coupling $\{X_n^1, X_n^2\}$ evolve jointly in time under the same update function ϕ (set of transition rules) and realized randomness U. The coupling $\{X_n^1, X_n^2\}$ must observe the following properties:

1. *X_n^1 and X_n^2 behave like X_n in their limiting distributions*:

$$X_n^1 \to f,\ X_n^2 \to f \text{, as } n \to \infty. \tag{1.36}$$

2. *Once X_n^1 and X_n^2 move to the same state they will evolve jointly and identically from then on into the future*:

$$\text{If } X_t^1 = X_t^2 \text{ then } X_{t+i}^1 = X_{t+i}^2\ \forall\, i > 0. \tag{1.37}$$

The event at time *m* where both chains merge to form a single chain is called coalescence. Coalescence occurs if, and only if, there is a non-zero probability that both chains, regardless of their initial states, merge into a single state in finite time. The time *m* at which coalescence occurs is known as the coupling time (C_t).

Definition 1.9: Coupling Time

C_t is a random finite time associated with the coalescence of a coupling:

$$C_t = \min\{t \geq 1 : \phi(\{X_{t-1}^1, X_{t-1}^2\}, U_t) \to X_t\ \}\ . \tag{1.38}$$

We can generalize from the simple pair-wise coupling $\{X_n^1, X_n^2\}$ to subset coupling and complete coupling. Subset coupling uses some collection of states from *D*, while complete coupling arises when we couple all states in *D*, when *D* is discrete and finite. If *D* is continuous we must couple all subsets, where the union of all subsets is equal to *D*. The use of complete coupling is crucial to the application of CFTP.

The efficiency, by which we mean the rate at which the coupling converges to *f*, of any coupling construct is bounded by the coupling inequality (Kendall, 2005). The coupling inequality implies no coupling can be more efficient than the rate of convergence (or mixing) of the underlying MC. A "good" coupling construction may be characterized as "practical to implement" or "close to maximal". The second implies that one coupling construction may be more efficient than another. Maximal coupling refers to the situation where the coupling construction attains equality in the coupling inequality, and is often referred to as the Vasershtein coupling. Maximal couplings are available for any weakly ergodic MC (Connor and Kendall, 2007; Lindvall, 2002; Thorisson, 2000), but are not always practical to construct or compute explicitly, rarely being co-adapted. Co-adapted coupling is when the progress of X_n^1 is not dependent upon the future of X_n^2 and vice versa. This implies that the MCs X_n^1 and X_n^2 when viewed separately, are still Markovian but the joint process $\{X_n^1, X_n^2\}$ is not. Co-adapted coupling is the common coupling construction used in exact MCMC (Burdzy and Kendall, 2000). The computational cost of the coupling construction should also be minimal.

1.3.3 Perfect Backwards Simulation.

Propp and Wilson (1996) described how using repeated recursions into the past allow generation of exact samples from *f*. The idea is that if X_n were run from $-\infty$ (i.e. infinitely far in the past) to time 0, we will have surely converged to *f*. The nature of a MC means that the further we go into the past the smaller the influence of the initial state(s) on the state at time 0. This means we need only go back "far enough" into the past to ensure complete coupling has occurred by the time the chains reach time 0.

Theorem 1.1: (Propp and Wilson, 1996)

Any uniformly ergodic X_n *with a sufficiently large recursion into the past, such that complete coupling occurs before time 0, will produce an exact draw from f at time* 0.

Remark

The convergence rate of the MC must be non-negligible and the bound on the rate of convergence induced by the coupling construction must not be poor, or conversely, should be maximal or as close to maximal as possible.

The form of this theorem does ignore the practical consequences of using coupling and the Markov property to construct an exact sampling algorithm. Foss and Tweedie (1998) demonstrate the existence of a CFTP algorithm is equivalent to uniform ergodicity. They do this by showing that successful coalescence in CFTP occurs if, and only if, X_n is uniformly ergodic. This is because when X_n is uniformly ergodic the state space is small, so there exists a probability of coalescence at each step. The number of steps we must move into the past to ensure coalescence is called the backwards coupling time (B_t). For any given realization of random numbers U for updating the coupled chains, there is a set of possible backwards coupling times, where the smallest one is chosen for the sake of efficiency. Let D be discrete with m states.

Definition 1.10: Backwards Coupling Time

B_t is a random time that occurs post coalescence of the complete coupling, and is defined as:

$$B_t = \min\{\, t \geq 1 : \phi(\{X_{-t}^{1},\ldots,X_{-t}^{m}\},U_{-t+1},U_{-t+2},\ldots,U_{-1},U_0) \to X_0 \,\}\,. \tag{1.39}$$

We update all states simultaneously to detect complete coupling, so for some time in the past the MC started in all states have coalesced, and by time zero occupy a single state. The following properties apply to B_t:

1. *The backwards coupling time is always greater than the coupling time and has the same distribution.*

2. *Any time greater than the backwards coupling time is also a backwards coupling time for a given realization of U.*

Propp and Wilson (1996) noted that the random variables C_t and B_t have the same distribution and are regeneration times.

B_t is dependent only on U, and the common state at time 0 is independent of any starting state. Hence, the draw at time 0 is guaranteed to be an exact sample from f. Standard CFTP is applicable for any uniformly ergodic X_{n} provided we can detect coalescence for the complete coupling. In the case of a continuous state space CFTP can be described in terms of sets. The discrete states are replaced by sets that are non-over-lapping, and whose union is the entire state space. Provided the constructed MC moving between sets is uniformly ergodic, the usual CFTP construction will apply. Assume that D is discrete and finite with states $\{d_1,d_2,\ldots,d_m\}$ and note the random numbers u_t must be reused. CFTP may be implemented using the following pseudo code in Algorithm IV.

Algorithm IV: Coupling from the past.

```
Set: coalescence = false.
Set: $t = 0$.
While coalescence = false
    Set: $t = t - 1$.
    Generate: $u_{t+1}$
    For $i = \{1,\dots, m\}$
        Set: $x_t^i = d_i$
        For $j = \{t, t+1,\dots,-1\}$
            Set: $x_{j+1}^i = \phi(x_j^i, u_{j+1})$
    If $x_0^1 = x_0^2 = \cdots = x_0^m$
        Set: coalescence = true
    Else
        Set: coalescence = false
```

The most efficient sequence of backwards recursions, as indicated by Propp and Wilson (1996), is a double till overshoot scheme that doubles the previous number of recursive steps. In general, the chain does not reach stationarity by C_t, hence taking the state at time C_t does not produce exact draws from *f.* This can be seen in Figure 1.2 for a simple random walk on $D = \{1,2,\dots,20\}$. In particular, the update function is move up with probability 0.5 and move down with probability 0.5, with reflections at the boundary.

The construction implies that the chains can only coalesce at the boundaries of *D,* so if we sample the state at C_t, we will only get the values 1 and 20. This is clearly incorrect as the true *f* of this simple random walk is the discrete uniform distribution on *D*. The last practical issue for CFTP is that an impatient user may induce bias by terminating runs that take a long time to coalesce. However methods such as Fill's algorithm (Fill, 1998) can avoid this problem.

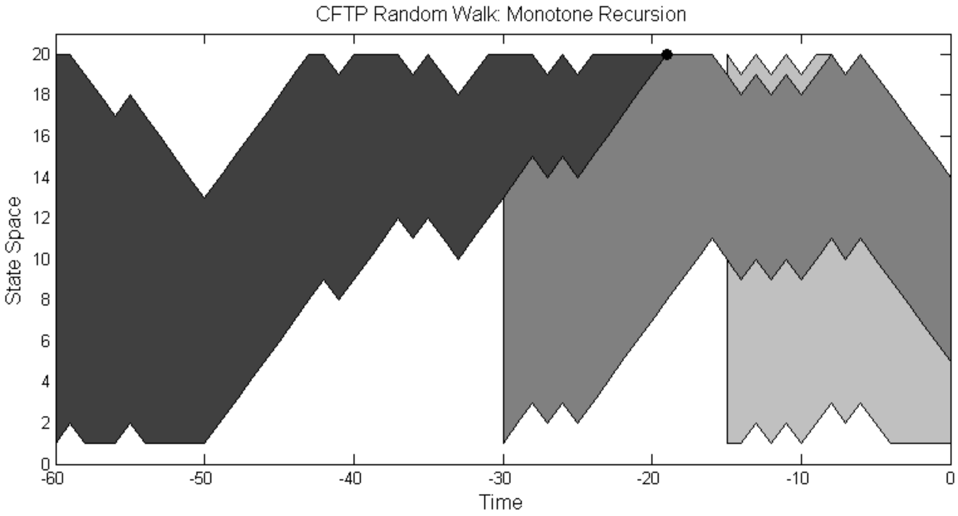


Figure 1.2. CFTP for a simple random walk on D = {1,...,20}. After 3 successive recursions (-15, -30, -60) , all chains have coalesced at the black dot (state 20) at *t* = -19.

1.3.4 Perfect Forwards Simulation

There is another class of exact simulation methods that utilize the fact that any uniformly ergodic MC satisfies the minorization condition, allowing the use of regeneration times to draw exact samples. Brooks *et al* (2006) show that any CFTP algorithm can be converted into a forwards algorithm due to the very fact it is uniformly ergodic. They build on work done by Hobert and Robert (2004) for estimating the minorization parameter ε, and demonstrate this conversion by taking advantage of regenerations. Simulated tempering involves constructing a MC that transitions between various levels of a heated (flatter) target distribution to aid mixing. It was Møller and Nicholls (1999) who initially observed that sampling from the "hottest" distribution introduced regenerations into the MC. This was used within CFTP to produce exact samples. We now recount a brief version of the theorem from Brooks *et al* (2006).

Theorem 1.2: (Brooks *et al*, 2006)

If we have a uniformly ergodic MC X_n (or equivalently the state space D is (m, ε, q) small) with initial probability distribution $X_0 \sim q$, some $\varepsilon > 0$, and residual kernel R^m, then starting X_n in q, running for a random number of iterations t ~ Geometric(*ε*) *independently of X_n, and updating according to the residual kernel will produce an exact draw from f.*

Remark

This requires that *q* and the residual kernel R^m are easy to sample from, and that *ε* must be non-negligible as is the case for any perfect sampling method. Note that generally R^m is easy to sample from when $m = 1$.

This theorem encompasses methods such as the multi-gamma sampler (Murdoch and Green, 1998), read-once CFTP (Wilson, 2000), and the catalytic coupler (Breyer and Roberts, 2000). Fundamental to the theorem is the mixture representation of *f* and the residual kernel. The theorem represented in algorithmic form is the well known splitting construction of Nummelin (1984), and Athreya and Ney (1978). This leads to an algorithm for obtaining exact draws from *f* (Algorithm V), note that here $m = 1$.

Algorithm V: Perfect Forward Sampling:

```
Independently simulate $x_0 \sim q$ and $t \sim$ Geometric($\varepsilon$).
If $t = 1$
    Set: $x = x_0$
Else
    For $i = \{1,...,t-1\}$
        Generate: $x_{i+1} \sim R(x_i, \cdot\ )$
    Set: $x = x_t$
```

Most forward perfect sampling algorithms may be viewed as intricate elaborations of this rather simple algorithm. Notice this algorithm requires no coupling mechanism to generate exact samples. The practicality of this approach revolves around the estimate of ε, and the ability to generate samples according to the residual kernel.

The read-once CFTP also follows from the above representation and can be extended without too much difficulty to unbounded state spaces. The perfect forward simulated tempering algorithm of Brooks *et al* (2006) uses all the elements of Theorem 1.2 along with a dominating chain on the random walk between temperature levels.

1.3.5 Monotonicity and Anti-Monotonicity

It is evident that for extremely large *D* detecting complete coupling is computationally intensive. Bounding conditions are particularly useful as they can be used detect complete coupling using a simple pair-wise coupling. For a more general discussion of bounding chains see Huber (2004). We can simplify the CFTP algorithm by using two extreme chains that bound all others:

$$X^U \succeq X \succeq X^L . \tag{1.40}$$

This requires the update function to have a monotone structure, and we must be able to identify the starting values from which to begin the upper and lower chains. Such chains may also be of use in forwards simulation. Dominating chains are a different form of bounding condition that are required for perfect sampling with geometrically ergodic MCs (Kendall, 2004), and as such we do not go into any detail in this thesis. The definitions of monotonicity and a related property, anti-monotonicity are as follows:

Definition 1.11: Monotone Update Functions

The update function ϕ is monotone if for some partial ordering $\preceq$ of D:

$$X_t^1 \preceq X_t^2 \Rightarrow \phi(X_t^1, U_{t+1}) \preceq \phi(X_t^2, U_{t+1}) . \tag{1.41}$$

This means the chains run from the states x^U and x^L, act as upper and lower bounding chains that sandwich the chains started in the states between x^U and x^L. Quite often x^U and x^L are the minimal and maximal states of *f*. Anti-monotonicity implies the reverse condition of monotonicity.

Definition 1.12: Anti-Monotone Update Functions

The update function ϕ is anti-monotone if for some partial ordering $\preceq$ of D:

$$X_t^1 \preceq X_t^2 \Rightarrow \phi(X_t^1, U_{t+1}) \succeq \phi(X_t^2, U_{t+1}) \tag{1.42}$$

Each bounding chain is updated based on the current state of the other bounding chain, i.e. the upper chain will use the current state of the lower chain to update, and vice versa. Each anti-monotone chain is by itself not Markovian, but the joint process and the single chain after coalescence are. A useful introduction to anti-monotone systems can be found in Haggstrom and Nelander (1998). The monotone CFTP algorithm for both monotone and anti-monotone update functions is given by the pseudo code in Algorithm VI. Assume that the upper chain is run from the maximum state of *D* (d_m), and the lower chain is run from the minimum state of *D* (d_1).

Algorithm VI: (Anti)-Monotone CFTP.

```
Set: coalescence = false.
Set: $t = 0$.
While coalescence = false
    Set: $t = t -1$
    Set: $x_t^U = d_m$, and $x_t^L = d_1$
    Generate: $u_{t+1}$
    For $j = \{t, t+1,\ldots,-1\}$
        Set $x_{t+1}^U = \phi(x_t^U, u_{t+1})$ [For anti-monotone: $x_{t+1}^U = \phi(x_t^L, u_{t+1})$]
        Set $x_{t+1}^L = \phi(x_t^L, u_{t+1})$ [For anti-monotone: $x_{t+1}^L = \phi(x_t^U, u_{t+1})$]
    If $x_0^U = x_0^L$
        Set: coalescence = true
    Else
        Set: coalescence = false
```

In later chapters, we will discuss the monotone CFTP versions of the Gibbs sampler and the IMH algorithm for BVS. We cover the monotone Gibbs MC in Chapter 2, and the exact sampling IMH algorithm in Chapter 5. The IMH and slice sampler algorithms both produce monotone MCs, (Schneider and Corcoran, 2004; Mira *et al*, 2001). The monotonicity of the Gibbs MC will depend primarily upon the structure and order of the conditional system for which it is specified. Figure 1.2 has been deliberately chosen to show the monotonicity of the update using a simple random walk (RW). The edges represent the chains started from states 1 and 20, at the boundaries of the state space. We need only monitor these two chains to detect complete coupling and to use monotone CFTP.

Other perfect sampling algorithms related to monotone methods partition the state space so the update for each partition is monotone, such as multi-gamma coupler (Murdoch and Green, 1998). Auxiliary variables may also be used to induce monotonicity in the update function with the simplest example of this the slice sampler (Mira *et al*, 2001). Recent work (Cai, 2005) has shown how to use non-monotone CFTP with a summary state. This approach constructs a general non-monotone version of CFTP for application to area-interaction point processes and birth-death processes. In particular, a single chain is constructed that can be used to monitor the certain (sets of the state space that have coalesced), and uncertain (sets of the state space yet to coalesce) parts of the chain. An exact draw under CFTP is assured when the uncertain part of X_n has vanished. This requires defining the entire state space as a union of the certain and uncertain parts. This approach requires a high degree of problem specific tailoring, so for more explicit details we refer the reader to the article (Cai, 2005). Monotone CFTP is similar in that the coalescence prior to time 0 means that by time 0, the uncertain part of *D* has vanished. Murdoch (2000) also shows it is possible to use mixtures of chains to improve the amenability of perfect sampling. Provided certain conditions are met the IMH sampler is uniformly ergodic. In particular, it is possible to use hybrid chains for perfect sampling on an unbounded states space. As an example, Murdoch uses a combined RW IMH chain to sample exactly from a Cauchy distribution.

For good introductions to CFTP, perfect sampling, and monotone and anti-monotone chains see any of the following: Dimakos (2001), Kendall (2005), Propp and Wilson

(1999), Thonnes (2000), Givens and Hoeting (2005: Chapter 8), Roberts and Casella (2004: Chapter 13), and (Lee, 2008).

1.4 Perfect Sampling and Bayesian Statistics

After discussing the basic principles involved in constructing perfect sampling methods we now review the literature for applications in Bayesian statistics. From the literature, it is clear that perfect sampling has abundant use in some areas, such as statistical physics, stochastic geometry, and spatial statistics.

In the case of spatial statistics great advancements have been made, and is an area where perfect sampling is now the sampling method of choice. Applications centre around Poisson processes, birth and death processes, queuing models, and area-interaction processes with both positive and negative attractions (Cai and Kendall, 2002; Ferrari *et al*, 2002; Kendall, 1997, 1998; Kendall and Møller, 2000; Haggstrom *et al*, 1999; Thonnes, 1999, 2000; Tweedie and Corcoran, 2001). This general applicability stems from a "cross-over" trick (Kendall and Møller, 2000). For example, using two bounding chains for a birth, a point may only be created in the lower chain if it passes the test in the upper chain, and vice versa. This produces chains that bound the state space and so CFTP can be used. However, these applications generally speaking do not involve Bayesian statistics. There are also abundant examples of applications of perfect sampling in spatial statistics and stochastic geometry. In particular, the Potts model, q-coloring graphs, lattices, the Ising model, dimer models, ice-dimer models, and Markov random fields, are among the common applications (Huber, 1998; Kendall and Thonnes, 1999).

Attempts have been made to extend perfect sampling to simple conditionally specified models using distributions such as the auto-gamma, auto-Poisson, and auto-negative-binomial distributions. The most common application is the auto-gamma model that is used to describe the pump data set (Murdoch and Green, 1998; Møller, 1999; Breyer and Roberts, 2000). The pump data set records the counts (**s**) of failures for 10 pump systems in a nuclear plant, along with the operation time (**t**) for each system. The auto-gamma model, priors and posteriors are:

$$\text{Model: } \mathbf{s}_k \sim \textbf{Poisson } (\lambda_k \mathbf{t}_k)\,.$$

$$\text{Priors: } f(\lambda_k) \sim \mathbf{Ga}(a,b)\text{, and } b \sim \mathbf{Ga}(r,v)\text{, with } a\text{, } r \text{ and } v \text{ fixed.} \tag{1.43}$$

$$\text{Posteriors: } f(\lambda_k \mid b, \mathbf{s}_k) \sim \mathbf{Ga}(a+\mathbf{s}_k, b+\mathbf{t}_k)\text{, and } f(b \mid \boldsymbol{\lambda}, \mathbf{S}) \sim \mathbf{Ga}(r+10a, v+\sum \lambda_k)\,,$$

where $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_k)$ and $\mathbf{S} = (\mathbf{s}_1,\ldots,\mathbf{s}_k)$. Murdoch and Green (1998) use the multi-gamma coupler with Gibbs sampling to sample from the posteriors by obtaining bounds through a restriction of the priors. Møller (1999) again uses the Gibbs sampler, but relies on methods from point processes to introduce bounding chains. Breyer and Roberts (2000) also with a Gibbs sampler, apply the catalytic coupler to sample from the posterior distributions for b and $\boldsymbol{\lambda}$.

Another area of interest has been finite mixture models. A k-component mixture model for an observation ($\mathbf{x}$) is:

$$\text{Model: } \mathbf{x} \sim \sum_{i=1}^{k} p_i f_i(\mathbf{x} \mid \boldsymbol{\theta}_i)\,.$$

$$\text{Priors: } p \sim \text{Dirichlet}(\alpha_1,\ldots,\alpha_k) \text{ where } \alpha_1,\ldots,\alpha_k > 0 \text{ and } \boldsymbol{\theta} \sim f(\boldsymbol{\theta}_h) \text{ where } \boldsymbol{\theta}_h \text{ are the hyper-parameters for the prior if required.} \tag{1.44}$$

Set version CFTP has been applied to sample from the posterior for simple 2- and 3-component mixtures where only the mixing proportions are unknown and all other parameters are assumed to be known (Hobert *et al*, 1999). The applicability of perfect slice sampling and the catalytic coupler were investigated assuming only the number of components in the mixture was known (Casella *et al*, 2002).

Perfect backwards simulated tempering, which is a monotone CFTP-like approach with a dominating process, has been used to generate exact samples from the posterior distributions of parameters for a generalized linear model (GLM) describing flour beetle mortality (Møller and Nicholls, 1999). The parameters (μ, σ, m), are assigned a normal, an inverse gamma, and a gamma distribution as priors. The choice of hyper-parameters in these priors is relatively weak. A second application to modeling radio carbon dating data has also been investigated, where the data (date) are modeled as piece-wise Gaussian, and the unknown mean parameter (for each date) represents the unknown true date. In this case there are 14 observations and so we must estimate the means $\theta_1,\dots,\theta_{14}$. These values all fall within an interval $[L,U]$ which is itself contained within a larger (time) interval $[A,B]$, and L and U are considered unknown. Letting $\boldsymbol{\theta} = (\theta_1,\dots,\theta_{14})$ and $x = (L, U, \boldsymbol{\theta})$ the state space of the joint posterior for L, U and $\boldsymbol{\theta}$ is $\{x : A \le L < U \le B, \boldsymbol{\theta} \in [L,U]^{14}\}$. The perfect sampling algorithm is constructed to sample exactly from the joint posterior for L, U and $\boldsymbol{\theta}$.

Two separate applications of exact MCMC to multinomial type data have also been conducted. The first, from Green and Murdoch (1999), uses a perfect RWMH algorithm to sample from the posterior for multinomial data using dirichlet priors. In particular, they sample the allele frequencies of the ABO blood group (p, q and r) defined on the unit cube. The second approach uses the perfect forwards simulated tempering method from Brooks *et al* (2006) for band analysis data. A group of Mallards have bands attached at an early age, and the bands are retrieved upon the demise of the Mallard within a certain time period. The likelihood is a product multinomial function with parameters (a, b, λ). The parameters a and b represent probabilities where $b = (1\text{-}a)$. In this case, an estimate of the likelihood is used to bound the convergence rate of the MC. The bound is then used to generate geometric times for use in a forward perfect sampling algorithm using a simulated tempering scheme. The appropriate number of tempered distributions is obtained with the hottest distribution constructed to ensure the bounding RW will coalesce in finite time. From this, perfect samples are obtained from the posterior distributions of the parameters.

An application of the perfect forwards simulated tempering method from Brooks *et al* (2006) to autoregressive time series is also described but not implemented. We also note that in Carvalho and Corcoran (2005), perfect simulation has been used to find the stationary distribution of autoregressive conditional heteroscedastic (ARCH) models, but not for a Bayesian analysis. Murdoch and Meng (2001) apply an auxiliary variable augmented version of CFTP and the read-once CFTP algorithm to sampling from mixtures of normal and t distributions, which are used as priors for Bayesian analysis.

The Bayesian linear model, the primary interest in this work, has had spectacular success when a monotone Gibbs MC is available. The main indication from the literature is that an orthogonal predictor matrix is the required condition for a monotone Gibbs MC. In

Chapter 2, we will attempt to provide further insights into this. Most applications involving curve and surface fitting use orthogonal decompositions of the original measurements via wavelets or radial basis functions (Ambler and Silverman, 2004; Holmes and Denison, 2002; Clyde *et al*, 1996; Holmes and Mallick, 1998, 2003; Lee *et al*, 2005). In all of these examples, standard monotone CFTP using Gibbs MCs is used to simulate from the posterior for model probabilities to facilitate variable selection and model averaging. Two attempts so far have been made to move beyond the orthogonal restriction. The first uses a bounded form of monotone CFTP with IMH (Schneider and Corcoran, 2004) on the joint space for β and γ. The second, known as the Gibbs coupler by Huang and Djuric (2002), is a support set coupling technique that requires known variance and regression coefficients. The underlying Gibbs MC is not monotone, but bounds on the support of each component can be derived. These bounds are then applied in the usual monotone CFTP fashion. In using the original predictor matrix and a Gibbs Markov chain, the two most relevant methods from the literature are the Swendsen-Wang algorithm (Huber 2003, Nott and Green, 2004), and the Catalytic coupler (Breyer and Roberts, 2000). The Swendsen-Wang case for BVS (Nott and Green, 2004) uses a method of introducing an auxiliary variable to treat the correlation structure as a spatial field with interactions along edges. Despite this relation to the Ising model and the use of block updates in the Gibbs sampler, the update structure of the Gibbs MC is not monotone, however, there is a bounding chain available for the Swendsen-Wang algorithm (Huber, 2003). This bounding chain is not compatible with the BVS implementation of Nott and Green (2004), and so does not help facilitate perfect sampling for posterior model probabilities. Note that for this setting if the predictor matrix was in fact orthogonal, then no correlation exists between the predictors so that the spatial representation is unnecessary.

The catalytic coupler (Breyer and Roberts, 2000) uses a rather more complex construction to check for coalescence using random maps and a basin of attraction. This requires introducing some distribution $b(\gamma)$, such that the ratio $b(\gamma)/f(\gamma)$ is bounded. If this condition is satisfied it is possible to construct the random map update according to the required constraints, and then the basin of attraction can be used to check for coalescence. In our case because we have no standard form for the posterior mass function of model probabilities, finding a bound for $b(\gamma)/f(\gamma)$ is practically impossible without extensive investigation of the posterior! This highlights the issues related to finding useful bounds for practical perfect sampling. The use of exact IMH for the marginal posterior model probabilities suffers from the need to find an efficient enough bound for detecting coalescence provided it exists.

1.5 Summary

The γ-formulation in Bayesian linear regression allows the Bayesian statistical framework to be extended to model selection. Further, the ability to create a posterior distribution of model probabilities provides a true Bayesian approach to incorporate model selection uncertainty into statistical analysis via model averaging. The posterior for γ when available in closed form, requires a stochastic sampling method to generate sample points when the number of predictors is large. In Bayesian statistics the common approach to do this is to use MCMC that generates approximate dependent samples from the desired distribution using the Metropolis-Hastings algorithm. MCMC requires diagnostics to determine an initial run of sample points to be discarded as burn-in. The remaining sample points are used for inference under the assumption the sample points after burn-in are in equilibrium. Exact or perfect sampling removes the need for burn-in assessment by generating i.i.d. sample points exactly according to the required distribution.

The use of exact sampling comes at the cost of requiring practical bounds on convergence, coupling mechanisms and increased computer resources (run time and memory). Exact MCMC has found great application for problems in statistical physics, stochastic geometry, and spatial statistics. Attempts have been made to generalize methods and improve their use however, for the most part exact MCMC methods remain very specific to the inference problem they are applied.

For BVS, more exotic forms of exact sampling are not possible due to the lack of available bounds. The monotone Gibbs sampler has been the best approach in terms of implementation and speed. However, the predictor matrix must be orthogonal which means variable selection is no longer possible for an existing set of non-orthogonal predictors. The use of exact IMH has also been investigated and has had little success due to similar problems with finding a bound much in the same way the more complex methods are not feasible.

CHAPTER 2

MONOTONICITY

"*The practicing Bayesian is well advised to become friends with as many numerical analysts as possible*."

- Prof. J. Berger, 1985

In this chapter sufficient conditions are explored for the construction of a monotone Gibbs MC to sample from $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$. We narrow our investigation by considering Gaussian errors and prior distributions common to the literature that ensure closed form expressions for $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$.

2.1 Monotonicity and Gibbs

To establish sufficiency for a monotone Gibbs MC for sampling from $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$ $\mathbf{X}$ must be orthogonal or orthonormal. Further, we require the marginal posterior for $\boldsymbol{\gamma}$ be available in closed form up to a normalizing constant. Assuming the predictor matrix is orthogonal, there are a number of choices within the context of the linear regression model for the error distribution, priors, and hyper-parameters therein. Under these constraints we will consider some general choices in accordance with the literature.

We have already noted requiring $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$ in closed form limits the choice of the error distribution which we assume is Gaussian. The most general form of priors assuming Gaussian errors are the conjugate priors for $\boldsymbol{\beta}_{\gamma}$ and σ^2. Beyond the conjugate class of priors default choices such as Jeffreys prior are also available. Within these cases we must also consider the choice of hyper-parameters. Hyper-parameters are generally independent of γ however, some choices may depend upon γ given a sensible justification for doing so. It is also more likely that hyper-parameters will depend upon γ when using an EB approach compared to a fully Bayesian one.

Recall the likelihood function

$$f(\mathbf{y} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}_\gamma, \sigma^2, \mathbf{X}) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma)^T (\mathbf{y} - \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma) \right\}, \qquad (2.1)$$

and let $f(\boldsymbol{\beta}_\gamma, \sigma^2 \mid \boldsymbol{\gamma}, \mathbf{X})$ be the joint prior for $\boldsymbol{\beta}_\gamma$, σ^2 and $f(\boldsymbol{\gamma})$ the prior for γ. The marginal posterior distribution of γ is

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) = f(\boldsymbol{\gamma}) \int\int f(\mathbf{y} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}_\gamma, \sigma^2, \mathbf{X}) f(\boldsymbol{\beta}_\gamma, \sigma^2 \mid \boldsymbol{\gamma}, \mathbf{X}) d\boldsymbol{\beta}_\gamma d\sigma^2 \,. \qquad (2.2)$$

The posterior for γ is proportional to the conditional density for γ, i.e.

$$\Pr(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X}) \propto f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}), \qquad (2.3)$$

and the conditional density used in the Gibbs sampler is

$$\begin{aligned} \Pr(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X}) &= \frac{\widetilde{\Pr}(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})}{\widetilde{\Pr}(\gamma_i = 0 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X}) + \widetilde{\Pr}(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})} \\ &= \left[1 + \frac{\widetilde{\Pr}(\gamma_i = 0 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})}{\widetilde{\Pr}(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})} \right]^{-1} \end{aligned}, \qquad (2.4)$$

where $\boldsymbol{\gamma}_{-i} = (\gamma_1, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots \gamma_k)$ and $\widetilde{\Pr}(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})$ is the un-normalized probability. We now recount a theorem from Dimakos (2001) for the component-wise partial order: $\boldsymbol{\gamma}^{(1)} \preceq \boldsymbol{\gamma}^{(2)}$ if, and only if, $\gamma_i^{(1)} \leq \gamma_i^{(2)}$ for all *i*.

Theorem 2.1: Monotone Gibbs MC for Variable Selection

The Gibbs MC is monotone for the component-wise partial order if $\Pr(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})$ *is increasing in* $\boldsymbol{\gamma}_{-i}$.

Thus, considering (2.4) we require:

$$\Pr(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}^{(1)}, \mathbf{y}, \mathbf{X}) \leq \Pr(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}^{(2)}, \mathbf{y}, \mathbf{X}), \tag{2.5}$$

or equivalently using (2.4) and (2.3):

$$\frac{f(\boldsymbol{\gamma}^{(1)} : \boldsymbol{\gamma}_{i=0}^{(1)} = 1 \mid \mathbf{y}, \mathbf{X})}{f(\boldsymbol{\gamma}^{(1)} : \boldsymbol{\gamma}_{i=1}^{(1)} = 1 \mid \mathbf{y}, \mathbf{X})} \geq \frac{f(\boldsymbol{\gamma}^{(2)} : \boldsymbol{\gamma}_{i=0}^{(2)} = 1 \mid \mathbf{y}, \mathbf{X})}{f(\boldsymbol{\gamma}^{(2)} : \boldsymbol{\gamma}_{i=1}^{(2)} = 1 \mid \mathbf{y}, \mathbf{X})}. \tag{2.6}$$

It is this inequality we will use to check the sufficient conditions for a monotone Gibbs MC for the priors and hyper-parameters explored in this chapter.

2.2 Uniform Prior for $\boldsymbol{\gamma}$

We now explore the fully conjugate, Zellner's and Jeffreys priors assuming a uniform prior for $\boldsymbol{\gamma}$.

2.2.1 Conjugate Priors

The normal likelihood suggests a Gaussian form for $\boldsymbol{\beta}|\sigma^2$ and an inverse gamma form for σ^2. These are the conjugate choice of priors for the linear regression model. Conjugate priors are a common choice in Bayesian statistics. For exponential family distributions conjugate priors are particularly useful, allowing straight forward integration to obtain marginal posterior distributions. Let $f(\boldsymbol{\gamma}) \propto 1$ and the joint prior for $\boldsymbol{\beta}$ and σ^2 conditional on $\boldsymbol{\gamma}$ with hyper-parameters $\widetilde{\boldsymbol{\beta}}_\gamma, \mathbf{V}_\gamma$, a, b, be

$$f(\boldsymbol{\beta}_\gamma, \sigma^2 \mid \boldsymbol{\gamma}) = \mathbf{N}_{q_\gamma+1}(\widetilde{\boldsymbol{\beta}}_\gamma, \sigma^2 \mathbf{V}_\gamma^{-1})\ \mathbf{IG}(a,b). \tag{2.7}$$

$\mathbf{N}_p$ denotes the multivariate normal distribution of dimension p, **IG** is the inverse gamma distribution and $q_\gamma = \boldsymbol{\gamma}^T\boldsymbol{\gamma} - 1$. The marginal posterior for $\boldsymbol{\gamma}$ is

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) \propto |\mathbf{V}_\gamma^{-1}|^{-0.5} |(\mathbf{V}_\gamma^*)^{-1}|^{0.5}\ (2b + \mathbf{y}^T\mathbf{y} - (\boldsymbol{\beta}_\gamma^*)^T (\mathbf{V}_\gamma^*)^{-1} (\boldsymbol{\beta}_\gamma^*) + \widetilde{\boldsymbol{\beta}}_\gamma^T \mathbf{V}_\gamma \widetilde{\boldsymbol{\beta}}_\gamma)^{-\frac{n}{2}-a} \tag{2.8}$$

where $\mathbf{V}_\gamma^* = (\mathbf{V}_\gamma + \mathbf{X}_\gamma^T \mathbf{X}_\gamma)$, and $\boldsymbol{\beta}_\gamma^* = (\mathbf{X}_\gamma^T \mathbf{y} + \mathbf{V}_\gamma \widetilde{\boldsymbol{\beta}}_\gamma)$. The joint posterior for $\boldsymbol{\beta}$ and σ^2 conditional on $\boldsymbol{\gamma}$ is

$$f(\boldsymbol{\beta}_\gamma, \sigma^2 \mid \boldsymbol{\gamma}, \mathbf{y}, \mathbf{X}) = \mathbf{N}_{q_\gamma+1}[(\mathbf{V}_\gamma^*)^{-1}(\boldsymbol{\beta}_\gamma^*), \sigma^2(\mathbf{V}_\gamma^*)^{-1}] \\ \mathbf{IG}\left[\frac{n}{2}+a, b+\frac{\mathbf{y}^T\mathbf{y}-(\boldsymbol{\beta}_\gamma^*)^T(\mathbf{V}_\gamma^*)^{-1}(\boldsymbol{\beta}_\gamma^*)+\widetilde{\boldsymbol{\beta}}_\gamma^T\mathbf{V}_\gamma\widetilde{\boldsymbol{\beta}}_\gamma}{2}\right], \tag{2.9}$$

with posterior expectations

$$\mathbb{E}[\boldsymbol{\beta}_\gamma \mid \boldsymbol{\gamma}, \sigma^2, \mathbf{y}, \mathbf{X}] = (\mathbf{V}_\gamma^*)^{-1}(\boldsymbol{\beta}_\gamma^*), \tag{2.10}$$

and

$$\mathbb{E}[\sigma^2 \mid \boldsymbol{\gamma}, \mathbf{y}, \mathbf{X}] = \frac{2b+\mathbf{y}^T\mathbf{y}-(\boldsymbol{\beta}_\gamma^*)^T(\mathbf{V}_\gamma^*)^{-1}(\boldsymbol{\beta}_\gamma^*)+\widetilde{\boldsymbol{\beta}}_\gamma^T\mathbf{V}_\gamma\widetilde{\boldsymbol{\beta}}_\gamma}{n+2a-2}. \tag{2.11}$$

Using (2.4) and (2.8) the update probability for the Gibbs sampler is

$$\frac{\widetilde{\Pr}(\gamma_i = 0 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})}{\widetilde{\Pr}(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})} =$$

$$\frac{|(\mathbf{V}_{\gamma_i=0}^*)^{-1}|^{\frac{1}{2}}|\mathbf{V}_{\gamma_i=0}^{-1}|^{-\frac{1}{2}}}{|(\mathbf{V}_{\gamma_i=1}^*)^{-1}|^{\frac{1}{2}}|\mathbf{V}_{\gamma_i=1}^{-1}|^{-\frac{1}{2}}}\left[\frac{2b+\mathbf{y}^T\mathbf{y}-(\boldsymbol{\beta}_{\gamma_i=1}^*)^T(\mathbf{V}_{\gamma_i=1}^*)^{-1}(\boldsymbol{\beta}_{\gamma_i=1}^*)+\widetilde{\boldsymbol{\beta}}_{\gamma_i=1}^T\mathbf{V}_{\gamma_i=1}\widetilde{\boldsymbol{\beta}}_{\gamma_i=1}}{2b+\mathbf{y}^T\mathbf{y}-(\boldsymbol{\beta}_{\gamma_i=0}^*)^T(\mathbf{V}_{\gamma_i=0}^*)^{-1}(\boldsymbol{\beta}_{\gamma_i=0}^*)+\widetilde{\boldsymbol{\beta}}_{\gamma_i=0}^T\mathbf{V}_{\gamma_i=0}\widetilde{\boldsymbol{\beta}}_{\gamma_i=0}}\right]^{\frac{n}{2}+a}. \tag{2.12}$$

Under the component-wise partial order we require (2.12) to be decreasing in $\boldsymbol{\gamma}_{-i}$. Assuming $\mathbf{V}$ is diagonal and $\mathbf{X}$ is orthogonal, for $\boldsymbol{\gamma}^{(1)}$ we obtain

$$\frac{\widetilde{\Pr}(\gamma_i^{(1)} = 0 \mid \boldsymbol{\gamma}_{-i}^{(1)}, \mathbf{y}, \mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(1)} = 1 \mid \boldsymbol{\gamma}_{-i}^{(1)}, \mathbf{y}, \mathbf{X})} =$$

$$\sqrt{\frac{v_{ii}+\mathbf{X}_i^T\mathbf{X}_i}{v_{ii}}}\left[1-\frac{(\boldsymbol{\beta}_i^*)^T(v_{ii}+\mathbf{X}_i^T\mathbf{X}_i)^{-1}(\boldsymbol{\beta}_i^*)-\widetilde{\boldsymbol{\beta}}_i^T v_{ii}\widetilde{\boldsymbol{\beta}}_i}{2b+\mathbf{y}^T\mathbf{y}-(\boldsymbol{\beta}_{\gamma_i=0}^{(1)*})^T(\mathbf{V}_{\gamma_i=0}^{(1)*})^{-1}(\boldsymbol{\beta}_{\gamma_i=0}^{(1)*})+\widetilde{\boldsymbol{\beta}}_{\gamma_i=0}^{(1)T}\mathbf{V}_{\gamma_i=0}^{(1)}\widetilde{\boldsymbol{\beta}}_{\gamma_i=0}^{(1)}}\right]^{(n/2)+a}, \tag{2.13}$$

and similarly for $\boldsymbol{\gamma}^{(2)}$:

$$\frac{\widetilde{\Pr}(\gamma_i^{(2)} = 0 \mid \boldsymbol{\gamma}_{-i}^{(2)}, \mathbf{y}, \mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(2)} = 1 \mid \boldsymbol{\gamma}_{-i}^{(2)}, \mathbf{y}, \mathbf{X})} = \tag{2.14}$$

$$\sqrt{\frac{v_{ii}+\mathbf{X}_i^T\mathbf{X}_i}{v_{ii}}}\left[1-\frac{(\boldsymbol{\beta}_i^*)^T(v_{ii}+\mathbf{X}_i^T\mathbf{X}_i)^{-1}(\boldsymbol{\beta}_i^*)-\widetilde{\boldsymbol{\beta}}_i^T v_{ii}\widetilde{\boldsymbol{\beta}}_i}{2b+\mathbf{y}^T\mathbf{y}-(\boldsymbol{\beta}_{\boldsymbol{\gamma}_i=0}^{(2)*})^T(\mathbf{V}_{\boldsymbol{\gamma}_i=0}^{(2)*})^{-1}(\boldsymbol{\beta}_{\boldsymbol{\gamma}_i=0}^{(2)*})+\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_i=0}^{(2)T}\mathbf{V}_{\boldsymbol{\gamma}_i=0}^{(2)}\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_i=0}^{(2)}}\right]^{(n/2)+a}.$$

In (2.13) and (2.14) the square-root term and the numerator $(\boldsymbol{\beta}_i^*)^T(v_{ii}+\mathbf{X}_i^T\mathbf{X}_i)^{-1}(\boldsymbol{\beta}_i^*)-\widetilde{\boldsymbol{\beta}}_i^T v_{ii}\widetilde{\boldsymbol{\beta}}_i$ are constant for γ_i. However, we cannot determine which denominator is smaller due to the addition of the quadratic term $\widetilde{\boldsymbol{\beta}}_{\gamma_i=0}^T\mathbf{V}_{\boldsymbol{\gamma}_i=0}\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_i=0}$. This means subjective choices of $\widetilde{\boldsymbol{\beta}}$ are insufficient for a monotone Gibbs MC as we cannot confirm (2.6).

Let $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ = 0, then (2.8) simplifies to

$$f(\boldsymbol{\gamma}\mid\mathbf{y},\mathbf{X})\propto|\mathbf{V}_{\boldsymbol{\gamma}}^{-1}|^{-0.5}|(\mathbf{V}_{\boldsymbol{\gamma}}^*)^{-1}|^{0.5}\,(2b+\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{X}_{\boldsymbol{\gamma}}(\mathbf{V}_{\boldsymbol{\gamma}}^*)^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{y})^{-\frac{n}{2}-a}, \qquad (2.15)$$

where $\mathbf{V}_{\boldsymbol{\gamma}}^*=(\mathbf{V}_{\boldsymbol{\gamma}}+\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})$. The Gibbs update probability (2.4) follows as

$$\frac{\widetilde{\Pr}(\gamma_i=0\mid\boldsymbol{\gamma}_{-i},\mathbf{y},\mathbf{X})}{\widetilde{\Pr}(\gamma_i=1\mid\boldsymbol{\gamma}_{-i},\mathbf{y},\mathbf{X})}=$$

$$\frac{|(\mathbf{V}_{\boldsymbol{\gamma}_i=0}^*)^{-1}|^{1/2}|\mathbf{V}_{\boldsymbol{\gamma}_i=0}^{-1}|^{-1/2}}{|(\mathbf{V}_{\boldsymbol{\gamma}_i=1}^*)^{-1}|^{1/2}|\mathbf{V}_{\boldsymbol{\gamma}_i=1}^{-1}|^{-1/2}}\left[\frac{2b+\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{X}_{\boldsymbol{\gamma}_i=1}(\mathbf{V}_{\boldsymbol{\gamma}_i=1}^*)^{-1}\mathbf{X}_{\boldsymbol{\gamma}_i=1}^T\mathbf{y}}{2b+\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{X}_{\boldsymbol{\gamma}_i=0}(\mathbf{V}_{\boldsymbol{\gamma}_i=0}^*)^{-1}\mathbf{X}_{\boldsymbol{\gamma}_i=0}^T\mathbf{y}}\right]^{(n/2)+a}. \qquad (2.16)$$

For the component-wise partial order (2.16) must be decreasing in $\boldsymbol{\gamma}_{-i}$. Assuming **V** is diagonal and **X** is orthogonal, for $\boldsymbol{\gamma}^{(1)}$:

$$\frac{\widetilde{\Pr}(\gamma_i^{(1)}=0\mid\boldsymbol{\gamma}_{-i}^{(1)},\mathbf{y},\mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(1)}=1\mid\boldsymbol{\gamma}_{-i}^{(1)},\mathbf{y},\mathbf{X})}=$$

$$\sqrt{\frac{v_{ii}+\mathbf{X}_i^T\mathbf{X}_i}{v_{ii}}}\left[1-\frac{\mathbf{y}^T\mathbf{X}_i(v_{ii}+\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{X}_i^T\mathbf{y}}{2b+\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{X}_{\boldsymbol{\gamma}_i=0}^{(1)}(\mathbf{V}_{\boldsymbol{\gamma}_i=0}^{(1)}+\mathbf{X}_{\boldsymbol{\gamma}_i=0}^{(1)T}\mathbf{X}_{\boldsymbol{\gamma}_i=0}^{(1)})^{-1}\mathbf{X}_{\boldsymbol{\gamma}_i=0}^{(1)T}\mathbf{y}}\right]^{(n/2)+a}, \qquad (2.17)$$

and for $\boldsymbol{\gamma}^{(2)}$:

$$\frac{\widetilde{\Pr}(\gamma_i^{(2)} = 0 \mid \boldsymbol{\gamma}_{-i}^{(2)}, \mathbf{y}, \mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(2)} = 1 \mid \boldsymbol{\gamma}_{-i}^{(2)}, \mathbf{y}, \mathbf{X})} = \sqrt{\frac{v_{ii} + \mathbf{X}_i^T \mathbf{X}_i}{v_{ii}}} \left[1 - \frac{\mathbf{y}^T \mathbf{X}_i (v_{ii} + \mathbf{X}_i^T \mathbf{X}_i)^{-1} \mathbf{X}_i^T \mathbf{y}}{2b + \mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{X}_{\boldsymbol{\gamma}_i=0}^{(2)} (\mathbf{V}_{\boldsymbol{\gamma}_i=0}^{(2)} + \mathbf{X}_{\boldsymbol{\gamma}_i=0}^{(2)T} \mathbf{X}_{\boldsymbol{\gamma}_i=0}^{(2)})^{-1} \mathbf{X}_{\boldsymbol{\gamma}_i=0}^{(2)T} \mathbf{y}} \right]^{(n/2)+a} . \tag{2.18}$$

The square-root and numerator ($\mathbf{y}^T \mathbf{X}_i (v_{ii} + \mathbf{X}_i^T \mathbf{X}_i)^{-1} \mathbf{X}_i^T \mathbf{y}$) terms in (2.17) and (2.18) are constant when updating the i^{th} component. The denominator is smaller for $\boldsymbol{\gamma}^{(2)}$ as the sum of squares is increasing in the number of predictors. Thus, the term is smaller than for $\boldsymbol{\gamma}^{(1)}$ so it follows (2.17) ≥ (2.18) confirming (2.6).

2.2.2 Zellner's Prior

Zellner’s G-prior (1986) avoids specification of the covariance structure and requires the choice of only one hyper-parameter $c > 0$. This has become a standard prior specification for model selection. It is simpler to deal with than the conjugate regime while retaining all the marginalization properties. The choice of c in Zellner’s prior will have the greatest impact on our posterior inference when using a flat prior for $\boldsymbol{\gamma}$. c can be interpreted as a measure of how much information is contained in the prior relative to the likelihood. If $c = 2$, then the prior has 50% weight relative to the data. c is a scale parameter as it has positive support and is used as a variance inflation parameter for $\mathbf{X}^T\mathbf{X}$.

When specifying a value for c, small values indicate a strong prior, while large values indicate a weak prior. Model selection using $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$ will behave similarly to the BIC for $c = n$ and the RIC for $c = k^2$ (Kass and Wasserman, 1995; Foster and George, 1994; Liang *et al*, 2008). EB procedures for Zellner’s prior either estimate c for every model (local EB), or for all models (global EB) by maximizing the corresponding likelihood functions for c. The MLE is then used as an estimate of c in Zellner’s prior. (Clyde and George, 2004; George and Foster, 2000; Hansen and Yu, 2001).

Let $f(\boldsymbol{\gamma}) \propto 1$ and the joint prior for $\boldsymbol{\beta}_{\gamma}$ and σ^2 be

$$f(\boldsymbol{\beta}_{\gamma}, \sigma^2 \mid \boldsymbol{\gamma}, \mathbf{X}, c) \propto \mathbf{N}_{q_\gamma+1}(\widetilde{\boldsymbol{\beta}}_{\gamma}, c\sigma^2 (\mathbf{X}_{\gamma}^T \mathbf{X}_{\gamma})^{-1}) \, (\sigma^2)^{-1}, \tag{2.19}$$

with hyper-parameters: $\widetilde{\boldsymbol{\beta}}_\gamma, c$. The marginal posterior given by (2.2) is

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}, c) \propto (c+1)^{-\frac{q_\gamma+1}{2}} \left(\mathbf{y}^T\mathbf{y} - \frac{1}{c+1}(c\mathbf{y}^T\mathbf{H}_\gamma \mathbf{y} - \widetilde{\boldsymbol{\beta}}_\gamma^T \mathbf{X}_\gamma^T \mathbf{X}_\gamma \widetilde{\boldsymbol{\beta}}_\gamma + 2\mathbf{y}^T\mathbf{X}_\gamma \widetilde{\boldsymbol{\beta}}_\gamma) \right)^{-n/2}, \quad (2.20)$$

where $\mathbf{H}_\gamma = \mathbf{X}_\gamma(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma^T$. We can deduce from the derivation of (2.20) that the joint posterior for $\boldsymbol{\beta}_\gamma$ and σ^2 is

$$\begin{aligned} f(\boldsymbol{\beta}_\gamma, \sigma^2 \mid \boldsymbol{\gamma}, \mathbf{y}, \mathbf{X}) = \mathbf{N}_{q_\gamma+1}&\left(\frac{c}{c+1}\hat{\boldsymbol{\beta}}_\gamma + \frac{1}{c+1}\widetilde{\boldsymbol{\beta}}_\gamma, \frac{c\sigma^2}{c+1}(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1} \right)\ldots \\ \mathbf{IG}&\left(\frac{n}{2}, \frac{1}{2}\left[\mathbf{y}^T\mathbf{y} - \frac{1}{c+1}(c\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y} - \widetilde{\boldsymbol{\beta}}_\gamma^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma\widetilde{\boldsymbol{\beta}}_\gamma + 2\mathbf{y}^T\mathbf{X}_\gamma\widetilde{\boldsymbol{\beta}}_\gamma) \right] \right), \end{aligned} \quad (2.21)$$

with posterior expectations

$$\mathbb{E}[\boldsymbol{\beta}_\gamma \mid \boldsymbol{\gamma}, \sigma^2, \mathbf{y}, \mathbf{X}] = \frac{c}{c+1}\hat{\boldsymbol{\beta}}_\gamma + \frac{1}{c+1}\widetilde{\boldsymbol{\beta}}_\gamma, \quad (2.22)$$

and

$$\mathbb{E}[\sigma^2 \mid \boldsymbol{\gamma}, \mathbf{y}, \mathbf{X}] = \frac{\mathbf{y}^T\mathbf{y} - (1/(c+1))(c\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y} - \widetilde{\boldsymbol{\beta}}_\gamma^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma\widetilde{\boldsymbol{\beta}}_\gamma + 2\mathbf{y}^T\mathbf{X}_\gamma\widetilde{\boldsymbol{\beta}}_\gamma)}{n-2}. \quad (2.23)$$

where $\hat{\boldsymbol{\beta}}_\gamma = (\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma^T\mathbf{y}$. Using (2.20) and (2.4) the Gibbs update probability is

$$\begin{aligned} &\frac{\widetilde{\Pr}(\gamma_i = 0 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})}{\widetilde{\Pr}(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})} = \\ &\sqrt{(c+1)}\left[\frac{\mathbf{y}^T\mathbf{y} - \widetilde{c}_1\,\mathbf{y}^T\mathbf{H}_{\gamma_i=1}\mathbf{y} + \widetilde{c}_2\widetilde{\boldsymbol{\beta}}_{\gamma_i=1}^T\mathbf{X}_{\gamma_i=1}^T\mathbf{X}_{\gamma_i=1}\widetilde{\boldsymbol{\beta}}_{\gamma_i=1} - \widetilde{c}_3\mathbf{y}^T\mathbf{X}_{\gamma_i=1}\widetilde{\boldsymbol{\beta}}_{\gamma_i=1}}{\mathbf{y}^T\mathbf{y} - \widetilde{c}_1\,\mathbf{y}^T\mathbf{H}_{\gamma_i=0}\mathbf{y} + \widetilde{c}_2\widetilde{\boldsymbol{\beta}}_{\gamma_i=0}^T\mathbf{X}_{\gamma_i=0}^T\mathbf{X}_{\gamma_i=0}\widetilde{\boldsymbol{\beta}}_{\gamma_i=0} - \widetilde{c}_3\mathbf{y}^T\mathbf{X}_{\gamma_i=0}\widetilde{\boldsymbol{\beta}}_{\gamma_i=0}} \right]^{n/2}, \end{aligned} \quad (2.24)$$

where $\widetilde{c}_1 = c/(c+1)$, $\widetilde{c}_2 = 1/(c+1)$ and $\widetilde{c}_3 = 2/(c+1)$. From Theorem 2.1 we require (2.24) to be decreasing in $\boldsymbol{\gamma}_{-i}$ for the component-wise partial order. Now suppose $\boldsymbol{\gamma}^{(1)} \preceq \boldsymbol{\gamma}^{(2)}$ then we require:

$$\frac{\widetilde{\Pr}(\gamma_i^{(1)} = 0 \mid \boldsymbol{\gamma}_{-i}^{(1)}, \mathbf{y}, \mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(1)} = 1 \mid \boldsymbol{\gamma}_{-i}^{(1)}, \mathbf{y}, \mathbf{X})} =$$

$$\sqrt{(c+1)}\left[1 - \frac{\widetilde{c}_1\, \mathbf{y}^T \mathbf{H}_i \mathbf{y} - \widetilde{c}_2 \widetilde{\boldsymbol{\beta}}_i^T \mathbf{X}_i^T \mathbf{X}_i \widetilde{\boldsymbol{\beta}}_i + \widetilde{c}_3 \mathbf{y}^T \mathbf{X}_i \widetilde{\boldsymbol{\beta}}_i}{\mathbf{y}^T \mathbf{y} - \widetilde{c}_1\, \mathbf{y}^T \mathbf{H}_{\boldsymbol{\gamma}_i=0}^{(1)} \mathbf{y} + \widetilde{c}_2 \widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_i=0}^{(1)T} \mathbf{X}_{\boldsymbol{\gamma}_i=0}^{(1)T} \mathbf{X}_{\boldsymbol{\gamma}_i=0}^{(1)} \widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_i=0}^{(1)} - \widetilde{c}_3 \mathbf{y}^T \mathbf{X}_{\boldsymbol{\gamma}_i=0}^{(1)} \widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_i=0}^{(1)}}\right]^{n/2} \tag{2.25}$$

to be greater than or equal to

$$\frac{\widetilde{\Pr}(\gamma_i^{(2)} = 0 \mid \boldsymbol{\gamma}_{-i}^{(2)}, \mathbf{y}, \mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(2)} = 1 \mid \boldsymbol{\gamma}_{-i}^{(2)}, \mathbf{y}, \mathbf{X})} =$$

$$\sqrt{(c+1)}\left[1 - \frac{\widetilde{c}_1\, \mathbf{y}^T \mathbf{H}_i \mathbf{y} - \widetilde{c}_2 \widetilde{\boldsymbol{\beta}}_i^T \mathbf{X}_i^T \mathbf{X}_i \widetilde{\boldsymbol{\beta}}_i + \widetilde{c}_3 \mathbf{y}^T \mathbf{X}_i \widetilde{\boldsymbol{\beta}}_i}{\mathbf{y}^T \mathbf{y} - \widetilde{c}_1\, \mathbf{y}^T \mathbf{H}_{\boldsymbol{\gamma}_i=0}^{(2)} \mathbf{y} + \widetilde{c}_2 \widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_i=0}^{(2)T} \mathbf{X}_{\boldsymbol{\gamma}_i=0}^{(2)T} \mathbf{X}_{\boldsymbol{\gamma}_i=0}^{(2)} \widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_i=0}^{(2)} - \widetilde{c}_3 \mathbf{y}^T \mathbf{X}_{\boldsymbol{\gamma}_i=0}^{(2)} \widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_i=0}^{(2)}}\right]^{n/2}. \tag{2.26}$$

It is clear we cannot determine if (2.25) ≥ (2.26) due to the terms $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_i=0}^T \mathbf{X}_{\boldsymbol{\gamma}_i=0}^T \mathbf{X}_{\boldsymbol{\gamma}_i=0} \widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_i=0}$ and $\mathbf{y}\mathbf{X}_{\boldsymbol{\gamma}_i=0} \widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_i=0}$ in the denominator. Thus, as in the fully conjugate case we cannot use a subjective choice of $\widetilde{\boldsymbol{\beta}}$ if we wish to construct a monotone Gibbs MC when the predictor matrix is orthogonal. Before dismissing this approach entirely we consider the following example.

EXAMPLE 2.1

When using an informative prior it is possible to minimize the Kullback-Leibler distance by choosing $\widetilde{\boldsymbol{\beta}}$ for the full model and from this information projecting the equivalent choices for sub-models as

$$\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = (\mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{X}_{\boldsymbol{\gamma}})^{-1} \mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{X} \widetilde{\boldsymbol{\beta}}, \tag{2.27}$$

There is some debate over the proper specification of the covariance matrix for the prior on $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, however, this does not apply when assuming **X** is orthogonal. The priors are

$$f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2 \mid \boldsymbol{\gamma}, \mathbf{X}, c) \propto \mathbf{N}_{q_\gamma+1}((\mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{X}_{\boldsymbol{\gamma}})^{-1} \mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{X} \widetilde{\boldsymbol{\beta}}, c\sigma^2 (\mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{X}_{\boldsymbol{\gamma}})^{-1})\ (\sigma^2)^{-1}, \tag{2.28}$$

and $f(\boldsymbol{\gamma}) \propto 1$. By substitution of (2.27) into (2.20) the posterior for $\boldsymbol{\gamma}$ becomes:

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}, c) \propto (c+1)^{-\frac{q_\gamma+1}{2}} \left(\mathbf{y}^T\mathbf{y} - \frac{c}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y} + \frac{1}{c+1}\tilde{\mathbf{y}}^T\mathbf{H}_\gamma\tilde{\mathbf{y}} - \frac{2}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\tilde{\mathbf{y}}) \right)^{-n/2} \tag{2.29}$$

where $\tilde{\mathbf{y}} = \mathbf{X}\tilde{\boldsymbol{\beta}}$. We note we can re-arrange (2.29) to

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}, c) \propto (c+1)^{-\frac{q_\gamma+1}{2}} \left(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y} + \frac{1}{c+1}(\mathbf{y}-\tilde{\mathbf{y}})^T\mathbf{H}_\gamma(\mathbf{y}-\tilde{\mathbf{y}}) \right)^{-n/2} \tag{2.30}$$

Using (2.30) and (2.4) the Gibbs update probability is

$$\frac{\widetilde{\Pr}(\gamma_i = 0 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})}{\widetilde{\Pr}(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})} = \sqrt{(c+1)} \left[\frac{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_{\gamma_i=1}\mathbf{y} + \tilde{c}_2(\mathbf{y}-\tilde{\mathbf{y}})^T\mathbf{H}_{\gamma_i=1}(\mathbf{y}-\tilde{\mathbf{y}})}{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_{\gamma_i=0}\mathbf{y} + \tilde{c}_2(\mathbf{y}-\tilde{\mathbf{y}})^T\mathbf{H}_{\gamma_i=0}(\mathbf{y}-\tilde{\mathbf{y}})} \right]^{n/2}. \tag{2.31}$$

Again we find ourselves with the addition of a quadratic term which will be increasing in q_γ. This is much the same as in the case of the conjugate and Zellner's case with a general choice of informative prior for $\tilde{\boldsymbol{\beta}}$.

■

Let $\tilde{\boldsymbol{\beta}}_\gamma = 0$ then (2.20) becomes:

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}, c) \propto (c+1)^{-\frac{q_\gamma+1}{2}} \left(\mathbf{y}^T\mathbf{y} - \frac{c}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y} \right)^{-n/2}, \tag{2.32}$$

and from (2.4) the Gibbs update probability is

$$\frac{\widetilde{\Pr}(\gamma_i = 0 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})}{\widetilde{\Pr}(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})} = \sqrt{(c+1)} \left[\frac{\mathbf{y}^T\mathbf{y} - \tilde{c}_1\mathbf{y}^T\mathbf{H}_{\gamma_i=1}\mathbf{y}}{\mathbf{y}^T\mathbf{y} - \tilde{c}_1\mathbf{y}^T\mathbf{H}_{\gamma_i=0}\mathbf{y}} \right]^{n/2}. \tag{2.33}$$

For the Gibbs MC to be monotone (2.33) must be decreasing in $\boldsymbol{\gamma}_{-i}$, for the component-wise partial order $\boldsymbol{\gamma}^{(1)} \preceq \boldsymbol{\gamma}^{(2)}$. For $\boldsymbol{\gamma}^{(1)}$ we obtain

$$\frac{\widetilde{\Pr}(\gamma_i^{(1)}=0\,|\,\boldsymbol{\gamma}_{-i}^{(1)},\mathbf{y},\mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(1)}=1\,|\,\boldsymbol{\gamma}_{-i}^{(1)},\mathbf{y},\mathbf{X})}=\sqrt{(c+1)}\left[1-\frac{\widetilde{c}_1\mathbf{y}^T\mathbf{H}_i\mathbf{y}}{\mathbf{y}^T\mathbf{y}-\widetilde{c}_1\mathbf{y}^T\mathbf{H}_{\gamma_i=0}^{(1)}\mathbf{y}}\right]^{n/2}, \tag{2.34}$$

and for $\boldsymbol{\gamma}^{(2)}$:

$$\frac{\widetilde{\Pr}(\gamma_i^{(2)}=0\,|\,\boldsymbol{\gamma}_{-i}^{(2)},\mathbf{y},\mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(2)}=1\,|\,\boldsymbol{\gamma}_{-i}^{(2)},\mathbf{y},\mathbf{X})}=\sqrt{(c+1)}\left[1-\frac{\widetilde{c}_1\mathbf{y}^T\mathbf{H}_i\mathbf{y}}{\mathbf{y}^T\mathbf{y}-\widetilde{c}_1\mathbf{y}^T\mathbf{H}_{\gamma_i=0}^{(2)}\mathbf{y}}\right]^{n/2}. \tag{2.35}$$

All terms in (2.30) and (2.31) are equal except for the term dependent upon the partial order, $\mathbf{y}^T\mathbf{H}_{\gamma_i=0}\mathbf{y}$. Because $\mathbf{y}^T\mathbf{H}_{\gamma_i=0}^{(1)}\mathbf{y} \leq \mathbf{y}^T\mathbf{H}_{\gamma_i=0}^{(2)}\mathbf{y}$, (2.35) has the smaller term in the denominator and is smaller overall so, (2.34) ≥ (2.35) as required.

For extreme choices of c the residual sums of squares (RSS) with the $c/(c+1)$ shrinkage factor asymptotically tends to the standard residual sum of squares:

$$\lim_{c\to\infty}(\mathbf{y}^T(\mathbf{I}_n-\frac{c}{c+1}\mathbf{H}_{\boldsymbol{\gamma}})\mathbf{y})\to(\mathbf{y}^T(\mathbf{I}_n-\mathbf{H}_{\boldsymbol{\gamma}})\mathbf{y}). \tag{2.36}$$

However the posterior distribution for $\boldsymbol{\gamma}$ will degenerate as

$$\lim_{c\to\infty}(c+1)^{-(q_\gamma+1)/2}\to 0. \tag{2.37}$$

This implies the (c + 1) term tends to zero at a rate dependent on q_γ. This means the null model ($q_\gamma = 0$) has the slowest rate of approach to 0, and so in the limit of $c \to \infty$ will become the most probable model in the posterior despite any evidence to the contrary. This phenomenon is referred to as Bartlett's Paradox (Bartlett, 1957) and this limiting behavior also applies to the BF.

EXAMPLE 2.2 Outlier Detection

From the results in Smith *et al* (1996) a form of weighted least squares for outlier detection is possible. An augmented form of Zellner's prior is proposed:

$$f(\boldsymbol{\beta}_{\boldsymbol{\gamma}},\sigma^2\,|\,\boldsymbol{\gamma},\mathbf{W},\mathbf{X})\propto\mathbf{N}_{q_\gamma+1}(0,c\sigma^2(\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{W}^{-1}\mathbf{X}_{\boldsymbol{\gamma}})^{-1})\,(\sigma^2)^{-1}, \tag{2.38}$$

where

$$\mathbf{W} = \begin{bmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_n \end{bmatrix}, \tag{2.39}$$

and where $w_i \in \{1, \kappa\}$ and c and κ are hyper-parameters and κ is a threshold for detecting outliers. The joint posterior for $\boldsymbol{\gamma}$ and $\mathbf{W}$ is then obtained as

$$f(\boldsymbol{\gamma}, \mathbf{W} \mid \mathbf{y}, \mathbf{X}) \propto f(\boldsymbol{\gamma}) f(\mathbf{W}) \int\int f(\mathbf{y} \mid \boldsymbol{\beta}_\gamma, \sigma^2, \boldsymbol{\gamma}, \mathbf{X}) f(\boldsymbol{\beta}_\gamma, \sigma^2 \mid \boldsymbol{\gamma}, \mathbf{W}, \mathbf{X})\, d\boldsymbol{\beta}_\gamma d\sigma^2, \tag{2.40}$$

The posteriors for $\boldsymbol{\gamma}$ and $\mathbf{W}$ are

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}, \mathbf{W}, c) \propto (c+1)^{-\frac{q_\gamma}{2}} \left(\mathbf{y}^T \mathbf{W}^{-1} \mathbf{y} - \frac{c}{c+1} \mathbf{y}^T \mathbf{W}^{-1} \mathbf{X}_\gamma (\mathbf{X}_\gamma^T \mathbf{W}^{-1} \mathbf{X}_\gamma)^{-1} \mathbf{X}_\gamma^T \mathbf{W}^{-1} \mathbf{y} \right)^{-\frac{n}{2}}, \tag{2.41}$$

and

$$f(\mathbf{W} \mid \mathbf{y}, \mathbf{X}, \boldsymbol{\gamma}, c) \propto$$

$$f(\mathbf{W}) |\mathbf{W}|^{-0.5} \left(\mathbf{y}^T \mathbf{W}^{-1} \mathbf{y} - \frac{c}{c+1} \mathbf{y}^T \mathbf{W}^{-1} \mathbf{X}_\gamma (\mathbf{X}_\gamma^T \mathbf{W}^{-1} \mathbf{X}_\gamma)^{-1} \mathbf{X}_\gamma^T \mathbf{W}^{-1} \mathbf{y} \right)^{-n/2}. \tag{2.42}$$

A Gibbs sampler can be used to sample from the posterior density of $\boldsymbol{\gamma}$ and **W**, by noting that

$$\begin{aligned} \Pr(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X}, \mathbf{W}, c) &\propto f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}, \mathbf{W}, c), \\ \Pr(w_i = \kappa \mid \boldsymbol{\gamma}, \mathbf{y}, \mathbf{X}, \mathbf{W}_{-i}, c) &\propto f(\mathbf{W} \mid \mathbf{y}, \mathbf{X}, \boldsymbol{\gamma}, c). \end{aligned} \tag{2.43}$$

where $\boldsymbol{\gamma}_{-i} = (\gamma_1, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots \gamma_k)$ and $\mathbf{W}_{-i} = diag(w_1, \ldots, w_{i-1}, w_{i+1}, \ldots w_k)$. Conditional upon a fixed sequence of weights the Gibbs MC will not be monotone. This is easiest to see when considering (2.38), even if **X** is orthonormal $(\mathbf{X}_\gamma^T \mathbf{W}^{-1} \mathbf{X}_\gamma)$ will not be diagonal when at least one w_i is not equal to the rest. A partial order for monotonicity is also unavailable for **W** and (2.42). This is because we need to express the hat matrix as a sum over observations, as **W** is updated for each observation. This is not possible, and so exact sampling using a monotone CFTP Gibbs sampler for (2.42) is not possible. However, if the number of observations is small (say $n < 30$) it would be possible to use full CFTP with the Gibbs sampler.

■

Because Zellner's prior for $\boldsymbol{\beta}_\gamma \mid \sigma^2$ is a special case of conjugate prior with $\mathbf{V} = \mathbf{X}^T\mathbf{X}/c$, by analogy it is clear that using Zellner's prior with a fully conjugate prior for σ^2 will be monotone provided $\tilde{\boldsymbol{\beta}}_\gamma = 0$.

2.2.3 Jeffreys Prior

The common alternative to avoid any hyper-parameter specification is Jeffreys prior. Jeffreys prior is given by the square root of the determinant of Fisher information, $\mathbf{I}(\boldsymbol{\theta})$:

$$f(\boldsymbol{\theta}) \propto |\mathbf{I}(\boldsymbol{\theta})|^{1/2} \tag{2.44}$$

It is re-parameterization invariant meaning if we transform the parameter, the prior for the transformed parameter is still Jeffreys prior. Because of the relation to the Fisher information, when there is large information, we minimize the influence of the prior such that it is as non-informative as possible. Priors like Jeffreys are considered a default procedure and in practice should be used when we have a lot of data and few parameters, i.e. when the likelihood will be very sharply peaked. Jeffreys prior does not satisfy the likelihood principle, is improper, may lead to indeterminate BF, and for proper Bayesians has little subjective justification with respect to prior information. Jeffrey noted in the multi-dimensional case ad hoc adjustments to the prior were required and stressed these priors for use in the uni-dimensional case as they may lead to incoherence or paradoxes in the multi-dimensional case. For a more detailed discussion of Jeffreys prior see Roberts (2001, Chapter 3). Following the derivation in Appendix B, Jeffreys true prior is

$$f(\boldsymbol{\beta}_\gamma, \sigma^2 \mid \boldsymbol{\gamma}, \mathbf{X}) \propto |\mathbf{X}_\gamma^T \mathbf{X}_\gamma|^{+\frac{1}{2}} (\sigma^2)^{-\left(\frac{q_\gamma+1}{2}+1\right)}. \tag{2.45}$$

where $q_\gamma = \boldsymbol{\gamma}^T\boldsymbol{\gamma} - 1$. Wasserman (2000) suggested an add-on adjustment to Jeffreys prior:

$$f(\boldsymbol{\beta}_\gamma, \sigma^2 \mid \boldsymbol{\gamma}, \mathbf{X}) \propto (2\pi)^{-\left(\frac{q_\gamma+1}{2}\right)} \mid \mathbf{X}_\gamma^T \mathbf{X}_\gamma \mid^{+\frac{1}{2}} (\sigma^2)^{-\left(\frac{q_\gamma+1}{2}+1\right)}, \tag{2.46}$$

and in similar spirit we propose an adjusted form, with arbitrary penalty $p > 0$:

$$f(\boldsymbol{\beta}_\gamma, \sigma^2 \mid \boldsymbol{\gamma}, \mathbf{X}) \propto (p)^{-\frac{q_\gamma}{2}} \mid \mathbf{X}_\gamma^T \mathbf{X}_\gamma \mid^{+\frac{1}{2}} (\sigma^2)^{-\left(\frac{q_\gamma+1}{2}+1\right)}. \tag{2.47}$$

We will use Jeffreys prior to refer to (2.47) and refer to Jeffreys true prior to distinguish (2.45) from (2.47). Let $f(\boldsymbol{\gamma}) \propto 1$ and the prior for $\boldsymbol{\beta}_\gamma$ and σ^2 be (2.47), then following (2.2) the marginal posterior for $\boldsymbol{\gamma}$ is

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) \propto \left(\frac{p}{2\pi}\right)^{-\frac{q_\gamma}{2}} (\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})^{-\frac{n}{2}}, \tag{2.48}$$

where $\mathbf{H}_\gamma = \mathbf{X}_\gamma(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma^T$. The joint posterior for $\boldsymbol{\beta}_\gamma$ and σ^2 is

$$f(\boldsymbol{\beta}_\gamma, \sigma^2 \mid \boldsymbol{\gamma}, \mathbf{y}, \mathbf{X}) = \mathbf{N}_{q_\gamma+1}(\hat{\boldsymbol{\beta}}_\gamma, \sigma^2(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1})\ \mathbf{IG}\left(\frac{n}{2}, \frac{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}}{2}\right), \tag{2.49}$$

with posterior expectations

$$\mathbb{E}[\boldsymbol{\beta}_\gamma \mid \boldsymbol{\gamma}, \sigma^2, \mathbf{y}, \mathbf{X}] = \hat{\boldsymbol{\beta}}_\gamma, \tag{2.50}$$

and

$$\mathbb{E}[\sigma^2 \mid \boldsymbol{\gamma}, \mathbf{y}, \mathbf{X}] = (\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})/(n-2). \tag{2.51}$$

Following (2.4) the probability for the update of the Gibbs sampler is

$$\frac{\widetilde{\Pr}(\gamma_i = 0 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})}{\widetilde{\Pr}(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})} = \sqrt{\frac{p}{2\pi}}\left[\frac{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_{\gamma_i=1}\mathbf{y}}{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_{\gamma_i=0}\mathbf{y}}\right]^{n/2}. \tag{2.52}$$

For the component-wise partial order $\boldsymbol{\gamma}^{(1)} \preceq \boldsymbol{\gamma}^{(2)}$ we require:

$$\frac{\widetilde{\Pr}(\gamma_i^{(1)}=0\,|\,\boldsymbol{\gamma}_{-i}^{(1)},\mathbf{y},\mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(1)}=1\,|\,\boldsymbol{\gamma}_{-i}^{(1)},\mathbf{y},\mathbf{X})}=\sqrt{\frac{p}{2\pi}}\left[1-\frac{\mathbf{y}^T\mathbf{H}_i\mathbf{y}}{\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{H}_{\gamma_i=0}^{(1)}\mathbf{y}}\right]^{n/2}, \tag{2.53}$$

to be greater than or equal to

$$\frac{\widetilde{\Pr}(\gamma_i^{(2)}=0\,|\,\boldsymbol{\gamma}_{-i}^{(2)},\mathbf{y},\mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(2)}=1\,|\,\boldsymbol{\gamma}_{-i}^{(2)},\mathbf{y},\mathbf{X})}=\sqrt{\frac{p}{2\pi}}\left[1-\frac{\mathbf{y}^T\mathbf{H}_i\mathbf{y}}{\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{H}_{\gamma_i=0}^{(2)}\mathbf{y}}\right]^{n/2}. \tag{2.54}$$

Because $\mathbf{y}^T\mathbf{H}_{\gamma_i=0}^{(1)}\mathbf{y} \le \mathbf{y}^T\mathbf{H}_{\gamma_i=0}^{(2)}\mathbf{y}$ it follows that (2.53) ≥ (2.54), so using an adjusted Jeffreys prior with an orthogonal predictor matrix is sufficient for a monotone Gibbs MC. We note that in general for a Jeffreys prior of the form:

$$\mathrm{C}\,(\sigma^2)^{-\alpha}\text{, where C is some constant w.r.t to } \sigma^2, \tag{2.55}$$

α must be of the form:

$$\alpha=\frac{q_\gamma+1}{2}+d\text{, where } d\ge 1, \tag{2.56}$$

in order to ensure monotonicity. Notice the posteriors for $\boldsymbol{\beta}$ and σ^2 for Zellner's prior with $\widetilde{\boldsymbol{\beta}}_\gamma$ = 0 become the same as those for Jeffreys prior in the limit $c \to \infty$. The penalty terms $(c+1)^{-(q_\gamma+1)/2}$ and $(p/2\pi)^{-q_\gamma/2}$ in the marginal posterior distribution of γ for Zellner’s and Jeffreys priors respectively are equivalent when $p = 2\pi(c+1)$.

2.3 Non-uniform priors for γ

The choice of prior for γ is an important component in determining the distribution of mass for the marginal posterior for γ. From (2.2) it is possible to assess the effect of priors for γ independent of the priors for $\boldsymbol{\beta}_\gamma$ and σ^2. Specifically, if the choice of priors for $\boldsymbol{\beta}_\gamma$, σ^2 and hyper-parameters are sufficient with an orthogonal predictor matrix for monotonicity, then f(γ) need only be constant for the i-th component or observe the component-wise partial order to preserve monotonicity.

2.3.1 Bernoulli class

Let the priors for β_γ and σ^2 be Zellner's prior (2.19) with $\widetilde{\boldsymbol{\beta}}_\gamma = 0$ and the prior for γ be the Bernoulli distribution, i.e.

$$f(\boldsymbol{\gamma}) = \prod_{i=1}^{k} \omega_i^{\gamma_i}(1-\omega_i)^{1-\gamma_i}. \tag{2.57}$$

with hyper-parameters $\omega_1,...,\omega_k$. With (2.57) and (2.19) the posterior for $\boldsymbol{\gamma}$ is

$$f(\boldsymbol{\gamma} \mid \mathbf{y},\mathbf{X}) \propto (c+1)^{-(q_\gamma+1)/2}(\mathbf{y}^T\mathbf{y} - \widetilde{c}_1\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})^{-n/2}\prod_{i=1}^{k}\omega_i^{\gamma_i}(1-\omega_i)^{1-\gamma_i}. \tag{2.58}$$

The update probability (2.4) for the Gibbs sampler is

$$\frac{\Pr(\gamma_i = 0 \mid \boldsymbol{\gamma}_{-i},\mathbf{y},\mathbf{X})}{\Pr(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i},\mathbf{y},\mathbf{X})} = \frac{\sqrt{(c+1)}(1-\omega_i)}{\omega_i}\left[1 - \frac{\widetilde{c}_1\mathbf{y}^T\mathbf{H}_i\mathbf{y}}{\mathbf{y}^T\mathbf{y} - \widetilde{c}_1\mathbf{y}^T\mathbf{H}_{\gamma_i=0}\mathbf{y}}\right]^{n/2}. \tag{2.59}$$

Thus, (2.59) according to the component-wise partial order $\boldsymbol{\gamma}^{(1)} \preceq \boldsymbol{\gamma}^{(2)}$ must be decreasing in $\boldsymbol{\gamma}_{-i}$. For $\boldsymbol{\gamma}^{(1)}$ we obtain:

$$\frac{\widetilde{\Pr}(\gamma_i^{(1)} = 0 \mid \boldsymbol{\gamma}_{-i}^{(1)},\mathbf{y},\mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(1)} = 1 \mid \boldsymbol{\gamma}_{-i}^{(1)},\mathbf{y},\mathbf{X})} = \frac{\sqrt{(c+1)}(1-\omega_i)}{\omega_i}\left[1 - \frac{\widetilde{c}_1\mathbf{y}^T\mathbf{H}_i\mathbf{y}}{\mathbf{y}^T\mathbf{y} - \widetilde{c}_1\mathbf{y}^T\mathbf{H}^{(1)}_{\gamma_i=0}\mathbf{y}}\right]^{n/2}, \tag{2.60}$$

and for $\boldsymbol{\gamma}^{(2)}$:

$$\frac{\widetilde{\Pr}(\gamma_i^{(2)} = 0 \mid \boldsymbol{\gamma}_{-i}^{(2)},\mathbf{y},\mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(2)} = 1 \mid \boldsymbol{\gamma}_{-i}^{(2)},\mathbf{y},\mathbf{X})} = \frac{\sqrt{(c+1)}(1-\omega_i)}{\omega_i}\left[1 - \frac{\widetilde{c}_1\mathbf{y}^T\mathbf{H}_i\mathbf{y}}{\mathbf{y}^T\mathbf{y} - \widetilde{c}_1\mathbf{y}^T\mathbf{H}^{(2)}_{\gamma_i=0}\mathbf{y}}\right]^{n/2}. \tag{2.61}$$

We have already shown Zellner's case with $\widetilde{\boldsymbol{\beta}}_\gamma = 0$ is monotone (2.34 and 2.35) and the additional $(1-\omega_i)/\omega_i$ is a constant for both $\boldsymbol{\gamma}^{(1)}$ and $\boldsymbol{\gamma}^{(2)}$, so (2.60) ≥ (2.61) and monotonicity follows. Let $\omega_1,...,\omega_k = \tau$, then we obtain the constant Bernoulli prior for $\boldsymbol{\gamma}$.

$$f(\boldsymbol{\gamma}) = \prod_{i=1}^{k}\tau^{\gamma_i}(1-\tau)^{1-\gamma_i}\text{, or equivalently } f(q_\gamma) \sim \mathit{Binomial}(\tau,k). \tag{2.62}$$

Using (2.62) $(1-\tau)/\tau$ replaces $(1-\omega_i)/\omega_i$ in (2.60) and (2.61) and again monotonicity follows. A common extension to the constant Bernoulli prior (2.62) is to use the conjugate beta hyper-prior for τ.

$$f(\boldsymbol{\gamma} \mid \tau) = \tau^{q_\gamma}(1-\tau)^{k-q_\gamma}\text{, and } f(\tau; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\tau^{a-1}(1-\tau)^{b-1}, \tag{2.63}$$

where a and b are hyper hyper-parameters and Γ is the gamma function. Because of the conjugate relationship it is straight forward to integrate out τ.

$$\begin{aligned} f(\boldsymbol{\gamma} \mid a, b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int \tau^{a+q_\gamma-1}(1-\tau)^{k+b-q_\gamma-1}\,d\tau \\ &= \frac{\Gamma(a+b)\Gamma(a+q_\gamma)\Gamma(k+b-q_\gamma)}{\Gamma(a)\Gamma(b)\Gamma(k+a+b)} \end{aligned} \tag{2.64}$$

Then with the result in (2.64) we obtain, up to proportionality (which requires a and b are independent of $\boldsymbol{\gamma}$), the Beta-Bernoulli prior:

$$f(\boldsymbol{\gamma} \mid a, b) \propto \Gamma(a+q_\gamma)\Gamma(k+b-q_\gamma). \tag{2.65}$$

Let the priors for $\boldsymbol{\beta}_\gamma$ and σ^2 be Zellner's prior (2.19) with $\tilde{\boldsymbol{\beta}}_\gamma = 0$, and the prior for $\boldsymbol{\gamma}$ be the Beta-Bernoulli prior. The marginal posterior for $\boldsymbol{\gamma}$ (2.2) is

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) \propto (c+1)^{-(q_\gamma+1)/2}(\mathbf{y}^T\mathbf{y} - \tilde{c}_1\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})^{-n/2}\Gamma(a+q_\gamma)\Gamma(b+k-q_\gamma), \tag{2.66}$$

which leads to the update probability for the Gibbs sampler:

$$\frac{\Pr(\gamma_i = 0 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})}{\Pr(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})} = \frac{\sqrt{(c+1)}(C/q_{\gamma_i=0} - 1)}{(a/q_{\gamma_i=0} + 1)}\left[\frac{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\tilde{c}_1\mathbf{H}_{\gamma_i=1}\mathbf{y}}{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\tilde{c}_1\mathbf{H}_{\gamma_i=0}\mathbf{y}}\right]^{n/2}, \tag{2.67}$$

where $C = b + k - 1$. According to theorem 2.1 (2.67) must be decreasing in $\boldsymbol{\gamma}_{-i}$ for the required component-wise partial. For $\boldsymbol{\gamma}^{(1)}$:

$$\frac{\widetilde{\Pr}(\gamma_i^{(1)}=0\,|\,\boldsymbol{\gamma}_{-i}^{(1)},\mathbf{y},\mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(1)}=1\,|\,\boldsymbol{\gamma}_{-i}^{(1)},\mathbf{y},\mathbf{X})} = \sqrt{(c+1)}\,\frac{C/q^{(1)}_{\gamma_i=0}-1}{a/q^{(1)}_{\gamma_i=0}+1}\left[1-\frac{\mathbf{y}^T\widetilde{c}_1\,\mathbf{H}_i\mathbf{y}}{\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\widetilde{c}_1\,\mathbf{H}^{(1)}_{\gamma_i=0}\mathbf{y}}\right]^{n/2}, \tag{2.68}$$

and for $\boldsymbol{\gamma}^{(2)}$:

$$\frac{\widetilde{\Pr}(\gamma_i^{(2)}=0\,|\,\boldsymbol{\gamma}_{-i}^{(2)},\mathbf{y},\mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(2)}=1\,|\,\boldsymbol{\gamma}_{-i}^{(2)},\mathbf{y},\mathbf{X})} = \sqrt{(c+1)}\,\frac{C/q^{(2)}_{\gamma_i=0}-1}{a/q^{(2)}_{\gamma_i=0}+1}\left[1-\frac{\mathbf{y}^T\widetilde{c}_1\,\mathbf{H}_i\mathbf{y}}{\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\widetilde{c}_1\,\mathbf{H}^{(2)}_{\gamma_i=0}\mathbf{y}}\right]^{n/2}. \tag{2.69}$$

Again all terms are known to permit monotonicity except for

$$\frac{C/q^{(1)}_{\gamma_i=0}-1}{a/q^{(1)}_{\gamma_i=0}+1}\overset{?}{\geq}\frac{C/q^{(2)}_{\gamma_i=0}-1}{a/q^{(2)}_{\gamma_i=0}+1}, \tag{2.70}$$

which we must show observes the required partial order. Both sides of (2.70) are the same except for q_γ where: $q^{(1)}_{\gamma_i=0}\leq q^{(2)}_{\gamma_i=0}$, so we can describe both sides as a single function of q_γ, i.e.

$$f(q_\gamma)=\frac{C/q_\gamma-1}{a/q_\gamma+1}. \tag{2.71}$$

This implies for monotonicity $f(q_\gamma^{(1)})\geq f(q_\gamma^{(2)})\;\forall\; q_\gamma^{(1)}\leq q_\gamma^{(2)}$ which is the definition of a decreasing function, which requires $f'(q_\gamma)\leq 0$.

$$f'(q_\gamma)=-\frac{C+a}{(a+q_\gamma)^2}. \tag{2.72}$$

Thus, because $f'(q_\gamma)<0$ it is a strictly decreasing function and is true for any choice of a and b. It then follows that (2.68) ≥ (2.69) and so provides a monotone Gibbs MC with an orthogonal predictor matrix. The most common choice of $a = b$ is 1 which is a uniform prior for τ.

2.3.2 Truncated Poisson Prior

Another prior for $\boldsymbol{\gamma}$ which may be utilized is the truncated Poisson distribution, i.e.

$$f(\boldsymbol{\gamma} \mid \lambda) \propto \binom{k}{q_\gamma}^{-1} e^{-\lambda} \frac{\lambda^{q_\gamma}}{q_\gamma!} \mathrm{I}_{\{0,\ldots,k\}}(q_\gamma). \tag{2.73}$$

With Zellner's prior (2.19) with $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ = 0 for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and (2.73) the prior for $\boldsymbol{\gamma}$, we obtain the posterior (2.2):

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) \propto (c+1)^{-(q_\gamma+1)/2} (\mathbf{y}^T\mathbf{y} - \tilde{c}_1 \mathbf{y}^T \mathbf{H}_{\boldsymbol{\gamma}} \mathbf{y})^{-n/2} \binom{k}{q_\gamma}^{-1} e^{-\lambda} \frac{\lambda^{q_\gamma}}{q_\gamma!}. \tag{2.74}$$

The update probability (2.4) for the Gibbs sampler is

$$\frac{\Pr(\gamma_i = 0 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})}{\Pr(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})} = \frac{\sqrt{(c+1)}(k - q_{\gamma_i=0})}{\lambda} \left[\frac{\mathbf{y}^T\mathbf{y} - \tilde{c}_1 \mathbf{y}^T \mathbf{H}_{\boldsymbol{\gamma}_i=1} \mathbf{y}}{\mathbf{y}^T\mathbf{y} - \tilde{c}_1 \mathbf{y}^T \mathbf{H}_{\boldsymbol{\gamma}_i=0} \mathbf{y}} \right]^{n/2}. \tag{2.75}$$

which gives:

$$\frac{\tilde{\Pr}(\gamma_i^{(1)} = 0 \mid \boldsymbol{\gamma}_{-i}^{(1)}, \mathbf{y}, \mathbf{X})}{\tilde{\Pr}(\gamma_i^{(1)} = 1 \mid \boldsymbol{\gamma}_{-i}^{(1)}, \mathbf{y}, \mathbf{X})} = \frac{\sqrt{(c+1)}(k - q_{\gamma_i=0}^{(1)})}{\lambda} \left[1 - \frac{\tilde{c}_1 \mathbf{y}^T \mathbf{H}_i \mathbf{y}}{\mathbf{y}^T\mathbf{y} - \tilde{c}_1 \mathbf{y}^T \mathbf{H}_{\boldsymbol{\gamma}_i=0}^{(1)} \mathbf{y}} \right]^{n/2}, \tag{2.76}$$

and

$$\frac{\tilde{\Pr}(\gamma_i^{(2)} = 0 \mid \boldsymbol{\gamma}_{-i}^{(2)}, \mathbf{y}, \mathbf{X})}{\tilde{\Pr}(\gamma_i^{(2)} = 1 \mid \boldsymbol{\gamma}_{-i}^{(2)}, \mathbf{y}, \mathbf{X})} = \frac{\sqrt{(c+1)}(k - q_{\gamma_i=0}^{(2)})}{\lambda} \left[1 - \frac{\tilde{c}_1 \mathbf{y}^T \mathbf{H}_i \mathbf{y}}{\mathbf{y}^T\mathbf{y} - \tilde{c}_1 \mathbf{y}^T \mathbf{H}_{\boldsymbol{\gamma}_i=0}^{(2)} \mathbf{y}} \right]^{n/2}. \tag{2.77}$$

Now because $q_{\gamma_i=0}^{(1)} \leq q_{\gamma_i=0}^{(2)}$ then:

$$\frac{\sqrt{(c+1)}(k - q_{\gamma_i=0}^{(1)})}{\lambda} \geq \frac{\sqrt{(c+1)}(k - q_{\gamma_i=0}^{(2)})}{\lambda}, \tag{2.78}$$

and so it follows that (2.76) ≥ (2.77). Thus, the use of the truncated Poisson prior for $\boldsymbol{\gamma}$ provides a monotone Gibbs MC. Consider the more general case where we specify a prior on q_γ. We then need to make the required adjustment in order to move to the $\boldsymbol{\gamma}$ space and derive the Gibbs update function.

Assume a prior on q_γ not involving the binomial coefficient, we get the following posterior:

$$f(\boldsymbol{\gamma} \mid \mathbf{y},\mathbf{X}) \propto (c+1)^{-(q_\gamma+1)/2}(\mathbf{y}^T\mathbf{y}-\tilde{c}_1\mathbf{y}^T\mathbf{H}_{\boldsymbol{\gamma}}\mathbf{y})^{-n/2}\binom{k}{q_\gamma}^{-1} f(q_\gamma), \tag{2.79}$$

which under Zellner's prior gives the ratio in the Gibbs sampler as

$$\frac{\Pr(\gamma_i=0 \mid \boldsymbol{\gamma}_{-i},\mathbf{y},\mathbf{X})}{\Pr(\gamma_i=1 \mid \boldsymbol{\gamma}_{-i},\mathbf{y},\mathbf{X})}=\frac{\sqrt{(c+1)}(k-q_{\gamma_i=0})}{(q_{\gamma_i=0}+1)}\frac{f(q_{\gamma_i=0})}{f(q_{\gamma_i=1})}\left[\frac{\mathbf{y}^T\mathbf{y}-\tilde{c}_1\mathbf{y}^T\mathbf{H}_{\boldsymbol{\gamma}_i=1}\mathbf{y}}{\mathbf{y}^T\mathbf{y}-\tilde{c}_1\mathbf{y}^T\mathbf{H}_{\boldsymbol{\gamma}_i=0}\mathbf{y}}\right]^{n/2}. \tag{2.80}$$

This leads to the monotonicity inequality

$$\frac{\widetilde{\Pr}(\gamma_i^{(1)}=0 \mid \boldsymbol{\gamma}_{-i}^{(1)},\mathbf{y},\mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(1)}=1 \mid \boldsymbol{\gamma}_{-i}^{(1)},\mathbf{y},\mathbf{X})}=\frac{\sqrt{(c+1)}(k-q_{\gamma_i=0}^{(1)})}{(q_{\gamma_i=0}^{(1)}+1)}\frac{f(q_{\gamma_i=0}^{(1)})}{f(q_{\gamma_i=1}^{(1)})}\left[1-\frac{\tilde{c}_1\mathbf{y}^T\mathbf{H}_i\mathbf{y}}{\mathbf{y}^T\mathbf{y}-\tilde{c}_1\mathbf{y}^T\mathbf{H}_{\boldsymbol{\gamma}_i=0}^{(1)}\mathbf{y}}\right]^{n/2}, \tag{2.81}$$

and

$$\frac{\widetilde{\Pr}(\gamma_i^{(2)}=0 \mid \boldsymbol{\gamma}_{-i}^{(2)},\mathbf{y},\mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(2)}=1 \mid \boldsymbol{\gamma}_{-i}^{(2)},\mathbf{y},\mathbf{X})}=\frac{\sqrt{(c+1)}(k-q_{\gamma_i=0}^{(2)})}{(q_{\gamma_i=0}^{(2)}+1)}\frac{f(q_{\gamma_i=0}^{(2)})}{f(q_{\gamma_i=1}^{(2)})}\left[1-\frac{\tilde{c}_1\mathbf{y}^T\mathbf{H}_i\mathbf{y}}{\mathbf{y}^T\mathbf{y}-\tilde{c}_1\mathbf{y}^T\mathbf{H}_{\boldsymbol{\gamma}_i=0}^{(2)}\mathbf{y}}\right]^{n/2}. \tag{2.82}$$

Because $q_{\gamma_i=0}^{(1)} \le q_{\gamma_i=0}^{(2)}$, $q_{\gamma_i=1}^{(1)} \le q_{\gamma_i=1}^{(2)}$ and k is fixed this implies:

$$\frac{k-q_{\gamma_i=0}^{(1)}}{q_{\gamma_i=1}^{(1)}} \ge \frac{k-q_{\gamma_i=0}^{(2)}}{q_{\gamma_i=1}^{(2)}}. \tag{2.83}$$

so provided $f(q_\gamma)$ is decreasing in q_γ then monotonicity is indeed preserved. It seems that priors for γ are passive with respect to monotonicity. The choice of prior for γ can at best preserve the monotonicity of the likelihood, but never induce a non-monotone likelihood to be a monotone posterior. Any augmentation of other priors in order to cancel with terms in the exchangeable class of priors works because we recover the underlying form of the independence prior. Hence we may as well use an independence prior. From the previous section, any case for an orthogonal **X** which is monotone for Zellner's prior will also be monotone for the cases of conjugate and Jeffreys priors.

2.3.3 Integration over *c*

When *c* is treated as an additional parameter we may assign a hyper-prior for *c*, to obtain a joint posterior for *c* and γ. Let the prior be *f*(γ) and the hyper-prior for *c* be *f*(*c*), then the joint posterior for γ and *c* is

$$f(\boldsymbol{\gamma}, c \mid \mathbf{y}, \mathbf{X}) \propto f(\mathbf{y} \mid \boldsymbol{\gamma}, \mathbf{X}, c) f(\boldsymbol{\gamma}) f(c) . \tag{2.84}$$

The marginal posterior distributions for either γ or *c* are then

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) \propto f(\boldsymbol{\gamma}) \int f(\mathbf{y} \mid \boldsymbol{\gamma}, \mathbf{X}, c) f(c) \, dc \ , \tag{2.85}$$

and

$$f(c \mid \mathbf{y}, \mathbf{X}) \propto f(c) \sum_{\gamma} f(\mathbf{y} \mid \boldsymbol{\gamma}, \mathbf{X}, c) f(\boldsymbol{\gamma}) \ . \tag{2.86}$$

If *c* is assumed to have positive integer support {1, 2, ... } then (2.85) gives the special case:

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) \propto f(\boldsymbol{\gamma}) \sum_{c=1}^{\infty} f(\mathbf{y} \mid \boldsymbol{\gamma}, \mathbf{X}, c) f(c) \ . \tag{2.87}$$

So we can perform inference on γ after integration over *c*, which may or may not lead to closed-form expressions. We can however, simplify by returning to an EB approach to determine the most likely value of *c* using the marginal posterior for *c*. For common choices of priors for *c* see Table 2.1.

The use of a flat prior for *c* means the marginal posterior is an equally weighted sum of all marginal posteriors of γ over the specified range of *c*. Such an option will not yield a proper posterior for *c*. The prior for *c* must decrease to 0 quickly enough as $c \to \infty$, in order to ensure the variance is finite and the posterior for *c* is proper. Celeux *et al*, (2006) use a compatibility approach to create a posterior distribution for γ that could be integrated over *c* to produce the marginal distribution for variable selection. Note that the power of 1 for *c* in the compatibility prior could be replaced with some other $a > 0$, in similar fashion to the hyper-*G* and hyper-*G*-*n* priors.

Table 2.1 Choices of hyper-prior for the hyper-parameter *c*.

Class	Prior
Compatibility Prior	$f(c) \propto c^{-1} \; \mathbb{I}_{\{1,2,3,\ldots\}}(c)$
Hyper-G prior	$f(c \mid a) = \frac{a-2}{2}(1+c)^{-a/2} \; \mathbb{I}_{[0,\infty)}(c),\, a > 2$
Hyper-G-n prior	$f(c \mid a) = \frac{a-2}{2n}(1+c/n)^{-a/2} \; \mathbb{I}_{[0,\infty)}(c),\, a > 2$
Zellner-Siow	$f(c) = \frac{e^{-n/2c}}{c^{3/2}} \; \mathbb{I}_{[0,\infty)}(c)$

In practice, the prior of *c* has finite support so that integrating out *c* involves summation for a range of *c* = {1,2,3,…..,$c_{\lim}$} for some specified upper limit. This bears an interesting relation to the fact that past a certain point the posterior will only ever select the model with no predictor (intercept only), the null model. It seems that a practical upper limit could be set with the idea in mind of minimizing the summation over a huge number of marginal posteriors for $\boldsymbol{\gamma}$ that only ever select the null model. The second is that for inference on parameters we no longer have the posterior marginal distributions for $\boldsymbol{\beta}$ and σ^2 available in closed form. We can however, compute the corresponding expectations, for example see Celeux *et al* (2006). This also means we now have a near automatic procedure except that the user must define the upper limit on the summation over *c*. In that work they propose a new family of priors for *c* called hyper-*G* priors, and hyper *G-n* priors (Celeux *et al*, 2006; Liang *et al*, 2008; Zellner and Siow, 1980).

Normally, for $\widetilde{\boldsymbol{\beta}} = 0$,

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) \propto f(\boldsymbol{\gamma}) \int (c+1)^{-\frac{q_\gamma + 1}{2}} \left(\mathbf{y}^T \mathbf{y} - \frac{c}{c+1} \mathbf{y}^T \mathbf{H}_\gamma \mathbf{y} \right)^{-n/2} f(c) dc \,. \qquad (2.88)$$

Let the prior for *c* be

$$f(c) \propto (c+1)^{-1}, \qquad (2.89)$$

an improper version of the hyper-G prior with a = 2. The work by Liang *et al* (2008) demonstrates this integration leads to a Gaussian hyper-geometric function which requires an approximation and as such confirming monotonicity for this approach is extremely difficult.

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) \propto (q_\gamma + 1)^{-1}\, {}_2F_1(n/2, 1; (q_\gamma + 3)/2, \mathbf{y}^T \mathbf{H}_\gamma \mathbf{y} / \mathbf{y}^T \mathbf{y}), \tag{2.90}$$

where:

$${}_2F_1(a, b; c, z) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{(c)_k\, k!} z^k \text{ , for |z| < 1, where } (x)_k = \Gamma(x+k)/\Gamma(x) \tag{2.91}$$

Typically however, the function is truncated and summed to a large value. Confirming monotonicity, even sufficiency, is not straight forward if possible. Notice that for Jeffreys prior, we cannot perform a similar integration like we do for c in Zellner's prior with $\tilde{\boldsymbol{\beta}}_\gamma$ = $\hat{\boldsymbol{\beta}}_\gamma$. It is incorrect to interpret p as a scale parameter like c.

2.4 Empirical Bayes

We now look at EB methods for specifying hyper-parameters. In particular, we consider the conjugate and Zellner's case with $\tilde{\boldsymbol{\beta}}_\gamma$ = $\hat{\boldsymbol{\beta}}_\gamma$ and then two examples of fully specified EB conjugate priors for $\boldsymbol{\beta}_\gamma$ and σ^2. Finally we return to integration over c in the EB setting.

2.4.1 An empirical Bayes choice for $\tilde{\boldsymbol{\beta}}_\gamma$

Using the conjugate priors (2.7) let $f(\gamma) \propto 1$, $\tilde{\boldsymbol{\beta}}_\gamma$ = $\hat{\boldsymbol{\beta}}_\gamma$, and note $\mathbf{X}_\gamma^T \mathbf{y} = (\mathbf{X}_\gamma^T \mathbf{X}_\gamma)\hat{\boldsymbol{\beta}}_\gamma$ the posterior (2.8) can be expressed as

$$\begin{aligned} &f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) \propto |\mathbf{V}_\gamma^{-1}|^{-0.5} |(\mathbf{V}_\gamma^*)^{-1}|^{0.5} \\ &\quad (2b + \mathbf{y}^T\mathbf{y} - (\mathbf{V}_\gamma \tilde{\boldsymbol{\beta}}_\gamma + \mathbf{X}_\gamma^T \mathbf{y})^T (\mathbf{V}_\gamma + \mathbf{X}_\gamma^T \mathbf{X}_\gamma)^{-1} (\mathbf{V}_\gamma \tilde{\boldsymbol{\beta}}_\gamma + \mathbf{X}_\gamma^T \mathbf{y}) + \tilde{\boldsymbol{\beta}}_\gamma^T \mathbf{V}_\gamma \tilde{\boldsymbol{\beta}}_\gamma)^{-\frac{n}{2} - a}. \end{aligned} \tag{2.92}$$

Using the identity:

$$(\mathbf{V}_\gamma + \mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1} = (\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1} - (\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}(\mathbf{V}_\gamma^{-1} + (\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1})^{-1}(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}, \qquad (2.93)$$

(2.92) becomes:

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) \propto \mid \mathbf{V}_\gamma^{-1} \mid^{-0.5} \mid (\mathbf{V}_\gamma^*)^{-1} \mid^{0.5} (2b + \mathbf{y}^T\mathbf{y} - \ldots$$
$$[\tilde{\boldsymbol{\beta}}_\gamma^T \mathbf{V}_\gamma \tilde{\boldsymbol{\beta}}_\gamma + \hat{\boldsymbol{\beta}}_\gamma^T \mathbf{X}_\gamma^T \mathbf{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma + (\tilde{\boldsymbol{\beta}}_\gamma - \hat{\boldsymbol{\beta}}_\gamma)^T (\mathbf{V}_\gamma^{-1} + (\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1})^{-1} (\tilde{\boldsymbol{\beta}}_\gamma - \hat{\boldsymbol{\beta}}_\gamma)] + \tilde{\boldsymbol{\beta}}_\gamma^T \mathbf{V}_\gamma \tilde{\boldsymbol{\beta}}_\gamma)^{-\frac{n}{2}-a}, \qquad (2.94)$$

which simplifies to

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) \propto \mid \mathbf{V}_\gamma^{-1} \mid^{-0.5} \mid (\mathbf{V}_\gamma^*)^{-1} \mid^{0.5} (2b + \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})^{-\frac{n}{2}-a}. \qquad (2.95)$$

where $\mathbf{V}_\gamma^* = (\mathbf{V}_\gamma + \mathbf{X}_\gamma^T\mathbf{X}_\gamma)$. The Gibbs update probability (2.4) is

$$\frac{\widetilde{\Pr}(\gamma_i = 0 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})}{\widetilde{\Pr}(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})} = \frac{\mid (\mathbf{V}_{\gamma_i=0}^*)^{-1} \mid^{1/2} \mid \mathbf{V}_{\gamma_i=0}^{-1} \mid^{-1/2}}{\mid (\mathbf{V}_{\gamma_i=1}^*)^{-1} \mid^{1/2} \mid \mathbf{V}_{\gamma_i=1}^{-1} \mid^{-1/2}} \left[\frac{2b + \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_{\gamma_i=1}\mathbf{y}}{2b + \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_{\gamma_i=0}\mathbf{y}}\right]^{(n/2)+a}. \qquad (2.96)$$

For the component-wise partial order (2.97) must be decreasing in $\boldsymbol{\gamma}_{-i}$. Taking $\boldsymbol{\gamma}^{(1)}$ and assuming $\mathbf{V}$ is diagonal we obtain:

$$\frac{\widetilde{\Pr}(\gamma_i^{(1)} = 0 \mid \boldsymbol{\gamma}_{-i}^{(1)}, \mathbf{y}, \mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(1)} = 1 \mid \boldsymbol{\gamma}_{-i}^{(1)}, \mathbf{y}, \mathbf{X})} = \sqrt{\frac{v_{ii} + \mathbf{X}_i^T\mathbf{X}_i}{v_{ii}}} \left[1 - \frac{\mathbf{y}^T\mathbf{H}_i\mathbf{y}}{2b + \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_{\gamma_i=0}^{(1)}\mathbf{y}}\right]^{(n/2)+a}, \qquad (2.97)$$

and for $\boldsymbol{\gamma}^{(2)}$:

$$\frac{\widetilde{\Pr}(\gamma_i^{(2)} = 0 \mid \boldsymbol{\gamma}_{-i}^{(2)}, \mathbf{y}, \mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(2)} = 1 \mid \boldsymbol{\gamma}_{-i}^{(2)}, \mathbf{y}, \mathbf{X})} = \sqrt{\frac{v_{ii} + \mathbf{X}_i^T\mathbf{X}_i}{v_{ii}}} \left[1 - \frac{\mathbf{y}^T\mathbf{H}_i\mathbf{y}}{2b + \mathbf{y}^T\mathbf{y} - \mathbf{y}^\mathbf{T}\mathbf{H}_{\gamma_i=0}^{(2)}\mathbf{y}}\right]^{(n/2)+a}. \qquad (2.98)$$

We require (2.97) ≥ (2.98) and all values are the same except for $\mathbf{y}^T\mathbf{H}_{\gamma_i=0}\mathbf{y}$ where $\mathbf{y}^T\mathbf{H}_{\gamma_i=0}^{(2)}\mathbf{y} \geq \mathbf{y}^T\mathbf{H}_{\gamma_i=0}^{(1)}\mathbf{y}$ so indeed (2.97) ≥ (2.98).

Using Zellner's prior (2.19) let $f(\gamma) \propto 1$, $\tilde{\boldsymbol{\beta}}_\gamma = \hat{\boldsymbol{\beta}}_\gamma$, then $\mathbf{y}^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma = \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}$ and $\tilde{\boldsymbol{\beta}}_\gamma^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma = \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}$, reducing (2.20) to

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}, c) \propto (c+1)^{-\frac{q_\gamma+1}{2}} (\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma \mathbf{y})^{-n/2} \tag{2.99}$$

The Gibbs update probability (2.4) is

$$\frac{\widetilde{\Pr}(\gamma_i = 0 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})}{\widetilde{\Pr}(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})} = \sqrt{(c+1)}\left[\frac{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_{\gamma_i=1}\mathbf{y}}{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_{\gamma_i=0}\mathbf{y}}\right]^{n/2}. \tag{2.100}$$

The monotone Gibbs MC exists when (2.100) is decreasing in $\boldsymbol{\gamma}_{-i}$. For $\boldsymbol{\gamma}^{(1)}$ becomes:

$$\frac{\widetilde{\Pr}(\gamma_i^{(1)} = 0 \mid \boldsymbol{\gamma}_{-i}^{(1)}, \mathbf{y}, \mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(1)} = 1 \mid \boldsymbol{\gamma}_{-i}^{(1)}, \mathbf{y}, \mathbf{X})} = \sqrt{(c+1)}\left[1 - \frac{\mathbf{y}^T\mathbf{H}_i\mathbf{y}}{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_{\gamma_i=0}^{(1)}\mathbf{y}}\right]^{n/2}, \tag{2.101}$$

and similarly for $\boldsymbol{\gamma}^{(2)}$:

$$\frac{\widetilde{\Pr}(\gamma_i^{(2)} = 0 \mid \boldsymbol{\gamma}_{-i}^{(2)}, \mathbf{y}, \mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(2)} = 1 \mid \boldsymbol{\gamma}_{-i}^{(2)}, \mathbf{y}, \mathbf{X})} = \sqrt{(c+1)}\left[1 - \frac{\mathbf{y}^T\mathbf{H}_i\mathbf{y}}{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_{\gamma_i=0}^{(2)}\mathbf{y}}\right]^{n/2}. \tag{2.102}$$

All values are the same where $\mathbf{y}^T\mathbf{H}_{\gamma_i=0}^{(2)}\mathbf{y} \geq \mathbf{y}^T\mathbf{H}_{\gamma_i=0}^{(1)}\mathbf{y}$ so that (2.101) ≥ (2.102).

2.4.2 Fully empirical conjugate priors

We now move to investigate an example in the literature of fully EB priors for $\boldsymbol{\beta}_\gamma$ and σ^2, and then an alternative.

EXAMPLE 2.3 Empirically based priors.

Cripps *et al* (2006) propose empirically based priors for $\boldsymbol{\beta}$ and σ^2 for use in variable selection for the Bayesian linear regression model which depend on $\boldsymbol{\gamma}$. The prior they propose for $\boldsymbol{\beta}_\gamma$ conditional on σ^2 is the same as Zellner's prior with $\widetilde{\boldsymbol{\beta}}_\gamma = \hat{\boldsymbol{\beta}}_\gamma$. The prior for σ^2 is designed to be less informative compared to the marginal likelihood of σ^2, and to provide an unbiased estimate of the variance via the mode of an inverse gamma distribution. Cripps *et al* (2006) choose to keep the intercept in all models and use $1/\sqrt{n}$, and then centre the predictors so that

$$\mathbf{X}_\gamma^T \mathbf{X}_\gamma = \begin{bmatrix} 1 & 0 \\ 0 & \mathbf{A}_\gamma \end{bmatrix}. \tag{2.103}$$

Let the joint empirical prior for $\boldsymbol{\beta}_\gamma$ and σ^2 be

$$f(\boldsymbol{\beta}_\gamma \sigma^2 \mid \boldsymbol{\gamma}, \mathbf{X}) = \mathbf{N}_{q_\gamma+1}\left(\hat{\boldsymbol{\beta}}_\gamma, \sigma^2 \begin{bmatrix} c_1 & 0 \\ 0 & c_2(\mathbf{A}_\gamma)^{-1} \end{bmatrix}\right) \mathbf{IG}\left(\frac{\kappa}{2} - 1, \frac{\kappa(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma \mathbf{y})}{2(n - q_\gamma)}\right), \tag{2.104}$$

where κ, c_1 and c_2 are hyper-parameters. With $f(\boldsymbol{\gamma}) \propto 1$, the marginal posterior for $\boldsymbol{\gamma}$ is

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) \propto (c_1 + 1)^{-0.5} (c_2 + 1)^{-\frac{(q_\gamma - 1)}{2}} (n + \kappa - q_\gamma)^{-\frac{n+\kappa-2}{2}} \left(\frac{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}}{2(n - q_\gamma)}\right)^{-n/2}, \tag{2.105}$$

where Cripps *et al* (2006) use $c_1 = n^2$, $c_2 = n$, and $\kappa = 7$. We note $(n\text{-}q_\gamma)$ can be thought of as a degrees of freedom type term and that $f(\sigma^2)$ will have 2^k possible values of b. From (2.4) the ratio in the Gibbs sampler update probability follows as

$$\frac{\Pr(\gamma_i = 0 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})}{\Pr(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})} = \sqrt{(c_2 + 1)} \left(\frac{n + \kappa - q_{\gamma_i=1}}{n + \kappa - q_{\gamma_i=0}}\right)^{\frac{n+\kappa-2}{2}} \left(\frac{n - q_{\gamma_i=0}}{n - q_{\gamma_i=1}}\right)^{\frac{n}{2}} \left[\frac{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_{\gamma_i=1}\mathbf{y}}{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_{\gamma_i=0}\mathbf{y}}\right]^{n/2} \tag{2.106}$$

Let $\boldsymbol{\gamma}^{(1)} \preceq \boldsymbol{\gamma}^{(2)}$, then by theorem 2.1 we require:

$$\frac{\widetilde{\Pr}(\gamma_i^{(1)} = 0 \mid \boldsymbol{\gamma}_{-i}^{(1)}, \mathbf{y}, \mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(1)} = 1 \mid \boldsymbol{\gamma}_{-i}^{(1)}, \mathbf{y}, \mathbf{X})} = \sqrt{(c_2 + 1)} \left(\frac{n + \kappa - q_{\gamma_i=1}^{(1)}}{n + \kappa - q_{\gamma_i=0}^{(1)}}\right)^{\frac{n+\kappa-2}{2}} \left(\frac{n - q_{\gamma_i=0}^{(1)}}{n - q_{\gamma_i=1}^{(1)}}\right)^{\frac{n}{2}} \left[1 - \frac{\mathbf{y}^T\mathbf{H}_i\mathbf{y}}{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\,\mathbf{H}_{\gamma_i=0}^{(1)}\mathbf{y}}\right]^{n/2}, \tag{2.107}$$

to be greater than or equal to

$$\frac{\widetilde{\Pr}(\gamma_i^{(2)} = 0 \mid \boldsymbol{\gamma}_{-i}^{(2)}, \mathbf{y}, \mathbf{X})}{\widetilde{\Pr}(\gamma_i^{(2)} = 1 \mid \boldsymbol{\gamma}_{-i}^{(2)}, \mathbf{y}, \mathbf{X})} \tag{2.108}$$

$$=\sqrt{(c_2+1)}\left(\frac{n+\kappa-q^{(2)}_{\gamma_i=1}}{n+\kappa-q^{(2)}_{\gamma_i=0}}\right)^{\frac{n+\kappa-2}{2}}\left(\frac{n-q^{(2)}_{\gamma_i=0}}{n-q^{(2)}_{\gamma_i=1}}\right)^{\frac{n}{2}}\left[1-\frac{\mathbf{y}^T\mathbf{H}_i\mathbf{y}}{\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{H}^{(2)}_{\gamma_i=0}\mathbf{y}}\right]^{n/2}.$$

The very right hand term involving the residual sum of squares has already been shown to observe the component-wise partial order and $\sqrt{(c_2+1)}$ is a constant. Of the remaining terms, we require:

$$\left(\frac{n+\kappa-q_{\gamma_i=1}}{n+\kappa-q_{\gamma_i=0}}\right)^{\frac{n+\kappa-2}{2}} \qquad (2.109)$$

to be decreasing in $\boldsymbol{\gamma}_{-i}$, This term is decreasing in $\boldsymbol{\gamma}_{-i}$ as the numerator will always be smaller than the denominator noting the relation.

$$\left(\frac{n-q^{(2)}_{\gamma_i=0}}{n-q^{(2)}_{\gamma_i=1}}\right)^{\frac{n}{2}} \qquad (2.110)$$

This term is increasing in q_γ because the denominator is smaller than the numerator which violates the required component-wise partial order. This means we cannot determine monotonicity.

■

We now describe an analogue to the empirically based prior of Cripps *et al* (2006) using the MLE for σ^2. Following Cripps *et al* (2006) let the joint prior for $\boldsymbol{\beta}_\gamma$ and σ^2 be

$$f(\boldsymbol{\beta}_\gamma,\sigma^2\mid\boldsymbol{\gamma},\mathbf{X})=\mathbf{N}_{q_\gamma+1}(\hat{\boldsymbol{\beta}}_\gamma,c\sigma^2(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1})\ \mathbf{IG}\left(\frac{\kappa}{2}-1,\frac{\kappa(\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})}{2n}\right). \qquad (2.111)$$

The mode of this prior corresponds to the MLE for σ^2 and the posterior for σ^2 is

$$f(\sigma^2\mid\boldsymbol{\gamma},\mathbf{X},\mathbf{y})=\mathbf{IG}\left(\frac{n+\kappa-2}{2},\frac{(n+\kappa)(\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})}{2n}\right), \qquad (2.112)$$

which has the mode of the distribution equal to the MLE for σ^2. The posterior for $\boldsymbol{\gamma}$ with $f(\boldsymbol{\gamma}) \propto 1$ given by (2.2) is

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) \propto (c+1)^{-\frac{q_\gamma+1}{2}} (\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_{\boldsymbol{\gamma}}\mathbf{y})^{-\frac{n}{2}}. \tag{2.113}$$

Thus the posterior for $\boldsymbol{\gamma}$ is in fact the same as that for the Zellner's case with $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ and a non-informative prior on σ^2. Thus, it follows that the choice of priors (2.112) is sufficient for a monotone Gibbs MC with an orthogonal design matrix by the results above in (2.101) and (2.102).

2.4.3 Integration over *c*

Returning to the special case of $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ it turns out that the integration becomes:

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) \propto f(\boldsymbol{\gamma})(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_{\boldsymbol{\gamma}}\mathbf{y})^{-\frac{n}{2}} \int (c+1)^{-\frac{q_\gamma+1}{2}} f(c)dc\,. \tag{2.114}$$

From the priors in Table 2.1 with continuous support the easiest case to deal with is the hyper-G prior. The hyper-G-n and Zellner-Siow priors are much more difficult to integrate over *c* even in this simpler setting.

$$\int (c+1)^{-\frac{q_\gamma+1}{2}} (c+1)^{-\frac{a}{2}} dc = \frac{2}{q_\gamma + a - 1}. \tag{2.115}$$

This is straight forward to find as it is the reciprocal of the normalizing constant of the prior which is of the same form as the penalty term. Let $f(\boldsymbol{\gamma}) \propto 1$, then

$$\frac{\Pr(\gamma_i = 0 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})}{\Pr(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})} = \frac{(q_{\gamma_i=0} + a)}{(q_{\gamma_i=0} + a - 1)} \left[\frac{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_{\boldsymbol{\gamma}_i=1}\mathbf{y}}{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_{\boldsymbol{\gamma}_i=0}\mathbf{y}} \right]^{n/2}. \tag{2.116}$$

Thus to show monotonicity we only need to demonstrate that:

$$\frac{(q^{(1)}_{\gamma_i=0} + a)}{(q^{(1)}_{\gamma_i=0} + a - 1)} \geq \frac{(q^{(2)}_{\gamma_i=0} + a)}{(q^{(2)}_{\gamma_i=0} + a - 1)}. \tag{2.117}$$

This can be made clear be demonstrating that the ratio of the left-hand term over the right-hand term in (2.117) is ≥ 1. This is the case as by rearrangement we find:

$$\frac{(q^{(1)}_{\gamma_i=0}+a)(q^{(2)}_{\gamma_i=0}+a)-(q^{(1)}_{\gamma_i=0}+a)}{(q^{(1)}_{\gamma_i=0}+a)(q^{(2)}_{\gamma_i=0}+a)-(q^{(2)}_{\gamma_i=0}+a)} \geq 1\,. \tag{2.118}$$

All terms are equal except those subtracted and the larger term is subtracted in the denominator making the ratio > 1. Note that this result generalizes to any choice of *f*(γ) that allows monotonicity. To complete the approach we must be able to sample from $f(c\,|\,\boldsymbol{\gamma},\mathbf{y},\mathbf{X})$ and in this case:

$$f(c\,|\,\boldsymbol{\gamma},\mathbf{y},\mathbf{X}) = \frac{q_\gamma + a - 1}{2}(1+c)^{-(q_\gamma+a+1)/2}\,. \tag{2.119}$$

Finding the CDF:

$$F(c) = \frac{q_\gamma + a - 1}{2}\int_0^c (1+t)^{-(q_\gamma+a+1)/2}dt = \left[1-\left(\frac{1}{1+c}\right)^{(q_\gamma+a-1)/2}\right], \tag{2.120}$$

the inverse CDF is

$$F^{-1}(u) = (1-u)^{-2/(q_\gamma+a-1)} - 1\,, \tag{2.121}$$

so we may use the inverse CDF method to generate random variables from the posterior of *c* after generating exact samples of γ using a monotone Gibbs MC. Typically integrating over *c* will help remove the choice of this parameter from the posterior for γ which in turn helps with variable selection. As variable selection is not a valid option for inference in this context, integration over *c* in the posterior for γ is only useful if we can do the same to other posterior distributions such as the posterior predictive distribution.

2.5 Summary

In this chapter we have explored a number of possible choices of priors and hyper-priors and the associated hyper-parameters and hyper hyper-parameters. Figure 2.1 and Tables 2.2 and 2.3 summarize these findings. To ensure clarity in the following discussion the prior for σ^2 is IG(α,β) so a can be used for the prior on c, and the prior for τ is the Beta(r,s).

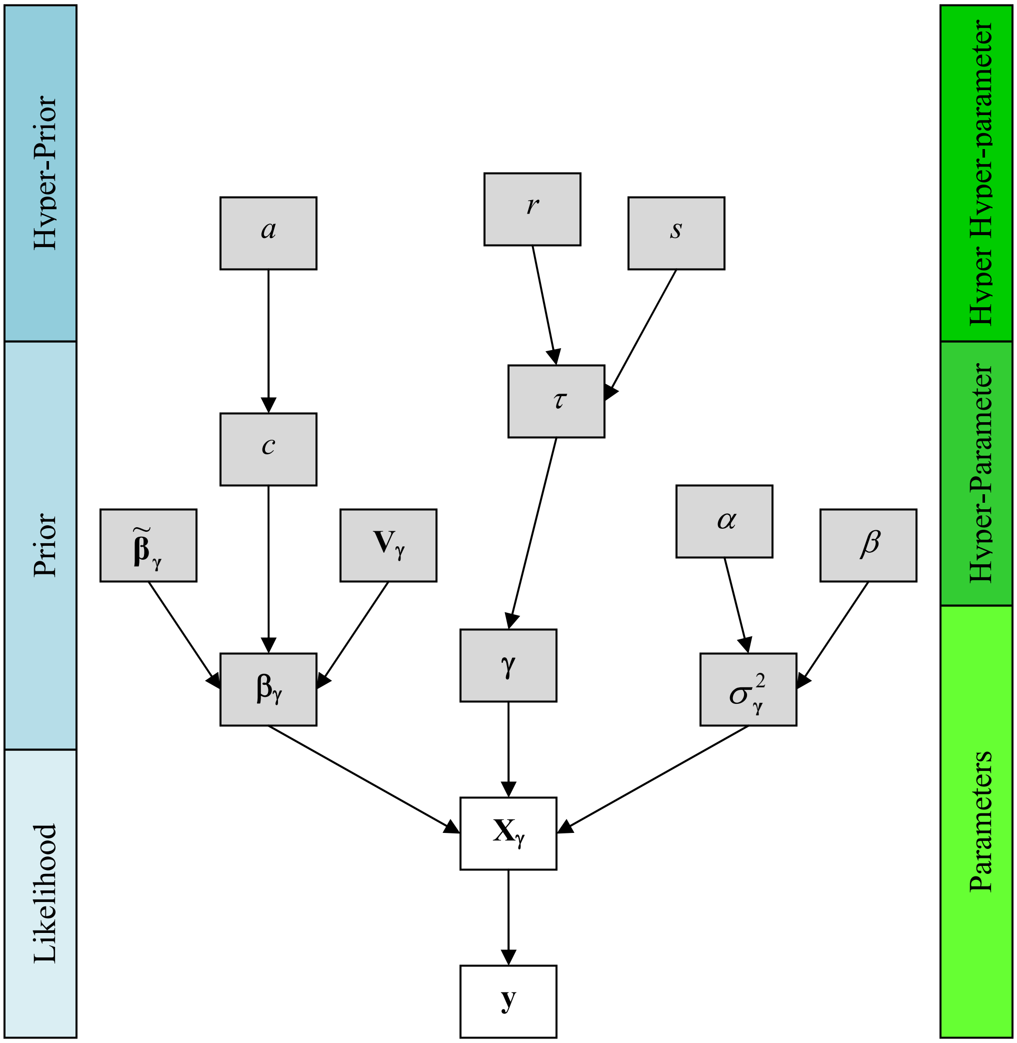


Figure 2.1 Diagram showing the paths of investigation for the sufficient conditions of a monotone Gibbs MC in the Bayesian normal linear regression model.

Assuming Gaussian errors and using the conjugate class of priors (including Zellner's prior) in a fully Bayes approach we must set $\tilde{\boldsymbol{\beta}} = 0$ to obtain a monotone Gibbs MC. This includes the Zellner's projection prior in example 2.1. We may choose $\tilde{\boldsymbol{\beta}}$ to be $\hat{\boldsymbol{\beta}}_\gamma$ as part of an empirical Bayes approach. In example 2.2 an extension using Zellner's prior for outlier detection does not permit a monotone Gibbs MC for posterior model probabilities or the posterior probabilities for outliers.

For the conjugate prior for σ^2 there is no justification for allowing the hyper-parameters to depend on $\boldsymbol{\gamma}$ in a fully Bayes approach. Any choice of α and β in a fully Bayes approach independent of $\boldsymbol{\gamma}$ will provide a monotone Gibbs MC with an orthogonal design matrix and an appropriate choice of prior for $\boldsymbol{\beta}$. In an EB approach it makes sense to allow α and β to depend on $\boldsymbol{\gamma}$ as we are using the data to estimate parameters. In example 2.3 we demonstrate that using the classical estimate of regression variance does not allow a monotone Gibbs MC, while using the MLE does. It should be noted that both these cases deliberately avoid having α depend upon $\boldsymbol{\gamma}$. This does not strictly have to be the case however, choosing α to depend on $\boldsymbol{\gamma}$ may prevent simplification of the ratio in the Gibbs sampler for confirming the sufficiency of monotonicity.

Table 2.2 Summary of Hyper-parameter conditions for the Conjugate Family of priors for monotonicity

	Description	**Condition(s)**
$\mathbf{V}_{\gamma}$	*The covariance matrix in the conjugate prior for $\boldsymbol{\beta}_{\gamma}$. Can be replaced with $\mathbf{X}^T\mathbf{X}$.*	Diagonal, positive definite.
$\tilde{\boldsymbol{\beta}}_{\gamma}$	*Prior estimates of the regression coefficients in the prior for $\boldsymbol{\beta}_{\gamma}$.*	0 (fully Bayes) or $\hat{\boldsymbol{\beta}}_{\gamma}$ (EB).
c	*A scale parameter in the variance term for the conjugate prior for $\boldsymbol{\beta}_{\gamma}$.*	Independent of $\boldsymbol{\gamma}$.
α	*The shape parameter in the conjugate prior for σ^2.*	Independent of $\boldsymbol{\gamma}$.
β	*The scale parameter in the conjugate prior for σ^2.*	Independent of $\boldsymbol{\gamma}$ or a function of $\mathbf{y}^T(\mathbf{I}_n - \mathbf{H}_{\gamma})\mathbf{y}$ (EB).

In a fully Bayes approach integration over c is possible however demonstrating monotonicity is far from straight forward. For an EB approach the integration over c is greatly simplified and for the case of the hyper-G prior we demonstrate monotonicity of the posterior for any choice of $a > 2$. We also show how the inverse CDF method may be used to generate samples from the posterior for c conditional on $\boldsymbol{\gamma}$. The adjusted Jeffreys prior we suggest with no hyper-parameters will always provide a monotone Gibbs MC provided the design matrix is orthogonal.

Priors for $\boldsymbol{\gamma}$ fall outside the rest of the investigation and we found that the Bernoulli class will give a monotone Gibbs MC. Further, integration over τ for the constant Bernoulli prior using the conjugate beta distribution will provide a monotone Gibbs MC for any choice of $r > 0$ and $s > 0$.

Table 2.3 Summary of Priors for γ and monotonicity for an orthogonal X.

Priors	**Conjugate***	**Zellner***	**Jeffrey**
Uniform	Yes (3)	Yes (1)	Yes (0)
Bernoulli	Yes ($k + 3$)	Yes ($k + 1$)	Yes (k)
Constant Bernoulli	Yes (4)	Yes (2)	Yes(1)
Beta-Bernoulli	Yes (5)	Yes (3)	Yes (2)
T-Poisson	No (4)	No (2)	No (1)
f*(q_γ)*	Yes( )	Yes( )	Yes( )

* Given the conditions for hyper-parameters in Table 2.2

** Provided the condition that $f(q_\gamma)$ is decreasing as q_γ increases.

( ) The number listed in brackets indicates the number of hyper-parameters required. For the conjugate and Zellner's prior we assume $\tilde{\boldsymbol{\beta}} = 0$ or $\hat{\boldsymbol{\beta}}_\gamma$, further, for the conjugate case we assume $\mathbf{V}$ as a single hyper-parameter. T-Poisson is the truncated Poisson. In the case of $f(q_\gamma)$ ( ) is necessarily left empty.

This of course requires that the posterior with a flat prior for $\boldsymbol{\gamma}$ is monotone to begin with. Any general prior for q_γ will permit a monotone Gibbs MC provided the prior probability is decreasing with increasing model size, this includes the example of the truncated Poisson prior.

From this we can tentatively make the following recommendations. In the conjugate class Zellner's seems the reasonable choice as it greatly reduces the number of hyper-parameters required by using the empirical covariance matrix, and we require $\widetilde{\boldsymbol{\beta}}$ = 0 and some choice of c. We may prefer Zellner's prior over Jeffreys prior when the likelihood is weak (small ratio of n:k), and may also depend on whether inference is more sensitive to c or $f(\gamma)$. When the sample size is large the adjusted version of Jeffreys prior should be a suitable choice given no requirement for hyper-parameter specification.

In an EB setting the suggested MLE based priors, or some variation therein, from example 2.3 are necessary to provide a monotone Gibbs MC. It should be noted that the posterior for γ under this setting is essentially the same as that for the adjusted Jeffreys prior. The Bernoulli class of priors for γ retain monotonicity, while non-flat priors for γ may only preserve and not induce monotonicity. Priors specified on q_γ can also produce monotone Gibbs MC provided the probability is decreasing in q_γ, an example of this is the truncated Poisson prior. Through-out this investigation it also became apparent that there does not appear to be any way to use a prior on γ to induce monotonicity when the posterior using a flat prior is not.

Future extensions to this work include investigating the Gibbs sampler constructed using the hyper-geometric function as in Liang *et al* (2008) , and using auxiliary variables to create a Gibbs sampler for γ when using an error distribution that is not Gaussian. The work in this chapter is an addition to current knowledge showing that an orthogonal predictor matrix is not the single and only sufficient requirement, for a monotone Gibbs MC. Finally, the results of this chapter also apply to wavelets applications, and any univariate non-linear regression where the series can be decomposed into a collection of orthogonal basis functions.

CHAPTER 3

ORTHOGONALITY

"*The goal is to transform data into information, and information into insight.*"

- Carly Fiorina, Hewlett Packard, 1999 - 2005

Having investigated and established a number of sufficient conditions for the construction of a monotone Gibbs Markov chain in the Chapter 2, we now provide some practical considerations of monotonicity and the inferential problems applicable when using **W**. Explaining **y** requires variable selection (**X** and γ) and determining the effect the predictors have on **y** through $\boldsymbol{\beta}_{\gamma}$. It is already well documented that linear regression using an orthogonal predictor matrix does not permit variable selection. This suggests our interest in γ should be for use in BMA. Thus, the focus for investigating the impact of orthogonalization is for the predictive modeling of **y** using BMA. This can also include tasks such as outlier detection. The issue of how well modeling **y** using **W** compares with that of **X** requires investigation.

3.1 Gibbs Update Probability

Recall the Gibbs update probability:

$$\Pr(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X}) = \left[1 + \frac{\widetilde{\Pr}(\gamma_i = 0 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})}{\widetilde{\Pr}(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})}\right]^{-1}, \tag{3.1}$$

where $\widetilde{\Pr}(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X}) \propto f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$, the un-normalized posterior. The analysis of a signal using wavelets is an example where monotone Gibbs CFTP has been used (Holmes and Denison, 2002). The signal is decomposed into a series of orthogonal or orthonormal basis functions, so the Gibbs MC is monotone. Perfect sampling is then used for model averaging and choosing a subset of basis functions for best explaining the observed signal.

The Gibbs coupler, presented by Huang and Djuric (2002), is an elegant perfect sampling method using support set coupling for model selection. This approach does not require **X** to be orthogonal (the actual Gibbs MC is not monotone), instead requiring $\boldsymbol{\beta}$ and σ^2 be known. Bounding chains on the support set $\{0,1\}$ are constructed for sequential updating of the support of γ_i to generate samples from $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$. When $\boldsymbol{\beta}$ and σ^2 are unknown, these bounding chains do not exist, preventing any extension of the Gibbs coupler to this setting.

Using simulated data where the predictor matrix **X** contains a correlation structure, and **W** is an orthogonalized version of **X**, we demonstrate the monotonicity of the Gibbs MC numerically for $k = 4$. Under Zellner's prior with $c = n$, we obtain the update probabilities $\Pr(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})$ in Table 3.1 for **X** and **W**. Table 3.2 contains the required component-wise partial orderings. With some inspection, it is apparent that the update probabilities for **W** satisfy Theorem 2.1, while those for **X** do not.

The consequences of the partial ordering have a natural interpretation in the context of linear regression. The partial ordering reflects the nested model structure shown in Figure 3.1, where the probability of adding a variable to a model is greater than adding that same variable to any of its sub-models. Other orderings might include a complete ordering through the decimal representation of $\boldsymbol{\gamma}$ or a partial ordering through q_γ. However, the decimal ordering does not translate into a natural model nesting structure and is therefore less useful. The ordering on q_γ also has a similar limitation.

Table 3.1 Gibbs Update Probabilities $\Pr(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X})$ for Fictitious Data, *k* = 4

$\gamma_{-i} \backslash i$			1		2		3		4	
0	**0**	**0**	0.3811	0.4423	0.4354	0.2115	0.2920	0.2556	0.4389	0.2448
0	**0**	**1**	0.3138	0.4512	0.2125	0.2116	0.2991	0.2571	0.4473	0.2462
0	**1**	**0**	0.2902	0.4539	0.4076	0.2116	0.2690	0.2556	0.2149	0.2448
0	**1**	**1**	0.2256	0.4636	0.2228	0.2116	0.3119	0.2571	0.2521	0.2463
1	**0**	**0**	0.3195	0.4425	0.3701	0.2116	0.2149	0.2646	0.3674	0.2515
1	**0**	**1**	0.3159	0.4514	0.2140	0.2117	0.2137	0.2667	0.3657	0.2535
1	**1**	**0**	0.2555	0.4541	0.3662	0.2117	0.2120	0.2646	0.2120	0.2515
1	**1**	**1**	0.2139	0.4638	0.2112	0.2117	0.2109	0.2668	0.2109	0.2536

□ Original Design Matrix (**X**) ▨ Orthogonal Design Matrix (**W**)

Table 3.2 Component-wise Partial Orderings for *k* = 4

dec(γ_{-i})	γ_{-i}	Consequence of partial ordering for Comparable States
0	[0 0 0]	≤ { [0 0 1], [0 1 0], [1 0 0], [0 1 1], [1 0 1], [1 1 0], [1 1 1] }
1	[0 0 1]	≤ { [0 1 1], [1 0 1], [1 1 1] }
2	[0 1 0]	≤ { [0 1 1], [1 1 0], [1 1 1] }
3	[0 1 1]	≤ { [1 1 1] }
4	[1 0 0]	≤ { [1 0 1], [1 1 0], [1 1 1] }
5	[1 0 1]	≤ { [1 1 1] }
6	[1 1 0]	≤ { [1 1 1] }
7	[1 1 1]	≥ { [0 0 0], [0 0 1], [0 1 0], [1 0 0], [0 1 1], [1 0 1], [1 1 0]}

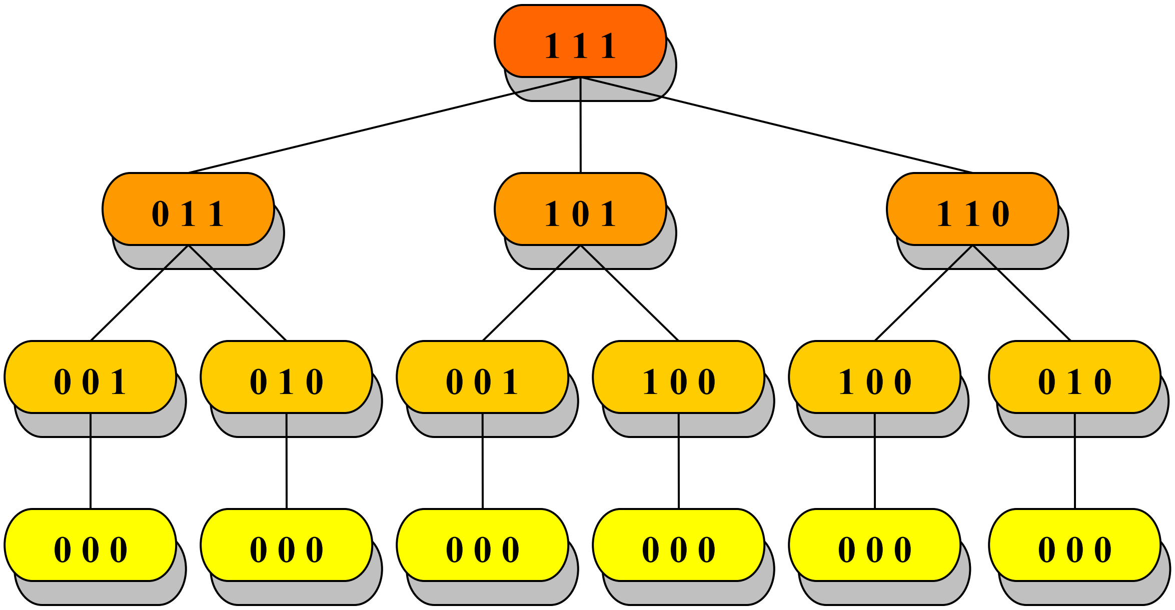


Figure 3.1 The nested model structure for model comparison in linear regression, and it's relation to the component-wise partial ordering of a monotone Gibbs Markov chain.

3.2 Orthogonality and $f(\gamma \mid y, X)$.

In practice, orthogonalization of **X** may arise for two reasons. The first is severe multi-colinearity where **X** contains strongly correlated predictors. This can result in poor numerical conditioning when calculating the inverse covariance matrix. An alternative is to remove variables that are strongly correlated with other variables, and then use a reduced design matrix. The second reason for orthogonalization is to reduce computation time. An orthogonal design matrix allows faster computation of the residual sum of squares. Any transformation should ideally retain as much of the correlation structure with the response as possible.

3.2.1 Transformation Methods.

We now detail the methods that we use to generate an orthogonal design matrix **W**. For fair comparisons between **X** and **W**, we construct **W** to have a constant in the first column for the intercept term. This is done by centering each of the k predictors in **X** by subtracting the column mean to obtain the centered version $\mathbf{X}_0$. An appropriate transform of $\mathbf{X}_0$ is used to create the orthogonal version $\mathbf{W}_0$. Each column is then divided by its inner product so that it has unit length. Finally, a column with the constant $1/\sqrt{n}$ is inserted to represent the intercept. This ensures no correlation with the other columns of $\mathbf{W}_0$ and so provides the orthonormal predictor matrix $\mathbf{W}$.

To summarize:

1. Take the predictor matrix with no intercept $\mathbf{X}$, and for $i = \{1,..,k\}$ compute $(\mathbf{X}_0)_i = \mathbf{X}_i - (1/n)\sum_{j=1}^{n}\mathbf{X}_{i,j}$. $\mathbf{X}_i$ is the i-th column of $\mathbf{X}$.
2. Use an appropriate transformation of the form: $\mathbf{W}_0 = \mathbf{X}_0\mathbf{A}$.
3. For $i = \{1,..,k\}$ compute $\mathbf{W}_i = (\mathbf{W}_0)_i / \sqrt{(\mathbf{W}_0)_i^T(\mathbf{W}_0)_i}$ and add a column of $1/\sqrt{n}$ for the intercept. $(\mathbf{W}_0)_i$ is the i-th column of $\mathbf{W_0}$, and similarly for $\mathbf{W}_i$.

We detail four methods for transforming $\mathbf{X}_0$ into $\mathbf{W}_0$. The first two of these methods are based on eigenvalue decompositions. These are generalized principal components (GPC) and the Lowdin transformation, an extension of singular value decomposition (SVD). The other two methods use the modified Gram-Schmidt (GS) procedure with different initial orderings of $\mathbf{X}_0$. For the SVD and GPC methods, $\mathbf{W}$ is invariant to re-ordering of $\mathbf{X}$. However, because the GS procedure is sequential $\mathbf{W}$ is not invariant to a re-ordering of $\mathbf{X}$.

1. General Principal Components (GPC)

Clyde *et al* (1996), and Holmes and Mallick (1998), use GPC to orthogonalize $\mathbf{X}$. We use the approach as described by Clyde *et al* for transforming $\mathbf{X}_0$. Let $\mathbf{U}$ be a matrix of eigenvectors, $\mathbf{\Lambda}$ be a diagonal matrix of eigenvalues and $\mathbf{D}$ be the diagonal of $\mathbf{X}_0^T\mathbf{X}_0$. The transformation is then:

$$\begin{aligned} \mathbf{R} &= \mathbf{D}^{-1/2}(\mathbf{X}_0^T\mathbf{X}_0)\mathbf{D}^{-1/2} \\ \mathbf{R} &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \\ \mathbf{W}_0 &= \mathbf{X}_0(\mathbf{D}^{-1/2}\mathbf{U}) \end{aligned} \tag{3.2}$$

Thus, $\mathbf{A}$ is $\mathbf{D}^{-1/2}\mathbf{U}$, $\mathbf{R}$ is a k x k matrix and $\mathbf{W}_0$ represents the principal components formed from $\mathbf{X}_0$. There can be directions that correspond to small eigenvalues which have a high degree of correlation with **y**, so choosing a subset based on the size of the eigenvalues can be misleading. However, it is important to note that because of model averaging this is not a concern.

2. Lowdin Transformation

The Lowdin transformation is an extension to the SVD. The Lowdin transformation is designed to minimize the distance (Frobenius norm) between the matrix **X** and the orthogonal version. The Frobenius norm is defined as the $\sqrt{(trace(\mathbf{X}^T\mathbf{X}))}$ which is a special case of the 2-norm for matrices. SVD is readily implemented by the `orth`(**X**) function in Matlab or the `svd`(**X**) function in R, and is:

$$\mathbf{X}_0 = \mathbf{USV}^T \tag{3.3}$$

where **U** is an *n* x *k* matrix with orthonormal columns, **S** is *k* x *k* diagonal matrix of singular values of $\mathbf{X}_0$, and **V** is *k* x *k* orthogonal matrix of right singular values of $\mathbf{X}_0$. **U** can serve as an orthogonal version of $\mathbf{X}_0$ with $\mathbf{A} = (\mathbf{SV}^T)^{-1}$, however because of the property of the Lowdin transformation if we are to use the SVD procedure it makes sense to use the Lowdin approach. The Lowdin transformation will construct:

$$\mathbf{W}_0 = \mathbf{UV}^T\,. \tag{3.4}$$

Which by the substitution of $\mathbf{U} = \mathbf{X}_0(\mathbf{SV}^T)^{-1}$ can be expressed as $\mathbf{W}_0 = \mathbf{X}_0(\mathbf{SV}^T)^{-1}\mathbf{V}^T$ so $\mathbf{A} = (\mathbf{SV}^T)^{-1}\mathbf{V}^T$. For further details see (Beaver, 2007).

3. The Modified Gram-Schmidt (GS) Procedure

For the centered predictor matrix $\mathbf{X}_0$ whose columns are in a prearranged order, let $\mathbf{x}_i$ denote the i^{th} column of $\mathbf{X}_0$, $\mathbf{w}_i$ be the i^{th} column of $\mathbf{W}_0$, and $\mathbf{w}_1 = \mathbf{x}_1 / \| \mathbf{x}_1 \|$ where $\| \mathbf{x}_1 \| = \sqrt{\mathbf{x}_1^T \mathbf{x}_1}$. The modified GS orthogonalization then proceeds as:

$$
\begin{aligned}
\mathbf{w}_2 &= \mathbf{x}_2 - \frac{\mathbf{x}_2 \cdot \mathbf{w}_1}{\| \mathbf{w}_1 \|^2}\mathbf{w}_1 \\
\mathbf{w}_3 &= \mathbf{x}_3 - \frac{\mathbf{x}_3 \cdot \mathbf{w}_1}{\| \mathbf{w}_1 \|^2}\mathbf{w}_1 - \frac{\mathbf{x}_3 \cdot \mathbf{w}_2}{\| \mathbf{w}_2 \|^2}\mathbf{w}_2 \\
\vdots &= \vdots \\
\mathbf{w}_k &= \mathbf{x}_k - \frac{\mathbf{x}_k \cdot \mathbf{w}_1}{\| \mathbf{w}_1 \|^2}\mathbf{w}_1 - \cdots - \frac{\mathbf{x}_k \cdot \mathbf{w}_{k-1}}{\| \mathbf{w}_{k-1} \|^2}\mathbf{w}_{k-1}
\end{aligned}
\tag{3.5}
$$

The modified approach is designed to reduce the numerical instability that can occur with the standard GS procedure. GS orthogonalization sequentially replaces each column in $\mathbf{X}_0$ with a rescaled version of the residuals resulting from the regression of that column on the preceding columns. Under the GS approach the **A** matrix is upper triangular. This may be useful because we can retain a rescaled version of the best explanatory variable from **X**. The main drawback to the GS method as highlighted by Clyde *et al* (1996) and Holmes and Mallick (1998), is the requirement to order the columns of **X** prior to transformation. Clyde *et al* (1996) and Holmes and Mallick (1998) do not discuss any methods for ordering the predictors, and so there is little analysis on the use of the GS method. We now describe the two methods we will use to order the columns of $\mathbf{X}_0$ prior to orthogonalization.

Method 1: Create $\mathbf{X}_0$ with the columns ordered in descending magnitude of correlation with **y**. The method will be abbreviated to GS_1.

Method 2: Order the predictors based on the magnitude of the correlation with **y** but also take into account the correlation structure of the resulting $\mathbf{X}_0$. In particular, we set the first predictor $\mathbf{x}_1$ as the one with the strongest correlation with **y**, and then choose the next predictor as the one that is both strongly correlated with **y** and weakly correlated with $\mathbf{x}_1$. This is repeated sequentially until all predictors are ordered. Suppose the first *j* predictors have been chosen and we must now choose the next one from the remaining predictors whose indices are contained in the set V. For $i \in \text{V}$, let $r(\mathbf{x}_i,\mathbf{y})$ and $r(\mathbf{x}_i,\mathbf{x}_j)$ be the correlations between one of these remaining predictors ($\mathbf{x}_i$) with **y** and with the *j*-th ordered predictor $\mathbf{x}_j$, respectively. Then the next ordered predictor is chosen as

$$\mathbf{x}_{j+1} = \underset{\{\mathbf{x}_i\}_{i\in V}}{\arg\min} \sqrt{|\, r(\mathbf{x}_i, \mathbf{x}_j)\,| + (1 - |\, r(\mathbf{x}_i, \mathbf{y}_i)\,|)}\ . \tag{3.6}$$

Thus, the next column in the ordered $\mathbf{X}_0$ is chosen as the predictor with the minimum distance (d) to the co-ordinates (0,0) using (3.6). This is repeated sequentially until all predictors are ordered.

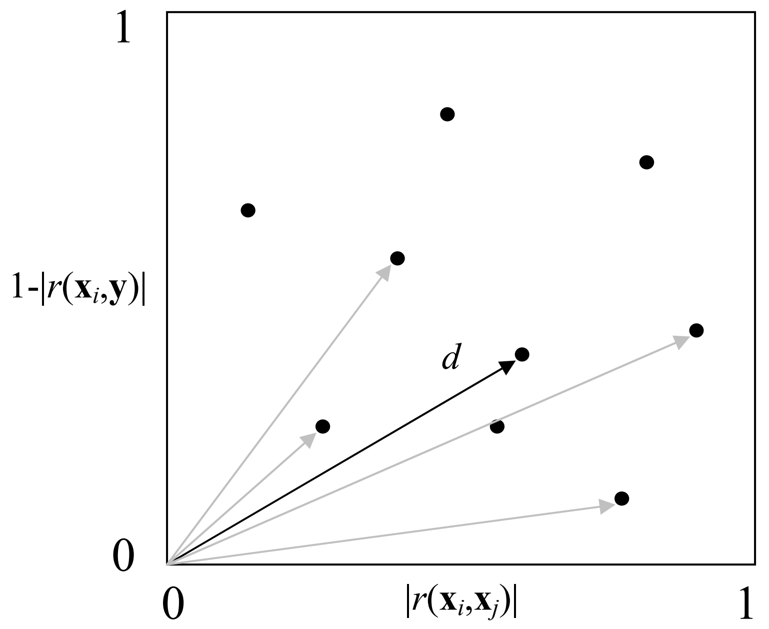

1
1-|r($\mathbf{x}_i$,$\mathbf{y}$)|
d
0
0
|r($\mathbf{x}_i$,$\mathbf{x}_j$)|
1

Figure 3.2 A graphical representation of Method 2 for ordering the predictors prior to orthogonalization by the Gram-Schmidt method.

This method will be abbreviated to GS_2. Both methods of ordering are justified in a heuristic sense, and stop short of the partial least squares (PLS) method detailed in Clyde *et al* (1996).

Two methods we do not review as discussed by Clyde *et al* (1996) are PLS and sliced inverse regression. Partial least squares uses an eigenvalue decomposition of the covariance matrix and then using $\mathbf{y}$, sequentially adjusts the columns of $\mathbf{X}$ to produce an orthogonal predictor matrix. Sliced inverse regression uses an eigenvalue decomposition of a weighted covariance matrix, created by dividing $\mathbf{y}$ into h slices and calculating a matrix of means. The resulting eigenvectors are then multiplied by the standardized version of **X** to create **W**. For studies to follow in this chapter the four methods GPC, Lowdin, GS_1 and GS_2 will be investigated.

3.2.2 Posteriors and Point Estimates.

While a number of orthogonal transformation methods may be very similar, they can lead to different posteriors for γ. The posteriors of **X** and **W** are not directly comparable due to the loss of interpretation for the predictors. The original posterior can be preserved by orthogonalizing for every γ however, this will not allow monotonicity. If we are concerned with inference about **y** then interpretability in the orthogonal space is not an issue. Further, if inference about **y** is not compromised by using **W** instead of **X**, then orthogonalization is an approach that will allow efficient perfect sampling. Thus, we may compare **X** and **W** based on the fitted response. To give an appreciation of these points we use the ozone data as an example. Details of the ozone dataset may be found in Appendix C.

Example: Ozone Data

In Figure 3.3 we plot the posterior distributions for γ using **X** and the four **W** methods for Zellner's prior with $c = n$ and $\tau = 0.5$. Figure 3.4 shows the BMA fitted **y** values for **X** and the **W** methods, and the true values for observations 20 to 40.

Not only are the posteriors of **X** and **W** not directly comparable, but because the **A** matrices in the transformation methods result in different mixing of **X** into the columns of **W**, the posteriors for **W** are also incomparable with one another. Thus, the distribution of mass can vary noticeably between different methods of orthogonalization. The posterior mass for the orthogonalization methods are less dispersed than for **X**, with noticeably less small probability models. In the case of the orthogonalization methods, the posterior mass appears the most concentrated for the GS_2 method, while the Lowdin method has two very similar separated patterns. The GPC and GS_1 methods occupy the upper half of the state space with different distributions of mass. It is clear that while the posteriors for **W** can vary noticeably in distribution from both **X** and other **W** methods, all provide suitable weights for obtaining a BMA fitted response, which from Figure 3.4 is comparable between all methods.

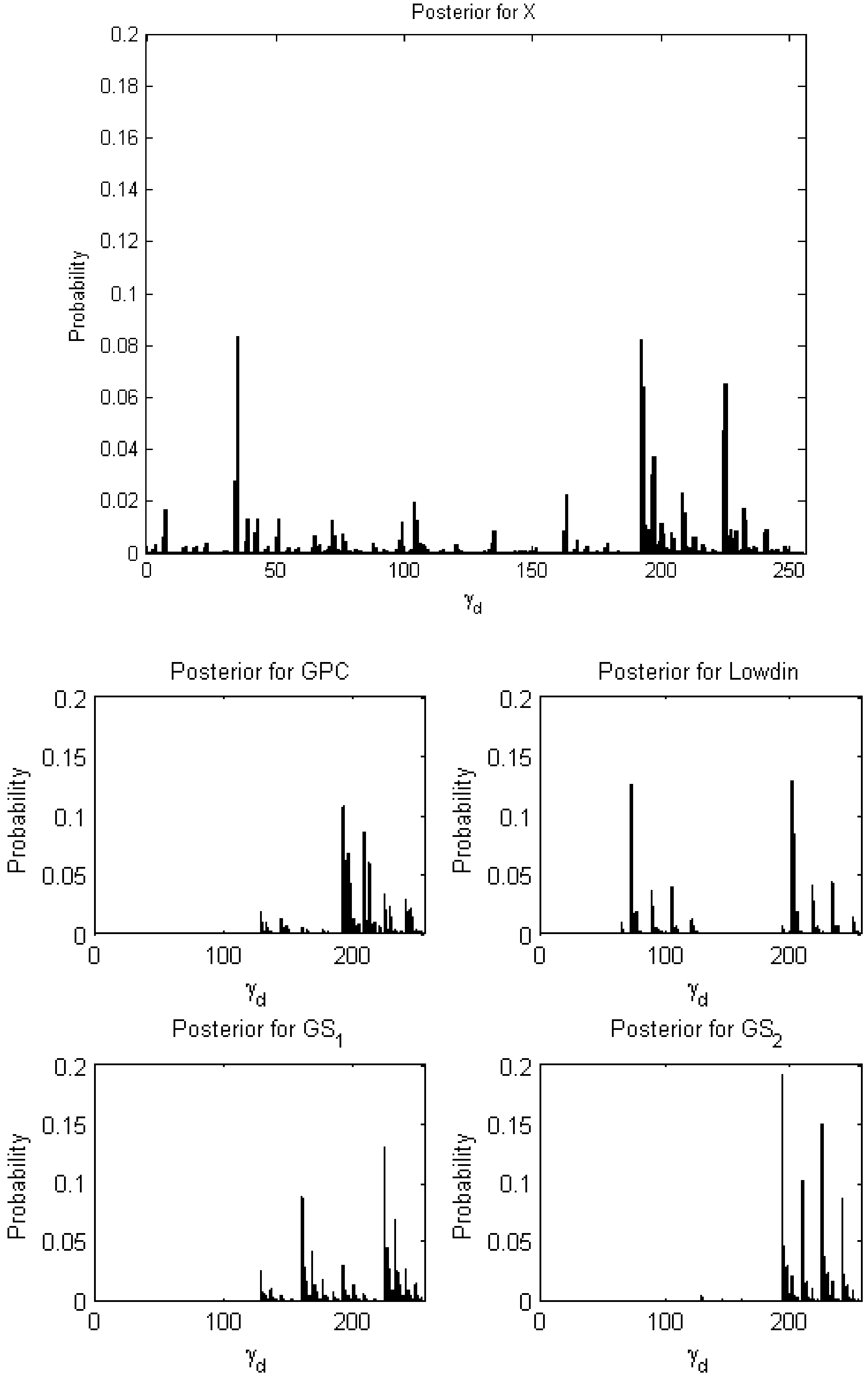


Figure 3.3 Posterior model probabilities using X and W for the ozone data, for Zellner's prior with $c = n$, a uniform prior for γ.

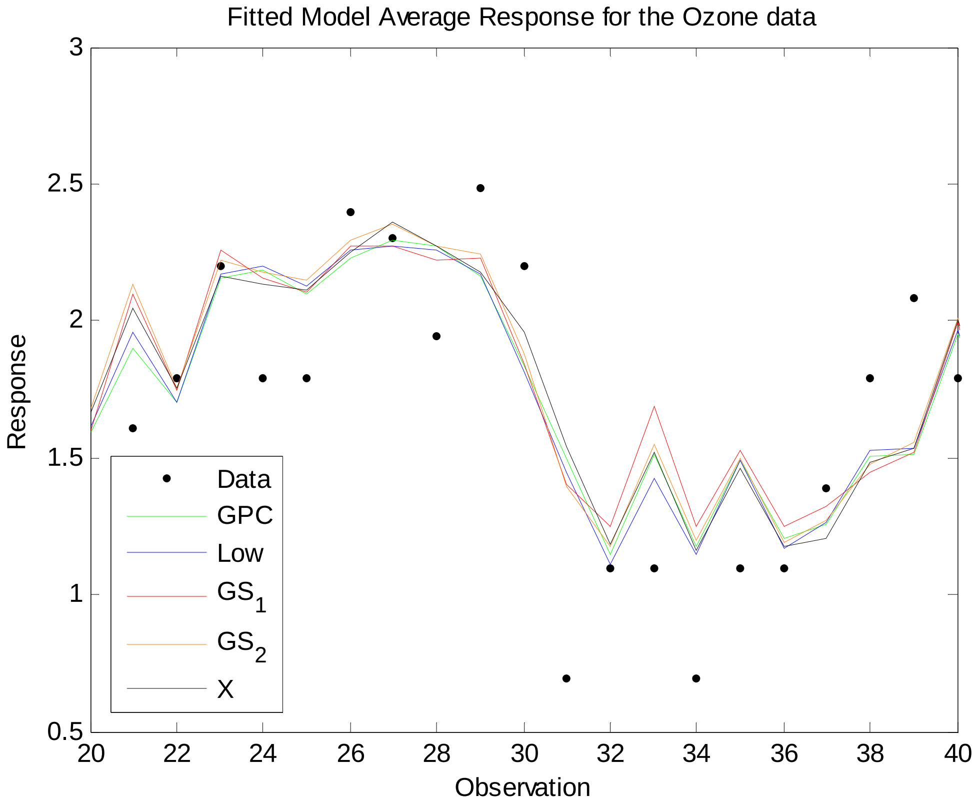


Figure 3.4 Plot of the fitted response using BMA with X and W for the ozone data. As above we use Zellner's prior with $c = n$, a uniform prior for γ.

As mentioned, previous work has indicated proper BVS is not possible with an orthogonal predictor matrix. However, it would be desirable to determine which of the predictors in the **X** space are important while working in the **W** space, by using a statistic to rank them. Since the transformation from $\mathbf{X}_0$ to $\mathbf{W}_0$ can be expressed as $\mathbf{W}_0 = \mathbf{X}_0\mathbf{A}$ for some k x k matrix **A**, we should be able to use $\mathbf{A}^{-1}$ to get back to **X** for inference on the relative importance of the predictors. Heuristically, some function of the MIP (Table 3.4) and $\mathbf{A}^{-1}$ should provide an estimate of the MIP for **X**. Table 3.3 shows the mixing coefficients (the **A** matrix) for all four methods for the ozone data. The columns represent the column of **W** and the numbered rows the columns of **X**. For example the first column for GPC indicates that the first column in **W** was dominated by the second column of **X**. For each of these columns we divided the values obtained by the largest absolute value so that all values fall between -1 and 1, the dominant predictor from **X** is indicated by -1 or 1.

Table 3.3 Mixing Coefficients for transforming $\mathbf{X}_0$ for the ozone data into $\mathbf{W}_0$ using the four methods of orthogonalization.

	X	Column of W							
		1	2	3	4	5	6	7	8
GPC	1	0	0.00054	-3.3E-05	-0.00045	0.006155	1	-0.55186	0.01615
	2	-1	0.025578	-0.02021	0.000153	-0.00299	0.003637	0.003279	-0.0005
	3	0	-0.02099	-0.01575	-1	0.3957	0.030988	0.061365	-0.00487
	4	0	0.00394	1	-0.11524	-0.00765	0.003127	-0.00022	-0.00218
	5	0	1	-0.00854	-0.20521	-0.03382	-0.10692	-0.01171	0.009055
	6	0	-0.00415	-0.00254	-0.13697	-1	0.055275	0.018478	-0.00445
	7	0	0.0008	0.000104	0.007678	0.008755	0.867895	1	-0.06572
	8	0	0	0	0	0	0.001062	0.00267	1
Low	1	1	-0.11047	-0.12068	0.027687	-0.6442	0.020333	-0.21418	0.008471
	2	-0.0006	0.532464	-0.00087	-0.01863	0.017826	0.003093	0.003514	-0.00042
	3	-0.0272	-0.03571	1	0.200588	0.298821	-0.24051	0.051891	-0.00352
	4	0.00190	-0.23114	0.060972	1	0.038669	0.017344	0.000221	-0.00212
	5	-0.0454	0.227672	0.093479	0.039796	1	0.041833	-0.03358	0.008304
	6	0.00774	0.213617	-0.40681	0.096513	0.226192	1	0.023839	-0.00363
	7	-0.3359	1	0.361656	0.005074	-0.74819	0.098226	1	-0.04203
	8	0.02551	-0.23028	-0.04711	-0.09313	0.3552	-0.02874	-0.08069	1
GS$_1$	1	0	0	-1	0.37826	-0.01026	-0.45081	0.270197	-0.41157
	2	1	0.004954	0.002244	0.002397	0.000455	0.008922	-0.00529	0.002204
	3	0	0	0	0	0	0	0	0.441922
	4	0	0	0	0	0	-0.31715	-0.02533	0.023198
	5	0	0	0	-0.13211	-0.00965	-0.01908	-0.02115	0.009704
	6	0	0	0	0	0	0	-0.63435	-0.34283
	7	0	1	0.892332	1	0.046404	0.91477	-1	1
	8	0	0	0	0	-1	1	0.410961	-0.28877
GS$_2$	1	0	1	0.054713	-0.73377	-0.52747	-0.27642	-0.35235	-0.41157
	2	1	0.002177	-0.00038	-0.00307	0.003917	0.004035	-0.00239	0.002204
	3	0	0	0	0	0	0	0	0.441922
	4	0	0	0	0	0	-0.23032	0.005164	0.023198
	5	0	0	0	0	0	0	0.133365	0.009704
	6	0	0	0	-0.92351	0.120546	-0.21438	0.046025	-0.34283
	7	0	0	0	0	1	0.150314	-0.90558	1
	8	0	0	1	1	0.034065	1	1	-0.28877

Because the GS method result in upper triangular **A** matrices, the situation is much simpler than the GPC or Lowdin method. While it appears that essentially each column in **W** is dominated by a unique column of **X**, some columns are dominated by a second. As an example the sixth column of **W** under the GPC transformation is dominated by the first column of **X** and the seventh column of **X** with a coefficient of 0.87. The Lowdin transformation also demonstrates a case where due to the high degree of correlation between the second and seventh predictors in **X**, the coefficient matrix actually has two columns dominated by the seventh predictor and none by the second. To ensure this behavior was specific to the ozone dataset for the reasons mentioned, the same calculation was performed for the physical dataset (Appendix C). The physical dataset exhibited the expected behavior of each column in **W** being dominated by a column from **X**. Note for the GPC method the first column of **W** is comparable to that of the GS methods. This is because of the negligible coefficients for all other columns of **X** except $\mathbf{X}_2$.

Table 3.4 The marginal inclusion probabilities for the posteriors in Figure 3.3

	Column of X/W							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
X	0.55	0.67	0.44	0.17	0.23	0.24	0.43	0.45
GPC	1	0.76	0.21	0.36	0.10	0.32	0.11	0.30
Low	0.40	1	0.21	0.20	0.87	0.12	0.12	0.31
GS_1	0.67	1	0.17	0.15	0.16	0.21	0.47	0.27
GS_2	0.94	1	0.18	0.13	0.12	0.28	0.10	0.34

Visual inspection of the **A** matrix and consideration of the MIP will provide an indication of important variables however, there appears to be no coherent way to perform BVS. When using an orthogonal predictor matrix, the magnitude of correlation determines the rank using the MIP. From Table 3.4 it is clear that the MIP in **W** pick out columns dominated by specific predictors. The first column of **W** under GPC is dominated by the second predictor in **X** and has a MIP of 1. The second column of **W** for the Lowdin transformation is dominated by the seventh predictor from **X**. The GS methods have been returned to the original ordering from **X**, and as a result both are dominated by a rescaled version of the second predictor from **X.** This is expected as the MIP for the first predictor in **W** represents the second predictor in **X** only. Thus, it seems when using an orthogonal transformation the focus of inference should be for **y** as BVS appears not to be possible.

When using an orthogonal predictor matrix the least squares estimate of the regression coefficients is simplified. Further, any column of **W** will give the same estimated regression coefficient irrespective of the other columns of **W** included in the model. The least squares estimate for the *i*-th predictor maybe obtained as

$$\hat{\beta}_i = \frac{\mathbf{W}_i^T \mathbf{y}}{\mathbf{W}_i^T \mathbf{W}_i}. \tag{3.7}$$

(3.7) simplifies further if **W** is orthonormal as $\mathbf{W}_i^T \mathbf{W}_i = 1$. When a predictor matrix is orthogonal the magnitude of the regression coefficients is dictated by the strength of correlation with **y**. The posterior for **β** under **W** has a fixed location regardless of γ, but the variance shrinks as more predictors are included. With a non-orthogonal predictor matrix both location and variance change with γ.

For the posterior of σ^2, the *a* parameter is typically independent of γ, and *b* is a function of the RSS for γ, resulting in a decreasing expectation (or estimate using the mode) for σ ^{2} as q_γ increases. This decrease follows the required partial order for monotonicity due to *b* involving the model dependent RSS term so that as more columns are included in the predictor matrix for the orthogonal case, the fit improves and as such the estimated variance reduces. This point relates to Example 2.3 from the previous chapter where the classical estimate of the variance cannot be guaranteed to follow the required partial order. We now provide an illustration of these facts using the ozone data and Jeffreys prior. For Jeffreys prior we have the following expectations for the regression coefficients and model variance:

$$\mathbb{E}[\boldsymbol{\beta}_\gamma \mid \boldsymbol{\gamma}, \mathbf{y}, \mathbf{X}] = \hat{\boldsymbol{\beta}}_\gamma, \tag{3.8}$$

and

$$\mathbb{E}[\sigma^2 \mid \boldsymbol{\gamma}, \mathbf{y}, \mathbf{X}] = (\mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{H}_\gamma \mathbf{y}) / (n-2). \tag{3.9}$$

The classical estimator of the variance is

$$\hat{\sigma}^2 = (\mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{H}_\gamma \mathbf{y}) / (n-k-1), \tag{3.10}$$

where *k* is the number of predictors. Beginning with the first column and an intercept the next three columns of **X** or **W** are added sequentially as indicated by the sequence of numbers in the column headings of Table 3.5. Table 3.5 records the estimates of (3.8)-(3.10) and **W** is obtained by the GPC method.

As discussed the estimate of $\boldsymbol{\beta}_1$ varies with **X** and does not change for **W**. The classical estimator of the variance shows both decreasing and increasing behavior as predictors are added. This is a clear indication that it cannot be guaranteed to follow the required partial order. Thus, for any posterior for σ^2, any estimate such as the mean or the mode that is in the form of the classical estimator for variance, will never produce a monotone Gibbs MC.

Table 3.5 Updating parameter estimates by adding one column at a time to X and W (GPC method).

Predictor Matrix	Quantity	Added predictors			
		1	1 2	1 2 3	1 2 3 4
X	$\mathbb{E}[\boldsymbol{\beta}_1]$	0.0305	0.0192	0.0220	0.0221
W (GPC)	$\mathbb{E}[\boldsymbol{\beta}_1]$	2.991	2.991	2.991	2.991
X	$\mathbb{E}[\sigma^2]$	0.2364	0.1839	0.1765	0.1746
	Classical $\hat{\sigma}^2$	0.2334	0.1863	0.1812	0.1816
W (GPC)	$\mathbb{E}[\sigma^2]$	0.2130	0.1930	0.1880	0.1787
	Classical $\hat{\sigma}^2$	0.2101	0.1954	0.1829	0.1858

3.3 W and X: A comparison

Variable selection is not possible so comparisons between a choice of **W** and **X** will rely on measures associated with the in-sample prediction of **y**. To this end, we provide plots of the residual sum of squares to investigate the concentration of posterior mass for γ when using **W**. This is accompanied by plots comparing model complexity and model competition for the posterior of γ. Finally, we use the DIC criterion extended to include BMA to provide a comparison of **X** and **W** for in-sample prediction. Zellner's prior and Jeffreys prior are investigated for a range of values of *c* and *p* respectively. We use the constant Bernoulli prior for γ and investigate a range of values of *τ*. We use four real data sets; ozone, physical, bodyfat, and crime, see Appendix C for details.

3.3.1 Shrinkage, Model complexity and Model Competition.

Figures 3.5 and 3.6 show the effect of orthogonalization on the residual sum of squares, and in particular, the separation in the model space. We use the following two measures

$$R_J = \frac{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}}{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\mathbf{F}\mathbf{y}}, \text{ and } R_1 = \frac{\mathbf{y}^T\mathbf{y} - 0.5\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}}{\mathbf{y}^T\mathbf{y} - 0.5\mathbf{y}^T\mathbf{H}_\mathbf{F}\mathbf{y}}, \tag{3.11}$$

which are proportional to the RSS term in the posterior for Jeffreys prior (R_J), and Zellner's prior with $c = 1$ (R_1). Both are divided by the corresponding RSS for the full model so that the minimum value is 1.

In comparing R_J and R_1 it is clear that while the patterns of separation are the same, the posterior based on R_1 will be flatter than for R_J. The shrinkage value of $c/(c+1)$ shrinks the sum of squares for each model towards zero, as the distribution when normalized will be flatter. This will also have the effect of increasing model competition, which we will elaborate on further in the work to follow.

The GS methods both seem to be relatively similar as both are dominated by the scaled predictor from **X** with the greatest correlation with **y**. The GPC method is similar in separation to the GS methods for the ozone and physical datasets. For the bodyfat and crime datasets, the GPC method produces a number of layers and minimal separation respectively. The Lowdin method is similar to the other methods for the ozone data. For the physical and crime datasets there is almost no separation at all, and for the bodyfat data there is a small amount of separation. This separation is what contributes to the shrinkage effect in the posterior for $\boldsymbol{\gamma}$ when using **W**. The move to **W** creates a separation of models based on fit, while the spread over q_γ is similar between **X** and **W**.

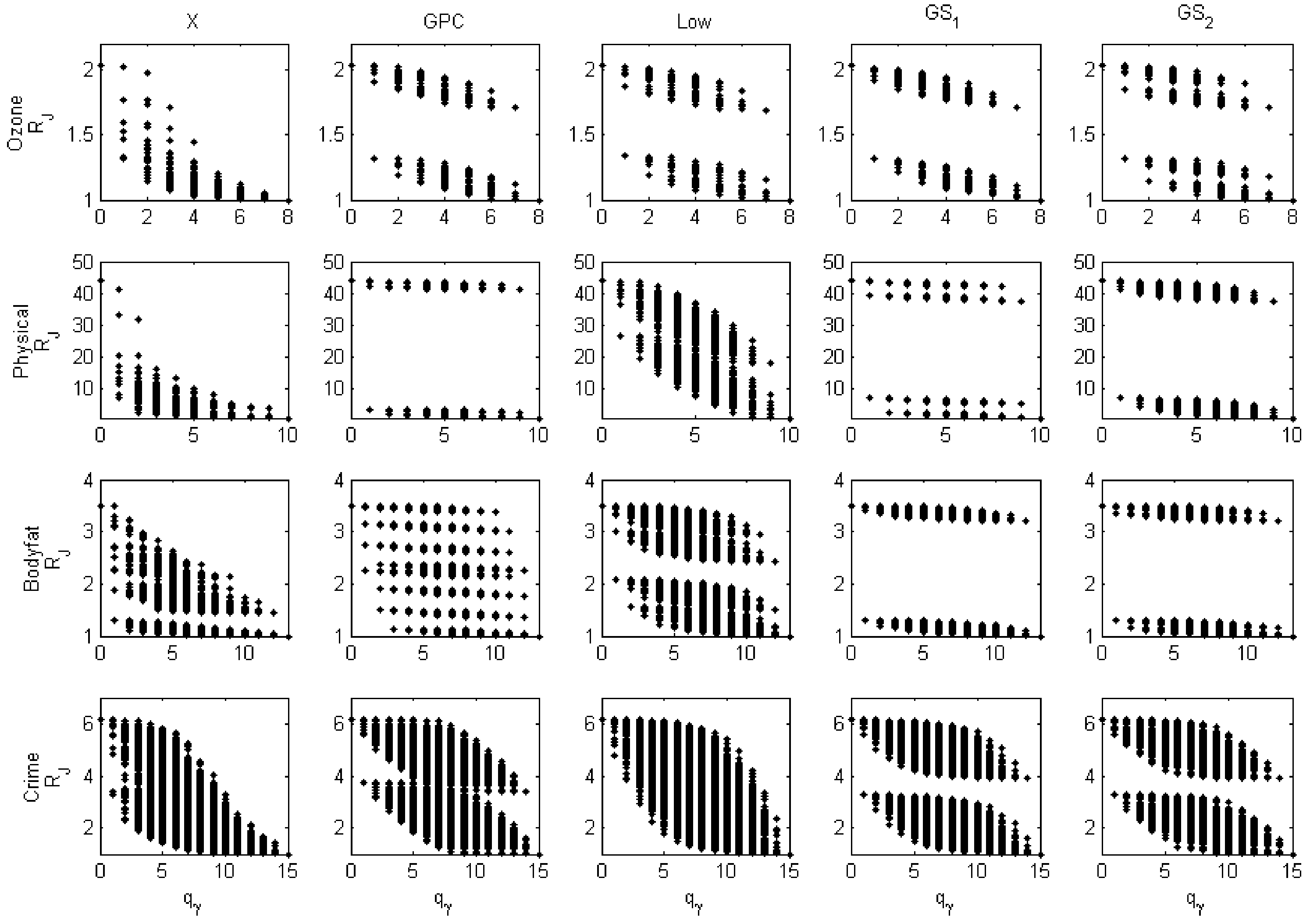


Figure 3.5 Plot of (R_J) for X, GPC, Lowdin. GS_1 and GS_2 for four real datasets.

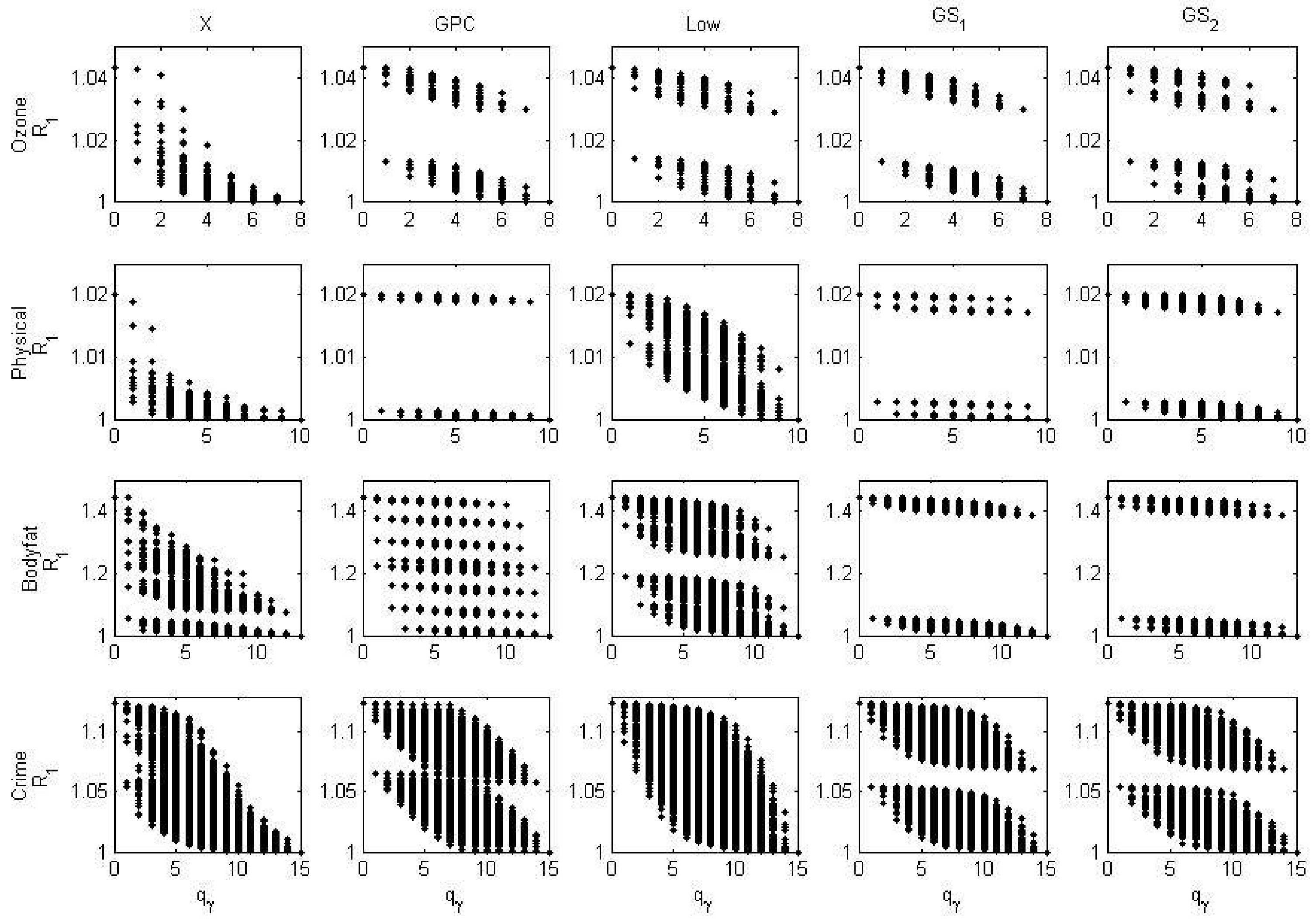


Figure 3.6 Plot of (R_1) for X, GPC, Lowdin. GS_1 and GS_2 for four real datasets.

While the shrinkage effect helps direct the posterior to a smaller group of suitable models, we do not wish this to be at the expense of increased model complexity. Ideally the number of predictors included in the **X** space and the number of columns used from **W** will be similar. Figures 3.7 and 3.8 show the expected model size for **X** and all **W** methods, for both Zellner's prior and Jeffreys prior over a range of τ and penalty (c or p). Recall the expected model size is defined as

$$f(q\,|\,\mathbf{y},\mathbf{X}) = \sum_{\{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}:\sum_{i=1}^{k}\gamma_i=q\}} f\;(\boldsymbol{\gamma}\,|\,\mathbf{y},\mathbf{X}), \tag{3.12}$$

and

$$\mathbb{E}[q] = \sum_{q=0}^{k} q f(q\,|\,\mathbf{y},\mathbf{X})\,. \tag{3.13}$$

We record the $\mathbb{E}[q_\gamma] = \mathbb{E}[q]$ -1 which indicates the number of predictors included, while omitting the contribution of the intercept which is common to all models. Thus, the result is between 0 and k. For the ozone data (k = 8) the expected model size for **X** and all **W** methods is similar over the range of τ with a fixed penalty for Zellner's prior and Jeffreys prior. Over the range of penalty values, the model size is lower for Zellner's prior for lower penalties due to the flattening effect of the $c/(c+1)$ term. Over the range of penalty all methods are very similar and the expected model size is decreasing as penalty increases. Notice that once the penalty exceeds *exp*(5), the expected model size for Zellner's prior and Jeffreys prior are very similar. For lower values of penalty and larger values of τ, the orthogonal methods move slightly above **X** for expected model size. The GS_1 and Lowdin methods obtain the highest expected model size compared to **X**. For lower values of τ and higher values of penalty the GS_1 and GPC methods obtain the smallest expected model sizes. Notice that the GS_2 method shows little departure from **X**. For the physical data for Zellner's prior the penalties for **X** and the **W** methods are similar. The GPC method stays the closest to **X** followed by both GS methods and the Lowdin transformation. The Lowdin method attains the smallest model size for lower values of τ and the greatest model size for larger values of τ. Over the range of c the Lowdin method moves above **X** at around *exp*(3.5), reaches a peak at *exp*(8), and moves below **X** around *exp*(9). Again the GS methods stay close to **X**, and the GPC method until *exp*(9) has the smallest expected model size compared to **X**. The Jeffreys prior demonstrates much different behavior. The Lowdin method for the entire range of τ has a much larger expected model size than the other methods. The GS_2 method also noticeably exceeds the expected model size for **X** over all values of τ. The GS_1 and GPC methods display the behavior observed for the ozone data over τ, where for lower values the expected model size is below that of **X**, and above for larger values of τ. Jeffreys prior over the range of penalty is again similar to Zellner's prior for values larger than approximately *exp*(5). The GS_1 and GPC methods are above **X** briefly for low penalty values and then move below **X** for larger values. The GS_2 method is similar, but takes much longer to move below **X**, and shows noticeable departure from **X** being well above $\mathbb{E}[q_\gamma]$ of **X** for the intermediate values of penalty. Much as in the case for Zellner's prior, the Lowdin method presents the most extreme expected model size being well above **X** until around *exp*(9).

At this point it serves to note that for smaller sample sizes (physical data), compared to larger sample sizes (ozone data), the behavior of the orthogonalization method may be less predictable or reliable. Clearly in this case the small sample size has been problematic for the Lowdin method. The increased model size will in part be due to the lack of separation in the residual sum of squares as demonstrated in Figures 3.7 and 3.8. The bodyfat dataset shows noticeable separation between all the methods. This dataset does have a large sample size, so the expected model size is not erratic, and for all methods the shape of change is very similar. For Zellner's prior and Jeffreys prior over τ, all methods increase and all **W** methods have a larger expected model size than **X**. The GS_2 method is the closest to **X**, followed by the GS_1 and GPC methods and finally, the Lowdin method. Over the range of penalty the same separation and ordering as noted, continues for Zellner's and Jeffreys prior with the expected model size decreasing with increasing penalty.

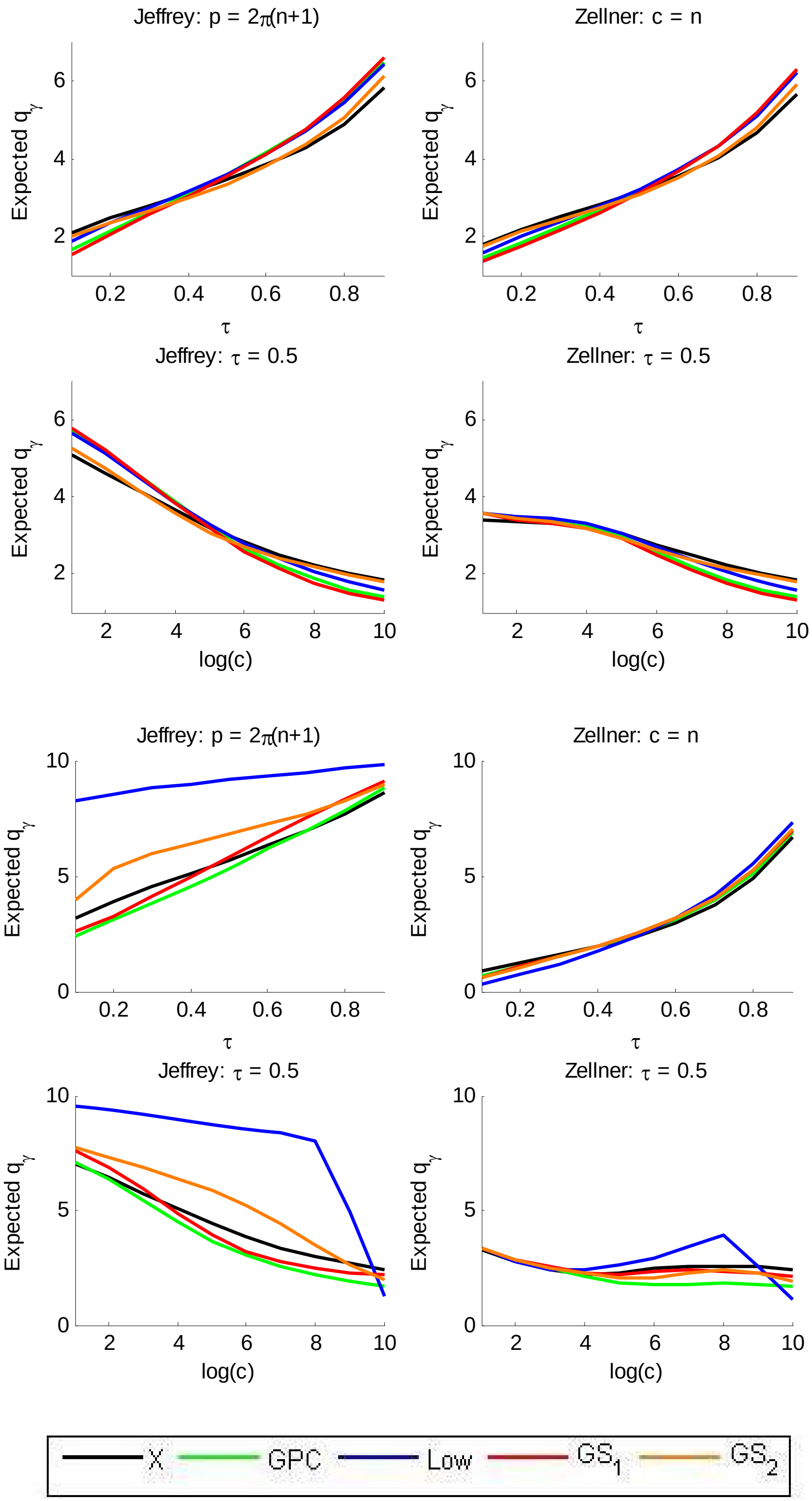


Figure 3.7 Expected Model Size: Top 4 panels: Ozone data with *k* = 8, and *n* = 80. Bottom 4 panels: Physical data with *k* = 10 and *n* = 27. For Jeffreys and Zellner's priors (*τ* = 0.5) the plots use the same scale on the axis where *p* = 2*π*(*c*+1).

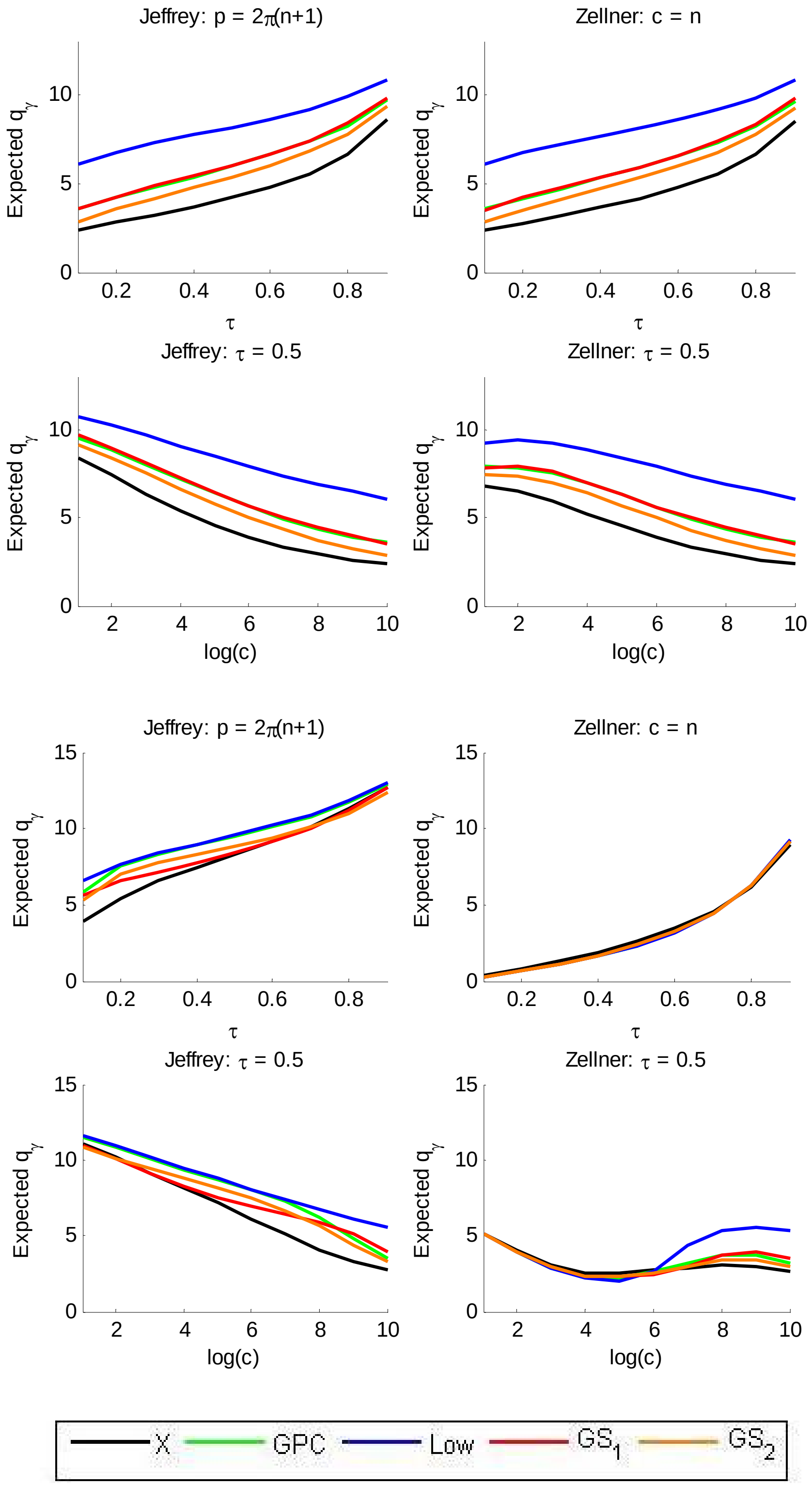


Figure 3.8 Expected Model Size: Top 4 panels: Bodyfat data with *k* = 13, and *n* = 250. Bottom 4 panels: Crime data with *k* = 15 and *n* = 47. For Jeffreys and Zellner's priors (τ = 0.5) the plots use the same scale on the axis where $p = 2\pi(c+1)$.

The crime data for Zellner's prior shows all methods to be very similar and increasing with τ. For Jeffreys prior over the range of τ, the Lowdin and GPC methods are the closest to E[q_γ] for **X** for larger values of τ, but are always above **X**. The GS methods are close to **X** and move below around $\tau = 0.5$. For Zellner's prior over the range of penalty the behavior is more dynamic compared to other datasets. The methods are all initially similar and then between *exp*(4) and *exp*(6) the orthogonal methods move below **X** with the Lowdin method attaining the smallest E[q_γ]. Above *exp*(6) the **W** methods move above **X** with the Lowdin method attaining the largest expected model size, and GS$_2$ staying the closest to **X**. Jeffrey's prior shows decreasing behavior over the range of penalty. The Lowdin and GPC methods stay strictly above **X**, while the GS methods for smaller values are just below **X** and then move above **X**. Recall the definition of model competition for $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$:

$$\mathrm{M}_{\alpha} = \min\left\{ j : \sum_{i=1}^{j} p_{(i)} \geq \alpha \right\}, \tag{3.14}$$

where $p_{(1)},\ldots,p_{(2^k)}$ are sorted model probabilities in decreasing order and $\alpha \in (0,1)$. If the posterior is a point mass then provided $\alpha < 1$, $M_\alpha = 0$, i.e. no model competition. We record the model competition for $\alpha = 0.99$, where we normalize M_α by dividing by (2^k). The ozone data for Zellner's prior and Jeffreys prior for τ displays a concave shape for model competition. In particular, the model competition for all values of τ for both priors is noticeably greater for **X**. The ordering of the **W** methods is consistent and the GS$_2$ method has the lowest model competition followed by the Lowdin transformation, the GPC method and finally GS$_1$. For the range of penalty for Zellner's prior, model competition is strictly decreasing with **X** again being the largest. For Jeffreys prior, model complexity increases slightly to a maximum around *exp*(4) and then decreases, while **X** provides the greatest model complexity. Over the range of penalty for both priors, the order of the **W** methods for increasing model complexity is the same as for τ; GS$_2$, Lowdin, GPC, and GS$_1$. For the physical data, the behavior between Zellner's prior and Jeffreys prior is very different. For Jeffreys prior over a range of τ and penalty the ordering of methods is the same.

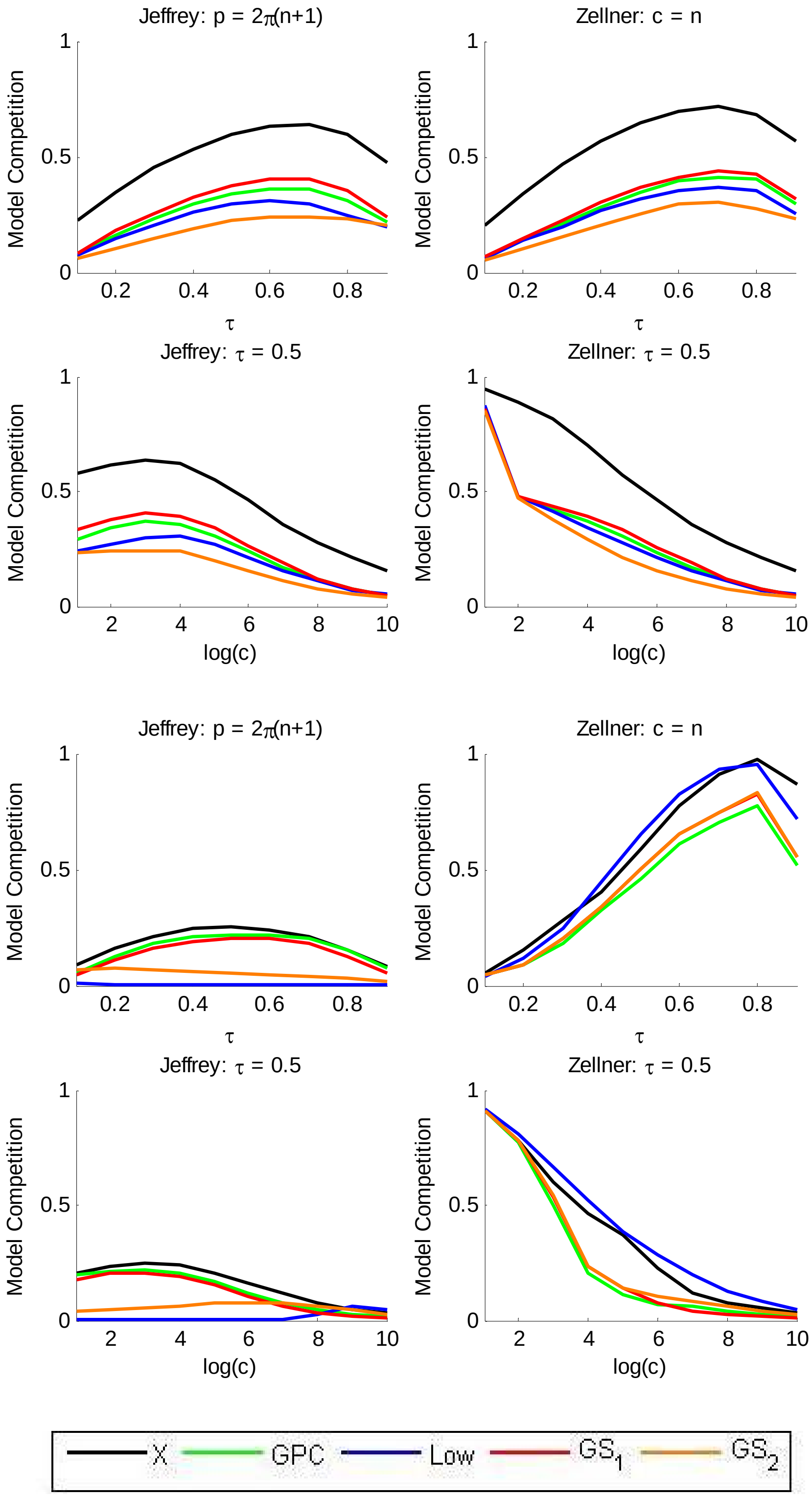


Figure 3.9 Model Competition: Top 4 panels: Ozone data with *k* = 8, and *n* = 80. Bottom 4 panels: Physical data with *k* = 10 and *n* = 27. For Jeffreys and Zellner's priors ($\tau = 0.5$) the plots use the same scale on the axis where $p = 2\pi(c+1)$.

The Lowdin method has the smallest model competition however, this is because from Figure 3.7 it is clear that the Lowdin method for most values of τ and penalty is close to favoring the full model. The profile of the GPC and GS_1 methods for model competition are the closest to **X** over the range of τ and penalty. The GS_2 method shows the lowest model competition next to the Lowdin method. For Zellner's prior the Lowdin method is very similar in model competition to **X** both over τ and the penalty. The GPC method for most values of τ and penalty attains the lowest model competition. The GS methods are intermediary to the Lowdin and GPC methods for model competition. Note that for the plot for Zellner's prior and τ, the GS methods are on top of each other. For the penalty the GS_1 method departs from the GS_2 method around $exp(5)$ and moves below GPC around $exp(6)$.

For the bodyfat data because n is large and used as the fixed penalty for choices of τ the model competition for Zellner's prior and Jeffreys prior are almost identical. Interestingly the Lowdin method has the lowest model competition, followed by GPC, GS_1 and GS_2. The Lowdin and GPC methods both stay below **X** for all values of τ, while for values less than 0.3 the GS methods move slightly above **X**. For the range of values for penalty the model competition for values larger than around $exp(5)$ appear similar. For Zellner's prior the model competition is much greater for smaller values of penalty, while for Jeffreys prior there is a maximum around $exp(2)$. For both priors, at large values of penalty ($>exp(7)$) the GS methods move above **X**, and the GPC and Lowdin methods remain very close to **X**. For Jeffreys prior the GS_2 method is closest to **X**, while the Lowdin method has the lowest model competition followed by the GS_1 and GPC methods. This behavior is similar for Zellner's prior except, the GS_1 method remains above GPC, and for values of penalty $< exp(1.5)$ moves above GS_2.

The crime data also shows noticeably different behavior between the priors. For choices of τ Jeffreys prior has a concave profile peaking around 0.6. The model competition for **X** is much greater than for the **W** methods, where GPC has the lowest model competition followed by the Lowdin method, GS_2 and GS_1. Over the range of penalty there is a peak around $exp(3)$, and while it is not clear due to the comparison with Zellner's prior, the ordering is the same as for over τ. For Zellner's prior and τ, **X** has the largest model competition and is noticeably larger than for Jeffreys prior.

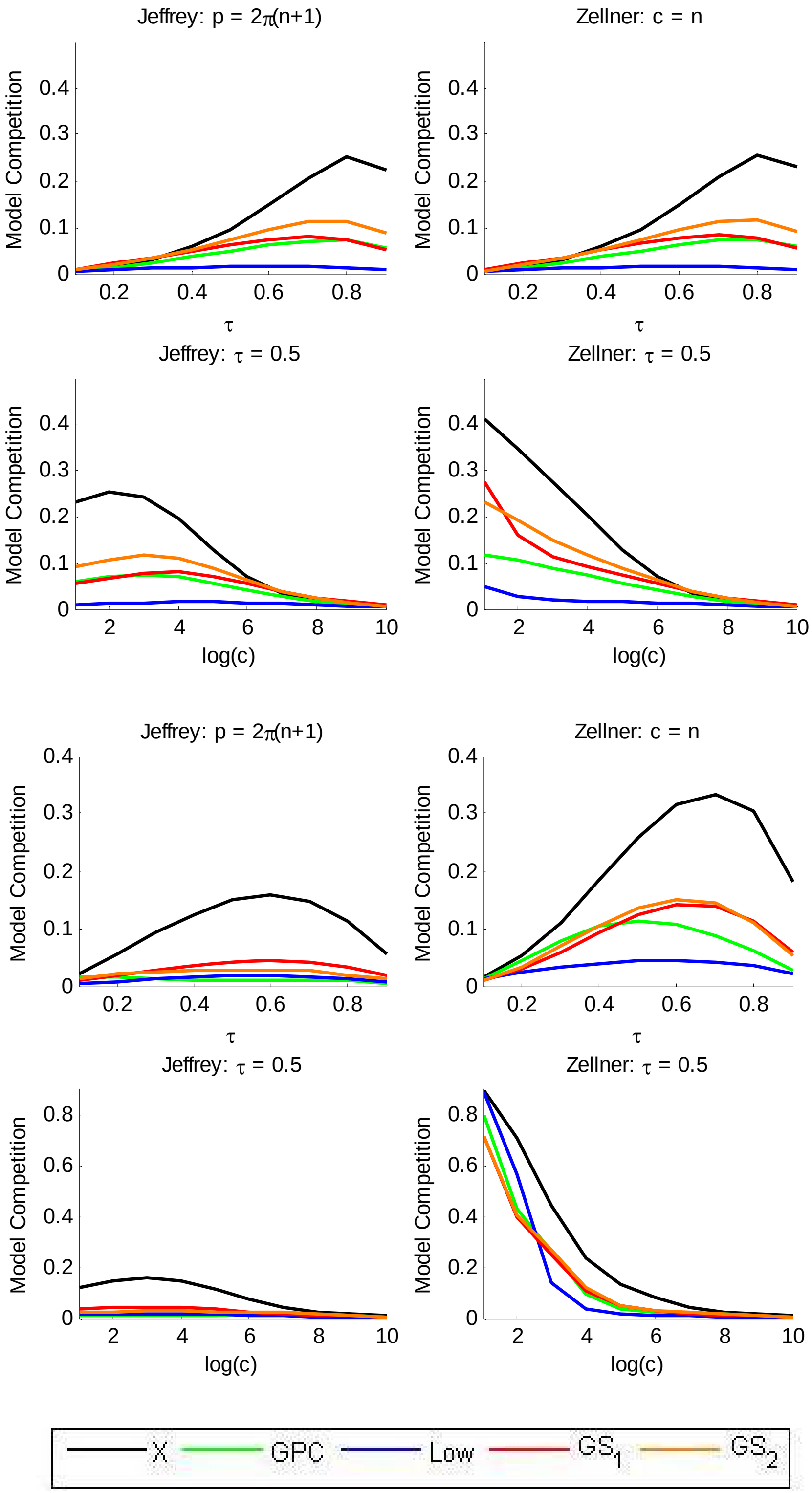


Figure 3.10 Model Competition: Top 4 panels: Bodyfat data with k = 13, and n = 250. Bottom 4 panels: Crime data with k = 15 and n = 47. For Jeffreys and Zellner's priors (τ = 0.5) the plots use the same scale on the axis where $p = 2\pi(c+1)$.

The Lowdin method attains the lowest model competition and for $\tau < 0.4$ the GS methods have lower model competition than GPC, and for $\tau > 0.4$ this behavior is reversed. Over the range of penalty the model competition between **X** and the **W** methods are not dissimilar. As with the other datasets model competition is decreasing with increasing penalty. The GS and GPC methods are similar, and for values of penalty less than *exp*(2.5) exhibit the lowest model competition followed by the Lowdin method. For values of penalty greater than *exp*(2.5) the Lowdin method has model competition lower than the GS and GPC methods. Finally, around *exp*(7) all **W** methods are very similar and by *exp*(8) this includes **X** also.

3.2.2 In Sample Prediction.

We now use the real datasets from above and the DIC criterion (Spiegelhalter *et al*, 2002) to compare the in-sample predictive ability of **X** and **W**. We now review DIC and its extension to include the model space. Posterior expected deviance generalizes naturally for integration over $\boldsymbol{\gamma}$:

$$\mathbf{DEV} = \overline{D}(\mathbf{y}) = \int -2\log f(\mathbf{y}\mid\boldsymbol{\theta})f(\boldsymbol{\theta}\mid\mathbf{y})d\boldsymbol{\theta} \text{ where } \boldsymbol{\theta} = (\boldsymbol{\beta},\sigma^2,\boldsymbol{\gamma}) \tag{3.15}$$

Extending this to assess BMA, an estimate of deviance is obtained by Monte Carlo simulation by simulating $\boldsymbol{\theta}_j = [\boldsymbol{\gamma}^j, (\sigma^2)^j, (\boldsymbol{\beta}_\gamma)^j]$ sequentially as outlined.

1. Generate: $\boldsymbol{\gamma}^1,\ldots,\boldsymbol{\gamma}^N \sim f(\boldsymbol{\gamma}\mid\mathbf{y},\mathbf{X})$
2. Then for $j = 1,..,N$ generate: $(\sigma^2)^j \sim f(\sigma^2\mid\boldsymbol{\gamma}^j,\mathbf{y},\mathbf{X})$, and $(\boldsymbol{\beta}_\gamma)^j \sim f(\boldsymbol{\beta}_\gamma\mid\boldsymbol{\gamma}^j,(\sigma^2)^j,\mathbf{y},\mathbf{X})$.

By letting $D(\mathbf{y},\boldsymbol{\theta}) = -2\log f(\mathbf{y}|\boldsymbol{\theta})$ we have $\overline{D}(\mathbf{y}) = \mathbb{E}[D(\mathbf{y},\boldsymbol{\theta})\mid\mathbf{y}]$. This means we can estimate the deviance and its precision as

$$\overline{D}(\mathbf{y}) \approx \frac{1}{N}\sum_{j=1}^{N} D(\mathbf{y},\boldsymbol{\theta}_j) \ , \ \mathbb{V}[\overline{D}(\mathbf{y})] \approx \frac{1}{N^2}\sum_{j=1}^{N}\left[D(\mathbf{y},\boldsymbol{\theta}_j) - \overline{D}(\mathbf{y})\right]^2 \tag{3.16}$$

For BMA using Zellner's prior, DIC can be estimated as

$$\mathrm{DIC}_\Gamma = p_d + \overline{D}(\mathbf{y}),$$
$$\text{where } p_d = \sum_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}} [\overline{D}_\gamma(\mathbf{y}) - D(\mathbf{y},(\hat{\boldsymbol{\beta}}_\gamma,\hat{\sigma}_\gamma^2))] f(\boldsymbol{\gamma} \mid \mathbf{y},\mathbf{X}). \tag{3.17}$$

where $\overline{D}_\gamma(\mathbf{y}) = \int -2\log f(\mathbf{y} \mid \boldsymbol{\theta},\boldsymbol{\gamma}) f(\boldsymbol{\theta} \mid \mathbf{y},\boldsymbol{\gamma}) d\boldsymbol{\theta}$ with $\boldsymbol{\theta} = (\boldsymbol{\beta},\sigma^2)$. $\hat{\boldsymbol{\beta}}_\gamma$ and $\hat{\sigma}_\gamma^2$ are the posterior expectations and may be replaced with other point estimates such as the median. Monte Carlo simulation is necessary for Zellner's prior and the required posteriors are

$$f(\boldsymbol{\gamma} \mid \mathbf{y},\mathbf{X},c) \propto (c+1)^{-\frac{q_\gamma+1}{2}} \left(\mathbf{y}^T\mathbf{y} - \frac{c}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y} \right)^{-n/2}, \tag{3.18}$$

where $\mathbf{H}_\gamma = \mathbf{X}_\gamma(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma^T$ and

$$f(\boldsymbol{\beta}_\gamma,\sigma^2 \mid \boldsymbol{\gamma},\mathbf{y},\mathbf{X}) = \mathbf{N}_{q_\gamma+1}\left(\frac{c}{c+1}\hat{\boldsymbol{\beta}}_\gamma, \frac{c\sigma^2}{c+1}(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1} \right) \mathbf{IG}\left(\frac{n}{2}, \frac{1}{2}\left[\mathbf{y}^T\mathbf{y} - \frac{c}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}) \right] \right). \tag{3.19}$$

If the sample size is large or the prior is very weak as in Jeffreys prior, we can estimate DIC_Γ as

$$\mathrm{DIC}_\Gamma = 2\mathbb{E}[q_\gamma] + 4 + \sum_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}} D(\mathbf{y},(\hat{\boldsymbol{\beta}}_\gamma,\hat{\sigma}_\gamma^2)) f(\boldsymbol{\gamma} \mid \mathbf{y},\mathbf{X}), \tag{3.20}$$

without the need for Monte Carlo simulation where again $\hat{\boldsymbol{\beta}}_\gamma$ and $\hat{\sigma}_\gamma^2$ are the required posterior expectations. The required posterior and expectations are

$$f(\boldsymbol{\gamma} \mid \mathbf{y},\mathbf{X}) \propto \left(\frac{p}{2\pi} \right)^{-\frac{q_\gamma}{2}} (\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})^{-\frac{n}{2}}, \tag{3.21}$$

and

$$\mathbb{E}(\boldsymbol{\beta}_\gamma \mid \sigma^2,\boldsymbol{\gamma},\mathbf{y},\mathbf{X}) = \hat{\boldsymbol{\beta}}_\gamma \text{ and } \mathbb{E}(\sigma^2 \mid \boldsymbol{\gamma},\mathbf{y},\mathbf{X}) = (\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})/(n-2). \tag{3.22}$$

Notice that (3.20) is equivalent to the model averaged AIC. For the most part, (3.17) and (3.20) will be close to minimum when the probability mass of the posterior degenerates to 1. This occurs when the posterior selects the model that is most likely according to DIC, and the expectations of the posterior distributions for $\boldsymbol{\beta}$ and σ^2 coincide with the maximum likelihood estimates. We compare all four methods of orthogonalization against **X** for a range of values in τ using the constant Bernoulli prior for $\boldsymbol{\gamma}$, considering a range of values of c and p for Zellner's and Jeffreys prior respectively. The results are done using the exact posterior for $\boldsymbol{\gamma}$. Thus, in the case of Jeffreys prior the results are exact. For Zellner's prior the DIC was estimated via Monte Carlo simulation with the standard deviation kept to at most 0.05. The calculation of DIC is implemented using the code Zellner.m and Jeffrey.m in Appendix D.

Figures 3.11 and 3.12 summarize the results of the DIC comparison. The grey band is the DIC value using **X** ±5, and represents a zone of indifference or equivalence. This is suggested to be a rough indication that there is no real difference between two different methods (Spiegelhalter *et al*, 2002). For the ozone data and Jeffreys prior, the only **W** method that falls outside the equivalence zone is GS_1 when $\tau < 0.3$ and the penalty is between *exp*(7) and *exp*(9). All other methods are comparable, and using DIC to rank the methods, GS_2 is consistently the closest to **X**. For Zellner's prior over τ all **W** methods are within ±5 of **X**, while the GS_1 method again falls outside the equivalence zone for penalties between *exp*(7) and *exp*(9). There is also very different behavior between Zellner's prior and Jeffreys prior for different choices of penalty due to the *c*/(*c*+1) term in $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$ for Zellner's prior. Thus, for the ozone data GS_2, GPC and the Lowdin methods appear suitable choices of orthogonalization method.

For the physical data using Jeffreys prior the GS_1 method slips above **X**+5 for τ between 0.2 and 0.3, while the GS_2 method goes above at the very lower limit around $\tau < 0.05$. For Jeffreys prior all methods except for the Lowdin method are within the equivalence zone. The Lowdin method remains above the equivalence zone until $\tau > 0.85$. In terms of penalty, Jeffreys prior for the GS_1 method is above **X**+5 from *exp*(4.5) to *exp*(6.5), and for values greater than *exp*(7) and *exp*(8) the GS_2 and Lowdin methods also go above **X**+5 respectively. The GPC method remains within the equivalence zone.

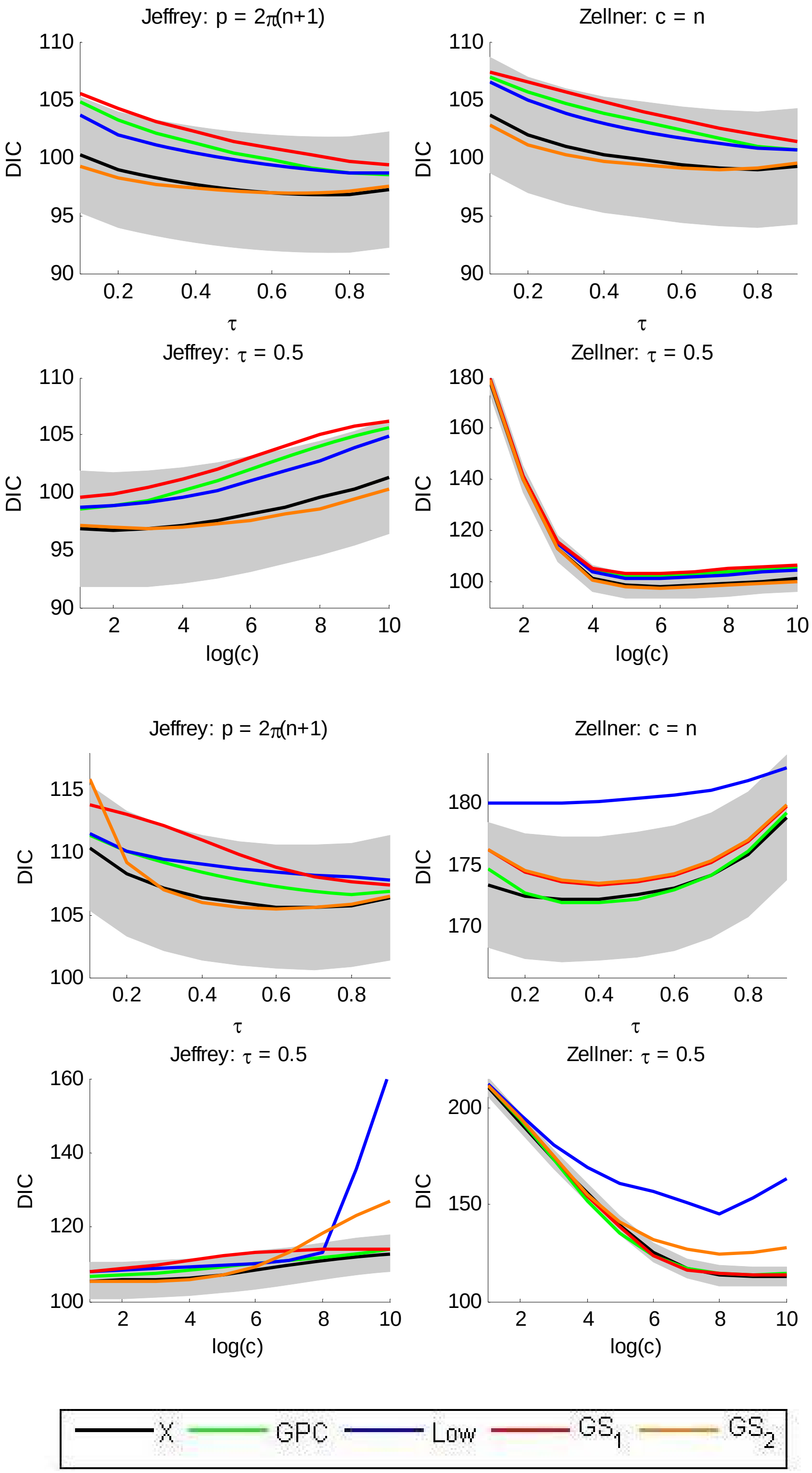


Figure 3.11 DIC: Top 4 panels: Ozone data with k = 8, and n = 80. Bottom 4 panels: Physical data with k = 10 and n = 27. Grey regions represent DIC of X ±5. For Jeffreys and Zellner's priors (τ = 0.5) the plots use the same scale on the axis where $p = 2\pi(c+1)$.

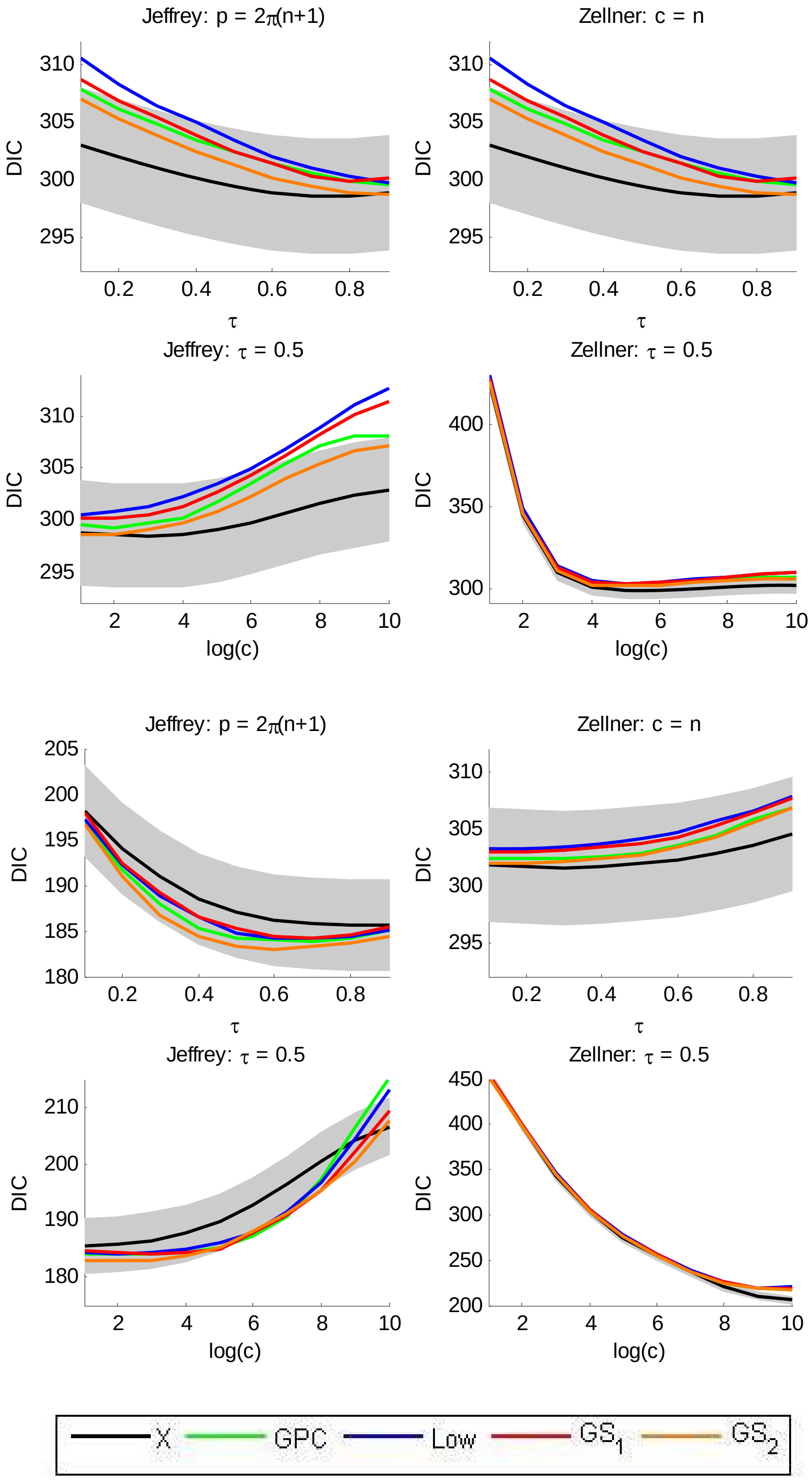


Figure 3.12 DIC: Top 4 panels: Bodyfat data with k = 13, and n = 250. Bottom 4 panels: Crime data with k = 15 and n = 47. Grey regions represent DIC of X ±5. For Jeffreys and Zellner's priors (τ= 0.5) the plots use the same scale on the axis where p = 2π(c+1).

For Zellner's prior over a range of penalty the GS_1 and GPC methods remain within the equivalence zone, while the GS_2 method moves above the equivalence zone after *exp*(5.6). The Lowdin method is above **X**+5 for penalty > *exp*(2.5) and shows dramatic increasing behaviour > *exp*(8). For the physical data the most consistent method appears to be the GPC method.

For the bodyfat data the sample size is large (n = 250), and so the DIC of Jeffreys and Zellner's priors for choices of τ is very similar. In both cases the GS_2 method is the closest to **X** followed by the GPC and GS_1 methods, and then finally the Lowdin method. Both the GS_2 and GPC method remain within +5 of **X** for all values of τ for Jeffreys prior and Zellner's prior. The GS_1 method is above **X**+5 for $\tau < 0.2$ and the Lowdin method is above **X**+5 for $\tau < 0.35$. For choices of penalty the GS_2 method stays within +5 of **X** for both Jeffreys prior and Zellner's prior. The next closest GPC, moves above **X** +5 for $c >$ *exp*(8) for Jeffreys prior and Zellner's prior. The GS_1 and Lowdin methods both move above **X** +5 for both Jeffreys prior and Zellner's prior when $c >$ *exp*(6). GS_2 and GPC appear to be the best suited **W** methods for the bodyfat data.

For the crime data and Jeffreys prior, a number of methods fall below **X**, but not outside the equivalence region. For choices of τ, the GS_2 method has the lowest values of DIC, followed by the GPC, GS_1 and Lowdin methods. For choices of penalty less than *exp*(4), the closest methods to **X** are the Lowdin and GPC followed by the GS methods. For choices of penalty greater than *exp*(7), the methods in order of descending DIC are; GPC, Lowdin, GS_1 and GS_2. In particular, around *exp*(9) the GPC and Lowdin methods move above **X** +5. For Zellner's prior the order of methods is similar to Jeffreys prior although, the values of DIC are above that of **X**. The closest method to **X** is GS_2 followed by GPC, GS_1 and the Lowdin transformation. Again all methods remain within the equivalence region. For choices of penalty under Zellner's prior the GS_2 method stays the closest to **X**, followed by the GPC, GS_1 and Lowdin methods. Around *exp*(8) all **W** methods move above the **X** +5 boundary. It is reasonable to suggest that all **W** methods are suitable for use with this dataset.

In general when the sample size is large and $c = n$ the value of DIC over values of τ are very similar. When n is small, Zellner's prior tends to larger values of DIC due to increased model competition. This behavior is also evident by the larger values of DIC for smaller values of *c*. Again the *c*/(*c*+1) is responsible for the increased model competition. Also notice that for Zellner's prior the values of DIC decrease more rapidly for the larger datasets, such as for the ozone and bodyfat datasets.

Overall it appears that the GS$_2$ method does remarkably well in attaining similar values of DIC compared to **X**. The GPC consistently performs well, which is no surprise, given its previous use in the literature. The GS$_1$ method performs moderately well however, the Lowdin method has turned out to be much less effective than hoped. Apart from the Lowdin case under Zellner's prior for the physical data, the **W** methods appear very competitive with **X**, except for some methods with extreme choices of τ (i.e. near 0 or 1), or extreme choices of penalty.

While DIC indicates comparable in-sample predictive ability, it is possible for the models to produce quite different predictions. Model checks can be performed using the posterior predictive distribution (PPD), which for Zellner's prior is

$$f(\tilde{\mathbf{y}} \mid \boldsymbol{\gamma}, \mathbf{y}, \mathbf{X}, \tilde{\mathbf{X}}, c) = \mathbf{T}\left(n, \boldsymbol{\mu}_\gamma, \frac{1}{n}[\mathbf{y}^T\mathbf{y} - \frac{c}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}]\boldsymbol{\Sigma}_\gamma \right), \tag{3.23}$$

where: $\boldsymbol{\Sigma}_\gamma = \left[\mathbf{I}_m + \frac{c}{c+1}\tilde{\mathbf{X}}_\gamma^T(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\tilde{\mathbf{X}}_\gamma \right]$, $\boldsymbol{\mu}_\gamma = \tilde{\mathbf{X}}_\gamma\left(\frac{c}{c+1}\hat{\boldsymbol{\beta}}_\gamma \right)$ and $\mathbf{H}_\gamma = \mathbf{X}_\gamma(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)\mathbf{X}_\gamma^T$.

For Jeffreys prior the PPD is

$$f(\tilde{\mathbf{y}} \mid \boldsymbol{\gamma}, \mathbf{y}, \mathbf{X}, \tilde{\mathbf{X}}) = \mathbf{T}\left(n, \tilde{\mathbf{X}}_\gamma\hat{\boldsymbol{\beta}}_\gamma, \frac{(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})}{n}[\mathbf{I}_m + \tilde{\mathbf{X}}_\gamma^T(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\tilde{\mathbf{X}}_\gamma] \right). \tag{3.24}$$

See Appendix B for the derivation of (3.23) and (3.24). For model checking we set $\tilde{\mathbf{X}}$ = **X**, and use the following quantities. Let F_i and F_i^{-1} be the corresponding posterior predictive CDF, and inverse CDF averaged over γ respectively for the *i*-th response. The probability of being more extreme than y$_i$ is

$$\min\{F_i(y_i), 1 - F_i(y_i)\}. \tag{3.25}$$

If (3.25) is very small for many observations then the model may be inadequate. Predictive coverage (PC) for an observation y$_i$ is defined as

$$\text{PC}_i = \mathbb{I}_{\left[F_i^{-1}(\alpha/2),F_i^{-1}(1-\alpha/2)\right]}(y_i)\,, \tag{3.26}$$

where α indicates the probability contained in the tails and $\mathbb{I}_\text{A}$ denotes the indicator function for the set A. Thus, PC indicates whether the (1-α) equal-tail PPD interval contains y$_i$. Predictive coverage can then be summarized for all n as $(1/n)\sum_{i=1}^{n}\text{PC}_i$. Finally, the probability of observing a more extreme value than a statistic of **y** such as the median, minimum, and maximum, can be used to check model adequacy. Let $\hat{\psi}$ be the observed statistic of **y**, and $\tilde{\psi}_1,\ldots,\tilde{\psi}_m$ represent m samples of this statistic generated by simulating m samples of **y** from the PPD averaged over **γ**. The probability of observing a more extreme value is:

$$\min\left\{\frac{1}{m}\sum_{i=1}^{n}\mathbb{I}_{(-\infty,\hat{\psi}]}(\tilde{\psi}_i),\frac{1}{m}\sum_{i=1}^{n}\mathbb{I}_{(\hat{\psi},\infty)}(\tilde{\psi}_i)\right\}. \tag{3.27}$$

Much like (3.25) a small value indicates model inadequacy. While the measures (3.25) – (3.27) cannot be used to categorically distinguish between competing methods, these checks do provide an indication of comparability. The results from the datasets above are extensive and as such are not provided in this thesis but may be obtained from the author if required. The code used for model checking is provided in Appendix D under ModelCheck.m.

Cases where the DIC of **X** and **W** were comparable were found to be similar using Model checking. Predictive coverage was typically the same, we used $\alpha = 0.05$, varying by only one or two observations. The tail probabilities (3.25) were typically acceptable and for some observations these values were similarly good or bad for **X** and **W**. In other cases **X** had better values of (3.25) for some observations compared to **W** and vice versa. The values of (3.27) were checked for the minimum, maximum and median. Again, the values were typically very similar between **X** and **W**. For cases where the DIC of the **W** methods moved outside the equivalence region, the model checking statistics indicated some difference between **X** and **W**. Specifically, the main difference was the proportion of low tail probabilities for observing a value more extreme than the observed response. Under **W** this proportion of values increased greatly when **W** moved well outside the equivalence region. In extreme cases when the posterior supported the null model due to extreme choices of penalty, the inference from model checking was poor. This agrees with the results of DIC. This extensive analysis while having not indicated **W** is better than **X** has definitely provided strong evidence that when modeling **y** using model averaging, an appropriate orthogonal transformation will provide equally comparable inference. These results are definitive for the datasets we have investigated however, there is no reason why we cannot assume these results will generalize to other real datasets analyzed using the Bayesian linear regression model.

3.3.3 Data splitting

As we have already discussed variable selection is not possible using **W**. However, according to DIC, the use of **W** instead of **X** for fitting the response using BMA is comparable and competitive. DIC is a measure for in-sample prediction however, there is also out-of-sample prediction to consider. The main problem with using **W** instead of **X** for out-of-sample prediction is the additional variability introduced by **A**. This extra variability results in out of sample prediction less accurate for **W** than for **X**, and the study in Cripps *et al* (2006) demonstrated this very point. Most studies use data splitting to determine out of sample predictive accuracy. Thus, a training sample is used to estimate regression coefficients and model weights, which are then used to fit a response with the remaining samples of **X**. Some distance measure between the fitted response and the remaining observations provides an indication of accuracy. While we haven't found any studies to support this point, it is reasonable to suggest that the larger the data partition for **X** then the greater the variation in **A** and consequently **W**. This in turn indicates that using leave one out cross-validation methods for detecting outliers may be less reliable. The conditional predictive ordinate (CPO) is a leave one out measure using the PPD for outlier detection. In (3.23) or (3.24) setting $\widetilde{\mathbf{y}} = y_i$ and $\widetilde{\mathbf{X}} = \mathbf{x}_i$ where $\mathbf{x}_i$ is the i-th row of **X**, and $\mathbf{y} = \mathbf{y}_{-i}$ and $\mathbf{X} = \mathbf{X}_{-i}$ the remaining data will give the CPO measure. This represents the probability of observing y_i conditional on the remaining data. A low value indicates a potential outlier. When using **X** it is possible to use MCMC output to estimate the CPO value without explicitly leaving out the i-th observation. From Gelfand and Dey (1994) let $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}_{\gamma}, \sigma^2_{\gamma})$ then

$$\mathrm{CPO}_i = f(y_i \mid y_{-i}, X_{-i}, X_i) = \left[\int \frac{1}{f(y_i \mid \boldsymbol{\theta})} f(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta} \right]^{-1}, \tag{3.28}$$

which can be approximated using MCMC output for $(\boldsymbol{\gamma}, \boldsymbol{\beta}_{\gamma}, \sigma_{\gamma}^2)$. This derivation relies on the conditioning on **X** as the retained values in **X** are the same, irrespective of the observation omitted. In constructing **W** we require **A** thus, this same property does not apply when using **W**. One could assume **A** does not change by constructing **A** from the full matrix. Either way, it is reasonable to assume for a large sample size and the minimum leave one out should provide the smallest variations in **A** so that hopefully outlier detection using CPO is comparable between **X** and **W**. This is clearly a direction for future research.

3.4 Summary

We have reviewed a numerical example of the update probabilities for the Gibbs sampler for **X** and **W** to demonstrate the difference between non-monotone and monotone orderings. This partial ordering follows the nested model structure for model comparison in linear regression.

We reviewed orthogonalization methods beginning with a common method known as generalized principal components. Following this, we recommended that if singular value decomposition is used to obtain an orthogonal predictor matrix, then the Lowdin transformation should be employed instead. This is because the Lowdin transformation is based on SVD and minimizes the distance between **X** and **W**, with respect to the L^2 matrix norm. Discussions from previous literature suggested the Gram-Schmidt transformation method might prove useful, but no methods for ordering **X** prior to transformation were given. We use the modified GS approach, and provided two methods for ordering the columns of **X** prior to transformation. We recommend a naïve method based on correlation with **y**, and another method partly inspired by partial least squares, obtaining an order which accounts for the correlation structure in the predictors and with **y**. This allowed us to use the GPC as a benchmark from the literature, and to trial three new methods in the context of model averaging with an orthogonal predictor matrix.

From this point we moved into numerical work. Using the ozone data set, we provided an example of the posterior distributions for γ, along with the MIP, the fitted model averaged response, and the matrices required to transform $\mathbf{X}_0$ into $\mathbf{W}_0$. This example demonstrated the posteriors can vary noticeably between the orthogonal methods, and are not directly comparable with each other or the posterior for **X**. The fitted model averaged values of **y** are also very similar between **X** and the orthogonal methods. While there is a relation between the mixing coefficients in **A** and the MIP probabilities of the columns of **W**, there appears to be no coherent way to use this information to obtain quantities that reflect the marginal inclusion probabilities in the **X** space. We then considered the effect of using **W** has on the posterior distributions of parameters such as $\boldsymbol{\beta}$ and σ^2. In particular, the partial ordering required for the Gibbs MC relies on certain properties of point estimates under the posterior distribution for $\boldsymbol{\beta}$ and σ^2.

The plots of the residual sum of squares for the four datasets; ozone, physical, bodyfat and crime, indicated the degree of shrinkage effect that can be obtained by moving to the **W** space. In particular the distinction between “poor” and “good” models becomes much clearer. As a result, the posterior model probabilities under **W** are more focused towards a particular subset of models. The difference between Zellner’s prior and Jeffreys prior with respect to *c* and *p* was demonstrated by the shrinkage term. Due to the $c/(c+1)$ term in the posterior for Zellner’s prior in the residual sum of squares in the posterior for γ, using Zellner’s prior flattens the distribution of the residual sum of squares for small values of *c*. We noted that while using **W** can focus the posterior mass, we would prefer this not be at the expense of using more predictors than **X**. Studies of the expected model size indicated the **X** and **W** are generally comparable for the average number of predictors used. The physical dataset did show a large difference between the Lowdin and the other orthogonalization methods, and the bodyfat data also showed some separation. The comparison of model competition provided confirmation of the shrinkage effect obtained by the residual sum of squares. It also again showed the difference, between Jeffreys and Zellner’s prior for smaller choices of penalty. Specifically, the model competition for small choices of penalty for Zellner’s prior is much greater than for Jeffreys prior, even for large sample sizes.

DIC using the plus or minus rule of 5 indicated that for the most part, the orthogonal methods were equivalent to using **X** for modeling the response. This provides good evidence that if we wish to model the response, and are not concerned with variable selection, then we may use a monotone CFTP Gibbs sampler to generate i.i.d. samples from $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$. It seems reasonable to suggest that the results and comparison are much more stable for larger sample sizes as indicated by the ozone and bodyfat data. The GPC method is well justified in its use in the literature, performing consistently well. Of the three new methods the GS_2 method proved to be very competitive and certainly on par with the GPC method. The GS_1 method did not perform particularly well in general, and the Lowdin transformation also performed rather poorly in terms of DIC compared to **X**. We also provided a brief discussion of out of sample prediction using **W**. Further research is required to determine if, under any circumstances, such as a large ratio of *n* to *k*, will allow **W** to be competitive for out-of-sample prediction for **y**. It also appears that unless simplifying assumptions are made, outlier detection using the CPO measure is not straight forward as it is when using **X**.

CHAPTER 4

GIBBS SAMPLING

"*A statistical analysis, properly conducted, is a delicate dissection of uncertainties, a surgery of suppositions*."

- M.J. Moroney

In this chapter we investigate the effect of sampling. Specifically, we compare the computational time for the standard and orthogonal Gibbs samplers and the perfect sampler. We monitor convergence of estimators per sample, and the convergence in distribution per sample and in real time. A larger simulation study is undertaken to compare the sampling methods for larger choices of k and n. We also explore factors affecting the BCT, which is heavily related to the efficiency of any monotone Gibbs CFTP. The factors affecting the BCT are information, and choices of hyper-parameters. Finally, we consider two larger datasets and generate samples using the three different approaches, analyzing the data as an analyst would in the real world. As part of this analysis we record cpu-time, exploration of $\boldsymbol{\Gamma}$, effective sample size, predictive coverage, DIC, MIP, and tail probabilities for the minimum and maximum of $\mathbf{y}$.

4.1 Algorithms and Computation

We now review the algorithms used in this chapter along with a discussion of computational aspects for calculating the Gibbs update probability. The computation of the hat matrix and/or the least square estimates requires the most computational effort due to the $(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}$ term. In the case of an orthonormal matrix there is no requirement to find this inverse as it is 1, and the hat matrix simplifies to $\mathbf{W}_\gamma\mathbf{W}_\gamma^T$. In general, for an orthogonal matrix $\mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T$ is a sum of the individual projection matrices $\mathbf{W}_i(\mathbf{W}_i^T\mathbf{W}_i)^{-1}\mathbf{W}_i^T$. The pseudo code for a monotone CFTP Gibbs sampler for $f(\boldsymbol{\gamma}\,|\,\mathbf{y},\mathbf{X})$ is given in Algorithm VII.

Algorithm VII: Monotone CFTP Variable Selection Gibbs Sampler.

```
Set: coalescence = false.
Set: $T$ = -1.
While coalescence = false
    Set: $T = 2T$
    Set: $(\boldsymbol{\gamma}^U)_T = \{1\}^k$, and $(\boldsymbol{\gamma}^L)_T = \{0\}^k$.
    For $t = \{T,\ldots,-1\}$
        For $i$ = 1,...,$k$.
            Generate: $(u_i)_{t+1} \sim \mathrm{U}(0,1)$.
            Compute: $\alpha_U = \Pr((\gamma_i^U)_{t+1} = 1\,|\,(\boldsymbol{\gamma}_{<i}^U)_{t+1},(\boldsymbol{\gamma}_{>i}^U)_t,\mathbf{y},\mathbf{X})$
            Compute: $\alpha_L = \Pr((\gamma_i^L)_{t+1} = 1\,|\,(\boldsymbol{\gamma}_{<i}^L)_{t+1},(\boldsymbol{\gamma}_{>i}^L)_t,\mathbf{y},\mathbf{X})$
            If $(u_i)_{t+1} \le \alpha_L$
                Set: $(\gamma_i^U,\gamma_i^L)_{t+1} = 1$.
            Else if $(u_i)_{t+1} > \alpha_U$
                Set: $(\gamma_i^U,\gamma_i^L)_{t+1} = 0$.
            Else
                Set: $(\gamma_i^U)_{t+1} = 1$, and $(\gamma_i^L)_{t+1} = 0$
    If $(\boldsymbol{\gamma}^U)_0 = (\boldsymbol{\gamma}^L)_0$
        Set: coalescence = true
    Else
        Set: coalescence = false
```

To manage the notation we have split $\boldsymbol{\gamma}_{-i}$ into those components already updated at time t, and those yet to be updated, so let $\boldsymbol{\gamma}_{<i}$ represent those components with indices in $\{1,...,i-1\}$ and $\boldsymbol{\gamma}_{>i}$ those components with indices in $\{i+1,...,k\}$. Note if i = 1 then there is no $\boldsymbol{\gamma}_{<i}$ and if i = k there is no $\boldsymbol{\gamma}_{>i}$. Recall we must reuse the random number u_t. To simplify the computations in the orthogonal case we can pre-compute:

$$\Pr(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X}) = \left[1 + \sqrt{(c+1)}\,\frac{1-\tau}{\tau}\left[1 - \frac{\tilde{c}_1 \mathbf{y}^T \mathbf{H}_i \mathbf{y}}{\mathbf{y}^T\mathbf{y} - \tilde{c}_1 \mathbf{y}^T \mathbf{H}_{\gamma_i = 0}\mathbf{y}}\right]^{n/2}\right]^{-1}, \qquad (4.1)$$

so in this case we can pre-compute $\sqrt{(c+1)}[(1-\tau)/\tau]$, $\mathbf{y}^T\mathbf{y}$, $\tilde{c}_1 = c/(c+1)$, n/2 and for each column of $\mathbf{W}$, $\mathbf{y}^T\mathbf{H}_i\mathbf{y} = \mathbf{y}^T\mathbf{W}_i\mathbf{W}_i^T\mathbf{y}$.

Algorithm VIII: Variable Selection Gibbs Sampler.

```
Set: $\boldsymbol{\gamma}_1 \in \boldsymbol{\Gamma}$
For $t$ = 1,…,$N$:
    For $i$ = 1,...,$k$.
        Generate: $u \sim U(0,1)$.
        Compute: $\alpha = \Pr((\gamma_i)_{t+1} = 1 \mid (\boldsymbol{\gamma}_{<i})_{t+1}, (\boldsymbol{\gamma}_{>i})_t, \mathbf{y}, \mathbf{X})$
        If $u \le \alpha$
            Set: $(\gamma_i)_{t+1} = 1$.
        Else
            Set: $(\gamma_i)_{t+1} = 0$
```

In the case of $\mathbf{X}$ we can perform as follows:

$$\Pr(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i}, \mathbf{y}, \mathbf{X}) = \left[1 + \sqrt{(c+1)}\,\frac{1-\tau}{\tau}\left[\frac{\mathbf{y}^T\mathbf{y} - \tilde{c}_1 \mathbf{v}_{\gamma_i=1}^T \mathbf{A}_{\gamma_i=1}^{-1}\mathbf{v}_{\gamma_i=1}}{\mathbf{y}^T\mathbf{y} - \tilde{c}_1 \mathbf{v}_{\gamma_i=0}^T \mathbf{A}_{\gamma_i=0}^{-1}\mathbf{v}_{\gamma_i=0}}\right]^{n/2}\right]^{-1} \qquad (4.2)$$

where again we pre-compute $\sqrt{(c+1)}[(1-\tau)/\tau]$, $\mathbf{y}^T\mathbf{y}$, $\tilde{c}_1 = c/(c+1)$, $n/2$, as well as $\mathbf{v} = \mathbf{X}^T\mathbf{y}$ and $\mathbf{A} = \mathbf{X}^T\mathbf{X}$.

Beyond this, any further speed in the case of **X** requires the use of additional methods for decomposing $\mathbf{X}^T\mathbf{X}$. Smith and Kohn (1996) recommend using update procedures for the Cholesky decomposition which is not easily implemented. Recent work by Eklund and Karlsson (2007), use a simulation study and found that increased speed can be obtained using Cholesky factorization and the sweep algorithm. We have employed all the usual tricks to improve the speed of updating for the standard Gibbs sampler, including using the centered matrix $\mathbf{X}_0$ which can help to reduce dependence between components, and improve the convergence of the Gibbs sampler.

4.2 Convergence and efficiency

Using the four datasets from the previous chapter, we now investigate the rate of convergence of the standard Gibbs sampler, the perfect version and the Gibbs sampler using **W**. To avoid the added complication of assessing burn-in for each standard Gibbs chain in **X** and **W**, we draw starting points according to the true posterior distribution, and use multiple chains. We use an orthogonalization method that performed well according to the DIC, which we have taken as the GS$_2$ method. We have used Jeffreys prior over Zellner's, mainly for computational simplicity. All work for this chapter was conducted on a stand-alone Compaq Presario 2500 laptop, running Mircosoft windows XP, with Maltab 7.0, using an Intel Celeron 2.60Ghz processor with 512 MB of RAM. This is done to help ensure some consistency in the recorded running times of the sampling methods used in this chapter.

4.2.1 Convergence in Distribution

The measure we use for convergence in distribution is

$$\sum_{i=1}^{2^k} | f_i - \hat{f}_i |, \tag{4.3}$$

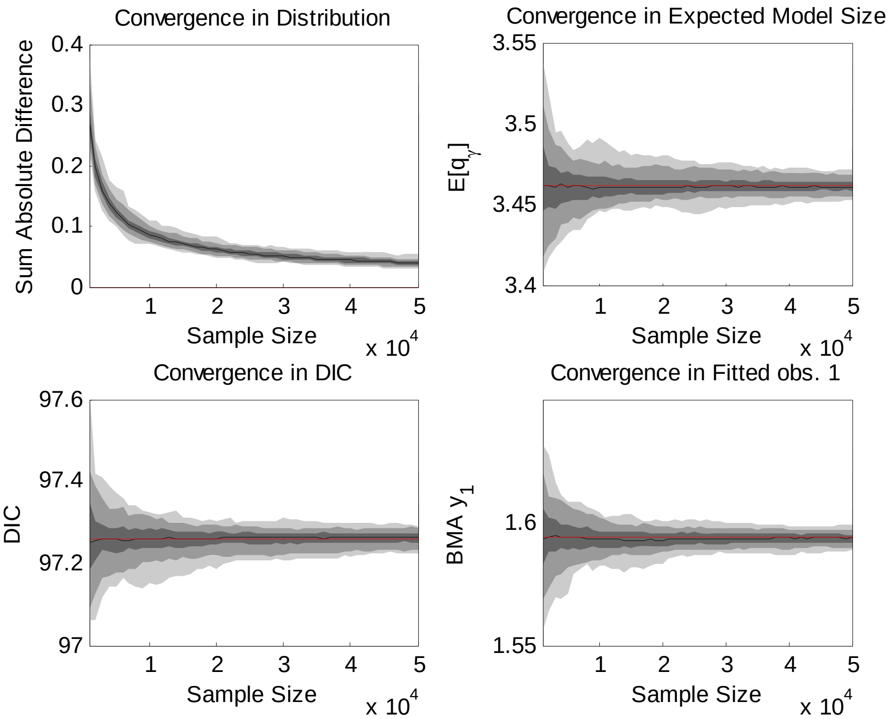
where f_i is the true posterior probability, and $\hat{f}_i$ is the estimated posterior probability, for state i. (4.3) is proportional to the total variation norm:

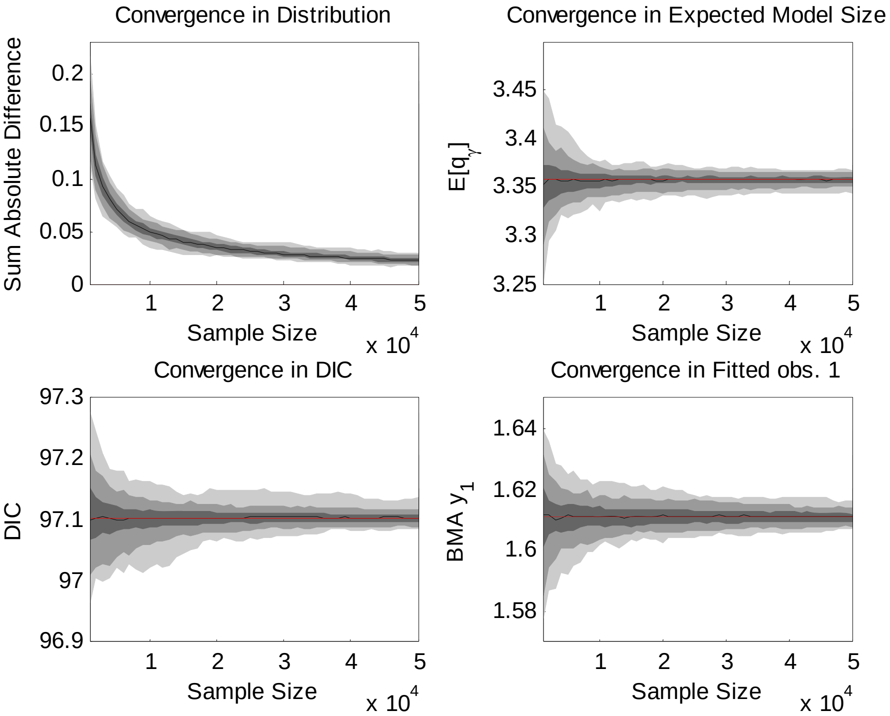
$$\| f-\hat{f}\|_{TV}=0.5\sum_{i=1}^{2^k}| f_i-\hat{f}_i |. \tag{4.4}$$

We also monitor the convergence of the expected number of predictors included, DIC, and the BMA fitted value for the first observation. The study was then run by selecting starting points according to the true posterior distribution for the standard Gibbs samplers for **X** and **W**. Then for each dataset for the three sampling methods, a chain of 50000 sample points was generated 100 times. For each chain the (4.3), $\mathbb{E}[q_\gamma]$, DIC, and BMA $\hat{y}_1$, is updated every 1000 iterations for every chain. Figures 4.1 – 4.4 show these results.

We can see that of the three methods, the rate of convergence in distribution and the other quantities is essentially equivalent for the exact sampler and the orthogonal Gibbs sampler. This is not to suggest that the orthogonal Gibbs sampler is generating i.i.d. sample points, but certainly updating every 1000 sample shows little difference in convergence. For the ozone, physical, and crime datasets, the orthogonal Gibbs sampler appears to be converging approximately twice as fast as the standard Gibbs sampler. The convergence rates per sample are much closer between the orthogonal and standard Gibbs sampler for the bodyfat data.

The fact that the orthogonal Gibbs sampler shows very little dependence can be confirmed by checking the auto-correlation function. As *k* increases, some small differences do appear between the orthogonal Gibbs sampler and exact sampling for convergence. This is due to the increased size of the state space. The most noticeable effect is a longer persistence in variability surrounding the estimates of $\mathbb{E}[q_\gamma]$, DIC and for the BMA $\hat{y}_1$. The standard Gibbs sampler clearly converges much more slowly than either of the methods in the **W** space. As a result so do the estimates of expected model size, DIC and BMA $\hat{y}_1$. The difference in the rate of convergence between **X** and **W**, is partly controlled by the difference in model competition or equivalently, how much more concentrated the posterior mass for **W** is compared to **X**. The bodyfat data has the closest values of model competition compared to **X**, while the crime data has the largest.


Convergence in Distribution
Sum Absolute Difference
0.4
0.3
0.2
0.1
0
1
2
3
4
5
Sample Size
x 10^4
Convergence in Expected Model Size
$E[q_\gamma]$
3.55
3.5
3.45
3.4
1
2
3
4
5
Sample Size
x 10^4
Convergence in DIC
DIC
97.6
97.4
97.2
97
1
2
3
4
5
Sample Size
x 10^4
Convergence in Fitted obs. 1
BMA y_1
1.6
1.55
1
2
3
4
5
Sample Size
x 10^4


Convergence in Distribution
Sum Absolute Difference
0.2
0.15
0.1
0.05
0
1
2
3
4
5
Sample Size
x 10^4
Convergence in Expected Model Size
$E[q_\gamma]$
3.45
3.4
3.35
3.3
3.25
1
2
3
4
5
Sample Size
x 10^4
Convergence in DIC
DIC
97.3
97.2
97.1
97
96.9
1
2
3
4
5
Sample Size
x 10^4
Convergence in Fitted obs. 1
BMA y_1
1.64
1.62
1.6
1.58
1
2
3
4
5
Sample Size
x 10^4

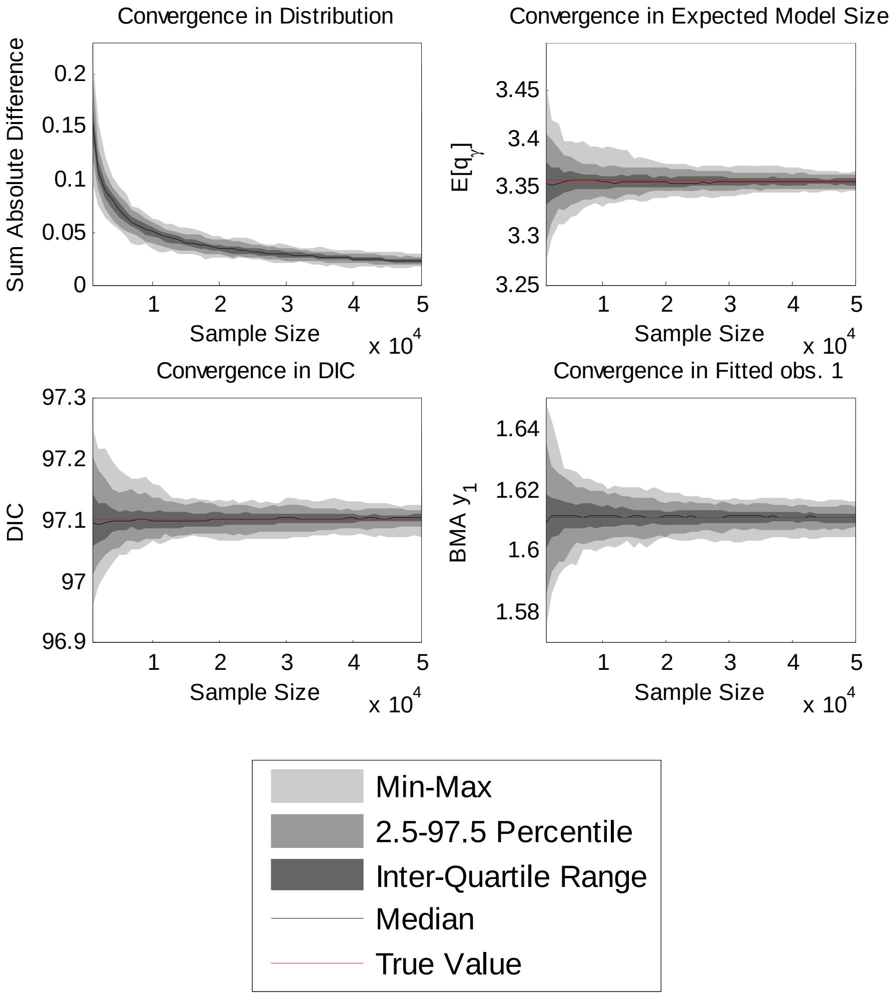
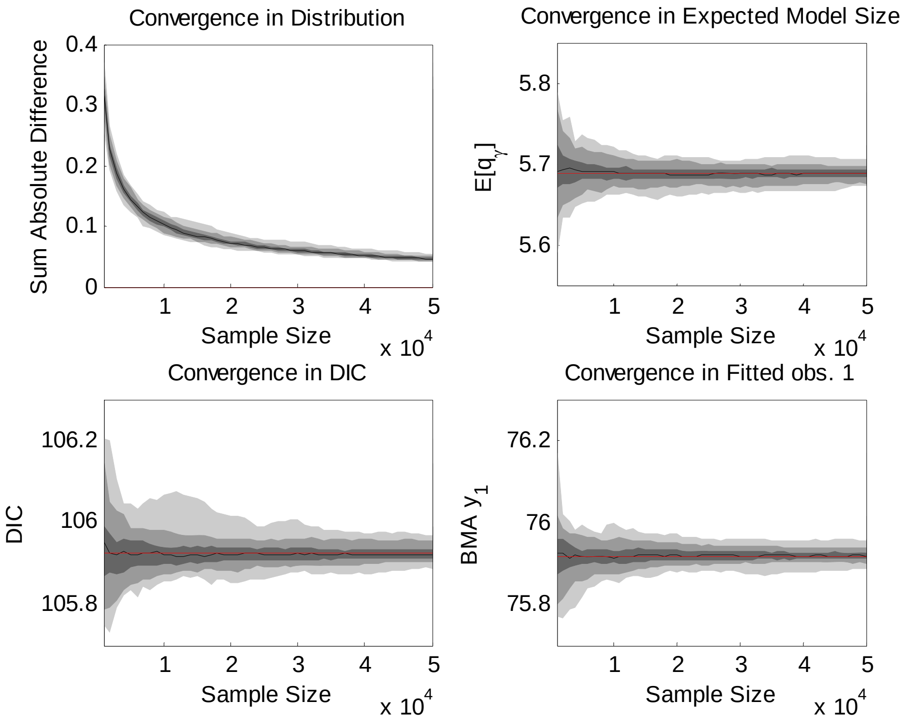
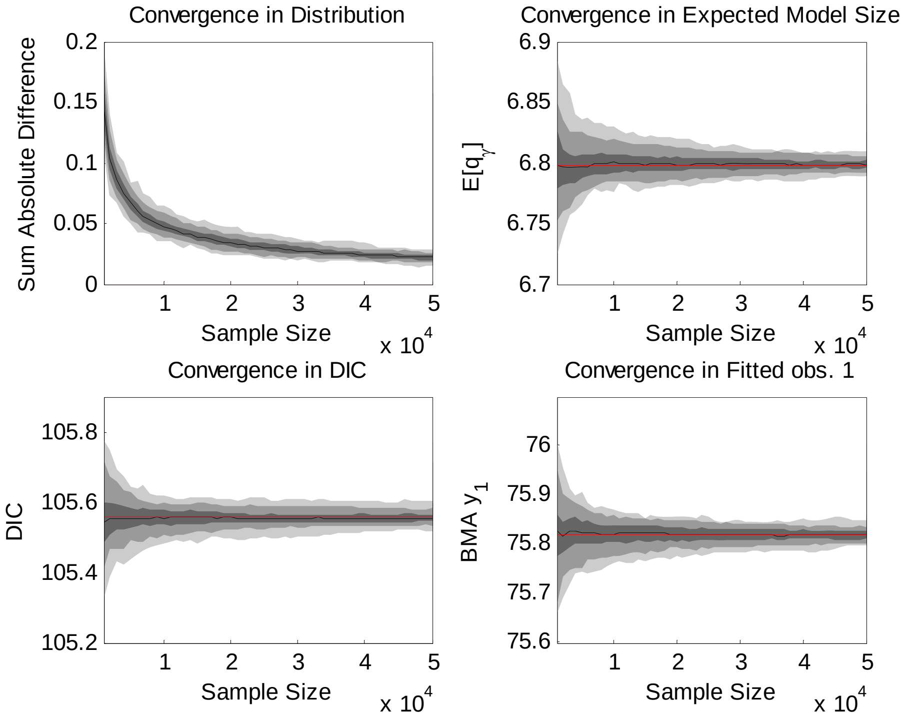


Figure 4.1 Convergence in distribution and various quantities for the ozone data using Jeffreys prior with a uniform prior for γ, p = 2π(n+1) and the GS$_2$ method for obtained W. Top 4 panels opposite is for the standard Gibbs sampler for X, bottom 4 panels opposite is for the monotone exact sampler using W, and the 4 panels above is for the standard Gibbs sampler with W.


Convergence in Distribution
Sum Absolute Difference
0.4
0.3
0.2
0.1
0
1
2
3
4
5
Sample Size
x 10^4
Convergence in Expected Model Size
E[q_γ]
5.8
5.7
5.6
1
2
3
4
5
Sample Size
x 10^4
Convergence in DIC
DIC
106.2
106
105.8
1
2
3
4
5
Sample Size
x 10^4
Convergence in Fitted obs. 1
BMA y_1
76.2
76
75.8
1
2
3
4
5
Sample Size
x 10^4


Convergence in Distribution
Sum Absolute Difference
0.2
0.15
0.1
0.05
0
1
2
3
4
5
Sample Size
x 10^4
Convergence in Expected Model Size
E[q_γ]
6.9
6.85
6.8
6.75
6.7
1
2
3
4
5
Sample Size
x 10^4
Convergence in DIC
DIC
105.8
105.6
105.4
105.2
1
2
3
4
5
Sample Size
x 10^4
Convergence in Fitted obs. 1
BMA y_1
76
75.9
75.8
75.7
75.6
1
2
3
4
5
Sample Size
x 10^4

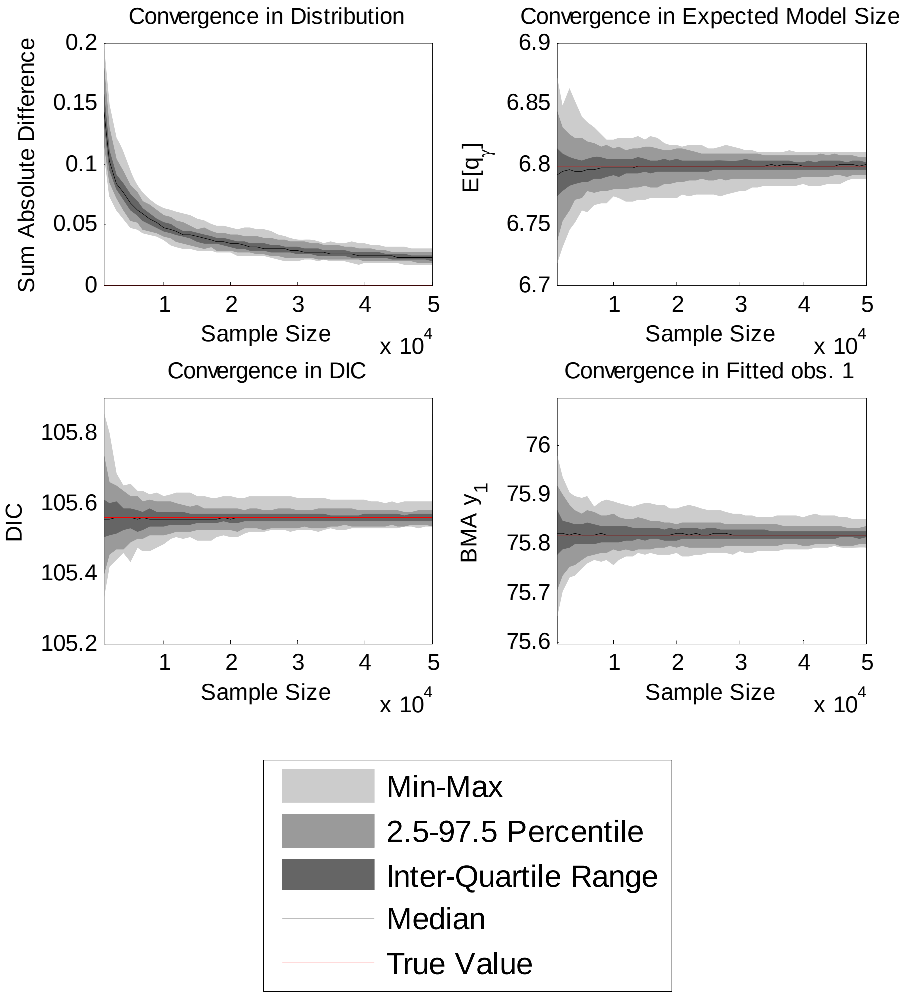
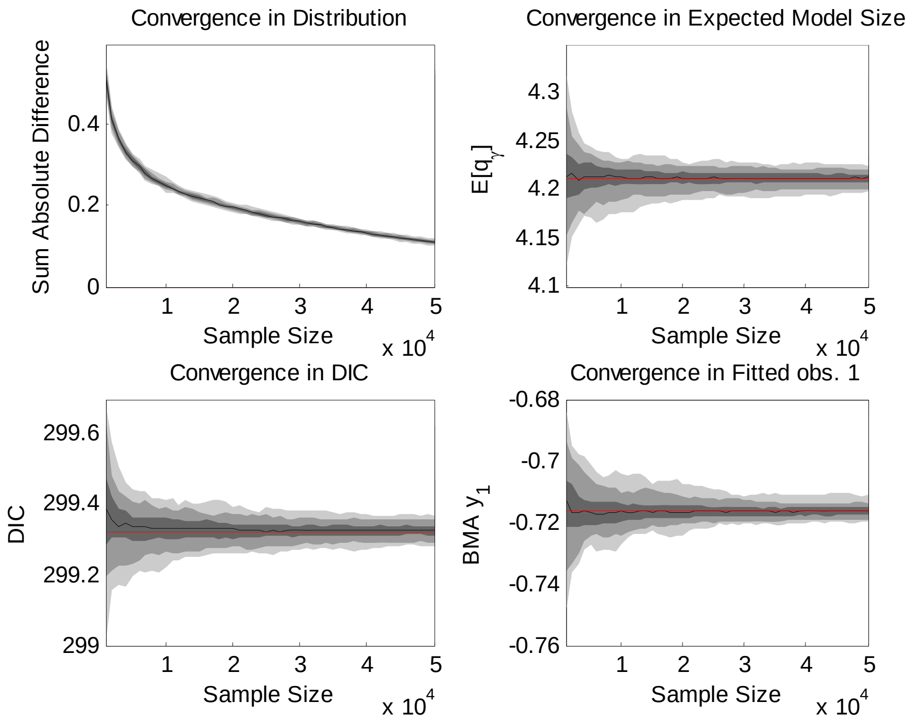
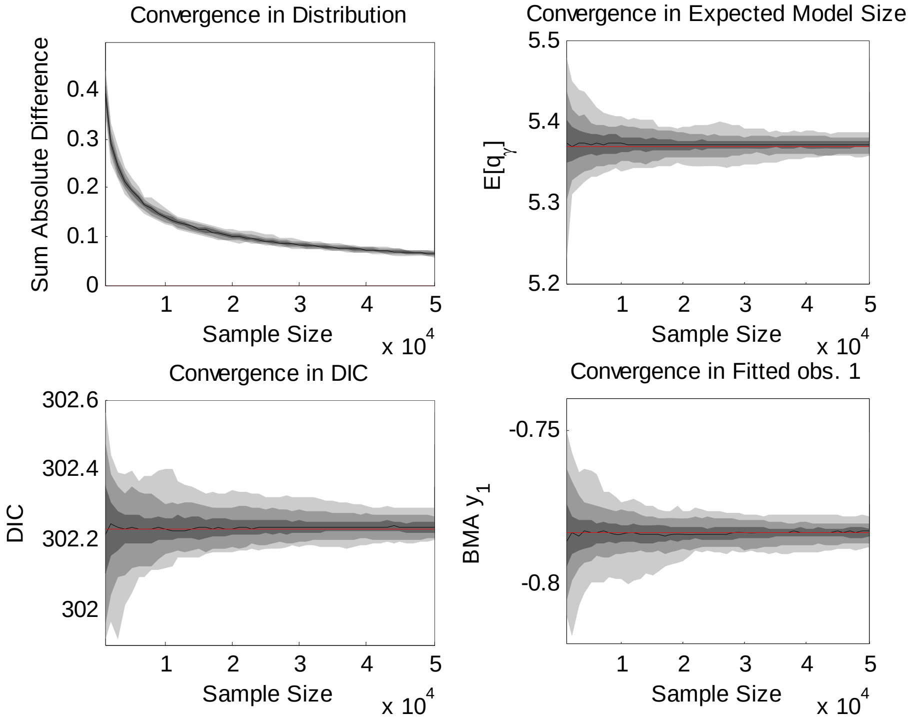


Figure 4.2 Convergence in distribution and various quantities for the physical data using Jeffreys prior with a uniform prior for γ, $p = 2\pi(n+1)$ and the GS$_2$ method for obtained W. Top 4 panels opposite is for the standard Gibbs sampler for X, bottom 4 panels opposite is for the monotone exact sampler using W, and the 4 panels above is for the standard Gibbs sampler with W.


Convergence in Distribution
Sum Absolute Difference
0.4
0.2
0
1
2
3
4
5
Sample Size
x 10^4
Convergence in Expected Model Size
$E[q_\gamma]$
4.3
4.25
4.2
4.15
4.1
1
2
3
4
5
Sample Size
x 10^4
Convergence in DIC
DIC
299.6
299.4
299.2
299
1
2
3
4
5
Sample Size
x 10^4
Convergence in Fitted obs. 1
BMA y_1
-0.68
-0.7
-0.72
-0.74
-0.76
1
2
3
4
5
Sample Size
x 10^4


Convergence in Distribution
Sum Absolute Difference
0.4
0.3
0.2
0.1
0
1
2
3
4
5
Sample Size
x 10^4
Convergence in Expected Model Size
$E[q_\gamma]$
5.5
5.4
5.3
5.2
1
2
3
4
5
Sample Size
x 10^4
Convergence in DIC
DIC
302.6
302.4
302.2
302
1
2
3
4
5
Sample Size
x 10^4
Convergence in Fitted obs. 1
BMA y_1
-0.75
-0.8
1
2
3
4
5
Sample Size
x 10^4

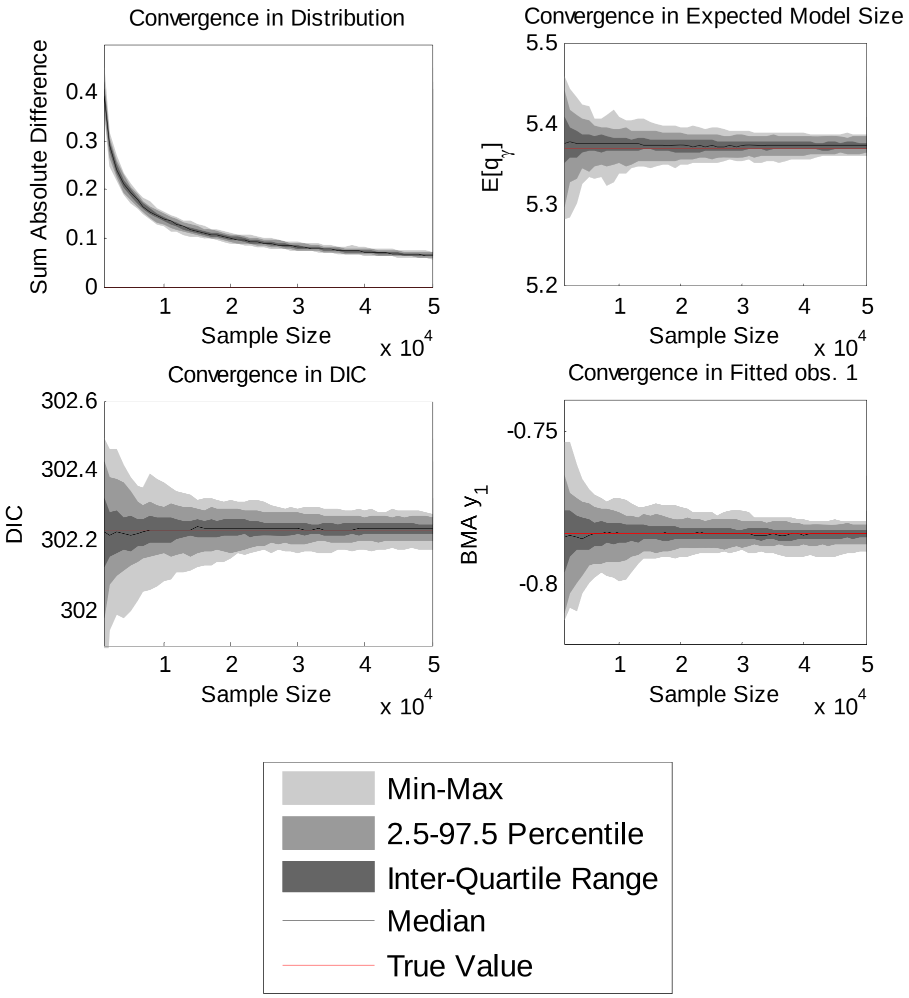
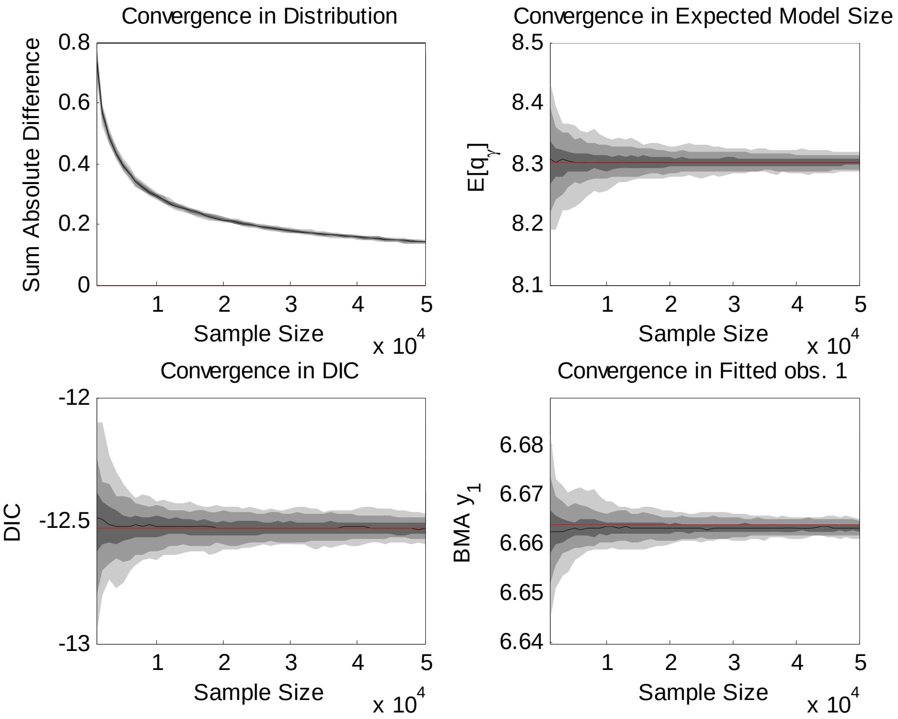
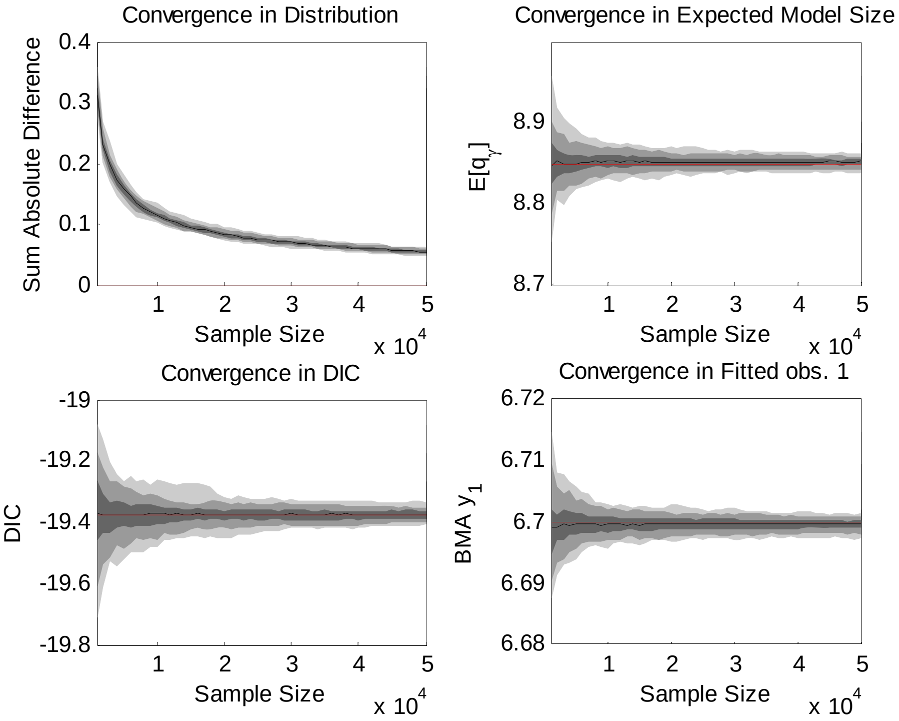


Figure 4.3 Convergence in distribution and fvarious quantities for the bodyfat data using Jeffreys prior with a uniform prior for γ, $p = 2\pi(n+1)$ and the GS$_2$ method for obtained W. Top 4 panels opposite is for the standard Gibbs sampler for X, bottom 4 panels opposite is for the monotone exact sampler using W, and the 4 panels above is for the standard Gibbs sampler with W.


Convergence in Distribution
Sum Absolute Difference
0.8
0.6
0.4
0.2
0
1
2
3
4
5
Sample Size
x 10^4
Convergence in Expected Model Size
E[q_γ]
8.5
8.4
8.3
8.2
8.1
1
2
3
4
5
Sample Size
x 10^4
Convergence in DIC
DIC
-12
-12.5
-13
1
2
3
4
5
Sample Size
x 10^4
Convergence in Fitted obs. 1
BMA y_1
6.68
6.67
6.66
6.65
6.64
1
2
3
4
5
Sample Size
x 10^4


Convergence in Distribution
Sum Absolute Difference
0.4
0.3
0.2
0.1
0
1
2
3
4
5
Sample Size
x 10^4
Convergence in Expected Model Size
E[q_γ]
8.9
8.8
8.7
1
2
3
4
5
Sample Size
x 10^4
Convergence in DIC
DIC
-19
-19.2
-19.4
-19.6
-19.8
1
2
3
4
5
Sample Size
x 10^4
Convergence in Fitted obs. 1
BMA y_1
6.72
6.71
6.7
6.69
6.68
1
2
3
4
5
Sample Size
x 10^4

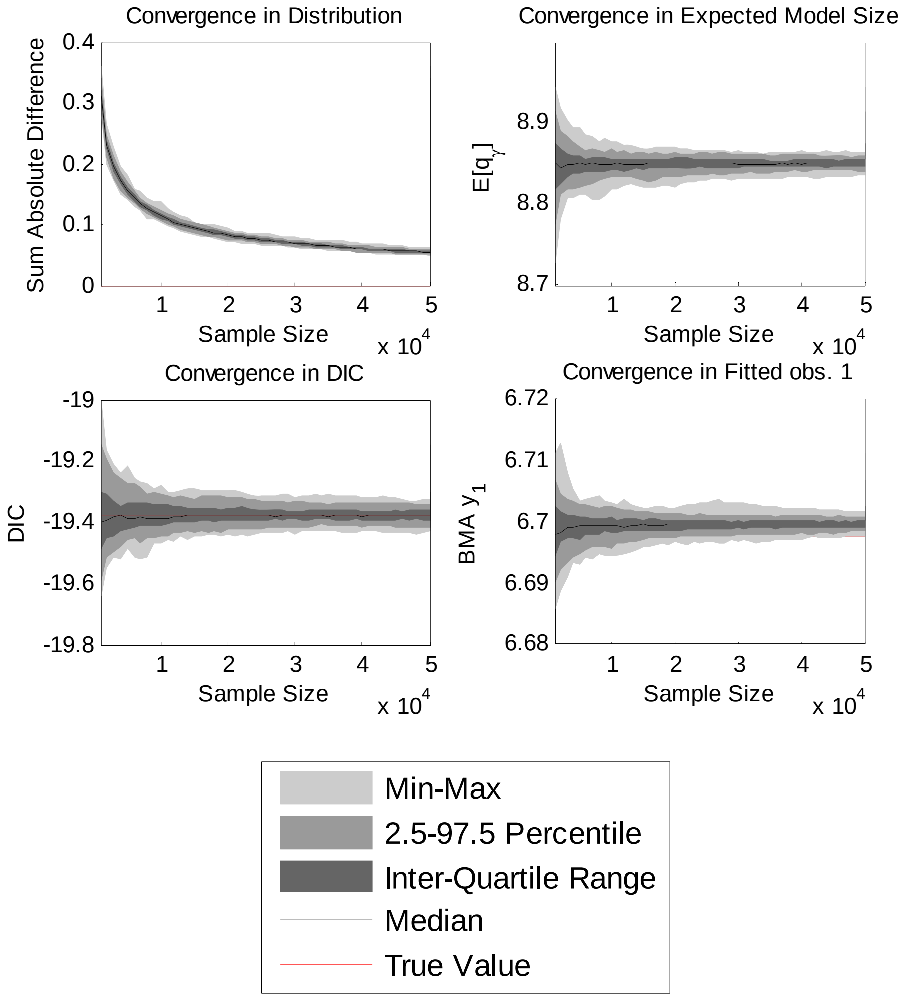


Figure 4.4 Convergence in distribution and various quantities for the crime data using Jeffreys prior with a uniform prior for γ, $p = 2\pi(n+1)$ and the GS$_2$ method for obtained W. Top 4 panels opposite is for the standard Gibbs sampler for X, bottom 4 panels opposite is for the monotone exact sampler using W, and the 4 panels above is for the standard Gibbs sampler with W.

The bodyfat data has the smallest difference in the rate of convergence of all datasets and the crime data has the largest. In all cases as k becomes larger and the state space increases in size, the value of (4.3) increases which is to be expected.

4.2.2 Convergence in real time

The results above indicate that on a per sample basis, the standard Gibbs sampler with an orthogonal predictor matrix and the exact sampler are by far the best approach, converging similarly to the posterior distribution for $\boldsymbol{\gamma}$. However, the reader may note that each approach requires different amounts of time to complete the required computation. Thus, a useful addition to the above investigation is to look at the convergence to $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$ in cpu-time. To this end, Figure 4.5 shows the median convergence statistic for each dataset for each of the three methods plotted against the cpu-time in seconds required to generate the number of samples used.

It is clear that after taking into account the computing time, the orthogonal Gibbs sampler is by far the best approach, as it has the superior convergence properties of an orthogonal design matrix and minimal computing time. For the ozone data the Gibbs sampler in $\mathbf{X}$ required 5.26 seconds of cpu-time per 1000 sample points, the orthogonal Gibbs sampler 1.65, and the exact sampler 5.57. For the physical data the standard Gibbs sampler used 6.93 seconds of cpu-time per 1000 sample points, while the orthogonal Gibbs sampler required 2.04, and the exact sampler 8.39. For the bodydata the Gibbs sampler with $\mathbf{X}$ used 8.70 seconds of cpu-time per 1000 sample points, the Gibbs sampler with $\mathbf{W}$ used 2.72, and the exact sampler used 9.36. Finally, for the crime data the standard Gibbs sampler required 11.31 seconds of cpu-time per 1000 sample points, the orthogonal Gibbs sampler 3.16, and the exact sampler 12.33. In general as k increases, more computing time is required and overall perfect sampling requires the greatest amount of computing time.

Interestingly the computing times between the perfect sampler and the Gibbs sampler using $\mathbf{X}$ are very similar. For these examples, the expected backwards coupling time is close to 2, although slightly larger for the physical and crime datasets where the greatest difference in computing time between the perfect sampler and the standard Gibbs sampler occur. Thus, from the study here, the orthogonal Gibbs sampler is approximately four times as fast as the standard Gibbs sampler.

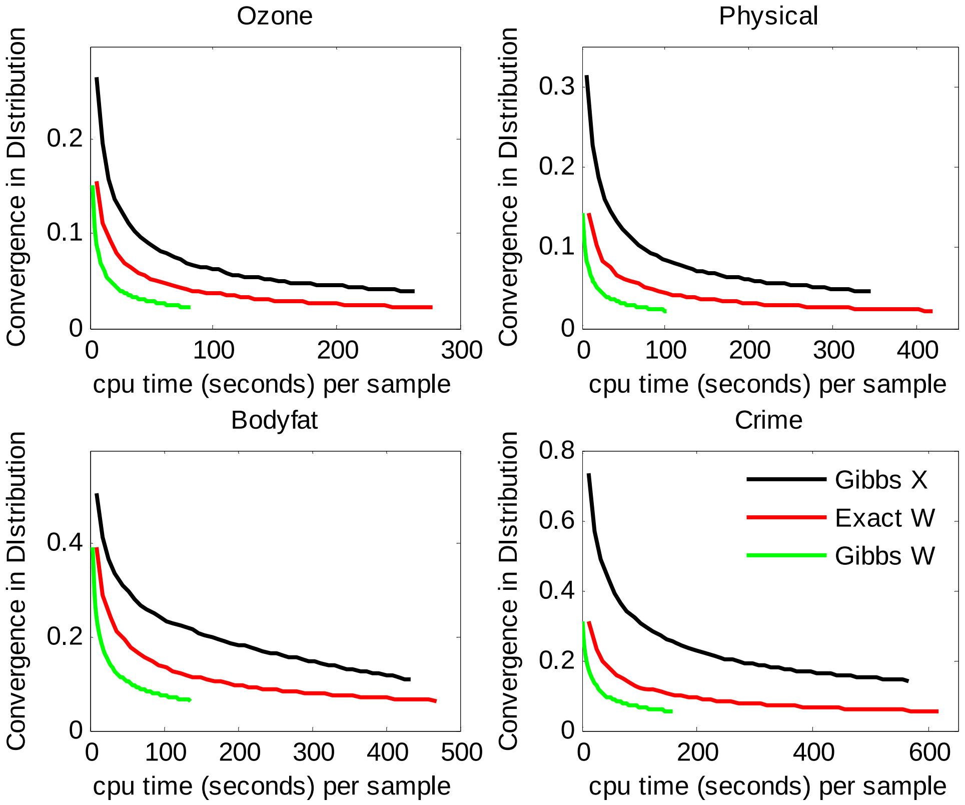


Figure 4.5 Convergence in cpu time for the four data sets.

As such, because the initial recursion of the perfect sampler is equivalent to four steps with the orthogonal Gibbs sampler, the exact sampler and standard Gibbs sampler are essentially equivalent in terms of computing time. The difference that has made the exact sampler take slightly longer than the standard Gibbs sampler is of course, those BCT which are greater than 2. This suggests if the BCT is extremely close to 2, then generating a single sample point for **X** is comparable to computing a single sample point using perfect sampling for **W**. Again we note it may be possible to improve upon the computation time of the Gibbs sampler using Cholesky updates, or the sweep algorithm. Clearly if the mean BCT was to increase dramatically, then the computational viability of the exact sampler compared to the Gibbs sampler in **X** may be called into question. This suggests knowledge of conditions that impact the BCT may prove useful in determining a choice between using the standard Gibbs sampler, and perfect sampling. Ultimately however, the standard Gibbs sampler using **W** will always be the fastest of the three methods. The performance of the orthogonal Gibbs sampler does bear consideration. Despite providing well approximated samples, the orthogonal Gibbs sampler nonetheless, requires a burn-in assessment. Thus, it seems prudent to recommend a hybrid approach between perfect sampling and standard MCMC. In particular, we use the monotone Gibbs CFTP to identify a starting point as time zero, and then allow the single Gibbs chain to continue forwards. This compromise takes full advantage of the reduced computing time of the orthogonal Gibbs sampler, and the use of exact sampling to remove the burn-in problem. The hybrid method will supplant the orthogonal Gibbs sampler in the following section.

4.2.3 Computational efficiency

To add to the analysis of computational time because the previous analysis involves relatively small datasets, we investigate the computational time involved in using a larger number of predictors and sample sizes. This study required simulation of multiple datasets. We use combinations of k = {30, 40, 50}, and n = {50, 100, 250, 500, 1000}. Within each choice of k the size of the true model was also varied according to v = {0.25k, 0.5k, 0.75k}. The simulated datasets did not have any extreme correlations introduced, as such, the correlation between predictors was typically between -0.4 and 0.4. For each case we generated 100 datasets and generated 50000 sample points. We use the standard Gibbs sampler on X (Gibbs), the hybrid method using the orthogonal Gibbs sampler (Hybrid) and the monotone CFTP Gibbs sampler (Exact). The median cpu-time to generate 1000 samples for each combination, along with the expected BCT for the exact sampler is recorded in Tables 4.1a-c. The focus is on computing time, so we do not include burn-in assessment for the standard Gibbs sampler. Clearly however, removing an initial run of values will increase the cost of the remaining sample points. Thus, the values reported are like a lower limit for the cpu-time per sample. The standard Gibbs sampler is started from randomly chosen values of $\boldsymbol{\gamma}$. Note we use the same priors and parameter specifications as in section 4.2.1.

Table 4.1a Computation time for the 3 sampling methods using simulated data, *k* = 30, *n* = 50, 100, 250, 500, 1000 and *v* = 7, 15, 23. The GS$_2$ method was used to orthogonalize the simulated data.

k	*n*	*v*	cpu-time (seconds)/1000 sample points		
			Exact (BCT)	**Gibbs**	**Hybrid**
30	**50**	7	70.50 (5.14)	22.44	5.51
		15	74.38 (5.44)	25.41	5.56
		23	65.34 (5.29)	29.91	5.62
	100	7	35.42 (3.31)	22.24	5.53
		15	36.31 (3.39)	25.37	5.60
		23	34.63 (3.36)	29.09	5.57
	250	7	28.64 (2.87)	22.72	5.67
		15	29.94 (2.93)	25.20	5.59
		23	28.03 (2.82)	29.89	5.62
	500	7	24.42 (2.43)	22.02	5.61
		15	25.13 (2.59)	25.60	5.69
		23	23.98 (2.36)	29.61	5.64
	1000	7	23.41 (2.07)	22.73	5.68
		15	24.37 (2.32)	25.41	5.78
		23	22.26 (2.11)	29.55	5.76

Table 4.1b Computation time for the 3 sampling methods using simulated data, *k* = 40, *n* = 50, 100, 250, 500, 1000 and *v* = 10, 20, 30. The GS$_2$ method was used to orthogonalize the simulated data.

k	*n*	*v*	cpu-time (seconds)/1000 sample points		
			Exact (BCT)	**Gibbs**	**Hybrid**
40	**50**	10	160.8 (9.43)	31.45	8.03
		20	164.9 (10.74)	34.91	8.05
		30	157.1 (9.46)	39.67	8.06
	100	10	65.02 (6.12)	31.83	8.06
		20	66.31 (5.41)	34.01	8.05
		30	62.63 (3.26)	39.86	8.06
	250	10	38.24 (3.40)	31.72	8.03
		20	38.94 (3.53)	34.03	8.10
		30	37.03 (3.44)	39.10	8.05
	500	10	31.42 (3.13)	31.74	8.09
		20	32.03 (3.32)	34.18	8.04
		30	29.98 (2.96)	39.59	8.08
	1000	10	25.41 (2.37)	31.59	8.05
		20	26.37 (2.40)	34.78	8.06
		30	24.26 (2.29)	39.82	8.09

Table 4.1c Computation time for the 3 sampling methods using simulated data, *k* = 50, *n* = 50, 100, 250, 500, 1000 and *v* = 12, 25, 38. The GS$_2$ method was used to orthogonalize the simulated data.

k	*n*	*v*	cpu-time (seconds)/1000 sample points		
			Exact (BCT)	**Gibbs**	**Hybrid**
50	**50**	12	269.5 (14.8)	43.12	10.11
		25	277.0 (16.3)	47.76	10.15
		38	265.8 (14.6)	51.22	10.18
	100	12	97.62 (7.40)	43.53	10.14
		25	99.01 (8.01)	47.92	10.16
		38	97.55 (7.43)	51.31	10.20
	250	12	46.97 (4.17)	43.08	10.17
		25	47.84 (4.34)	47.93	10.19
		38	46.11 (4.03)	51.11	10.21
	500	12	37.42 (3.21)	43.88	10.14
		25	38.03 (3.45)	47.89	10.15
		38	36.98 (3.33)	51.13	10.20
	1000	12	31.41 (2.33)	43.34	10.18
		25	33.37 (2.43)	47.45	10.19
		38	30.26 (2.34)	51.32	10.21

We can see the following patterns emerging. The hybrid sampler is unaffected by increasing sample size, and the number of predictors in the true model, but increases with increasing *k*. In particular, there is an approximate 2.2 second increase in cpu-time from *k* = 30 to 40, and from 40 to 50. The standard Gibbs sampler is also unaffected by increasing sample size however, it is affected by *k* and *v*. This is due to the calculation of the inverse covariance matrix. With increasing *k* this calculation requires more time and while sampling is proceeding, if the size of the true model *v* approaches the size of *k*, then on average the repeated calculation of the inverse requires more time. Compared to the orthogonal Gibbs sampler the standard Gibbs sampler for small *v* can require 2-3 times the required cpu-time, and for large *v* around 4-5 times the required cpu-time.

The patterns exhibited by the exact sampler require a bit more thought. The first is that with decreasing *n* and increasing *k*, the necessary computing increases due to an increase in BCT. The use of *v* complicates the relationship further. When *v* is close to *k* or 0, the BCT decreases reducing the amount of cpu-time required. However, the algorithm is also coded to take advantage of monotonicity so if the updated component is set to 1 for the lower chain, then the upper chain is also set to 1, without the need for computation. This means the cpu-time should also be decreasing with increasing *v* as more components will be 1 according to the lower chain. It is the interaction between these two aspects that results in the increase, and slightly larger decrease, in cpu-time as *v* increases towards *k*. The hybrid sampler is the most efficient method. Further, the results above agree with the previous results that the perfect sampler is competitive with the Gibbs sampler for **X** provided the BCT is close to 2. However, once burn-in is taken into account, the balance of efficiency will likely shift towards the exact sampler. This also provides an indication of how *k* and *n* affect the BCT namely, as the number of predictors increases or the sample size decreases towards *k* the BCT increases.

4.3 Backwards Coupling Time

With the computational and convergence aspects investigated, we now move to investigate the BCT, information, and the probability of coalescence, which are unique to exact sampling and affect the efficiency of the exact sampler.

4.3.1 Information and BCT

We now investigate how information, and choices of hyper-parameter, affect the backwards coupling time. Large BCT means the perfect sampler will be less computationally competitive with the orthogonal Gibbs sampler. In assessing information we use entropy. The measure of entropy for a univariate probability mass function is called the Shannon entropy. For the natural logarithm (*nat* units), the Shannon entropy of $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$ is

$$H(\boldsymbol{\gamma}) = -\sum_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}} f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) \log f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) . \tag{4.5}$$

We clarify now that as the measure of *H* becomes larger the entropy is increasing, which implies the uncertainty associated with the corresponding random variable is also increasing, corresponding to less information. We use the real datasets, and calculated (4.5) for Zellner's prior and Jeffreys prior for choices of *c* and *τ*, and we use all four methods for obtaining **W**. The results are shown in Figures 4.6-4.9.

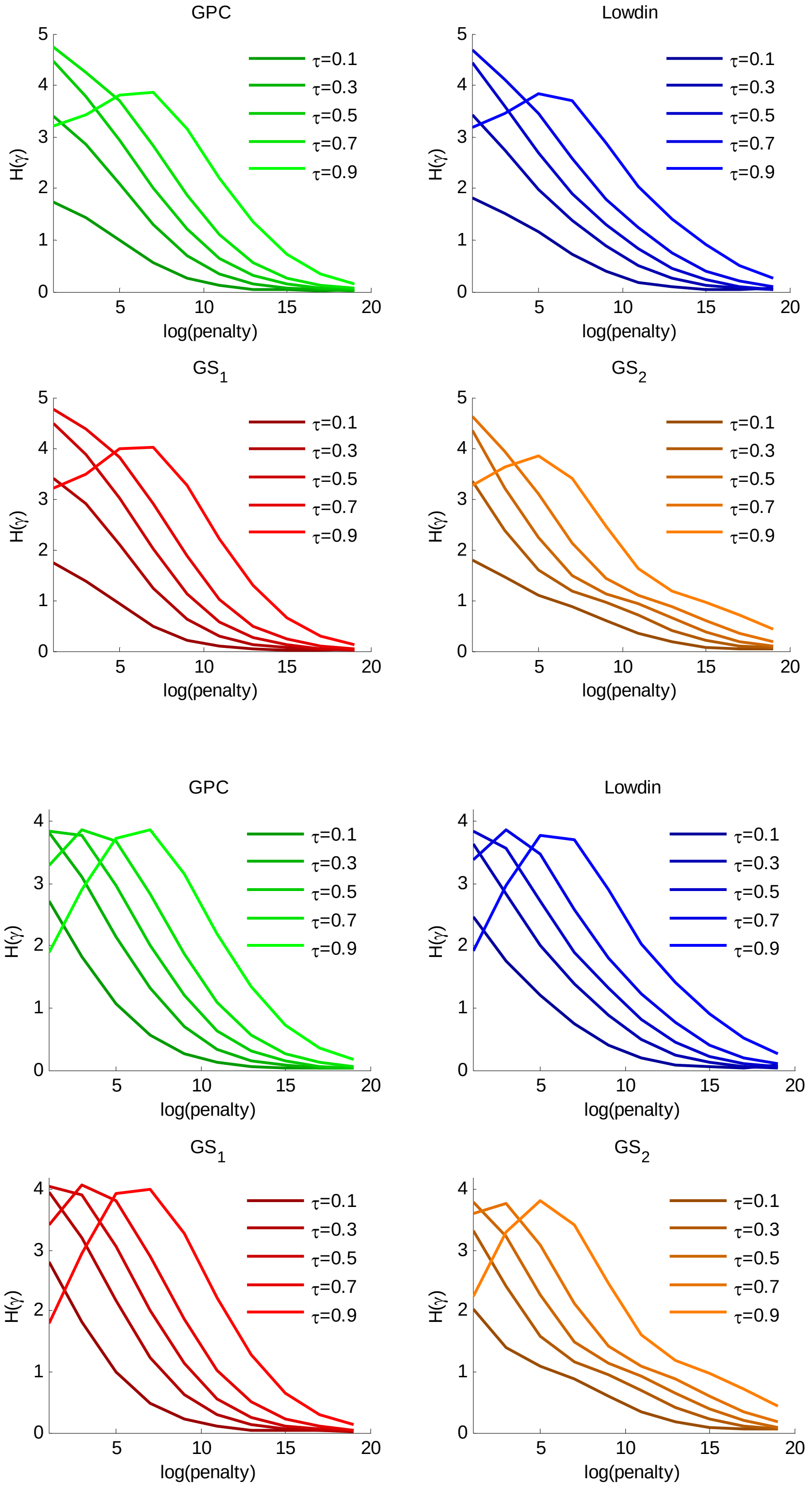


Figure 4.6 Entropy of $f(\gamma \mid \mathrm{y}, \mathrm{X})$ for the ozone data as penalty (*c* or *p*) increases for choices of *τ*. The top four panels are for Zellners' prior and the bottom four are for Jeffreys prior.

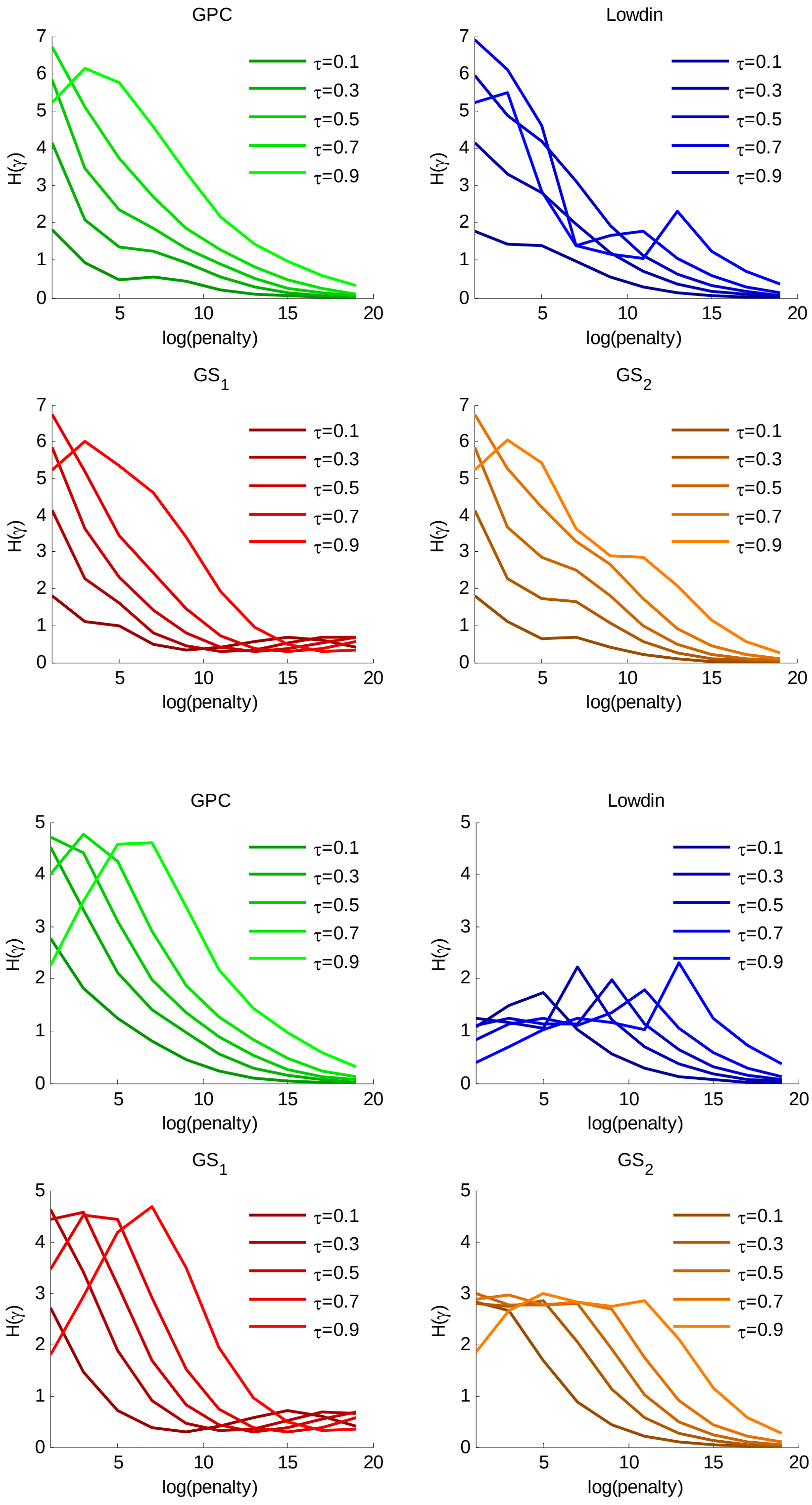


Figure 4.7 Entropy of $f(\gamma \mid \mathrm{y}, \mathrm{X})$ for the physical data as penalty (*c* or *p*) increases for choices of *τ*. The top four panels are for Zellners' prior and the bottom four are for Jeffreys prior.

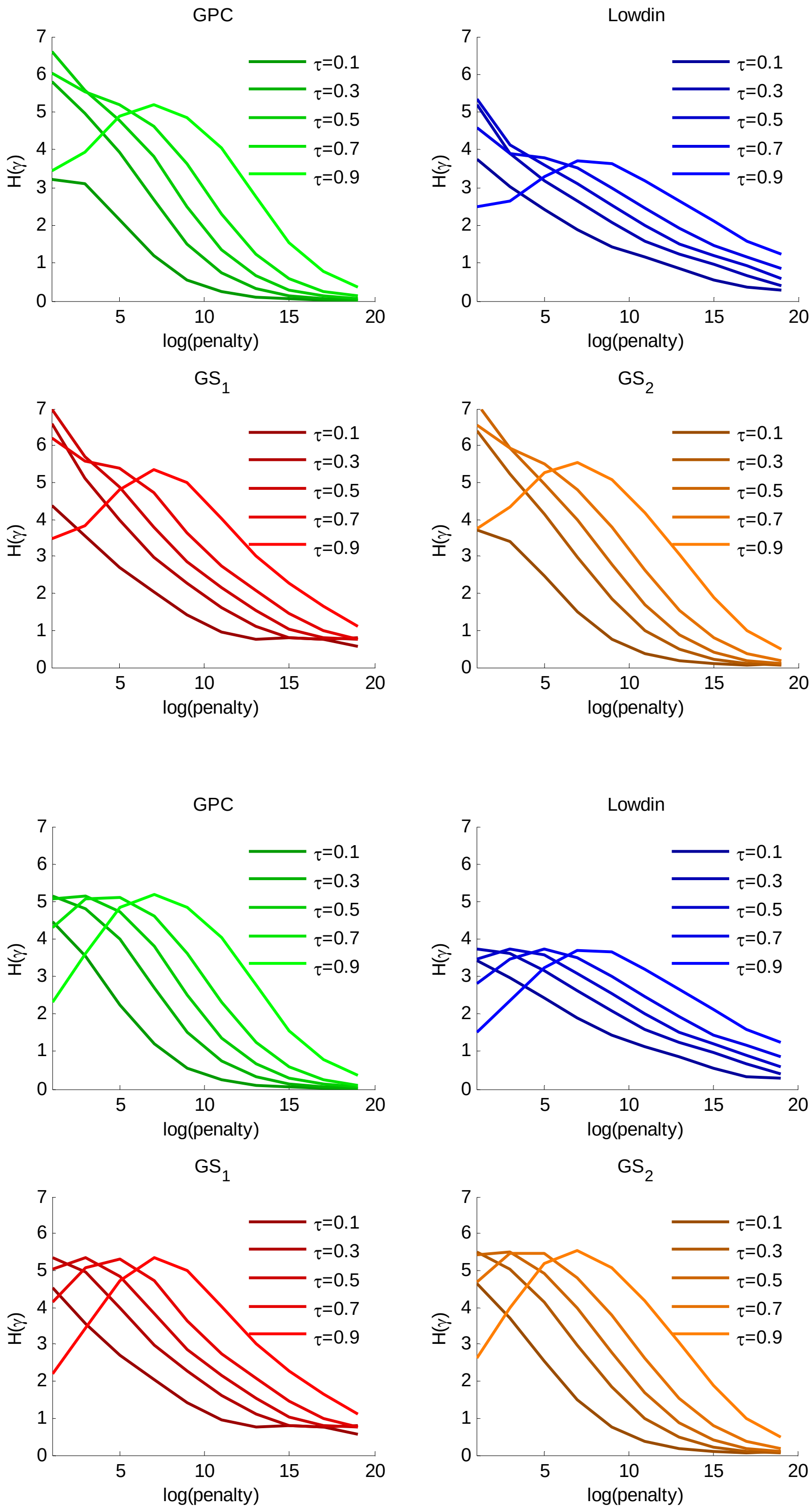


Figure 4.8 Entropy of $f(\gamma \mid \mathrm{y}, \mathrm{X})$ for the bodyfat data as penalty (*c* or *p*) increases for choices of τ. The top four panels are for Zellners' prior and the bottom four are for Jeffreys prior.

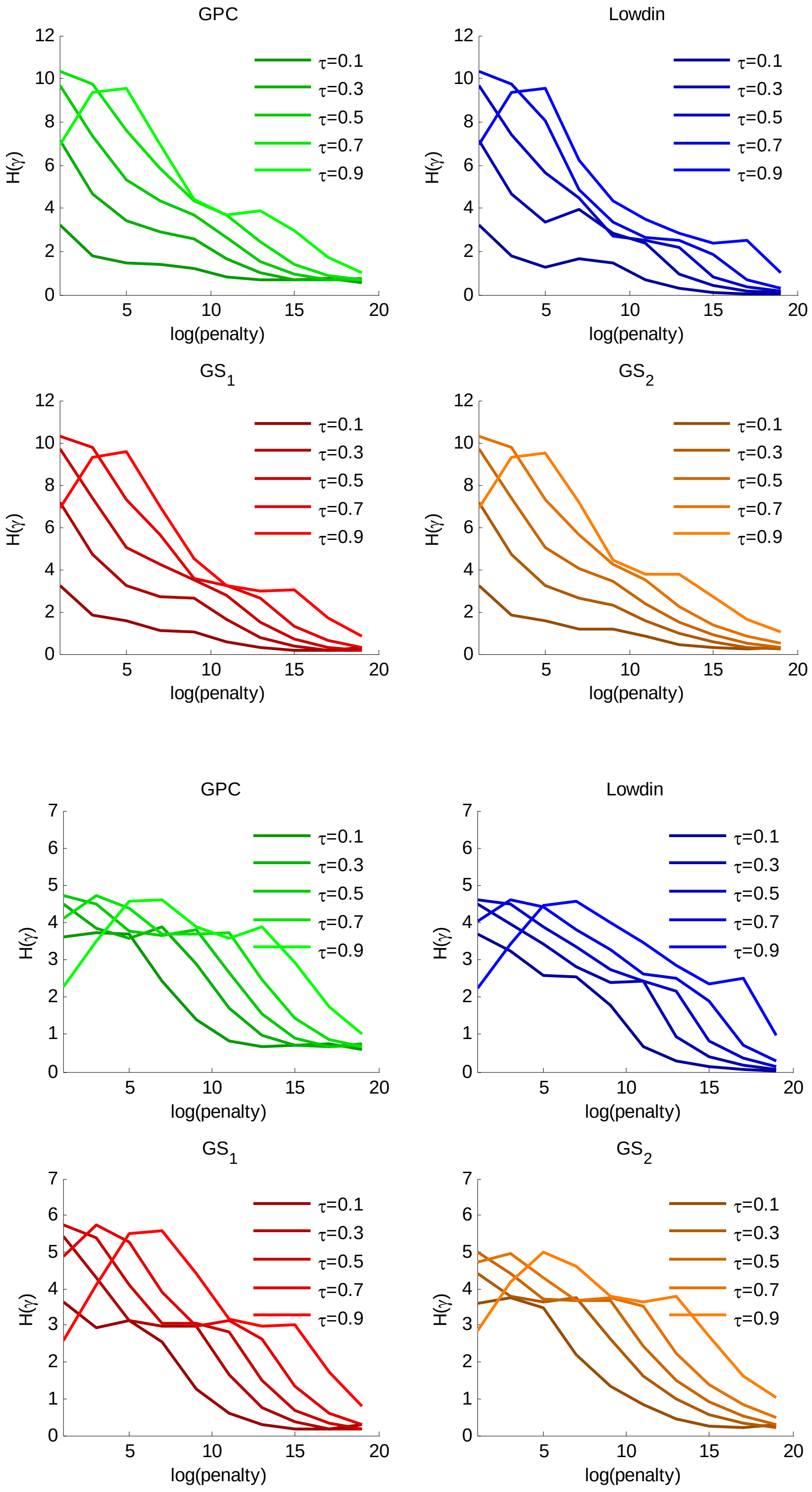


Figure 4.9 Entropy of $f(\gamma \mid y, X)$ for the crime data as penalty (*c* or *p*) increases for choices of τ. The top four panels are for Zellner's prior and the bottom four are for Jeffreys prior.

Figure 4.6 shows the entropy using (4.5) for the ozone data. The entropy decreases as c gets larger, and as τ increase the curve moves further to the right showing a peak in the extreme case of $\tau = 0.9$. Notice that the entropy curves are all extremely similar for all orthogonalization methods. The entropy curves for the physical data show similar behavior to that of the ozone data for choices of τ and penalty. However, the entropy shows increased variability, and some differences have appeared between the **W** methods. The most notable difference between the methods is the Lowdin, which shows much lower entropy for small values of penalty for both Zellner's prior and Jeffreys prior. The behavior of the entropy at lower values of penalty is most noticeable for Jeffreys prior.

For the bodyfat data again the Lowdin method stands out as having much less entropy over moderate choices of penalty than the other **W** methods. This applies to both Zellner's prior and Jeffreys prior and again, lower values of penalty for Zellner's prior produces more variability in the values of entropy. Similar changes in entropy to the previous datasets over τ and penalty are also exhibited. Finally, for the crime dataset while the entropy profiles appear slightly more erratic than for previous datasets, we observe similar profiles within each prior for all **W** methods. Again we see the flattening aspect for lower values of penalty for Jeffreys prior compared to Zellner's prior. It is clear the entropy is decreasing for extreme choices of τ and penalty, and that entropy is similar to the model competition statistic introduced in Chapter 1 and used in Chapter 3. Both measures provide some indication of the concentration of mass in the posterior, although the model competition measure has a more direct interpretation. With the indications of entropy in mind, we now investigate the BCT.

We do not have a closed form expression to provide the parameter of a geometric distribution to describe the distribution of backwards coupling times. We explore choices of τ with the penalty fixed at $c = n$ for Zellner's prior and $p = 2\pi(n+1)$ for Jeffreys as before, and choices of penalty with τ fixed at 0.5. From Figures 4.10 and 4.11 we can see that in general, the BCT varies between the datasets and **W** methods.

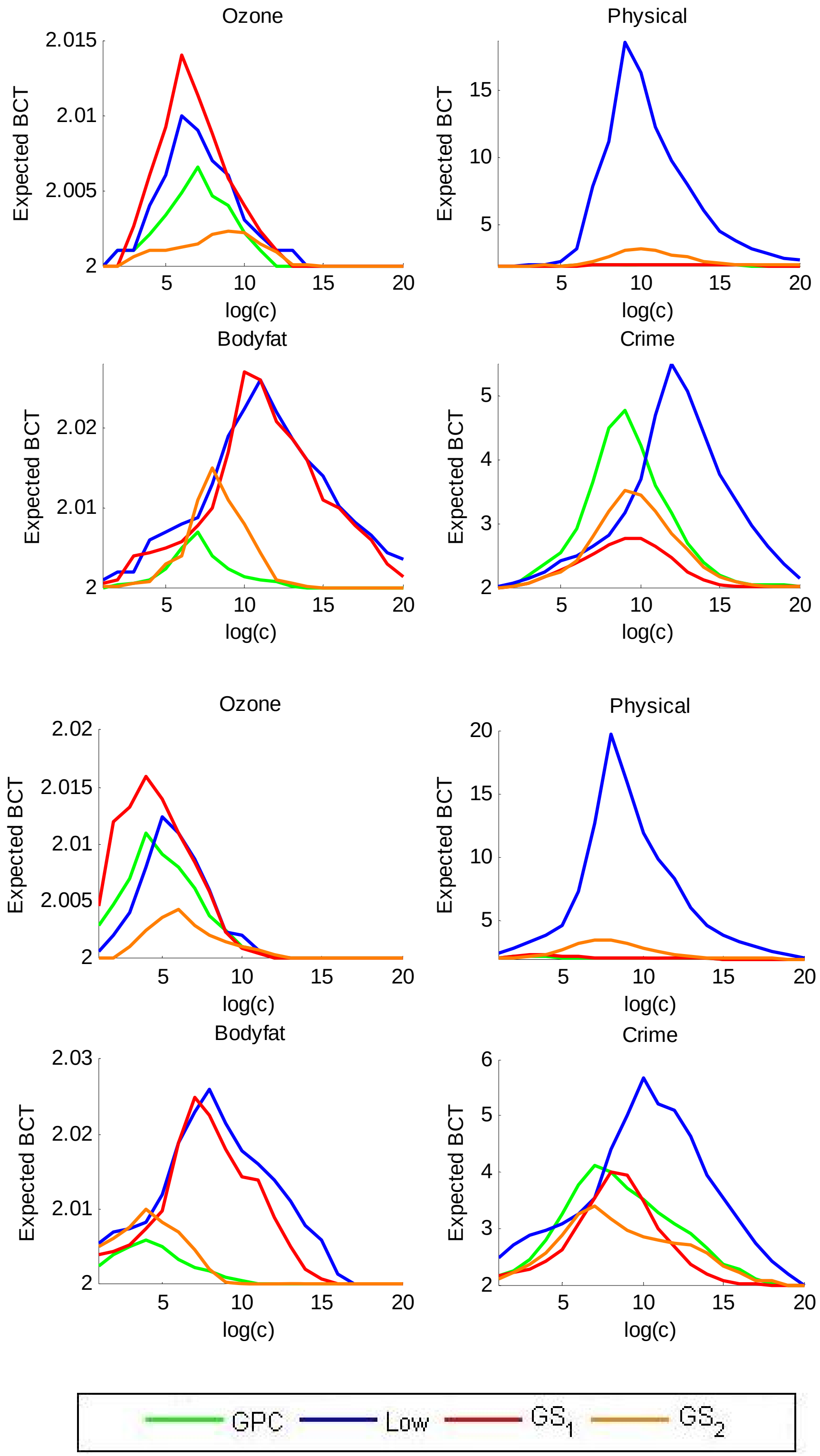


Figure 4.10 Estimated backwards coupling time over a range of penalty for four real datasets using all W methods for Zellner's prior (top four panels) and Jeffrey prior (bottom for panels). Notre for Jeffreys prior *c* = *p*/(2π)-1.

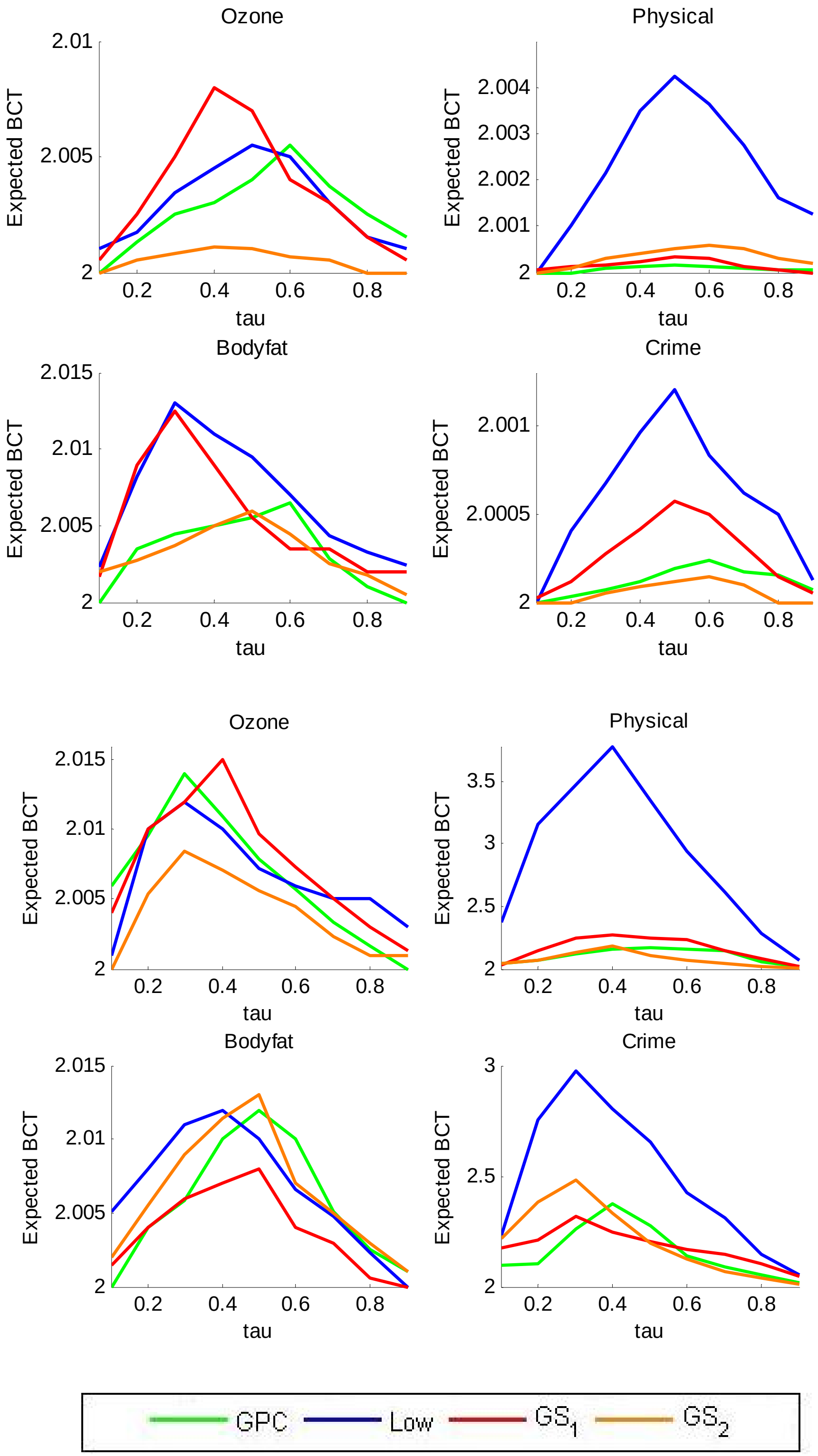


Figure 4.11 Estimated backwards coupling time over a range of τ for four real datasets using all W methods for Zellner's prior (top four panels) and Jeffrey prior (bottom for panels).

The differences can be attributed predominantly to sample size and information as the physical data has the smallest sample size (27) and ratio of n:k (2.7) followed by the crime data with 47 and 3.13. The larger datasets, ozone with n = 80 and a ratio of n:k of 10, and bodyfat with 250 and 19.2 both have BCT very close to 2. This indicates that for datasets with n much larger than k, the BCT of the perfect sampler will be minimal. In all cases for choices of τ and penalty the BCT displays convex behavior.

For choices of penalty with τ fixed at 0.5, we see the BCT increases as τ increase from 0 and then decrease again while approaching 1. This represents increasing and then decreasing model competition, which is increasing and then decreasing entropy. For choices of penalty, again as the penalty increases the BCT increases. As the penalty becomes more extreme, the BCT decreases. This is due to increasing and then decreasing model uncertainty or equivalently increasing and then decreasing entropy. Thus, it is possible to reduce the BCT of the exact sampler with extreme values of τ or penalty.

Comparing the orthogonalization methods we can see again these results are in line with the entropy and model uncertainty. For the ozone data, it appears that the GS_2 method predominantly has smallest BCT, followed by the GPC, Lowdin and GS_1 methods over both τ and c, for Zellner's prior and Jeffreys prior. For the physical data using Zellner's and Jeffreys prior and a choice of penalty, the largest BCT is attained by the Lowdin method, then GS_2, GS_1 and GPC. For Zellner's prior and Jeffreys prior over choices of τ the order is Lowdin method, then GS_1, GS_2 and GPC. For bodyfat data the Lowdin and GS_1 methods again are associated with the larger BCT, and the GS_2 and GPC methods smaller backwards coupling times. Finally, for the crime data the Lowdin method has the largest BCT, with the GS_1, GS_2 and GPC methods all relatively similar.

4.3.2 Coalescence

The BCT is related to the probability of coalescence in a single sweep of the Gibbs sampler. This in turn, is related to the sequence of probabilities for the coalescence of each component. The maximum probability of remaining undecided for the update of a single component is

$$\Pr(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i} = \{1,...,1\}) - \Pr(\gamma_i = 1 \mid \boldsymbol{\gamma}_{-i} = \{0,...,0\}) \tag{4.6}$$

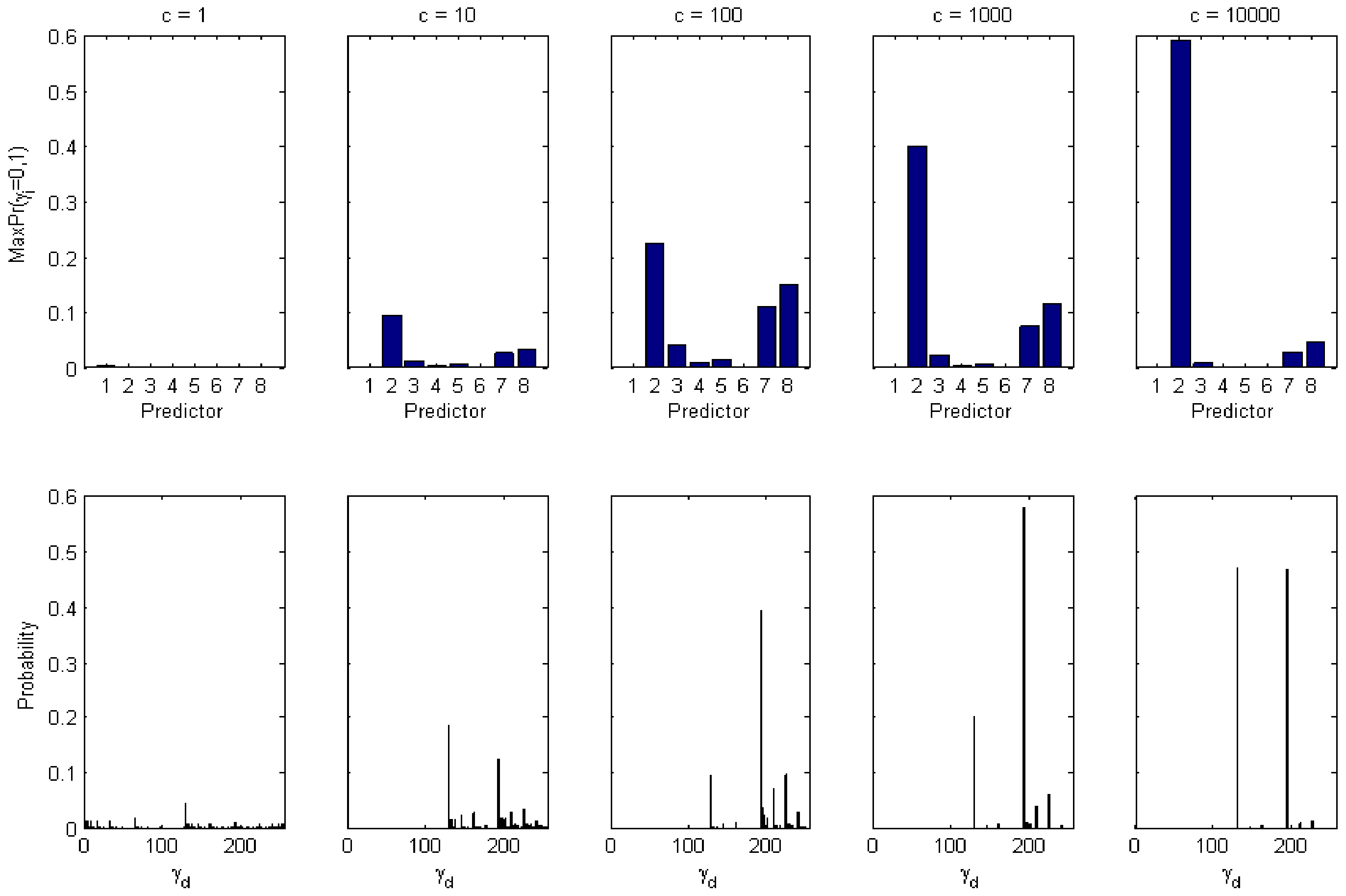


Figure 4.12a Probability of coalescence for each component for the ozone data using the GS$_2$ method. The posterior distribution is also provided to show the relation between the probability of coalescence and the distribution of mass for γ. Note $\tau = 0.5$.

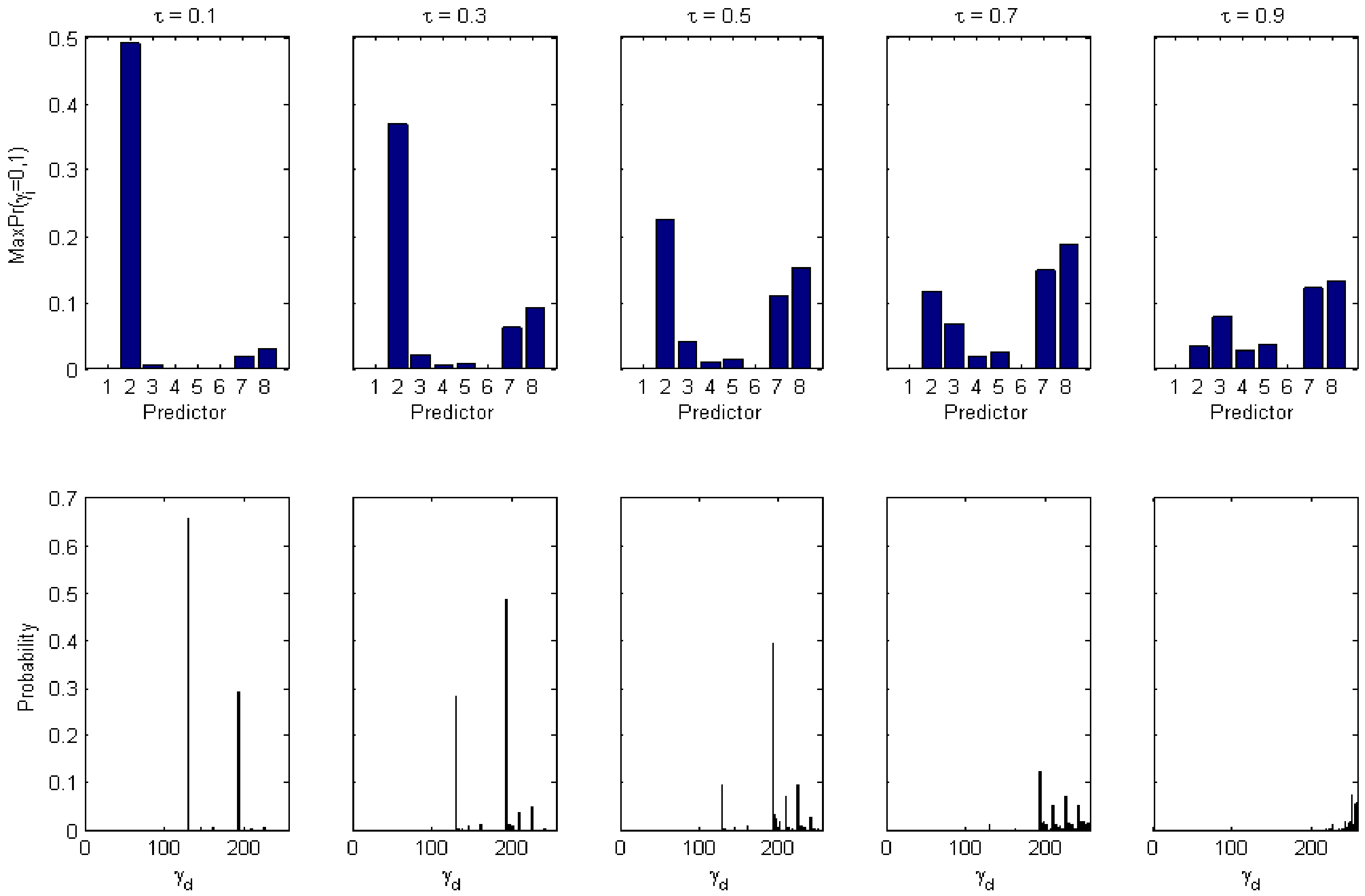


Figure 4.12b Probability of coalescence for each component for the ozone data using the GS$_2$ method. The posterior distribution is also provided to show the relation between the probability of coalescence and the distribution of mass for γ. Note *c* = *n*.

The greater (4.6) the less likely a predictor will couple to a single value, increasing the BCT. In a sense these distances also indicate the level of dependence between components. If the distance in (4.6) is zero for each component, then the components are independent and the posterior is a collection of independent Bernoulli random variables. Under orthogonal transformation these distances can become very small. However, the value of (4.6) can never be zero as this would require the residual sum of squares to be non-decreasing with the number of predictors, which is not true for Gaussian least squares. We also provide the posterior for $\boldsymbol{\gamma}$, to show the relation to the probability of remaining undecided.

As τ increases the posterior moves towards the full model, as this happens, the probability of remaining undecided becomes more evenly spread over the predictors. Of particular interest is when $\tau = 0.1$, or when c is large, both producing a posterior that is distinctly bimodal. The bimodality is a result of the penalty. The two competing models in decimal notation are 128 and 192, and correspond to the $\boldsymbol{\gamma}$ vectors {1 0 0 0 0 0 0 0}, and {1 1 0 0 0 0 0 0}, see Figures 4.12a-b. Thus, we observe a great detail of uncertainty surrounding the inclusion of the second predictor in **W**. This results in an increased expected BCT. This is curious as the previous results with entropy mean for this scenario the entropy would be minimal, i.e. assuming 2 competing models with equal probability (4.5) is approximately 0.7. However for this scenario the less overlap in terms of included predictors the greater the BCT thus, the BCT can be influenced by model competition even in the minimal case of two competing models. For lower values of τ and for intermediate values of penalty the BCT increases as the posteriors exhibit more model competition (greater entropy) and so as a result, the probabilities given by (4.6) increase.

Clearly the BCT is influenced by the amount of information and in particular, when k is large and n is small the BCT increases. Choices of hyper-parameters such as τ and penalty that maximize model competition will also drive up the BCT. Finally, bi-modality in the posterior for $\boldsymbol{\gamma}$ can also produce an increased BCT due to a large uncertainty about whether a predictor should be included or not.

4.4 Examples

We now attempt to emulate the conditions an analyst might face in the real world. We use two real datasets. The first is the body measurement data which records 21 body dimension measurements as well as age, weight, height, and gender on 507 individuals. Weight is the response variable. The second is the baseball dataset, which records 27 performance statistics for major league baseball players (excluding pitchers) in the 1991 season. The players salary in 1992 is the response variable. Further details can be found in Appendix C. We perform the analysis using **X** and the GS_2 orthogonalization method. We use the Jeffreys prior as n for both datasets is large, 507 and 337 respectively, and set $p = 2\pi(n+1)$ with a uniform prior on $\boldsymbol{\gamma}$.

For both datasets we run the standard Gibbs sampler on **X**, the hybrid method for **W**, and the perfect sampler for **W**, for 100000 iterations. The Gibbs sampler for **X** unlike the samplers using **W**, requires burn-in assessment. We used three sub chains of length 50000 to help assess convergence. We inspected the auto-correlation and partial auto correlation functions, while applying default methods such as Gelman and Rubin's, Geweke's, and Heidelberger and Welch's convergence diagnostics. These diagnostics can all be implemented routinely in the software R, using the CODA package. The burn-in was then discarded prior to inference. We record the expected model size, DIC, state space explored, predictive coverage for **y** (see 3.26), and tail probabilities for the minimum and maximum of **y**, using 1000 samples of **y** for each model sampled. These estimates are recorded in Table 4.2.

Plots of the 95% predictive regions for **y** along with the fitted values (first 50 observations) using model averaging for the baseball measurement data are given in Figure 4.14, and for the body data Figure 4.16. Plots of the estimated $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$ are given in Figures 4.13 and 4.15. It is clear that using the orthogonal transformation is comparable for inference about **y**, both from the recorded predictive coverage and from Figures 4.14 and 4.16. The predictive coverage is the same for the baseball data, and for the body measurement data.

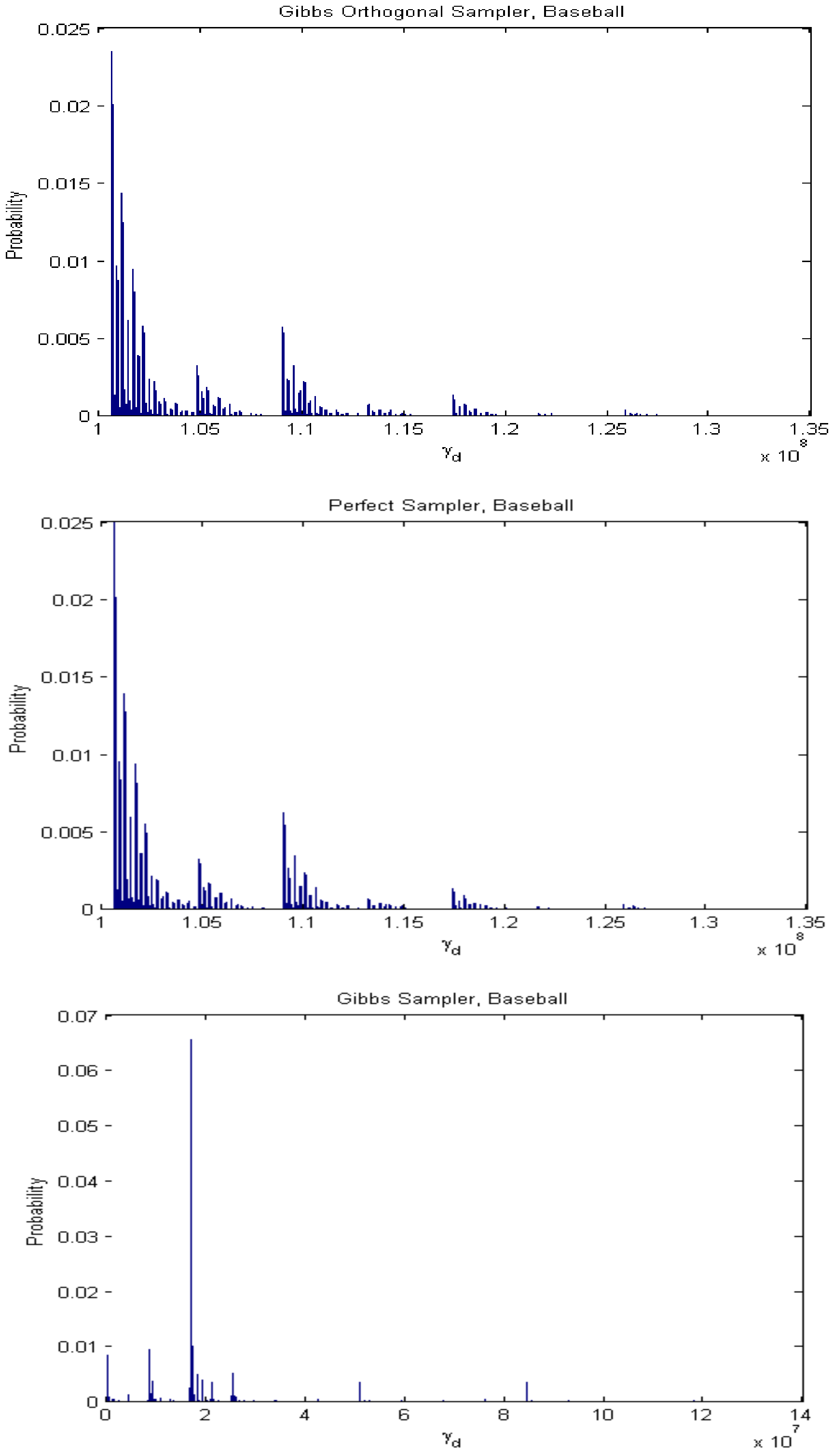
Table 4.2 Results for the Body measurement and Baseball datasets

DataSet	**Method**	$\mathbb{E}[q_\gamma]$	DIC	**$\|\Gamma\|$ (%)**	Min	Max	Coverage
Body $n/k \approx 25$	Gibbs **X**	9.07	270.76	9128 (0.054)	0.16	0.12	99.6%
	Hybrid **W**	12.2	273.25	8064 (0.048)	0.17	0.12	99.4%
	Exact **W**	12.2	273.43	8174 (0.049)	0.17	0.12	99.4%
Baseball $n/k \approx 12$	Gibbs **X**	7.45	685.02	17645 (0.013)	0.10	0.13	98.8%
	Hybrid **W**	6.30	686.46	21058 (0.016)	0.09	0.14	98.8%
	Exact **W**	6.30	686.38	21147 (0.016)	0.09	0.14	98.8%

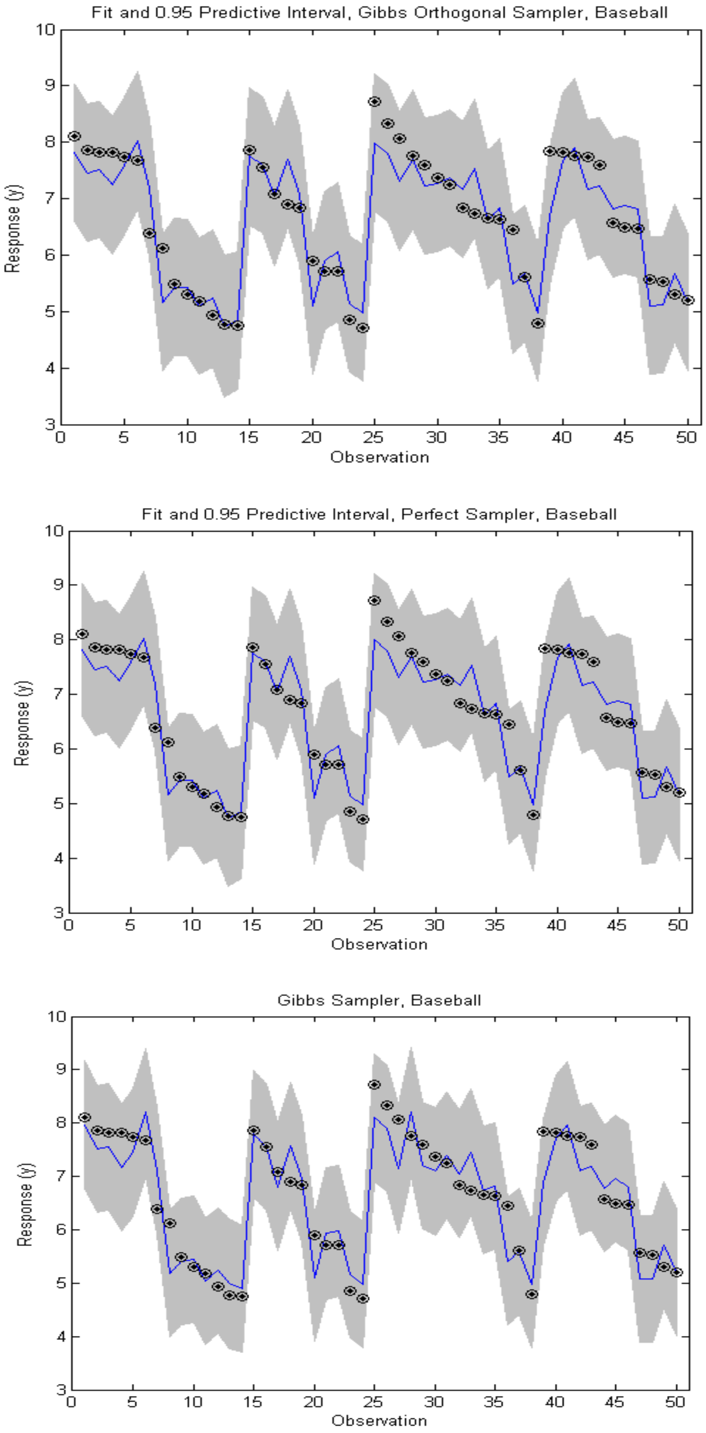
For both datasets the difference in predictive coverage between **X** and **W** is 1 observation, this is not overly concerning given the sample sizes involved. From Figures 4.13 and 4.15 the posteriors are the same for the perfect sampler and orthogonal Gibbs sampler. In the case of the body measurement data the posterior using **X** appears more concentrated than for **W**, however it is possible that 100000 samples was simply not enough to allow sufficient exploration of the state space. For the baseball data, a similar case is presented where the posterior for **X** has a larger concentration of mass at the mode for the standard Gibbs sampler. Again dependence and insufficient exploration may be responsible.

The expected model size under the posterior varies, for the body data the expected model size for **W** involves 3 more predictors than that for **X**. For the Baseball data, the expected model size equates to 1 more predictor for **X** than **W**. The number of models explored, shows that for the body data the Gibbs sampler explored more models than the orthogonal based methods, while for the Baseball example it explored much less. The values in brackets indicate the percentage of the entire state space explored, in all cases less than 0.1 of a percent of the space was explored. There was slightly more variation in the tail probabilities for the minimum and maximum however, all methods are close and indicate adequate performance using **W**.

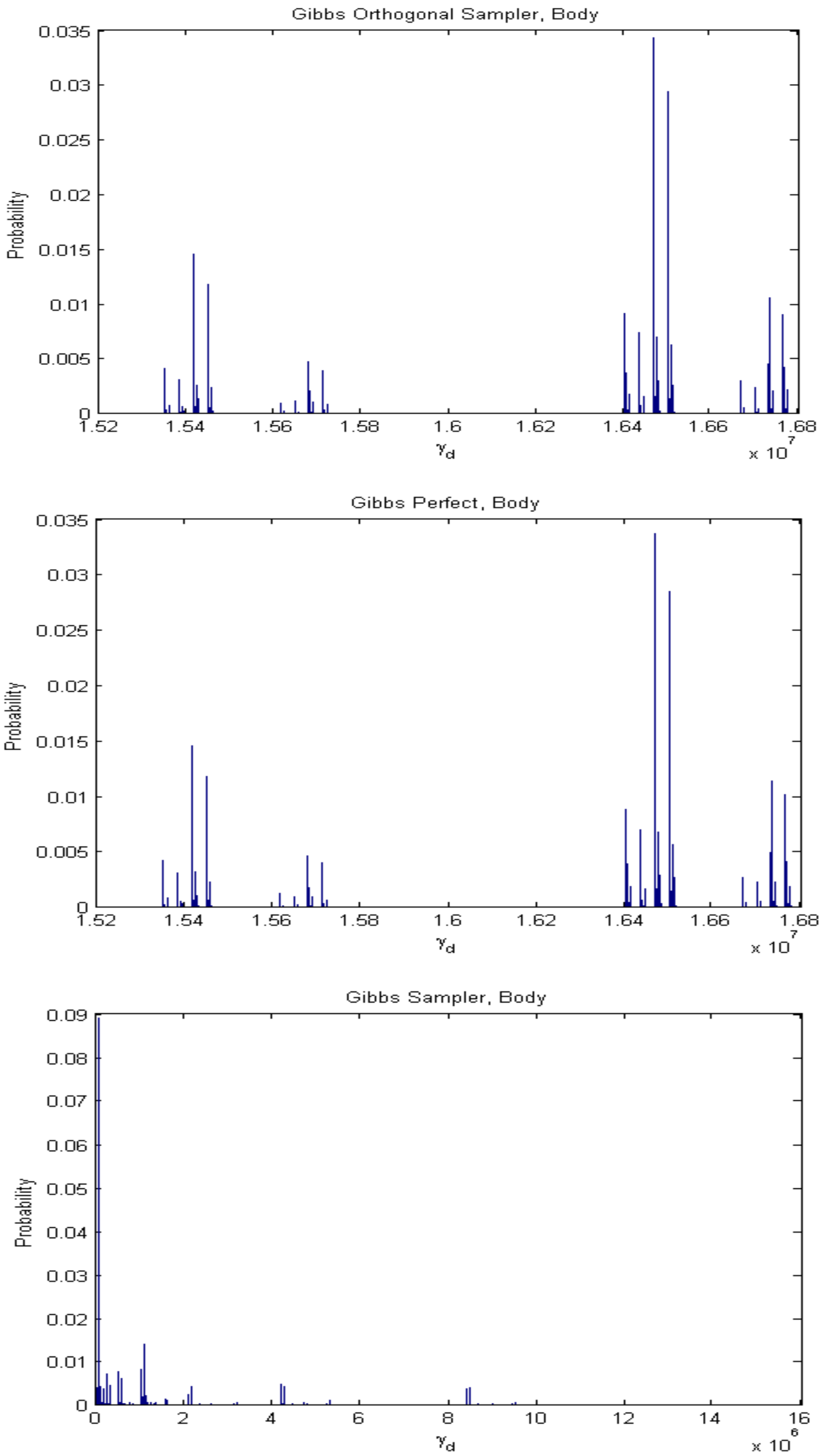
In the plots of the 95% predictive region around the fitted values shows that the response is fitted well after averaging over the number of models explored by each sampler. In particular in the case of each data set the differences in these plots are minor. The posteriors for γ under **W** for both data sets are almost identical for the orthogonal Gibbs and perfect samplers.



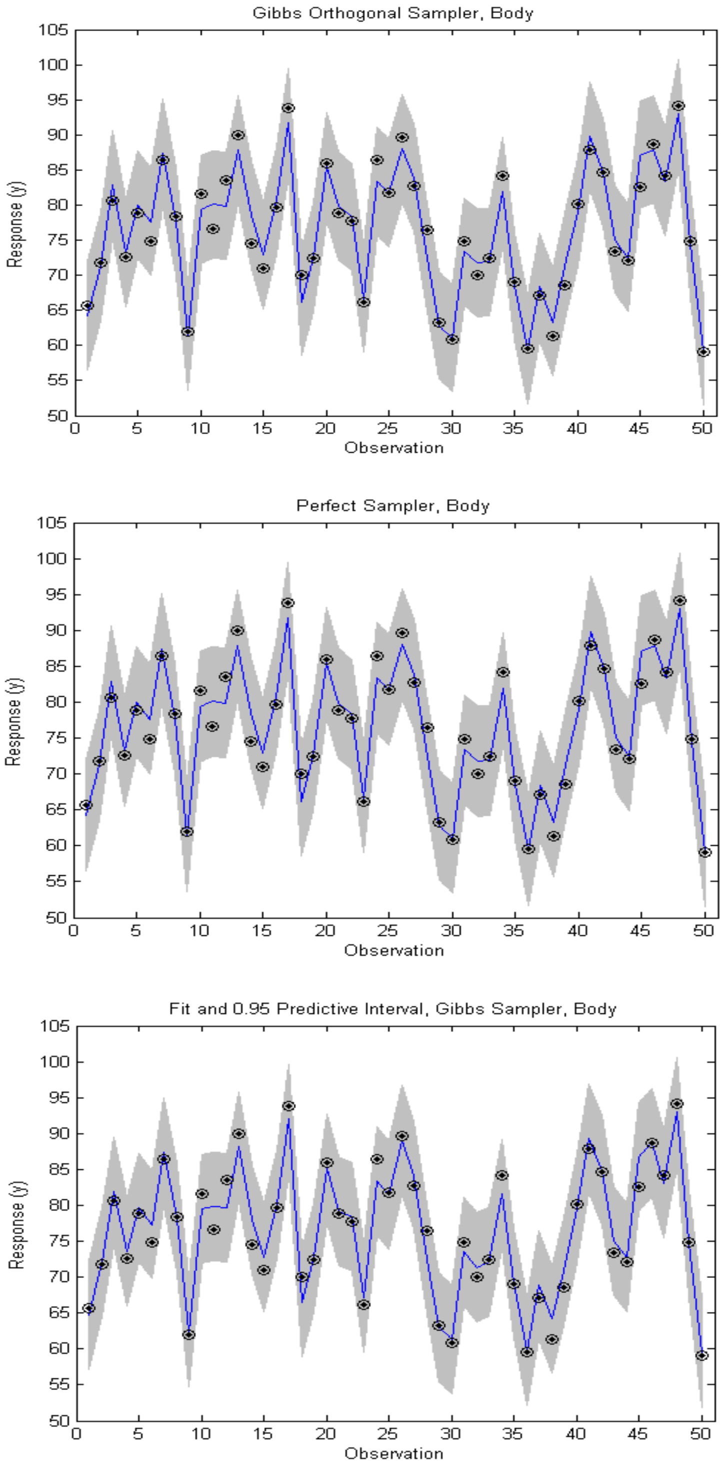
4.13 The estimated posterior for γ using the hybrid Gibbs sampler (top), exact sampler (middle) and the standard Gibbs sampler (bottom) for the Baseball data.



4.14 The fitted response (dot and circle), observed response (blue line) and 95% predictive interval using the hybrid Gibbs sampler (top), exact sampler (middle) and the standard Gibbs sampler (bottom) for the Baseball data.



4.15 The estimated posterior for γ using the hybrid Gibbs sampler (top), exact sampler (middle) and the standard Gibbs sampler (bottom) for the Body data.



4.16 The fitted response (dot and circle), observed response (blue line) and 95% predictive interval using the hybrid Gibbs sampler (top), exact sampler (middle) and the standard Gibbs sampler (bottom) for the Body data.

For the body data the posterior for **X** appears much less spread than for the posterior for **W**, and has a much more dominant peak at probability 0.09, while the mode for **W** is around 0.035. For the baseball data again the posterior using **X** is much more peaked and less spread than for **W**, with the maximum probability at around 0.07, compared to 0.025. In terms of efficiency for the body measurement data the Gibbs sampler generated 41.4 samples per second of cpu-time compared to 162 for the hybrid Gibbs sampler, and 40.1 for monotone Gibbs CFTP with a mean BCT of 2.025. For the baseball data the Gibbs sampler generated 37.9 samples per second of cpu-time compared to 153 for the hybrid Gibbs sampler and 38.2 for monotone Gibbs CFTP with a mean BCT of 2.030.

This comparison can be extended to include the sample size after burn-in for the Gibbs sampler using **X**, and the effective sample size for both the Gibbs sampler using **X** and the hybrid method. The burn-in for the Gibbs sampler in **X** was approximately 15000 for the body measurement data and 21000 for the baseball data. Thus for burn-in, the computational efficiency reduces to 35.2 and 32.21 samples per second of cpu-time respectively. Effective sample size equates the dependent sample to the equivalent amount of i.i.d. sample points. We use the function from the CODA package in R for calculating the effective sample size. For the samples generated by the hybrid method the effective sample was essentially the number of sample points generated. The effective sample size for the Gibbs sampler for **X** was approximately 79000 (out of 100000 or 79%) for the body measurement data and 71000 (out of 100000 or 71%) for the baseball data. Thus the efficiency of the standard Gibbs sampler reduces further to 32.7 and 26.9 samples per second of cpu-time for the body measurement and baseball datasets respectively. Thus, the hybrid method is clearly the most efficient followed by monotone Gibbs CFTP.

4.5 Summary

The orthogonal Gibbs sampler manages to attain similar levels of convergence compared to the monotone Gibbs CFTP sampler, due to a minimal amount of dependence in the Gibbs MC. This coupled with the improved computational efficiency of using an orthonormal predictor matrix, results in the orthogonal Gibbs sampler being the most efficient method per sample point, and for convergence in real time. For the minimal

BCT of 2, the Gibbs sampler using **X**, and monotone Gibbs CFTP, are comparable in terms of cpu-time per sample point. However, monotone Gibbs CFTP wins out due to the superior convergence in real time. While we did not go into details in order to keep the comparisons straight forward, other methods for obtaining the inverse covariance matrix, such as the Cholesky updating method, may provide a further decrease in the cpu-time for the Gibbs sampler. However, any advantage from this will likely be negated by the requirement to discard samples as burn-in. Thus, unless the BCT is extremely large, exact sampling will provide the greatest efficiency for convergence in real time, to the posterior distribution of model probabilities.

The BCT is a crucial factor in the comparison between the efficiency of monotone Gibbs CFTP, the Gibbs sampler in **X** and the hybrid method. Choices of τ and penalty for Zellner's prior and Jeffreys prior, were investigated for the BCT and compared against information in the posterior for $\boldsymbol{\gamma}$. Identifying the conditions under which BCT is close to 2 indicated that with n much larger than k, the exact sampler is most efficient. Orthogonalization methods such as the GS_2 and GPC methods, tended to have the smallest BCT. Choices of hyper-parameters than maximize model competition, large probabilities of remaining un-coalesced, and sample size close to k can all produce a large BCT. Finally, the real world examples re-iterated that the GS_2 method and orthogonalization in general, are well suited to modeling the response, especially with large n. Further, the best sampling method to use is either the hybrid or monotone Gibbs CFTP sampler.

CHAPTER 5

EXACT IMH AND REJECTION SAMPLING

> "*Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise*."
>
> \- Tukey, 1962

Compared to the Gibbs sampler, the independence Metropolis-Hastings (IMH) algorithm is assured monotonicity. This is because the update probabilities of the IMH MC have an inherent minimum. Knowing the point in the state space for which this minimum occurs, allows detection of complete coupling for the IMH chain. This means monotone CFTP is readily available and because no transformation of **X** is required, variable selection is possible.

Some unpublished work (Murray, 2004) has mentioned the relationship between IMH and rejection sampling. We review the argument that for any target (f) and proposal (q) distribution, the coalescence of the exact IMH sampler is also a rejection sampler f. This point may make the notion of an exact IMH sampler obsolete. The relationship between the IMH algorithm, rejection sampling and importance sampling was explored by Liu (1996) using a detailed eigenvalue analysis. Liu (1996) concluded that the IMH sampler is asymptotically as efficient as rejection sampling for estimating expectations. Of particular interest is that in discussing sample weights f_i/q_i, Liu (1996) noted that the largest value of this ratio is equivalent to the optimal choice of bound for an equivalent rejection sampler, thus indicating that perfect IMH is indeed rejection sampling. Liu

(1996) does not discuss perfect sampling which is unsurprising as this paper was published the same year as the CFTP approach of Propp and Wilson (1996).

In this chapter we begin by reviewing the exact IMH sampler, detecting coalescence and the relation to rejection sampling. This is followed by a discussion of how this relation removes the need for the backwards framework of CFTP and the relation to regeneration. The efficiency of IMH and rejection sampling is also discussed. We briefly review the variable selection method of Schneider and Corcoran (2004) and show it is rejection sampling. We then explore perfect sampling for $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$ for Zellner's prior and Jeffreys prior using exact IMH/rejection sampling. The common problem is the inability to establish effective bounds. The general idea will be to construct the proposal distribution to reduce *f*/*q* to a function of the residual sum of squares. We investigate the efficiency of such an approach for choices of penalty and τ for the constant Bernoulli prior for $\boldsymbol{\gamma}$. The second part of this chapter deals with Zellner's prior and generating exact samples from the posterior of the hyper-parameter *c* conditional on $\boldsymbol{\gamma}$. We investigate the efficiency of a standard approach, before moving to the interesting case of thinking about rejection sampling as exact IMH to improve efficiency.

5.1 Perfect Sampling with the IMH Sampler

5.1.1 Exact IMH

Let $\mathbf{x}$ be a vector of data and *a* be an unknown parameter. Assigning a prior to *a* ($f(a)$) by Bayes theorem the posterior $f(a \mid \mathbf{x}) \propto f(\mathbf{x} \mid a) f(a)$. Let *q*(*a*) be a proposal distribution for generating i.i.d. candidate values, with *q*(*a*) chosen to be heavier in the tails than the target density $f(a \mid \mathbf{x})$. The acceptance probability for moving from state *a* to a' for the independence Metropolis-Hastings sampler is

$$\alpha(a, a') = \min\left\{1, \frac{f(a' \mid \mathbf{x}) q(a)}{f(a \mid \mathbf{x}) q(a')}\right\} \text{ for } f(a \mid \mathbf{x}) q(a') > 0, \qquad (5.1)$$

and 1 if $f(a \mid \mathbf{x}) q(a') = 0$. The update function of a MC constructed using the IMH sampler is monotone according to the acceptance probabilities. The minimum state based on the ordering of α occurs at the maximum of the ratio between the posterior and proposal distributions:

$$\frac{f(a_m)}{q(a_m)} \geq \frac{f(a)}{q(a)} \text{ so that } \alpha(a_m, a) \leq \alpha(a_{-m}, a) \text{ for all } a \in \mathbf{A}, \tag{5.2}$$

where a_{-m} denotes all states in **A** excluding a_m. The starting state (a_m) for the exact IMH algorithm is subsequently defined as

$$a_m = \underset{a \in A}{\arg\max} \left\{ \frac{f(a \mid \mathbf{x})}{q(a)} \right\}. \tag{5.3}$$

This means if we can identify the maximum for the ratio of the posterior to the proposal for all a, we have identified the point at which the smallest acceptance probability occurs. This point can be identified by finding the derivative:

$$\frac{d}{da} \left\{ \frac{f(a \mid \mathbf{x})}{q(a)} \right\}, \tag{5.4}$$

setting equal to zero and solving for a. If no closed form solution exists optimization methods may be used to identify a_m numerically, provided (5.4) exists. We now discuss some special cases of (5.3). Let $q(a) = f(a)$ which requires that $f\,(a)$ be proper, then (5.3) becomes:

$$a_m = \underset{a \in A}{\arg\max} \left\{ \frac{f(\mathbf{x} \mid a) f(a)}{f(a)} \right\} = \underset{a \in A}{\arg\max} \{ f(\mathbf{x} \mid a) \} \tag{5.5}$$

Thus, if we use the prior as the proposal distribution we can simplify the search for the starting point. This approach is particularly useful when closed form expressions for maximum likelihood estimates are available. Let $q(a) \propto 1$ then

$$a_m = \underset{a \in A}{\arg\max} \left\{ \frac{f(a \mid \mathbf{x})}{1} \right\} = \underset{a \in A}{\arg\max} \{ f(a \mid \mathbf{x}) \}. \tag{5.6}$$

Thus, the starting point for the exact IMH algorithm using (5.6) is the mode of the posterior distribution. The IMH algorithm can also be bounded by finding

$$C \geq \left\{ \frac{f(a \mid \mathbf{x})}{q(a)} \right\} \text{ for all } a \in \mathbf{A}, \tag{5.7}$$

such that

$$\alpha(a_m, a) \geq \left\{ \frac{f(a \mid \mathbf{x})}{Cq(a)} \right\}. \tag{5.8}$$

If the bound is good enough and we can determine it, we do not require knowledge of a_m. However, all choices of bound will result in a perfect sampler less efficient than using a_m which is the optimal choice of C.

Under the partial ordering of acceptance probabilities, we need only monitor a path from a_m to assess the backwards coupling time. This is because the ordering is based on the acceptance probabilities and as such, the lower path will be the hardest state in the state space to move from. Hence, when the lower path accepts a move so will the upper path, or any other chain started from any other state of **A** resulting in coalescence. This means the lower path need only be run, reducing the computational effort to a single chain only. The IMH CFTP algorithm is given in Algorithm IX.

The exact IMH CFTP sampler moves backwards in time, until a move from a_m to another state is finally accepted indicating complete coupling. The algorithm then proceeds forwards until time zero to obtain an exact sample, reusing the random numbers and generated proposal values. For the bounded case, find C at **(1)** omit **(2)**, and the probability for detecting coalescence **(3)** is

$$\alpha_c = \min\left\{ 1, \frac{f(a_{t+1})}{Cq(a_{t+1})} \right\}. \tag{5.9}$$

After detecting coalescence the forward propagation of the IMH chain to time zero proceeds as normal. Work by Corcoran and Tweedie (2000) showed the distribution of backwards coupling times (T) for the exact IMH sampler is geometric:

$$f(T) = (1-\rho)^{T-1} \rho \, \mathbb{I}_{\{1,\ldots,\infty\}}(T) \text{ where } \rho = \frac{q(a_m)}{f(a_m \mid \mathbf{x})}. \tag{5.10}$$

Algorithm IX: Exact IMH CFTP sampler

```
Find initial minimum state $a_m$. (1)
Set: coalescence = false.
Set: $t = 0$.
While coalescence = false
    Set: $t = t -1$.
    Set: $a_t = a_m$. (2)
    Generate: $a_{t+1} \sim q(a)$ and $u_{t+1} \sim \mathrm{U}(0,1)$.
    Compute: $\alpha_c = \min\left\{ 1 , \frac{f(a_{t+1})q(a_m)}{f(a_m)q(a_{t+1})} \right\}$ (3)
    If $u_t \le \alpha_c$
        Set: coalescence = true.
        For $i = \{ t+2, t+3, ...., 0 \}$
            Compute: $\alpha = \min\left\{ 1 , \frac{f(a_i)q(a_{i-1})}{f(a_{i-1})q(a_i)} \right\}$.
            If $u_t < \alpha$
                Set $a_i = a_{i-1}$.
```

The mean of the geometric distribution is the expected backwards coupling time: $\mathbb{E}[T] = \rho^{-1}$. Using (5.10) requires q and f are known exactly and not just up to normalizing constant.

5.1.2 Coalescence and Rejection Sampling

It turns out the exact IMH CFTP sampler is in fact redundant, as the approach for detecting coalescence is rejection sampling. Let $P(a)$ be the probability of accepting the proposed point a in standard rejection sampling:

$$P(a) = \frac{f(a \mid \mathbf{x})}{Cq(a)} \quad . \tag{5.11}$$

(5.11) may be used to generate exact samples using Algorithm X provided $Cq(a) \geq f(a \mid \mathbf{x})$ for all $a \in \mathbf{A}$.

Algorithm X: Rejection Sampler

```
Find $C$ and set: $v = 0$.
While $v < n$:
        Generate: $a' \sim q(a)$ and $u \sim \mathrm{U}(0,1)$.
        Compute: $\alpha = \frac{f(a' \mid \mathbf{x})}{Cq(a')}$
        If $u \leq \alpha$
                Set: $v = v + 1$.
                Set: $a_v = a'$.
```

For detecting coalescence with the exact IMH sampler:

$$P(a) = \frac{f(a \mid \mathbf{x})}{q(a)} \frac{q(a_m)}{f(a_m \mid \mathbf{x})} \text{ ,} \tag{5.12}$$

where a_m is defined as in (5.3). We can re-write (5.12) as

$$P(a) = \frac{f(a \mid \mathbf{x})}{Cq(a)} . \tag{5.13}$$

where $C = f(a_m \mid \mathbf{x}) / q(a_m)$. Thus (5.11) and (5.13) are of the same form so it suffices to show that $Cq(a) \geq f(a \mid \mathbf{x})$ holds for all a. Taking $Cq(a) \geq f(a \mid \mathbf{x})$ and substituting C we have:

$$\frac{f(a_m \mid \mathbf{x})}{q(a_m)} q(a) \geq f(a \mid \mathbf{x}) , \tag{5.14}$$

which can be re-expressed as

$$\frac{f(a_m \mid \mathbf{x})}{q(a_m)} \geq \frac{f(a \mid \mathbf{x})}{q(a)} . \tag{5.15}$$

Notice that this is now equivalent to the partial order required for the IMH CFTP sampler which is clearly true given the definition of a_m. It is also clear that with respect to the rejection sampler this is the optimal choice of C, and is the minimum value such that $Cq(a) \geq f(a \mid \mathbf{x})$. In the case of bounded IMH the same relationship applies, except the value of C is no longer the optimal choice. Figure 5.1 makes this relationship clear.

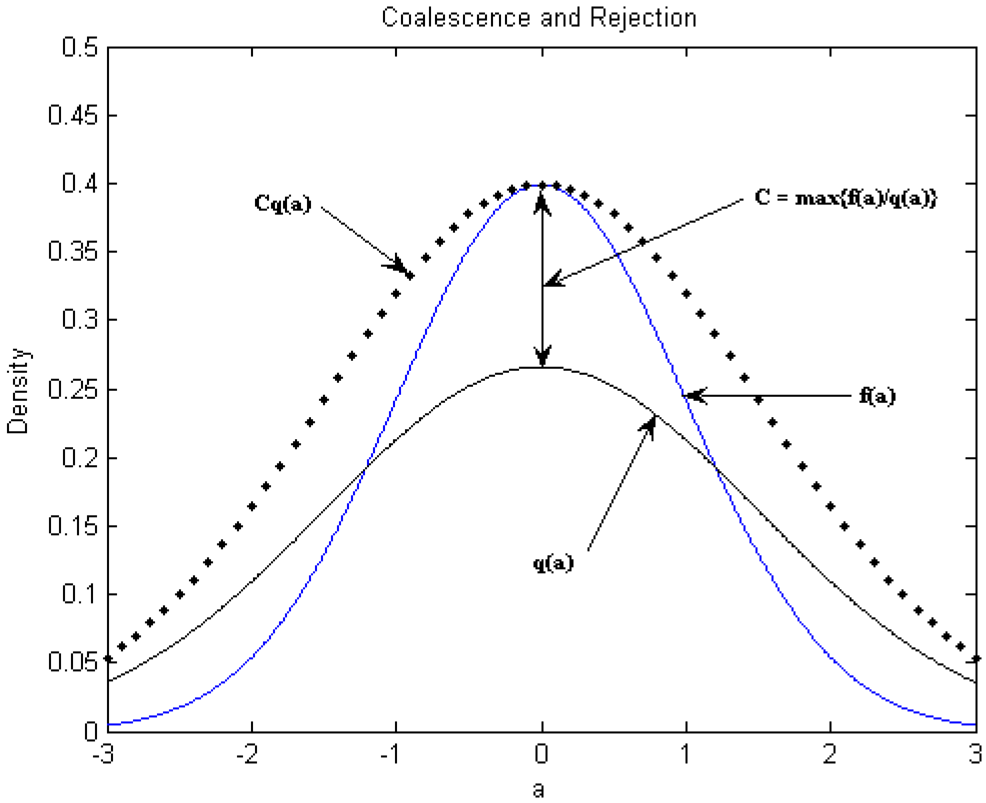


Figure 5.1 The maximum ratio of the target to proposal and the relationship to the optimum choice of *C* for a rejection sampler.

This can also be demonstrated with the following toy example.

Toy Example

Let $\mathbf{X} = \{1,2,3,4\}$ with the corresponding posterior probabilities $f(x)$ = [0.13 0.19 0.42 0.26], and let the proposal distribution be $q(x)$ = [0.15 0.35 0.25 0.25]. Clearly x_m is at x = 3 and the corresponding probabilities of accepting a move from x_m to any other x is [0.52 0.32 1 0.62]. Taking these probabilities and multiplying by the probability of proposing each x will indicate the relative proportions of sample points generated at coalescence. These values are [0.077 0.116 0.25 0.155]. Due to rounding it is not immediately obvious that these values are proportional to f however, dividing by the sum we obtain [0.13 0.19 0.42 0.26].

■

This along with the previous mathematical argument makes it very clear that the points generated at the detection of coalescence are indeed exact draws from f.

5.1.3 Forwards Simulation and Regeneration

The way we detect coalescence is actually a rejection sampler so that unlike traditional CFTP the point of coalescence is an exact draw from f. This means we may abandon the backwards coupling framework and use forward coupling as the point of coalescence is an exact draw. This provides along with the rejection sampler for generating i.i.d. sample points, the option to generate exact dependent samples by recording all sample points from an IMH chain after the first sample point generated by rejection. This is in essence a forwards coupling algorithm and eliminates the need to assess burn-in.

Using a simulated data set with $k = 10$ and $n = 100$ for Zellner's prior with $c = n$ and a uniform prior for γ we show the relation between coupling times, coalescence and regeneration. Figures 5.2a-c show a sequence of plots from an actual realization using the rejection sampling and coupled IMH chains run from all states.

From Figure 5.2a we can see the IMH chain started from every state has coalesced by the time the chain has turned blue. From the previous discussion this implies the point at which the chain turns blue is the point of coalescence and an exact draw from f. This point should coincide with a rejection sampler, using the same sequence of random numbers and proposed values, for the same choice of C. The blue chain also represents a sequence of exact dependent draws from f. In Figure 5.2b the red line shows how we can monitor a single chain, started from the state where the maximum of the f/q occurs, to detect coalescence.

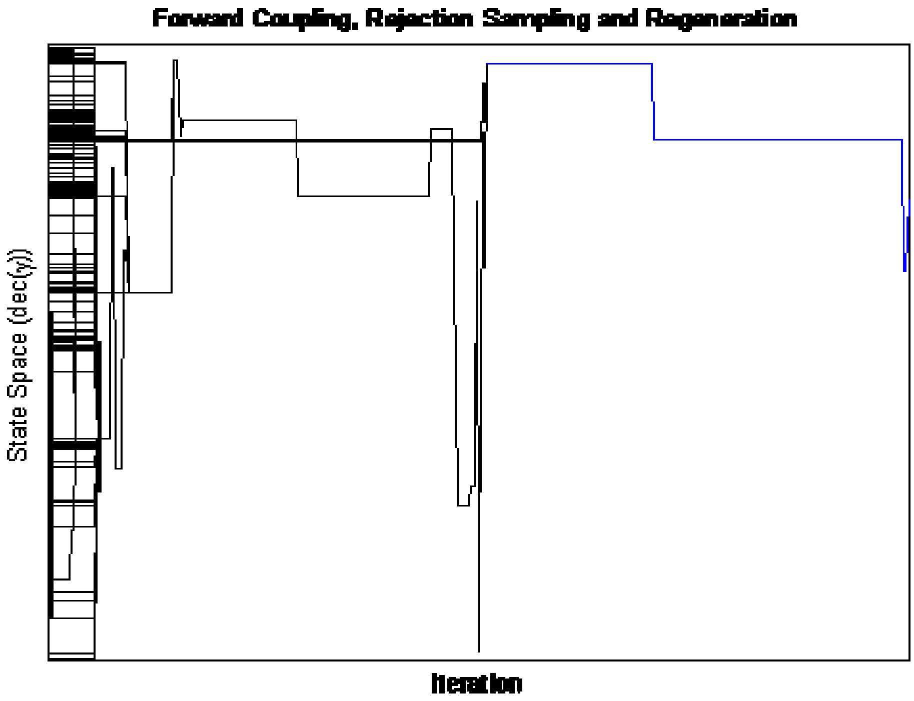


Figure 5.2a Forward IMH for generating exact dependent sample points.

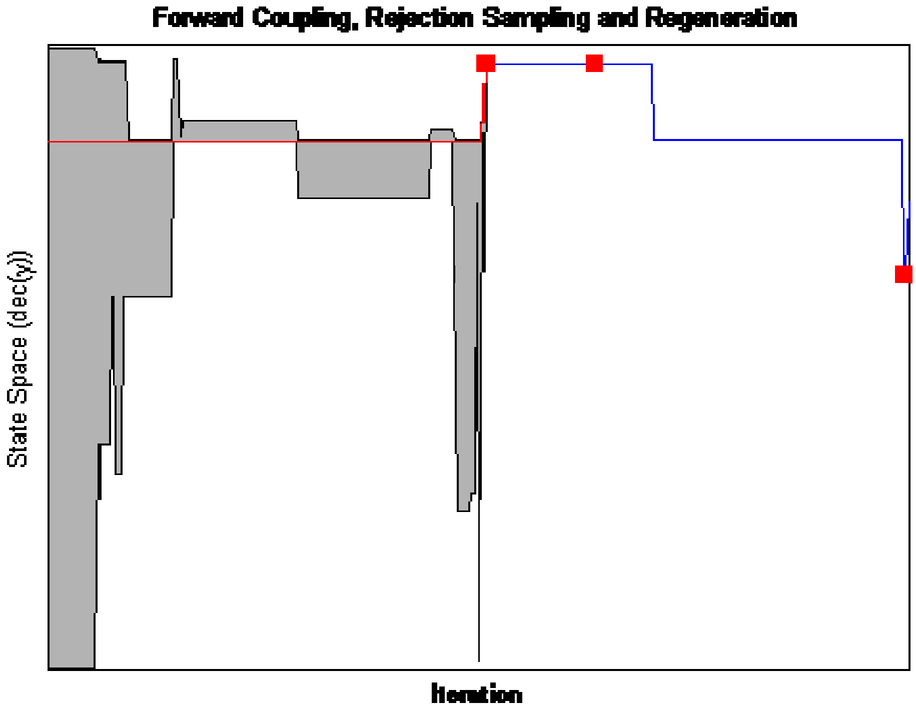


Figure 5.2b The rejection sampler for the same sequence of candidate values and uniform random numbers.

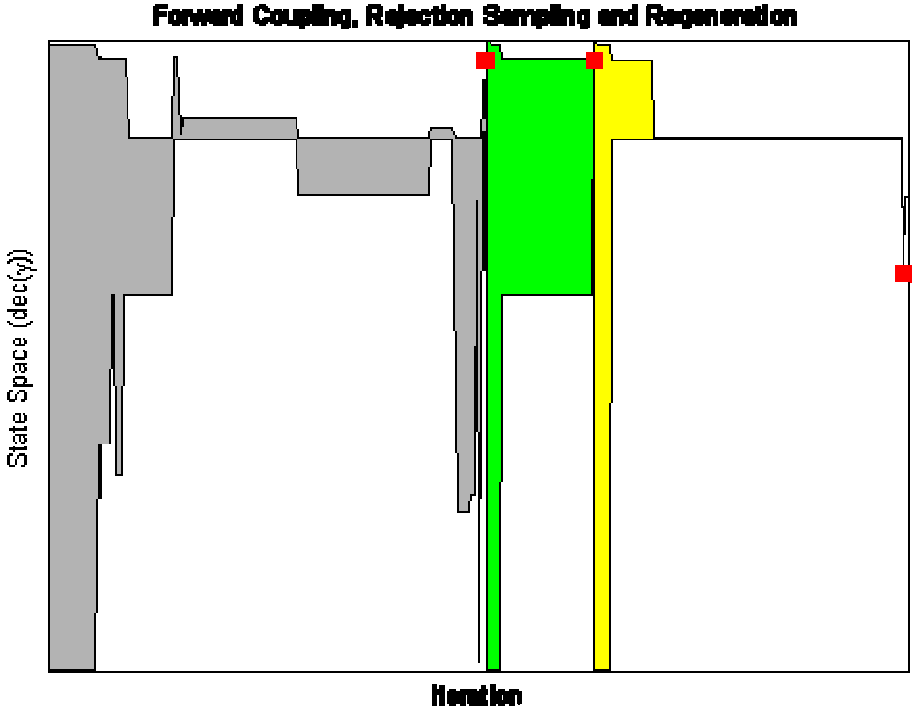


Figure 5.2c The relation between coupling, rejection sampling and independent tours for regeneration times.

More importantly, the sequence of red squares represents the three sample points generated by rejection sampling. The first red square clearly coincides with the coalescence of the IMH chain. Finally in Figure 5.2c, we can see that the process of complete coupling repeats itself so we can visualize the regeneration times of the MC. Each colored region represents a restart of the coupled IMH chain from all states. As expected, these coincide with the points produced by the rejection sampler. Each sequence between the rejection points represents a tour, and so indicates the rejection points as regenerations in the IMH chain. The beauty of this demonstration is that it ties together the ideas of IMH, rejection sampling and regeneration. We have often mentioned the coupling times and regeneration times are of the same distribution however, in this example, they have the exactly the same geometric distribution which is also the distribution of waiting times for the rejection sampler.

As mentioned, we can use rejection sampling to find a single point of coalescence and then run an IMH chain forward from that point on. This chain will be a dependent sequence of exact draws, as the starting point of the IMH chain is an exact draw under rejection sampling. The drawback to this, is the poor exploration of the state space and large dependence in the IMH chain when the proposal is dissimilar to the target. Thus, a large number of samples will be required to approximate the posterior distribution well despite an exact guarantee. This of course, is akin to an automatic procedure for assessing burn-in, but we obtain dependent sample points instead of i.i.d.

This brings us back to an interesting point from the previous chapter. Perfect sampling is efficient when the underlying MC is rapidly convergent to the distribution of interest. In terms of efficiency there is little difference between the exact IMH sampler and rejection sampling. The recursion to detect coalescence will come at the same computational cost as the rejection sampler. The IMH CFTP sampler however, requires storage of all random numbers for re-use in the forwards propagation after complete coupling. Thus, it seems reasonable to argue that because monotone IMH CFTP requires more memory than rejection sampling, then rejection sampling should be preferred.

5.1.4 The Joint Approach of Schneider and Corcoran (2004)

We now review the exact implementation of the bounded IHM algorithm by Schneider and Corcoran (2004) for BVS in Bayesian linear regression. They use a joint estimation approach for both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, so:

$$\beta_i = \gamma_i \theta_i \,, \tag{5.16}$$

where θ is the corresponding regression coefficient. This is to produce a mixture distribution prior for $\boldsymbol{\beta}$, so that for a given vector of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ can be recovered as the positions where elements of $\boldsymbol{\beta}$ are non-zero.

The likelihood function is

$$\ell\ (\boldsymbol{\beta}, z) = z^{\frac{n}{2}} \exp\left\{ -\frac{1}{2} z(\mathbf{y} - \sum_{i=1}^{q_\gamma} \beta_i \mathbf{X}_i)^T (\mathbf{y} - \sum_{i=1}^{q_\gamma} \beta_i \mathbf{X}_i) \right\}, \tag{5.17}$$

where $z = (1/\sigma^2)$. The priors used are

$$\boldsymbol{\beta} \mid z \sim N\left(\boldsymbol{\xi}, \frac{1}{z}\mathbf{V}\right),\ z \sim \mathbf{G}\left(\frac{v}{2}, \frac{v\lambda}{2}\right), \text{ and } f(\boldsymbol{\gamma}) \propto 1, \tag{5.18}$$

where $\boldsymbol{\xi}$, **V**, v and λ are hyper-parameters to be chosen and **G** is the gamma distribution. Choosing the proposal distributions to be:

$$\boldsymbol{\beta} \mid z \sim N\left(\boldsymbol{\xi}, \frac{1}{z}\mathbf{V}\right),\ z \sim \mathbf{G}\left(\frac{n+v}{2}, \frac{v\lambda}{2}\right), \text{ and } f(\boldsymbol{\gamma}) \propto 1, \tag{5.19}$$

then we require the maximum of the ratio:

$$\frac{\ell\ (\boldsymbol{\beta}, z)}{z^{n/2}} = \exp\left\{-\frac{1}{2}z(\mathbf{y} - \sum_{i=1}^{q_\gamma}\beta_i\mathbf{X}_i)^T(\mathbf{y} - \sum_{i=1}^{q_\gamma}\beta_i\mathbf{X}_i)\right\}. \tag{5.20}$$

The minimum bound is not available in closed form however, we know that exp(-x) where $x \geq 0$ will always be ≤ 1. Thus the bounded IMH algorithm is used with $C = 1$. To simulate samples of $\boldsymbol{\beta}$, q_γ elements are randomly selected and set equal to one, and then $\boldsymbol{\beta}$ is simulated from the required normal distribution. Their results show that while this approach works, it is computationally intensive requiring large BCT even for small values of k. Finally as we have shown in 5.1.2, this exact IMH sampler is a rejection sampler at the point of coalescence. This requires noting that because with (5.20) we have shown $C \geq f/q$ it then follows that $Cq \geq f$, as detailed above for the bound.

5.2 Variable Selection using exact IMH

The requirement to find the maximum of the posterior divided by the proposal (5.3), the likelihood (5.5), or the posterior (5.6) is a great hindrance to the use of exact IMH for generating exact samples from $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$. This is because it is a large discrete state space, and the starting point cannot be found without some examination of $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$, as the derivative for $\boldsymbol{\gamma}$ is unavailable. We introduce an approach for rejection sampling from $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$ for Zellner's prior and Jeffreys prior, and a special case using Jeffreys prior.

5.2.1 Zellner and Jeffreys

These approach we are about to outline will rely upon the fact that residual sum of squares is minimized when γ is the full model. While there is no straight forward manner in which to show this (such as with a derivative), the argument follows from the least squares optimization view point. In least squares, the optimization minimizes the sum of squares error (SSE) with respect to the regression coefficients $\boldsymbol{\beta}$. Thus,

$$\min_{\boldsymbol{\beta}} SSE(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})^2 \ , \tag{5.21}$$

where $\mathbf{X}_i$ is the *i*-th row of $\mathbf{X}$. Because this optimization is essentially unconstrained it is weakly smaller with an increasing number of predictors. This can be put down to the fact that the addition of any predictor, even with only minimal correlation, will provide some reduction in the SSE even if almost negligible. This property is perhaps better known as the reason for the R-square value being unsuitable for model selection. As the number of predictors included in the model increases the R-square does also. This suggests that if we can use *q* to reduce the posterior for γ to a function of the RSS, then the bound is the max{RSS} allowing the use of rejection sampling.

The target distribution using Zellner's prior is

$$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) \propto (c+1)^{-\frac{q_\gamma + 1}{2}} (\mathbf{y}^T \mathbf{y} - \tilde{c}_1 \mathbf{y}^T \mathbf{H}_\gamma \mathbf{y})^{-\frac{n}{2}} f(\boldsymbol{\gamma}) \, . \tag{5.22}$$

where $\tilde{c}_1 = c/(c+1)$ and $\mathbf{H}_\gamma = \mathbf{X}_\gamma (\mathbf{X}_\gamma^T \mathbf{X}_\gamma)^{-1} \mathbf{X}_\gamma^T$. With the proposal density $q(\gamma)$ we require:

$$\boldsymbol{\gamma}_m = \arg\max_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}} \left\{ \frac{(c+1)^{-\frac{q_\gamma + 1}{2}} (\mathbf{y}^T \mathbf{y} - \tilde{c}_1 \mathbf{y}^T \mathbf{H}_\gamma \mathbf{y})^{-n/2} f(\boldsymbol{\gamma})}{q(\boldsymbol{\gamma})} \right\}, \tag{5.23}$$

to perform perfect sampling with the IMH algorithm, or equivalently optimal rejection sampling. By setting $q(\gamma) = (c+1)^{-(q_\gamma + 1)/2} f(\gamma)$ (5.23) becomes:

$$\boldsymbol{\gamma}_m = \arg\max_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}} \{ (\mathbf{y}^T \mathbf{y} - \tilde{c}_1 \mathbf{y}^T \mathbf{H}_\gamma \mathbf{y})^{-n/2} \} \, . \tag{5.24}$$

Given the assertion in relation to (5.21) choosing γ to be the full model satisfies (5.24), and so provides the starting point for IMH and the bound necessary for rejection sampling. If we choose *f*(γ) to be the constant Bernoulli prior, then we can summarize the proposal distribution as

$$q(q_\gamma) \sim \binom{k}{q_\gamma}(c+1)^{-\frac{q_\gamma+1}{2}}\tau^{q_\gamma}(1-\tau)^{k-q_\gamma}\,, \tag{5.25}$$

which can be normalized by dividing by the sum. This proposal is also simple to obtain even when *k* is large. To propose candidate values of γ using (5.25) randomly choose q_γ according to (5.25) then choose q_γ positions in γ to be set equal to one. Of course (5.25) may be simplified further by taking $f(\boldsymbol{\gamma}) \propto 1$.

Taking a similar approach for Jeffreys prior the target distribution is

$$f(\boldsymbol{\gamma}\,|\,\mathbf{y},\mathbf{X}) \propto (p/2\pi)^{-q_\gamma/2}(\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})^{-\frac{n}{2}}f(\boldsymbol{\gamma}), \tag{5.26}$$

We require the starting point $\boldsymbol{\gamma}_m$ to detect complete coupling or use rejection sampling. Letting *q*(γ) be the proposal density, we require:

$$\boldsymbol{\gamma}_m = \arg\max_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}}\left\{\frac{(p/2\pi)^{-q_\gamma/2}(\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})^{-n/2}f(\boldsymbol{\gamma})}{q(\boldsymbol{\gamma})}\right\}. \tag{5.27}$$

Letting *q*(γ) = $(p/2\pi)^{-q_\gamma/2}f(\boldsymbol{\gamma})$,

$$\boldsymbol{\gamma}_m = \arg\max_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}}\{(\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})^{-n/2}\}. \tag{5.28}$$

(5.28) is again maximized when we use the full model.

With these two methods in mind we compute the efficiency of the rejection sampler exactly when *k* is small using:

$$\mathbb{E}[\mathrm{T}] = f(\boldsymbol{\gamma}_m)/q(\boldsymbol{\gamma}_m). \tag{5.29}$$

We investigate different values of τ and penalty for Zellner's prior and Jeffreys prior. The results are presented in Tables 5.1 and 5.2.

Table 5.1 Efficiency of rejection sampling for choices of τ

Data	Prior					τ				
(n,k)		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Ozone	Z							157	45.5	9.4
(80,8)	J								72.0	12.6
Physical	Z	21.7	11.2	7.12	4.97	3.65	2.76	2.12	1.65	1.29
(27,10)	J							988	150	16.4
Bodyfat	Z									1032
(250,13)	J									1108
Crime	Z	20.6	17.0	13.8	10.9	8.39	6.21	4.41	2.95	1.84
(47,15)	J									330

Table 5.2 Efficiency of rejection sampling for choices of penalty

Data	Prior			c/p		
(n,k)		1	5	10	50	100
Ozone	Z	1.43	5.46	17.0		
(80,8)	J	24.9	88.7	192		
Physical	Z	1.03	1.22	1.58	9.12	40.2
(27,10)	J	186				
Bodyfat	Z	12.9	429	2891		
(250,13)	J	1632	6414			
Crime	Z	1.02	1.19	1.50	9.53	61.3
(47,15)	J	6427				

Z is Zellner's prior, J is Jeffreys prior, and the grey boxes represent cases where the expected waiting time per sample point is greater than the size of the state space (2^k). For choices of τ, c is set equal to n, and for values of penalty, τ = 0.5. As with other comparisons we set $p = 2\pi(c+1)$.

The first major difficulty is the number of instances where the rejection sampler is impractical. This is not unsurprising as the rejection sampler or the IMH sampler will be less efficient when the prior which is the proposal conflicts strongly with the function of the RSS. This will generally be the case as the full model provides the bound and the prior/proposal is essentially a mechanism for penalizing the RSS. In the case of τ when it favors complex models, thus agreeing with the RSS, the efficiency is manageable. For

Jeffreys prior the smallest manageable value of τ is 0.7 for the physical dataset, 0.8 for the ozone data and 0.9 for the bodyfat and crime data. For Zellner's prior the bodyfat and ozone datasets need values of τ greater than 0.9 and 0.8 respectively. The physical and crime data sets have very manageable waiting times for all values of τ. This is because the sample size is small and with $c = n$, the shrinkage term $c/(c+1)$ results in a very flat function of the RSS.

For varying values of penalty the Jeffreys prior becomes inefficient for values between 10 and 50 for the ozone data, 1 and 5 for the physical and crime data, and between 5 and 10 for the bodyfat data. For Zellner's prior rejection sampling becomes inefficient between 10 and 50 for the ozone and bodyfat data, and between 100 and 500 for the physical and crime datasets. This definitely provides the indication that for Zellner's prior with reasonable choices of τ and small values of c is a more than adequate method to generate exact samples from $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$. Thus, when k is large and the sample size is small the rejection sampler with Zellner's prior and $c = n$ is a competitive choice for sampling from $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$.

For Jeffreys prior we may choose p to be 2π so that there is no penalty for model complexity letting the role fall explicitly to the choice of $f(\gamma)$. This allows the proposal distribution to simply be the prior for γ which we take to be the constant Bernoulli prior. As above we investigate the efficiency of choices of τ for the four real datasets. The results are presented in Table 5.3.

Table 5.3 Efficiency of rejection sampling for choices of τ for Jeffreys prior with $p = 2\pi$

Data(n,k)	τ								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Ozone (80,8)			78.7	28.3	12.5	6.44	3.65	2.24	1.46
Physical (27,10)				227	63.5	21.9	8.73	3.90	1.91
Bodyfat (250,13)			3316	593	141	41.1	13.9	5.29	2.21
Crime (47,15)				8888	1082	182	39.1	10.0	3.00

The grey squares again represent cases where the expected waiting time is greater than the size of the state space. Compared to the results from Tables 5.1 and 5.2, a larger range of values for τ are viable but still require larger values. The ozone and bodyfat data is reasonable for values of τ greater than 0.3, and 0.4 for the physical and crime datasets. We now use examples to illustrate the use of this rejection sampler.

5.2.2 Examples

We use the special case of Jeffreys prior above to illustrate the variable selection for the ozone data and Zellner's prior for the crime data and compare the MIP to the true values. We then proceed with an example using two larger datasets where we do not have the true values to compare against.

Example 1: Ozone Data and the Special Case of Jeffreys Prior.

For the ozone data set we use the special case of Jeffreys prior where we set $p = 2\pi$. Then using the constant Bernoulli prior we set $\tau = 0.3$. Using the rejection sampler, 1000 sample points were generated and the true MIP compared with those estimated from the sample. The results are shown in Table 5.4.

Table 5.4 True and Estimated MIP from the Rejection Sampler for the Ozone Data

Predictor	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
True	0.781	0.783	0.709	0.382	0.451	0.463	0.511	0.802
Estimated	0.779	0.782	0.711	0.382	0.449	0.462	0.511	0.801

The MIP are very well estimated using the 1000 exact sample points. The maximum waiting time was 558 with a mean of 83. While this provides a great illustration of the use of the rejection sampler in this context, the number of evaluations required far exceeds the size of the state space. This indicates, much in line with the results above, that variable selection with the rejection sampler is possible. However, there is the additional requirement that the number of function evaluations required to generate samples should be much less than the size of the state space. If this is not the case, then direct calculation of the exact posterior would be preferable.

Example 2: Crime Data and Zellner's Prior.

Using the crime data we set $c = n = 47$ and $\tau = 0.4$ for the constant Bernoulli prior. Using the rejection sampler 5000 sample points were generated. Table 5.5 shows the MIP estimated using the generated sample and the true MIP.

Table 5.5 True and Estimated MIP from the Rejection Sampler for the Crime Data

Predictor	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
True	0.106	0.102	0.113	0.249	0.235	0.096	0.097	0.098	0.153	0.090	0.096	0.131	0.127	0.126	0.093
Estimated	0.091	0.102	0.112	0.250	0.248	0.094	0.098	0.098	0.149	0.090	0.096	0.138	0.130	0.122	0.096

Even with only 5000 samples from a state space of size 2^{15} most estimates apart from that for the fifth predictor are within less than 0.01 of the true value. The mean BCT was 11.20 with a maximum of 79, and generated 55 samples per second of cputime.

Example 3: Larger *k* and Zellner's Prior.

In this example we do not have the true values to compare against. We use the larger datasets, body measurements and the baseball datasets from the examples at the end of previous chapter. This is to explore the observation that using Zellner's prior with moderate values of c and τ should allow perfect sampling to be feasible. Using the body measurement data first, there are 24 predictors and a sample of 350. We choose $c = 200$ which is almost half of $c = n$ and as above use $\tau = 0.4$. The rejection sampler was used to generate 1000 sample points and the estimated MIP are given in Table 5.6.

Table 5.6 Estimated MIP from the Rejection Sampler for the Body Measurement Data

Predictor	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
Estimated	0.082	0.109	0.102	0.113	0.115	0.124	0.091	0.130	0.093	0.195	0.289	0.348	0.146	0.378	0.203	0.140	0.225	0.203	0.175	0.094	0.128	0.084	0.276	0.127

The rejection sampler required 7.879 seconds of cpu-time per sample point or equivalently 0.127 sample points per second of cpu-time. The mean waiting time was 5288.7 with a maximum of 44590. In total the number of candidate points required to generate 1000 samples was 10569646, which is 31.5% of the size of the state space (2^{24}). The fact that the rejection sampler was actually feasible is related to how flat the posterior is. The flatness of the posterior is clearly reflected in the consistently low values of MIP, with the largest values at 0.348 and 0.378. Despite this, the MIP still provide an indication of relative importance. This is indicated by the predominant selection of predictors 11, 12, 14 and 23 which correspond to the predictors chest girth, waist girth, hip girth and height respectively. The selection of these predictors is no surprise as they represent measurements of large body structures and height. For the baseball data, we truncate the dataset to the first 50 sample points to provide a small *n* with large *k* in comparison to the large *n* and *k* of the body data, and again generate 1000 sample points with the rejection sampler. We use c = 50 and τ = 0.4. The estimated MIP are shown in Table 5.7.

Table 5.7 Estimated MIP from the Rejection Sampler for the Baseball Data

Predictor	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Estimated	0.075	0.091	0.221	0.192	0.125	0.100	0.144	0.192	0.117
Predictor	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
Estimated	0.108	0.099	0.091	0.958	0.877	0.103	0.080	0.090	0.085
Predictor	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**
Estimated	0.095	0.100	0.137	0.156	0.097	0.112	0.100	0.119	0.094

The rejection sampler required 6.292 seconds of cpu-time per sample point or equivalently 0.159 sample points per second of cpu-time. The mean waiting time was 3945 with a maximum of 29714. In total the number of candidate points required to generate 1000 samples was 3945158, which is 2.94% of the size of the state space, 2^{27}. Much like for the body measurement dataset, the MIP are mainly low values. However, there is a clear indication for predictors 13 (free agent) and 14 (arbitration) with MIP of 0.958 and 0.877 respectively. These predictors are significant as both represent cases where the player is in a position to negotiate either directly, or indirectly their salary.

These predictors also represent the cases where the player is in the best positions to market themselves to potential teams.

5.3 The Conditional Distribution of *c*

The parameter c in Zellner's prior has received a great deal of attention under the focus of variable selection (Liang *et al*, 2008; Celeux *et al*, 2007). However, another aspect of this process is for model averaging. If the weights for averaging are independent of c then so too must be the posterior distributions we average over, such as the posterior for $\boldsymbol{\beta}_\gamma$ or the posterior predictive distribution for $\mathbf{y}$. Given we can in some cases integrate out c in the posterior for $\boldsymbol{\gamma}$, we could then generate samples values of c conditional upon $\boldsymbol{\gamma}$ as follows:

$$\begin{gathered}\text{Generate: } f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}). \\ \text{Generate: } f(c \mid \boldsymbol{\gamma}, \mathbf{y}, \mathbf{X}). \\ \text{Generate: } f(\boldsymbol{\beta}_\gamma \mid \boldsymbol{\gamma}, c, \mathbf{y}, \mathbf{X}) \text{ or } f(\tilde{\mathbf{y}} \mid \boldsymbol{\gamma}, c, \mathbf{y}, \mathbf{X}, \tilde{\mathbf{X}}).\end{gathered} \tag{5.30}$$

Liang *et al* (2008) have already demonstrated that the integration over c results in a hyper-geometric function, however if no closed form is available then presumably some sort of Monte Carlo approximation can be employed if k is small enough. The simplest case is to consider when we do not need to sample from $\boldsymbol{\gamma}$ and can obtain the posterior explicitly. Clearly even with this possible integrating over c it still requires producing sample points from the posterior for c. We now detail an efficient rejection sampler for doing so. Using Zellner's prior the posterior for c conditional on $\boldsymbol{\gamma}$ is

$$f(c \mid \boldsymbol{\gamma}, \mathbf{y}, \mathbf{X}) \propto (c+1)^{-\frac{q_\gamma+1}{2}} (\mathbf{y}^T\mathbf{y} - \frac{c}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})^{-\frac{n}{2}} f(c). \tag{5.31}$$

Again taking the initial approach of setting $q(c) = f(c)$, and because for the penalty term the smallest c is preferred while the second term prefers c to infinity, we choose the hyper-G-n prior as it is a good match for the sparseness of (5.31) ignoring the contribution from the prior. The hyper-G-n prior with parameter $a > 2$ is

$$f(c \mid a) = \frac{a-2}{2n}(1 + c/n)^{-a/2} \ \mathbb{I}_{[0,\infty)}(c). \tag{5.32}$$

In order to have $q(c) = f(c)$ we must be able to simulate from $f(c)$. The CDF of (5.32) is

$$F(c) = \frac{a-2}{2n}\int_0^c (1+t/n)^{-a/2}dt = \left[1 - \left(\frac{n}{n+c}\right)^{(a-2)/2}\right], \tag{5.33}$$

and the inverse CDF is

$$F^{-1}(u) = n[(1-u)^{-2/(a-2)} - 1]. \tag{5.34}$$

Thus, we can use the inverse CDF method to generate samples from (5.31) for use in the rejection/exact IMH algorithm. The optimal bound via the maximum of the ratio of the target to the proposal can be obtained by differentiating:

$$\frac{f(c\,|\,\boldsymbol{\gamma},\mathbf{y},\mathbf{X})}{f(c)} \propto (c+1)^{-\frac{q_\gamma+1}{2}}\left(\mathbf{y}^T\mathbf{y} - \frac{c}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}\right)^{-\frac{n}{2}}, \tag{5.35}$$

with respect to c. The resulting derivative is

$$\begin{aligned} &-\left(\frac{q_\gamma+1}{2}\right)(c+1)^{-\frac{q_\gamma+3}{2}}\left(\mathbf{y}^T\mathbf{y} - \frac{c}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}\right)^{-\frac{n}{2}}\ldots \\ &-0.5(c+1)^{-\frac{q_\gamma+1}{2}}\left(\mathbf{y}^T\mathbf{y} - \frac{c}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}\right)^{-\frac{n+2}{2}} n\left(-\frac{\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}}{(c+1)^2}\right) \end{aligned}. \tag{5.36}$$

Setting (5.36) equal to zero and solving:

$$c_m = \frac{n\mathbf{y}^Y\mathbf{H}_\gamma\mathbf{y} - (q_\gamma+1)\mathbf{y}^T\mathbf{y}}{(q_\gamma+1)(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})}. \tag{5.37}$$

This means the starting point for detecting coalescence for the exact IMH sampler is (5.37) and the value of C the ensures $Cq(c) \geq f(c\,|\,\boldsymbol{\gamma},\mathbf{y},\mathbf{X})$ is (5.35) evaluated at (5.37). Note because we do not have the normalizing constant for (5.35) we cannot simply calculate the efficiency of the rejection sampler. The code for this rejection sampler is provided in Appendix D under cRejection1.m. We use this rejection sampler for the posterior for c conditional on the full model and the null model for the ozone, physical, bodyfat and crime datasets with a = 3 for $f(c)$. Ten thousand sample points were generated in each case and the density histogram plotted with the overlay of an inverse gamma density fitted by MLE (Figures 5.3 and 5.4). Table 5.8 records the minimum, mean and maximum waiting times for the rejection sampler.

The ozone data is much flatter for the null model than the full model. This is because the posterior of c according to (5.35) will become flatter as the RSS become larger. This is not always the case however as it is difficult to assess where the mode of the posterior will be. For the ozone data in both cases the fitted inverse gamma distribution does a good job, although it does appear to undercut the true density to the right of the mode where the density is rapidly decreasing into the right hand tail. The crime data show similar behavior to the ozone data as the distribution of c conditional on the null model is flatter than for the full. The inverse gamma approximations also perform well with similar undercutting of the estimated true density to the right of the mode.

The physical data shows different behavior and in fact the conditional density for c is flatter for the full model than the null model. This is because the mode for the full model is much further to the right than for the null model. For both the full and null model the inverse gamma approximation over-estimates the mode and undercuts at the right of the mode similarly to the ozone and crime datasets. For the bodyfat data the posterior of c conditional on the null model is flatter than that for the full model. The inverse gamma distribution for the null model over estimates the mode and undercuts the right hand tail more extensively than any of the other examples.

The inverse gamma distribution for the full model undercuts the mode and only slightly the right hand tail. Overall it does appear the inverse gamma distribution can provide a decent approximation to conditional posterior of c. The reason for this line of investigation will become clearer later in this section. It is also worth noting that the mode of the distributions above can be used to estimate c for a local empirical Bayes approach.

Table 5.8 shows the efficiency of the rejection sampler for each set of samples generated for Figure 5.3 and 5.4, where N = null model, F = full model. The efficiency is an artifact of how well the posterior and the proposal distribution agree. This means any distributions that are not flat or have a mode far from zero will be the least efficient.

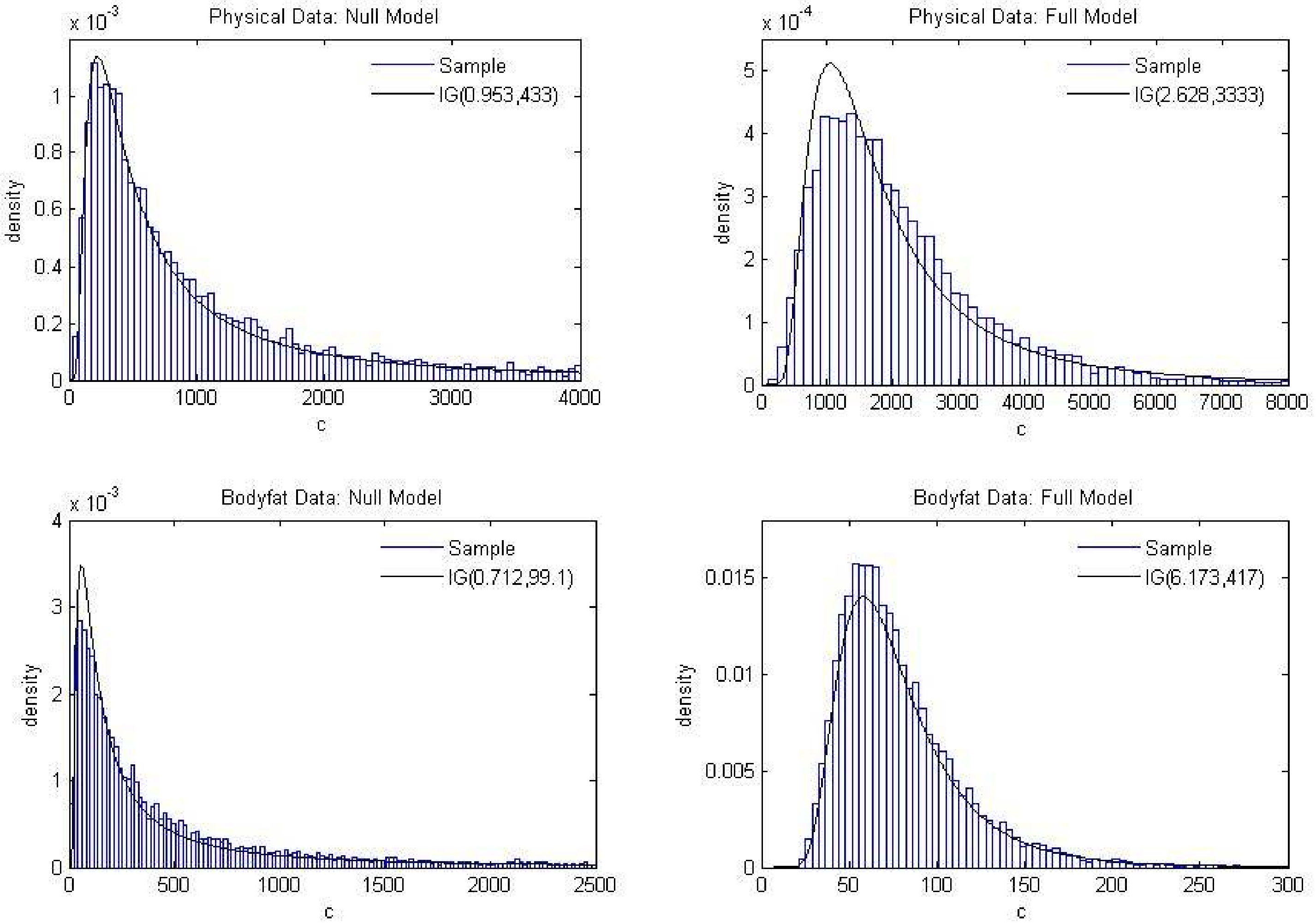


Figure 5.3 the posterior distribution of *c* conditional on the full and null models for the physical and bodyfat data. The parameters for an approximation to the posterior using an inverse gamma density are also shown.

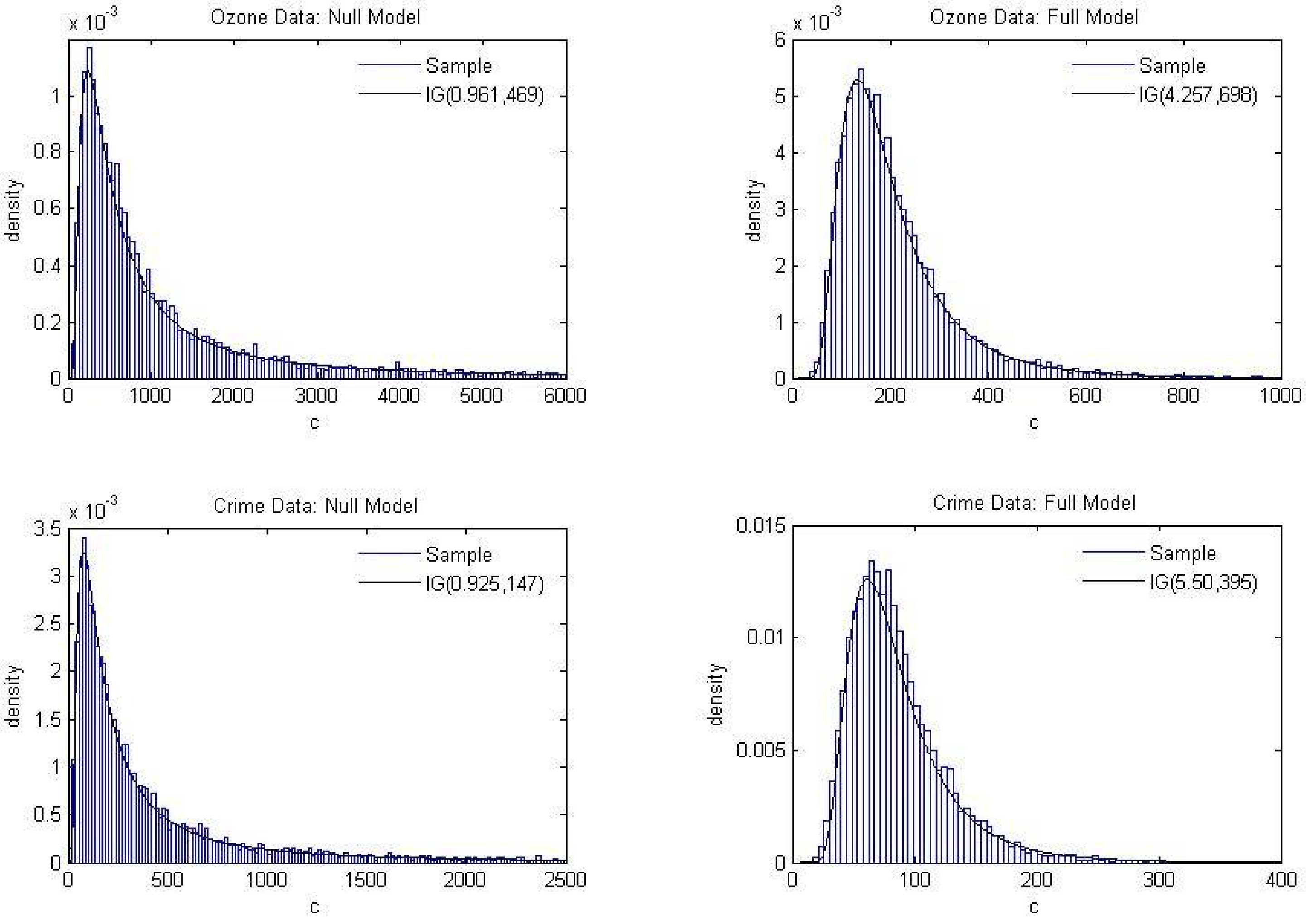


Figure 5.4 the posterior distribution of *c* conditional on the full and null models for the ozone and crime data. The parameters for an approximation to the posterior using an inverse gamma density are also shown.

Table 5.8 Efficiency of the Rejection Sampler for the Conditional Posterior of *c*

	Ozone		Physical		Bodyfat		U.S. Crime	
	N	F	N	F	N	F	N	F
Minimum	1	1	1	1	1	1	1	1
Mean	2.4	4.3	4.1	12.3	2.3	11.1	2.0	4.9
Maximum	17	34	36	107	16	87	16	45

Due to the increased flattening for the posterior conditional on the null model, the mean waiting time is smaller than for the full model unless the mode of the full model is much further from zero than for the null model. Overall the most efficient case for the null model is the crime data with a mean of 2 and a maximum of 16. The most costly under the null model is the physical data at 4.1 and 36. For the full model the most efficient is the ozone data with a mean waiting time of 4.3 and a maximum of 34. The least efficient is the physical data with 12.3 and 107. The comparison between efficiency for all cases is determined by how flat the distribution is, and where the mode is located. If the distribution is flat and has a mode close to zero the posterior agrees with the proposal and so the rejection sampler is efficient.

This approach was recommended for inference when integrating over c so we use the ozone data as an example.

Ozone Example

We now look at the difference in model averaging for integration over c by comparing $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}, c)$ and $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$. We use the ozone data and present the posterior distribution of observation 5 and $\boldsymbol{\beta}_9$. We chose a = 3 and τ = 0.45 to weakly penalize model complexity. The posterior for $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$ was obtained by Monte Carlo integration using simulation from $f(c)$ for every $\boldsymbol{\gamma}$.

Figure 5.5 shows $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}, c)$, $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$ and $f(c \mid \mathbf{y}, \mathbf{X})$. Comparing the two posteriors for $\boldsymbol{\gamma}$, integrating over c appears very similar but in this case produces less model uncertainty. This is because as c gets larger the posterior mass concentrates and then eventually moves towards the null model. Form the marginal posterior of c most of the density falls between 50 and 1500 so that most values of c are larger than c = n = 80 resulting in the more concentrated posterior.

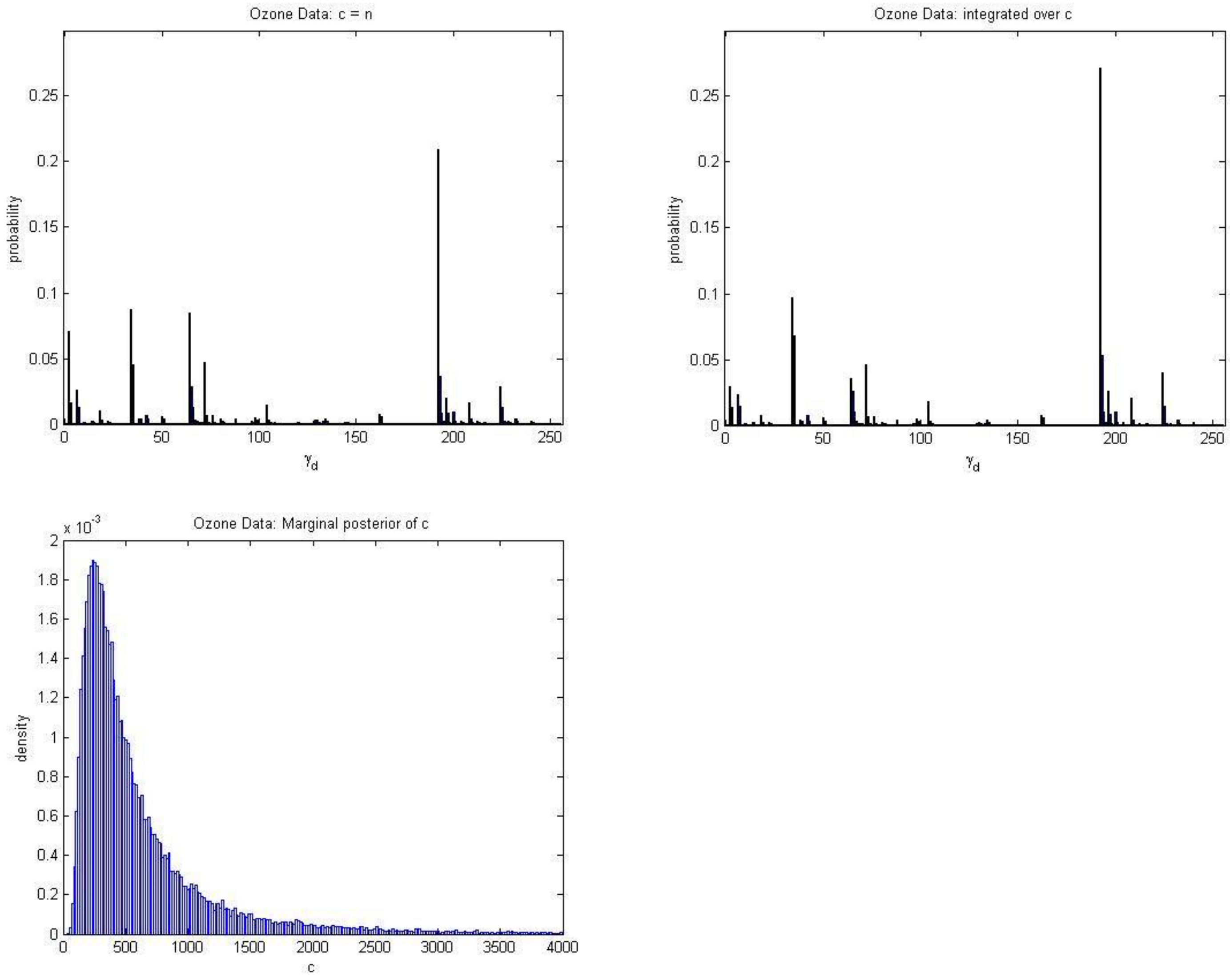


Figure 5.5 Posterior mass functions of γ for *c* = *n* and integrated over *c*. The marginal posterior for *c* is also shown.

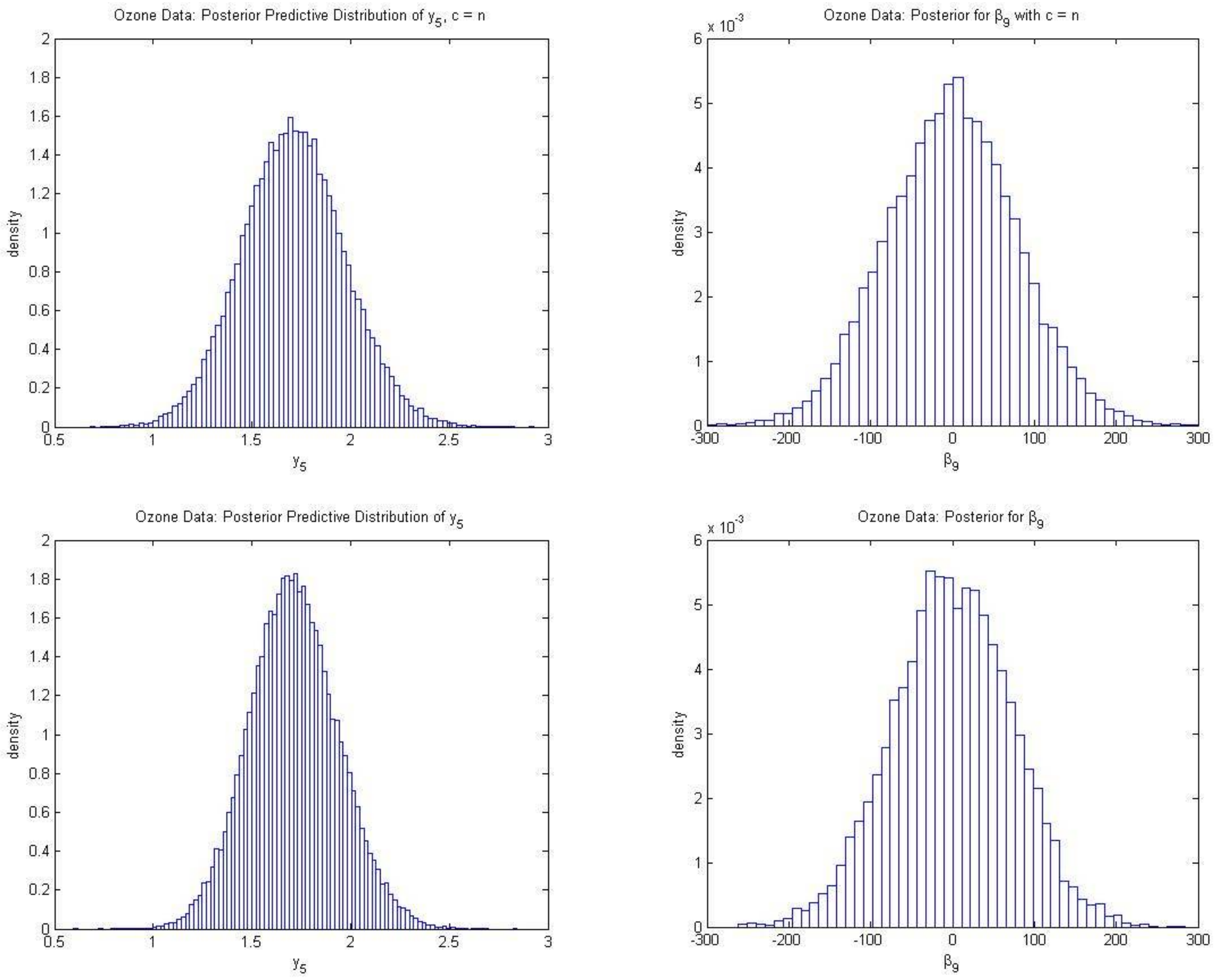


Figure 5.6 Posterior predictive density for observation 5, and posterior density of β_9 for $c = n$ (top) and integrated over c (bottom).

This means integrating over c is certainly a desirable approach for variable selection. Table 5.9 shows the MIP for the posteriors in Figure 5.5. Again we can see predictors 1, 2, 3, and 8 have increased MIP for $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$ compared to $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}, c)$, while predictors 4, 5, 6, and 7 have decreased MIP. Figure 5.6 shows the comparison of the posterior distributions of $\mathbf{y}_5$ and $\boldsymbol{\beta}_9$ for $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$ and $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}, c)$. Similarly to the results above we find the posteriors of $\mathbf{y}_5$ and $\boldsymbol{\beta}_9$ for $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$ exhibit less spread than the equivalent distributions with c = 80. In both cases the posterior of $\boldsymbol{\beta}_9$ is conditional upon an estimate of σ^2 which we take as the posterior expectation of the model averaged posterior for σ^2. The estimate of σ^2 is a function of the residual sum of squares term including the $c/(c+1)$ shrinkage term.

Table 5.9 Comparison of the Marginal Inclusion Probabilities for $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}, c)$ and $f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$.

Predictor	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}, c = n)$	0.44	0.64	0.28	0.09	0.15	0.13	0.40	0.25
$f(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$	0.52	0.67	0.31	0.07	0.13	0.12	0.37	0.27

Thus, because the values of c are predominantly larger when integrating over c the shrinkage term $c/(c+1)$ in the variance estimate will be closer to 1. This implies if most of the posterior density of c is supported by values larger than $c = n$, then the variance will on average be smaller according to the posterior distribution for σ^2. This also applies to the PPD where the variance is a function involving the RSS term with the shrinkage factor of $c/(c+1)$.

As noted earlier it appears from Figures 5.3 and 5.4 that the posterior for c can be reasonably approximated by an inverse gamma density. This suggests an inverse gamma proposal that matches the target well could make a very good proposal distribution. The idea is that in thinking about exact IMH we could increase the efficiency with suitable estimates, by finding the starting point as a function of the parameters of the proposal distribution. Such an approach provides a rejection sampler with the benefit of a direct way to think about adaptation to improve efficiency. Assuming an inverse gamma proposal with parameters v and w, and using hyper-G-n prior we can find c_m with:

$$\frac{df(c\,|\,\boldsymbol{\gamma},\mathbf{y},\mathbf{X})}{dc}=\frac{d}{dc}\frac{(c+1)^{-\frac{q_\gamma+1}{2}}\left(\mathbf{y}^T\mathbf{y}-\frac{c}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}\right)^{-\frac{n}{2}}(1+c/n)^{-a/2}}{c^{-(v+1)}\exp(-w/c)}, \tag{5.38}$$

which is:

$$\begin{aligned}\frac{df(c\,|\,\boldsymbol{\gamma},\mathbf{y},\mathbf{X})}{dc}=&\frac{c^{v+1}(c+1)^{-\frac{q_\gamma+1}{2}}\left(\frac{q_\gamma+1}{2}\right)\left(1+\frac{c}{n}\right)^{-a/2}}{\exp\left(-\frac{w}{c}\right)\left(\mathbf{y}^T\mathbf{y}-\frac{c}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}\right)^{n/2}}-0.5\ldots\\ &\frac{c^{v+1}(c+1)^{-\frac{q_\gamma+1}{2}}n\left(-\frac{\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}}{(c+1)}+\frac{c\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}}{(c+1)^2}\right)}{\exp\left(-\frac{w}{c}\right)\left(\mathbf{y}^T\mathbf{y}-\frac{c}{c+1}\mathbf{y}^{\mathbf{T}}\mathbf{H}_\gamma\mathbf{y}\right)^{(n+2)/2}\left(1+\frac{c}{n}\right)^{a/2}}-0.5\ldots\\ &\frac{ac^{v+1}(c+1)^{-\frac{q_\gamma+1}{2}}\left(1+\frac{c}{n}\right)^{-(a+2)/2}}{n\exp\left(-\frac{w}{c}\right)\left(\mathbf{y}^T\mathbf{y}-\frac{c}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}\right)^{n/2}}-\ldots\\ &\frac{(-v-1)c^{v+1}(c+1)^{-\frac{q_\gamma+1}{2}}\left(1+\frac{c}{n}\right)^{-a/2}}{\exp\left(-\frac{w}{c}\right)\left(\mathbf{y}^T\mathbf{y}-\frac{c}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}\right)^{n/2}}-\ldots\\ &\frac{wc^{v+1}(c+1)^{-\frac{q_\gamma+1}{2}}\left(1+\frac{c}{n}\right)^{-a/2}}{c^2\exp\left(-\frac{w}{c}\right)\left(\mathbf{y}^T\mathbf{y}-\frac{c}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}\right)^{n/2}}\end{aligned} \tag{5.39}$$

(5.39) can be arranged into a polynomial expression, hence there is more than one possible solution. Thus we must solve:

$$0=r_4c^4+r_3c^3+r_2c^2+r_1c+r, \tag{5.40}$$

where:

$$\begin{aligned}
r_4 &= (\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})(q_\gamma + a - 2v - 1) \\
r_3 &= (\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})(q_\gamma - 2vn + 2w) + (2\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})(a - 2v) + \ldots \\
&\quad (q_\gamma - n - 3)\mathbf{y}^T\mathbf{y} + 2\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y} \\
r_2 &= (2w - 2vn)(2\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}) + 2wn(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}) + (2-n)n\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y} + \ldots \\
&\quad (q_\gamma n - 3n + a - 2v - 2)\mathbf{y}^T\mathbf{y} \\
r_1 &= (w - n - vn + 2wn)2\mathbf{y}^T\mathbf{y} - 2wn\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y} \\
r &= 2wn\mathbf{y}^T\mathbf{y}
\end{aligned} \tag{5.41}$$

Restricting the solutions of (5.40) to the positive real line and then from the remaining candidate values taking the correct choice as the value that maximizes (5.38) we obtain c_m and the required bound C. The efficiency of this approach relies on the choice of v and w. This requires that estimation of v and w from a posterior sample converges to some value. Using the first approach with $q(c) = f(c)$ we generated 100 batches of 10000 sample points and for each we estimated v and w using MLE after 100, 200, 300,… and so on sample points. The results are shown in Figure 5.7 where we have chosen γ to be the full model. Using these estimated values of v and w we also recorded the corresponding value of c_m to be used in the new rejection sampler shown in Figure 5.8. Figure 5.8 also plots various quantiles of the posterior to show where the estimates of c_m fall in relation to the tails.

From Figure 5.7 the plot for each parameter v and w shows clear convergence towards values of approximately 4.3 and 700 respectively. The variability, ignoring the extreme values, in the estimates of v and w requires around 3000 sample points before stabilizing. From Figure 5.8 we can see that the value of c_m is converging to a point (approximately 290) just outside the 75th percentile of the posterior distribution. This agrees with the observation in figure 5.4 where the posterior appeared to dominate the proposal in the right hand tail close to the mode. Provided we have enough samples we can obtain a suitable approximation to the posterior using the estimated parameters. Using the estimated parameters from the ten thousand sample points generated as shown in Figures 5.3 and 5.4 we run the modified rejection sampler with the inverse gamma proposal.

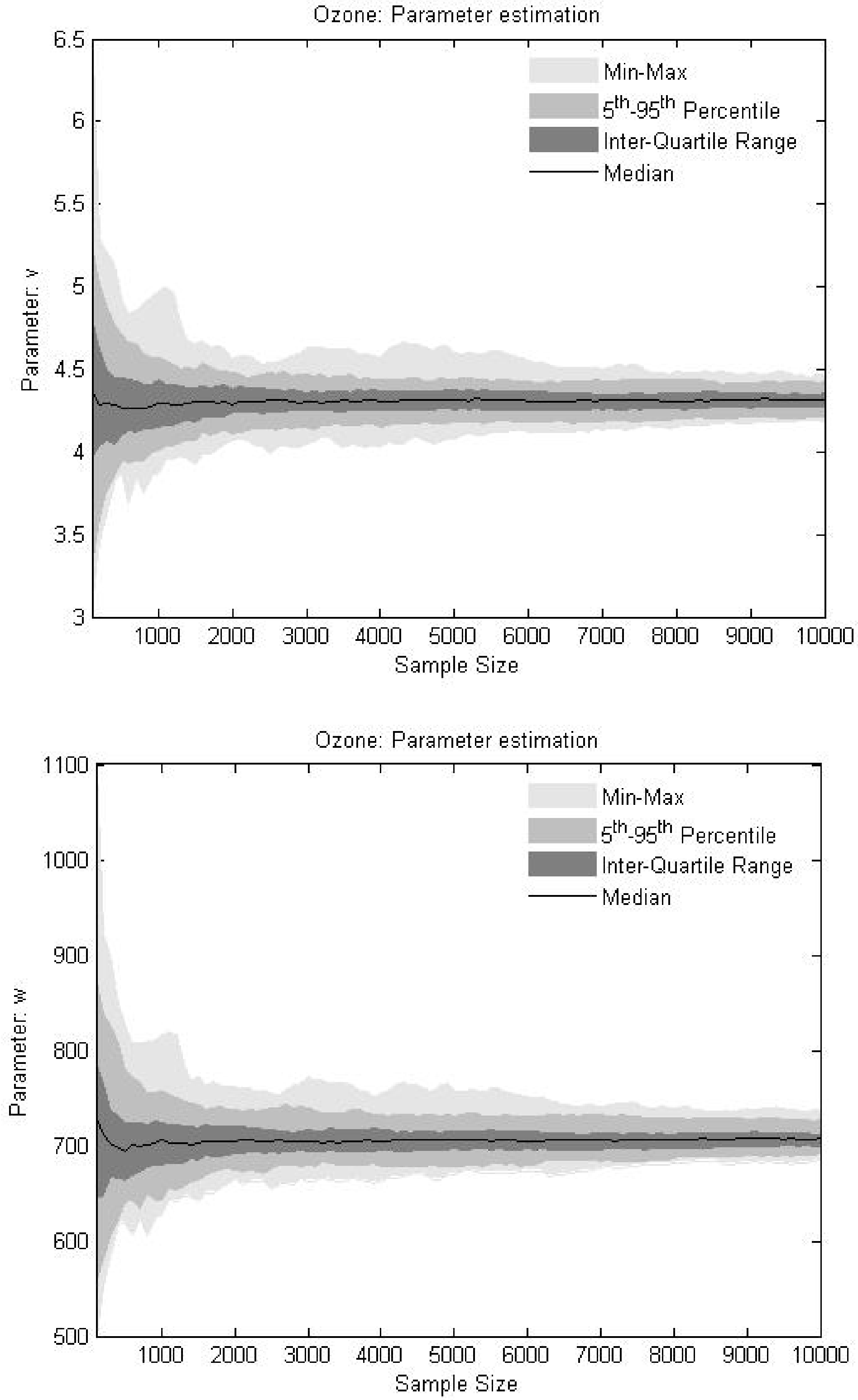


Figure 5.7 The convergence of parameter estimation for the inverse gamma approximation. γ is chosen to be the full model.

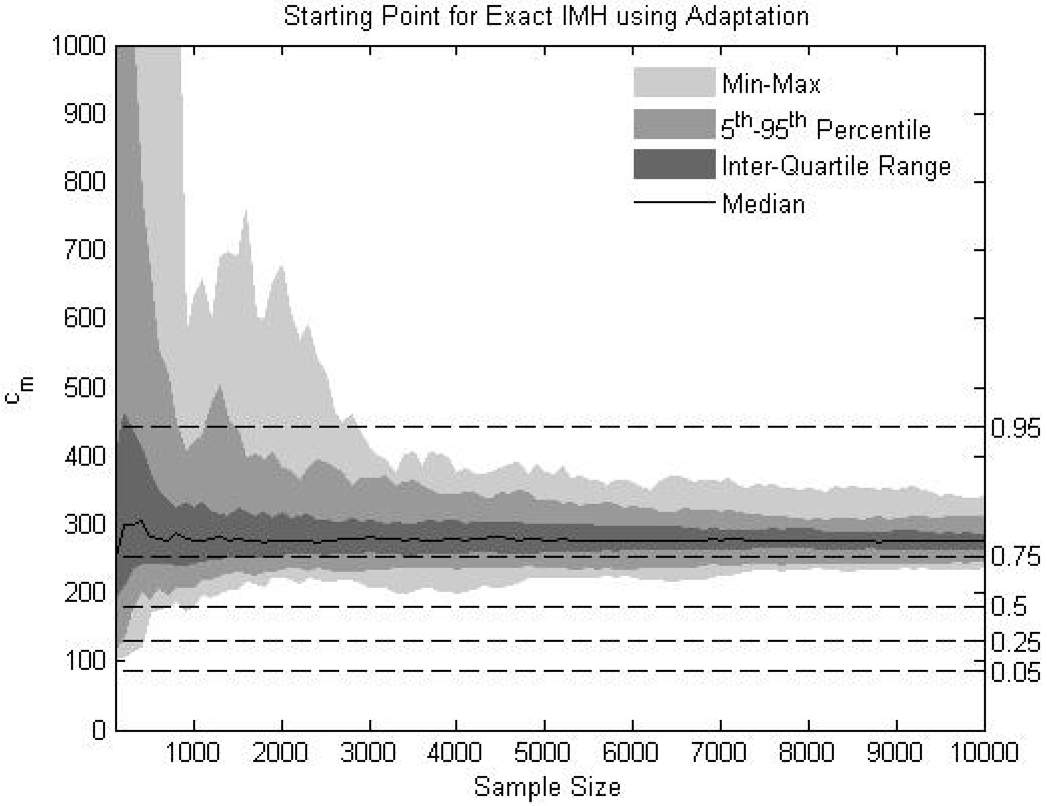


Figure 5.8 Convergence of the estimated value of c_m and the relation to the posterior distribution for c. γ is chosen to be the full model.

Table 5.10 Efficiency of the Rejection sampler with adapted values.

	Ozone		**Physical**		**Bodyfat**		**Crime**	
	N	**F**	**N**	**F**	**N**	**F**	**N**	**F**
Minimum	1	1	1	1	1	1	1	1
Mean	1.02	1.06	1.12	1.25	1.26	1.04	1.08	1.11
Maximum	3	4	5	8	6	4	4	5

Ten thousand sample points are generated and the efficiency of the rejection sampler is estimated for comparison with Table 5.10. The code for this rejection sampler is given in Appendix D under cRejection2.m.

Comparing Table 5.8 to Table 5.10 there is a dramatic increase in efficiency. For all datasets the mean waiting time is close to one. This indicates that the inverse gamma distribution is a suitable approximation so that the rejection sampler is close to i.i.d.

sampling but not quite. The largest waiting time is 8 and the smallest is 3. These results clearly indicate the benefit of thinking about rejection sampling and exact IMH as one and the same. The approach is straight forward and greatly improves the efficiency. The reader may be wondering why we have no results starting from a default proposal distribution such as the **IG**(1,1). The reason for this is in this application the efficiency and numerical stability of the solution is unfortunately very sensitive to the choice of parameters in the proposal. Specifically in order to have good efficiency and a suitable starting value the mode of the proposal and target must be similar otherwise the target will dominate the proposal far out in the right hand tail. This is undesirable as the proposal should have heavier tails than the target.

5.4 Summary

The use of the exact IMH sampler appears to be redundant as the detection of coalescence is a rejection sampler for the target distribution. This relationship was not unknown, and provides an elegant demonstration of the relation between rejection sampling and IMH as well as the relationship between BCT, coupling times and regeneration times. Ultimately, the rejection sampler should be the preferred option for exact sampling with exact IMH. The method of Schneider and Corcoran (2004) is also detecting coalescence using rejection sampling and so a rejection sampler could have been used without the need for the added forward propagation.

With these facts in mind, we then investigated the use of rejection sampling for the marginal posterior for γ. Noting that the RSS is weakly decreasing in the number of predictors by virtue of Gaussian least squares optimization, then reducing $f(\gamma \,|\mathbf{y}, \mathbf{X})$ to a function of the RSS only will allow us to find a bound for rejection sampling.

We provide the methodology to perform the reduction for both Zellner's and Jeffreys prior by choosing the proposal distribution to be a function of the penalty term in the posterior and $f(\gamma)$. The distribution of mass under the prior for γ and the strength of the penalty will have a marked impact on the efficiency of the rejection sampler. This is because IMH sampling and equivalently rejection sampling, will be inefficient when the proposal is at odds with the posterior. Thus, the more extreme the penalty, or the more mass under the prior supporting small models, the longer the expected waiting time for generating sample points. The results using the real datasets using the constant Bernoulli prior for γ supports this conclusion. Generally values of $\tau > 0.7$ and small values of penalty are required for the rejection sampler to be feasible. Further examples demonstrated that sampling under Jeffreys prior will typically be inefficient even with $p = 2\pi$.

Zellner's prior however is feasible when c is relatively small. The results using the crime data and larger datasets indicated that rejection sampling can indeed be efficient. For the moment efficiency refers the mean waiting time and number of proposals required. This is because the Gibbs sampler may have a faster rate of convergence in real time. This is clearly a comparison for future work.

In the example from Schneider and Corcoran (2004), a fictitious data set with $k = 3$, $n = 20$ and conjugate priors were used. The recorded minimum, maximum and mean backwards coupling time for that example was 2, 4257 and 543 respectively. This suggests that rejection sampling in this case is remarkably inefficient, so while appealing in theory clearly has limited appeal based on practicality. There is crucial distinction between their method and the rejection sampler used here. In our case if the required number of proposals is larger than 2^k, then direct calculation of the posterior should be preferred, however, this is not possible for the approach of Schneider and Corcoran (2004).

Having provided a way to generate exact samples for BVS from the marginal posterior of γ, it is reasonable to conclude from the results for the real datasets that Zellner's prior is preferred to Jeffreys prior when k is large and n is small. Using $c = n$ and no available prior information for β should allow a moderately efficient rejection sampler to be used for BVS as described.

The method of the exact IMH/rejection sampling is closely related to the use of perfect simulated tempering in both the backwards or forwards context. The use of these methods is a possibility for future research, although it is unlikely to prove much more efficient than the standard rejection sampler. This is because the waiting times will be the same as for the rejection sampler and will provide an estimate of the mean waiting time for the perfect forwards simulated tempering.

With the promising results of the rejection sampler for $f(\gamma\,|\mathbf{y}, \mathbf{X})$ we then moved to the case of generating exact samples from the marginal posterior for c conditional on γ. This bears relevance for model averaging when γ is integrated over c. The second motivation was to demonstrate how thinking of exact IMH and rejection sampler as one and the same provides a useful perspective for improving the efficiency of a rejection sampler.

The marginal posterior for c conditional for γ is typically quite flat as is the hyper-G-n prior. This means using the prior for c as the proposal distribution produces a small amount of conflict between the target and proposal and hence, an efficient rejection sampler. This provides a method of generating sample points and showed the posterior can be well approximated by an inverse gamma density. With this in mind, the next step was to find a starting point for the exact IMH, and hence optimal bound for rejection sampling as a function of the parameters of the proposal density.

The solution of the derivative was not available in a simple form and has multiple solutions. However, with the appropriate restrictions one of the root solutions to a polynomial function provided the optimal bound. Using samples generated by the first rejection sampler simulation results showed good convergence of the estimated parameters for the proposal distribution and consequently, the value of c_m for reasonable sample sizes. Consistent tail behavior may be an issue for future research as if the positive solutions are too far from mode of the posterior, then we are no longer dominating the target in the tails with the chosen proposal density. Further, future research will involve reviewing literature to find densities where the ideas discussed above can, and have yet to be applied, to various posterior densities in Bayesian analysis.

CHAPTER 6

DISCUSSION AND CONCLUSIONS

> "*Some problems are so complex, that you have to be highly intelligent and well-informed, just to be undecided about them*."
>
> - Laurence J. Peter

We now discuss and summarize the finding of this research. We break this discussion up into sections, following the aims outlined in the first chapter, and discuss each in turn. This is followed by a more general discussion, recommendations and an outline of future research.

Assuming an orthogonal predictor matrix, check the robustness in the construction of monotone Gibbs Markov chains to choices of priors and hyper-parameters.

The construction of a Gibbs monotone MC is not as straight forward as simply having an orthogonal predictor matrix. Previous work (Kuo and Mallick, 1998; Holmes and Denison, 2002) has mainly focused on the fully conjugate case, including Zellner's prior. Assuming $\mathbf{X}$ is orthogonal, the error distribution should be Gaussian, and the priors for $\boldsymbol{\beta}$ and σ^2 are restricted to the conjugate class, or non-informative type priors such as Jeffreys prior.

For the conjugate class of priors (including Zellner's) and a fully Bayesian approach, we cannot obtain a monotone Gibbs MC if we choose prior subjective values of $\boldsymbol{\beta}$. This constraint also includes the special case of Zellner's projection prior. Otherwise, with $\widetilde{\boldsymbol{\beta}}$

= 0, and any choice of *a* and *b* for the inverse gamma prior for σ^2, monotonicity is available with an orthogonal predictor matrix. Introducing weighted least squares for the purpose of detecting or guarding against outliers, produces a covariance matrix that is no longer orthogonal. Thus, methods such as those in Smith *et al* (1996) cannot be improved by perfect sampling with a monotone Gibbs MC.

For priors on $\boldsymbol{\gamma}$, the literature assumes either flat, or constant Bernoulli priors, which will indeed allow a monotone Gibbs Markov chain, provided the choices of prior for $\boldsymbol{\beta}$ and σ^2 do so. We extend this to the case of assuming a beta hyper-prior, for the choice of τ in the constant Bernoulli prior. For any choice of hyper hyper-parameters in the beta hyper-prior, a monotone Gibbs MC is possible provided other conditions are met as above. It is possible to specify a generic prior on the q_γ space, and obtain a monotone Gibbs MC provided $f(q_\gamma)$ does not involve the binomial coefficient, and is non-increasing for increasing q_γ. An example of this is to use a truncated Poisson distribution for q_γ.

Recent work by Cripps *et al* (2006), allowed the hyper-parameters of the prior for σ^2 to depend on $\boldsymbol{\gamma}$ thus, taking an EB approach to variable selection and model averaging. Specifically, Cripps *et al* (2006) choose hyper-parameters for $f(\sigma^2)$ to obtain the classical estimator for variance as the mode of the posterior distribution of σ^2. For the associated posterior for $\boldsymbol{\gamma}$ the Gibbs MC is not monotone. We then used an analogue of the approach of Cripps *et al* (2006) by using the MLE of the variance. The Gibbs MC is monotone for this choice. This highlights an interesting issue for EB methods in this context. Hyper-parameters can be chosen to produce an estimator of a certain form for a posterior distribution dependent on $\boldsymbol{\gamma}$. The estimator, such as the mean, can typically be expressed as a function of parameters for that posterior distribution. The dependence on $\boldsymbol{\gamma}$ requires these parameters must also observe the required partial order, otherwise monotonicity is unavailable. The shape parameter of the inverse gamma distribution typically cannot depend on $\boldsymbol{\gamma}$, as this prevents monotonicity.

Recent work on Zellner's prior involves assigning a prior to *c* (Celeux *et al*, 2007). If we assume the hyper-G prior of Liang *et al* (2008) and use the MLE of $\boldsymbol{\beta}$ in Zellner's prior, then a closed form expression for the posterior for $\boldsymbol{\gamma}$ integrated over *c* is available. This posterior with an orthogonal predictor matrix will permit a monotone Gibbs MC.

Determine the effect of using an orthogonal predictor matrix on inference using model averaging and the linear regression model.

Clyde *et al*, (1996) provide four methods of orthogonalization for use in transforming a non-orthogonal predictor matrix. The GS procedure was noted by Clyde *et al* (1996) and Holmes and Mallick (1998), however the requirement to order the columns of **X** before use deterred any attempt to use it in practice. Holmes and Mallick (1998) are the first to use an orthogonal transformation and then use the monotone Gibbs CFTP, for the Bayesian linear model for a standard regression problem. Holmes and Mallick (1998) provide little detail on suitable methods for orthogonalizing in the regression context. Thus, the kind of extensive analysis of Clyde *et al* (1996) and the application of perfect sampling with the monotone Gibbs sampler, have yet to be done together. In this context we applied three new methods of orthogonalization. The first, the Lowdin method, is invariant to the order of columns in the **X** space and is an extension of SVD. SVD is a way to obtain an orthogonal design matrix in its own right, and recent work by Beaver (2007) shows the Lowdin transformation minimizes the L^2 matrix norm between **X** and **W**. The other two methods use the modified GS transformation, and so depend on the order of the columns of **X**. We suggest an initial naïve method of ordering based on the correlation between **y** and **X**, and a slightly more complicated method, by taking account of the correlation between predictors. This approach was inspired by the partial least squares approach, which uses **y** in the orthogonalization of **X**.

Using **W** instead of **X** has a number of effects. The most obvious is the ability to access a monotone CFTP Gibbs sampler, per the sufficiency requirements established in Chapter 2 and discussed above. **W** also induces a larger disparity between a collection of good and bad models, as such, the posterior mass becomes more concentrated compared to the posterior using **X**. This shrinkage effect was evident in plots of the residual sum of squares, model competition, and entropy. The expected model size for γ does not appear to be vastly different for **W** compared to **X**. The DIC study and model checking provided strong evidence, that for most choices of the hyper-parameters τ, c and p, the use of **W** instead of **X** provides comparable performance for in-sample prediction. However, in using **W** we sacrifice variable selection, and possibly out-of-sample predictive ability. There was some suggestion that the use of orthogonal transformations may also be more reliable in large sample situations, or certainly when *n* is much larger than *k*. The use of **W** also provides faster computation of the posterior, and the Gibbs update probabilities. The use of **W** also aids convergence by reducing the dependence between the components in $\boldsymbol{\gamma}$. These points coupled with the additional concentration of mass in the posterior under **W**, suggests that a Gibbs MC in the **W** space, should display superior real time rates of convergence compared to the equivalent sampler using **X**. This is indeed what we observed.

The DIC and model checking studies indicated that the GPC, as found in previous work, was a consistent and reliable method for transformation. The GS_2 method also proved to be extremely consistent and reliable, offering a new method of orthogonalization for use in practice. The GS_1 and Lowdin transformation methods, were less reliable and as such, we would have to recommend some caution if using these methods. Clyde *et al* (1996) felt they could not recommend any particular orthogonal method they investigated, as an optimal choice. While we cannot suggest an optimal choice either, because of the more extensive empirical investigation, it seems based on these results that the GPC method is a reliable option, as is the GS_2 method.

From three versions of the Gibbs sampler; standard with the original predictor matrix, standard with an orthogonal predictor matrix and perfect with an orthogonal predictor matrix, determine which is the best choice according to computational efficiency and rate of convergence to the stationary distribution.

We compared monotone Gibbs CFTP using **W**, and the Gibbs sampler in the **X** space. We also considered the competitiveness of the standard Gibbs sampler using an orthogonal predictor matrix, which then gave rise to a hybrid method. There was no indication of a substantial difference based on inference between the orthogonal Gibbs sampler and the perfect version. The standard Gibbs sampler requires a similar amount of computing time to monotone Gibbs CFTP when the BCT is minimal. However, the monotone Gibbs CFTP will converge faster in real time. The comparison is complicated by burn-in for the Gibbs sampler and the incomparability of the posteriors for **X** and **W**. The requirement for burn-in will reduce the efficiency of generating usable sample points with the Gibbs sampler for **X**. The shrinkage effect allows the Gibbs sampler for **W** to converge faster than the equivalent sampler for **X**.

For the monotone Gibbs CFTP, we also spent time investigating factors affecting the BCT. This is because if the BCT increases, then monotone Gibbs CFTP becomes less computationally competitive with other sampling methods. Results suggest that decreasing information or increasing *k*, increases the BCT. Information is related to the ratio of the sample size to the number of predictors. If *n* is close to *k* a lot of uncertainty surrounds the inclusion of predictors, as a result, there is a lot of model competition in the posterior for γ which is indicated by entropy. This result can be affected by the choice of hyper-parameters. Hyper-parameter choices that reduce model competition and hence entropy, will reduce the BCT. These choices are typically extreme values such as 0 and 1 for τ, and small (<10) and extremely large ($>10^6$) for *c* or *p*. These extreme choices generally drive the posterior towards the null or full model. The exception to this rule is demonstrated by considering the maximum probability of remaining un-coalesced for the perfect sampler. The example we use demonstrates how even though extreme values of penalty reduce the entropy, the BCT can still be noticeable because of a large probability or remaining un-coalesced for a single component. This can occur when the posterior is transitioning towards the null model as the penalty increases. Bi-modality occurs as the mass shifts from the model with the two most important predictors, to a model containing only one. This can introduce a large probability of remaining un-coalesced for the predictor included in one model and not the other. This in turn can increase the BCT.

The orthogonal Gibbs sampler, proved to be more computationally efficient than the perfect sampler. What is interesting is the comparable convergence to the perfect sampler. In particular, the convergence was similar per sample point, and the minimal dependence in the orthogonal Gibbs chain can be shown with auto-correlation plots. This leads to an effective sample size for the examples we investigated, essentially equivalent to the number of samples generated. This means from a practical point of view, the dependent chain is essentially as good the i.i.d. sequence from monotone Gibbs CFTP. The convergence in real time provided even more support for the efficiency of the orthogonal Gibbs sampler. This evidence leads to the suggestion of using a combination of perfect sampling, and the standard orthogonal Gibbs sampler. This hybrid approach uses monotone Gibbs CFTP to detect a starting point at time zero, and then the single coalesced Gibbs chain is allowed to continue onwards from time zero. This removes the burn-in issue, but retains the maximum computational efficiency from using an orthogonal Gibbs sampler. Monotone Gibbs CFTP has the advantage of i.i.d. points and the hybrid method is almost as good with less computing time. Thus, we recommend either approach, as both are more efficient than the Gibbs sampler for $\mathbf{X}$, when modeling the response under model averaging. Ultimately the choice lies with the user however, we suggest monotone Gibbs CFTP should be used when the BCT is minimal and the hybrid method otherwise.

Provide further exploration of the application of the perfect sampling version of the independence Metropolis-Hastings algorithm for BVS.

Rejection sampling is the method by which we detect coalescence in monotone CFTP using the IMH sampler. This suggests that perfect sampling with IMH is a redundant concept, as there is no reason to use the recursive framework for exact IMH, when we can use rejection sampling whenever exact IMH is possible. The perfect sampler of Schneider and Corcoran (2004) for BVS is also a rejection sampler.

We manipulate the proposal distribution for IMH, so the required bound for rejection sampling may be found as a function of the residual sum of squares. For Zellner's prior with $\widetilde{\boldsymbol{\beta}}$ = 0 and Jeffreys prior, rejection sampling is only feasible for choices of hyper-parameters that minimize the difference between the posterior and proposal. Using Zellner's prior when n is close to k, and choosing $c = n$, the rejection sampler was efficient enough for practical use. The difference between the method of Schneider and Corcoran (2004) and our work, is their method was inefficient, requiring thousands of samples to generate one exact value for very small values of k, in their case 3. However, their approach is able to incorporate a prior choice for $\widetilde{\boldsymbol{\beta}}$. The rejection sampler we propose cannot.

The rejection sampler for the marginal posterior distribution for *c*, allowed a demonstration of how useful it can be to think of a rejection sampler as an IMH sampler. The approach provided a framework that allowed the efficiency of the rejection sampler, to be improved though adaptation of the proposal. Specifically, empirical estimates of the parameters for the proposal distribution can be adapted, as we generate samples from the posterior. Even though the rejection sampling using $f(c)$ as the proposal was quite efficient the adaptation improved the efficiency noticeably.

General comments

Holmes and Mallick (1998) were the first to apply perfect sampling to the Bayes linear model for a regression problem, and subsequent work by Holmes and Denison (2002) focuses on the wavelets case explicitly. The wavelets case decomposes the response into a series of orthogonal basis functions, which are then used as predictors to reconstruct the response. For regression we already have existing non-orthogonal predictors. This is the crucial difference between the linear regression problem and the wavelets regression problem. Further, for the wavelets case typically the series length (n) must be a power of 2, and the decomposition generates n separate basis functions as predictors. Thus, with $k = n$ we can expect the BCT to be larger for the wavelets case, compared to standard regression where $n >> k$. This may seem trivial and while BCT are often reported in previous work, no discussion is given on why the two may differ. In the work by Lee *et al* (2005) even though BCT were around 16 to 31 which seems long by comparison to the regression case, clearly this is not as extreme as it can get. This suggests that the analysis done on both sufficient conditions for monotonicity, BCT and the comparison of a hybrid orthogonal sampler compared to the perfect sampler, are all relevant to the wavelets case. In particular, the work we have done suggests we may improve the BCT with choices of hyper-parameters, and that the hybrid approach may also be useful.

A point of interest in this work is the additional investigation of Jeffreys true prior. We note ad hoc adjustments of this fashion have been done before by Wasserman (2000). The motivation behind use of the adjusted Jeffreys prior is to allow the sensible introduction of a suitable penalty term. Thus, if n is large, k is small and no subjective information is available, there should be little objection to the use of the adjusted Jeffreys prior. Note that such a prior may not be a reasonable suggestion for the wavelets work as $n = k$. In particular, the monotone Gibbs CFTP is well suited to cases with larger n and small k, and so we suspect would typically be employed with perfect sampling. We also note that the alternate Jeffreys prior performed well for inference when modeling $\mathbf{y}$, but did not provide an efficient rejection sampler for BVS for some reasonable choices of hyper-parameters.

Recommendations

The recommendations we make following the work of chapters 2, 3 and 4 are as follows. When modeling the response $\mathbf{y}$ orthogonalize using either the GPC or GS_2 method, then:

1. If k is large and $n >> k$ use perfect sampling and Jeffreys alternate prior as we have strong information from the data and a minimal BCT
2. If k is large and n is close to k use the hybrid approach to avoid the large BCT due to little information from the data. Use Zellner's prior with a standard choice for c such as $c = n$.

When choosing hyper-parameters, avoid choices that disrupt monotonicity, or choices that greatly increase the BCT for case 2 above.

1. For Zellner's prior set $\tilde{\boldsymbol{\beta}} = 0$.
2. For the conjugate prior for σ^2 avoid choices of a and b that result in the posterior mean or median corresponding to the classical estimator of variance.
3. Avoid extreme choices of τ, c or p.

From chapter 5:

1. Exact IMH is redundant as it is built on rejection sampling, thus the rejection sampler should always be used when ever exact IMH can.
2. When performing variable selection use the rejection sampler when n is close to k, k is large and use Zellner's prior with $c = n$, $\tilde{\boldsymbol{\beta}} = 0$. Choose the value of τ in the constant Bernoulli prior as required.

3. If a maximizable expression for the ratio of a posterior to a proposal distribution can be obtained, then we may use empirical estimates of those parameters as sampling progresses to improve the efficiency of rejection sampling.

Future Work

We now highlight some avenues for future research that have arisen as a consequence of the work presented in this thesis.

1. Determine if a monotone Gibbs sampler for error distributions other than the Gaussian, can be constructed for BMA in linear regression.
2. Alternative perfect sampling methods to those discussed in this thesis for joint posteriors. Joint posterior of interest in Bayesian linear regression may include for example, the joint posterior for variable selection and outlier detection.
3. Determine if out-of-sample prediction under suitable conditions, such as large n and an orthogonal design matrix, is competitive with out-of-sample prediction using $\mathbf{X}$.
4. Determine if simulated tempering can be more efficient than the rejection sampler in chapter 5.
5. Extend a rejection sampler (or other perfect sampling method) to the marginal posterior for $\boldsymbol{\gamma}$ for $\widetilde{\boldsymbol{\beta}} \neq 0$.
6. Conduct a literature review for suitable posteriors where solutions of the starting point for exact IMH, is available as a function of the parameters of the proposal distribution. This will allow the rejection sampler to be adapted for increased efficiency.
7. Investigate in detail, the relationship between rejection sampling and Fill's algorithm.
8. Examine if perfect sampling within a larger Gibbs chain (Metropolis within Gibbs for example) provides any improvement to convergence and/or inference.

In summary, from a theoretical standpoint perfect sampling is an elegant and appealing method to generate exact i.i.d. sample points from posterior distributions in Bayesian analysis. However, practical considerations complicate the desire to use perfect sampling as a standard tool. First of all, constructing a perfect sampling algorithm for a specific problem can be difficult at best, and impossible at worst. The rate of convergence of the underlying MC, dictates the distribution of geometric waiting times for generating perfect samples, and so may only be as efficient as the underlying MC. Perfect sampling algorithms typically require greater computing effort, both in terms of calculation and memory. The main disadvantage with MCMC is the requirement to assess the burn-in period, which perfect sampling removes the need for. Thus, perfect sampling with all its theoretical appeal is likely to be a sought after sampling method for some time to come. In recent times perfect sampling has received less attention with a shift towards adaptive MCMC methods; however, perfect sampling is a topic well worth the research, despite the potential difficulties.

APPENDIX

APPENDIX A: Probability Distributions

We omit the continuous uniform and discrete densities, and provide those probability distributions that are of particular relevance to the work in this thesis.

Bernoulli

Notation: $X \sim$ **Br**(p)

Form: $$p(x) = p^x (1-p)^{1-x} \mathbb{I}_{\{0,1\}}(x)$$

Conditions: $0 \le p \le 1$.

Beta

Notation: $X \sim$ **Be**(α, β)

Form: $$p(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \mathbb{I}_{[0,1]}(x)$$

Conditions: $\alpha, \beta > 0$.

Binomial

Notation: $X \sim$ **Bi**(n, p)

Form: $$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \mathbb{I}_{\{0,\ldots,n\}}(x)$$

Conditions: $0 \le p \le 1$, $n \in \{1,2,\ldots\}$.

Geometric

Notation: $X \sim$ **Ge**(p)

Form: $$p(x) = (1-p)^{x-1} p \mathbb{I}_{\{1,2,\ldots\}}(x)$$

Conditions: $0 \le p \le 1$.

Inverse-Gamma

Notation: $X \sim$ **IG**(α, β)

Form:

$$p(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp(-\beta / x) \mathbb{I}_{[0,+\infty)}(x)$$

Conditions: $\alpha, \beta > 0$.

Normal

Notation: $X \sim \mathbf{N_p}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$

Form:

$$p(\mathbf{x}) = (2\pi)^{-p/2} \,|\, \boldsymbol{\Sigma} \,|^{-0.5} \exp\left[-0.5(\mathbf{x}-\boldsymbol{\theta})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\theta})\right]$$

Conditions: $\boldsymbol{\Sigma}$ is a p by p positive-definite symmetric matrix; $\mathbf{x}, \boldsymbol{\theta} \in \mathbb{R}^{\mathrm{p}}$.

Poisson

Notation: $X \sim \mathbf{Po}(\lambda)$

Form:

$$p(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \mathbb{I}_{\{0,1,\ldots\}}(x)$$

Conditions: $\lambda > 0$.

Student-t

Notation: $X \sim \mathbf{St_p}(v, \boldsymbol{\theta}, \boldsymbol{\Sigma})$

Form:

$$p(\mathbf{x}) = \frac{\Gamma((v+p)/2)/\Gamma(v/2)}{\sqrt{|\,\boldsymbol{\Sigma}\,|}(v\pi)^{p/2}}\left[1 + \frac{(\mathbf{x}-\boldsymbol{\theta})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\theta})}{v}\right]^{-\left(\frac{v+p}{2}\right)}$$

Conditions: $v > 0$, $\boldsymbol{\Sigma}$ is a p by p positive-definite symmetric matrix; $\mathbf{x}, \boldsymbol{\theta} \in \mathbb{R}^{\mathrm{p}}$.

APPENDIX B: Derivation of Posteriors

B.1 Conjugate Priors:

Priors:

$$f(\boldsymbol{\beta}_\gamma,\sigma^2\,|\,\boldsymbol{\gamma})=\mathbf{N}_{q_\gamma+1}(\widetilde{\boldsymbol{\beta}}_\gamma,\sigma^2\mathbf{V}_\gamma^{-1})\ \mathbf{IG}(a,b)\text{, with }f(\boldsymbol{\gamma})\propto 1$$

Posterior for γ:

$$f(\boldsymbol{\gamma}\,|\,\mathbf{y},\mathbf{X})\quad =\int\!\!\int f(\boldsymbol{\gamma},\boldsymbol{\beta}_\gamma,\sigma^2\,|\,\mathbf{y},\mathbf{X})d\boldsymbol{\beta}_\gamma d\sigma^2$$

$$\propto\int\!\!\int f(\mathbf{y}\,|\,\boldsymbol{\gamma},\boldsymbol{\beta}_\gamma,\sigma^2,\mathbf{X})f(\boldsymbol{\beta}_\gamma\,|\,\boldsymbol{\gamma},\sigma^2,\mathbf{X})f(\sigma^2\,|\,\mathbf{X})f(\boldsymbol{\gamma})d\boldsymbol{\beta}_\gamma d\sigma^2$$

$$\propto\int\left\{\int\mathbf{N}_n(\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma,\sigma^2\mathbf{I}_n)\mathbf{N}_{q_\gamma+1}(\widetilde{\boldsymbol{\beta}}_\gamma,\sigma^2\mathbf{V}_\gamma^{-1})d\boldsymbol{\beta}_\gamma\right\}\mathbf{IG}(a,b)d\sigma^2$$

$$\propto(2\pi)^{-\frac{q_\gamma+1}{2}}\,|\,\mathbf{V}_\gamma^{-1}\,|^{-\frac{1}{2}}\ldots$$

$$\ldots\int\left\{\int\exp\left[-\frac{1}{2\sigma^2}\left((\mathbf{y}-\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma)^T(\mathbf{y}-\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma)+(\boldsymbol{\beta}_\gamma-\hat{\boldsymbol{\beta}}_\gamma)^T\mathbf{V}_\gamma(\boldsymbol{\beta}_\gamma-\hat{\boldsymbol{\beta}}_\gamma))\right)\right]d\boldsymbol{\beta}_\gamma\right\}\ldots$$

$$\ldots(\sigma^2)^{-\left(\frac{n+q_\gamma+1}{2}+a\right)-1}\exp\left[-\frac{b}{\sigma^2}\right]d\sigma^2$$

$$=(2\pi)^{-\frac{q_\gamma+1}{2}}\,|\,\mathbf{V}_\gamma^{-1}\,|^{-\frac{1}{2}}\int\left\{\int\exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{y}^T\mathbf{y}+\boldsymbol{\beta}_\gamma^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma-2\boldsymbol{\beta}_\gamma^T\mathbf{X}_\gamma^T\mathbf{y}+\boldsymbol{\beta}_\gamma^T\mathbf{V}_\gamma\boldsymbol{\beta}_\gamma\ldots\right.\right.\right.$$

$$\left.\left.\left.\ldots+\widetilde{\boldsymbol{\beta}}_\gamma^T\mathbf{V}_\gamma\widetilde{\boldsymbol{\beta}}_\gamma-2\boldsymbol{\beta}_\gamma^T\mathbf{V}_\gamma\widetilde{\boldsymbol{\beta}}_\gamma\right)\right]d\boldsymbol{\beta}_\gamma\right\}(\sigma^2)^{-\left(\frac{n+q_\gamma+1}{2}+a\right)-1}\exp\left[-\frac{b}{\sigma^2}\right]d\sigma^2$$

$$=(2\pi)^{-\frac{q_\gamma+1}{2}}\,|\,\mathbf{V}_\gamma^{-1}\,|^{-\frac{1}{2}}\int\left\{\int\exp\left[-\frac{1}{2\sigma^2}\left([\boldsymbol{\beta}_\gamma-(\mathbf{V}_\gamma+\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}(\mathbf{X}_\gamma^T\mathbf{y}+\mathbf{V}_\gamma\widetilde{\boldsymbol{\beta}}_\gamma)]^T\ldots\right.\right.\right.$$

$$\left.\left.\left.\ldots(\mathbf{V}_\gamma+\mathbf{X}_\gamma^T\mathbf{X}_\gamma)[\boldsymbol{\beta}_\gamma-(\mathbf{V}_\gamma+\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}(\mathbf{X}_\gamma^T\mathbf{y}+\mathbf{V}_\gamma\widetilde{\boldsymbol{\beta}}_\gamma)]\right)\right]d\boldsymbol{\beta}_\gamma\right\}\ldots$$

$$\ldots\exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{y}^T\mathbf{y}-(\mathbf{X}_\gamma^\mathbf{T}\mathbf{y}+\mathbf{V}_\gamma\widetilde{\boldsymbol{\beta}})^T(\mathbf{V}_\gamma+\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}(\mathbf{X}_\gamma^T\mathbf{y}+\mathbf{V}_\gamma\widetilde{\boldsymbol{\beta}})+\widetilde{\boldsymbol{\beta}}_\gamma^T\mathbf{V}_\gamma\widetilde{\boldsymbol{\beta}}_\gamma\right)\right]\ldots$$

$$\ldots(\sigma^2)^{-\left(\frac{n+q_\gamma+1}{2}+a\right)-1}\exp\left[-\frac{b}{\sigma^2}\right]d\sigma^2$$

$$=|\mathbf{V}_\gamma^{-1}|^{-\frac{1}{2}}|(\mathbf{V}_\gamma+\mathbf{X}_\gamma^T\mathbf{X})^{-1}|^{\frac{1}{2}}\int\Bigg\{(\sigma^2)^{-\left(\frac{n}{2}+a\right)-1}\exp\Bigg[-\frac{1}{\sigma^2}\Bigg(\frac{1}{2}(\mathbf{y}^T\mathbf{y}+\tilde{\boldsymbol{\beta}}_\gamma^T\mathbf{V}_\gamma\tilde{\boldsymbol{\beta}}_\gamma\cdots$$

$$\cdots-(\mathbf{X}_\gamma^T\mathbf{y}+\mathbf{V}_\gamma\tilde{\boldsymbol{\beta}}_\gamma)^T(\mathbf{V}_\gamma+\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}(\mathbf{X}_\gamma^T\mathbf{y}+\mathbf{V}_\gamma\tilde{\boldsymbol{\beta}}_\gamma)\Bigg)+b\Bigg]\Bigg\}d\sigma^2$$

$$\propto|\mathbf{V}_\gamma^{-1}|^{-\frac{1}{2}}|(\mathbf{V}_\gamma+\mathbf{X}_\gamma^T\mathbf{X})^{-1}|^{\frac{1}{2}}\cdots$$

$$\cdots\Big(2b+\mathbf{y}^T\mathbf{y}-(\mathbf{X}_\gamma^T\mathbf{y}+\mathbf{V}_\gamma\tilde{\boldsymbol{\beta}}_\gamma)^T(\mathbf{V}_\gamma+\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}(\mathbf{X}_\gamma^T\mathbf{y}+\mathbf{V}_\gamma\tilde{\boldsymbol{\beta}}_\gamma)+\tilde{\boldsymbol{\beta}}_\gamma^T\mathbf{V}_\gamma\tilde{\boldsymbol{\beta}}_\gamma\Big)^{-\frac{n}{2}-a}$$

Joint posterior for β and σ^2:

$$f(\boldsymbol{\beta}_\gamma,\sigma^2\,|\,\boldsymbol{\gamma},\mathbf{y},\mathbf{X})\quad=f(\boldsymbol{\beta}_\gamma\,|\,\sigma^2,\boldsymbol{\gamma},\mathbf{y},\mathbf{X})f(\sigma^2\,|\,\boldsymbol{\gamma},\mathbf{y},\mathbf{X})$$

$$=\mathbf{N}_{q_\gamma+1}[(\mathbf{V}_\gamma^*)^{-1}(\boldsymbol{\beta}_\gamma^*),\sigma^2(\mathbf{V}_\gamma^*)^{-1}]\,\mathbf{IG}\left[\frac{n}{2}+a,b+\frac{y^Ty-(\boldsymbol{\beta}_\gamma^*)^T(\mathbf{V}_\gamma^*)^{-1}(\boldsymbol{\beta}_\gamma^*)+\tilde{\boldsymbol{\beta}}_\gamma^T\mathbf{V}_\gamma\tilde{\boldsymbol{\beta}}_\gamma}{2}\right]$$

where: $\mathbf{V}_\gamma^*=(\mathbf{V}_\gamma+\mathbf{X}_\gamma^T\mathbf{X}_\gamma)$, and $\boldsymbol{\beta}_\gamma^*=(\mathbf{X}_\gamma^T\mathbf{y}+\mathbf{V}_\gamma\tilde{\boldsymbol{\beta}}_\gamma)$.

B.2 Zellner's Prior:

Priors:

$$f(\boldsymbol{\beta}_\gamma,\sigma^2\,|\,\boldsymbol{\gamma},\mathbf{X},c)\propto\mathbf{N}_{q_\gamma+1}(\tilde{\boldsymbol{\beta}}_\gamma,c\sigma^2(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1})\;(\sigma^2)^{-1},\text{ with }f(\boldsymbol{\gamma})\propto 1$$

Posterior for γ:

$$f(\boldsymbol{\gamma}\,|\,\mathbf{y},\mathbf{X},c)\quad=\int\!\!\int f(\boldsymbol{\gamma},\boldsymbol{\beta}_\gamma,\sigma^2\,|\,\mathbf{y},\mathbf{X},c)d\boldsymbol{\beta}_\gamma d\sigma^2$$

$$\propto\int\!\!\int f(\mathbf{y}\,|\,\boldsymbol{\gamma},\boldsymbol{\beta}_\gamma,\sigma^2,\mathbf{X})f(\boldsymbol{\beta}_\gamma\,|\,\boldsymbol{\gamma},\sigma^2,\mathbf{X},c)f(\sigma^2\,|\,\mathbf{X})f(\boldsymbol{\gamma})d\boldsymbol{\beta}_\gamma d\sigma^2$$

$$\propto\int\Big\{\int\mathbf{N}_n(\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma,\sigma^2\mathbf{I}_n)\mathbf{N}_{q_\gamma+1}(\tilde{\boldsymbol{\beta}}_\gamma,c\sigma^2(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1})d\boldsymbol{\beta}_\gamma\Big\}(\sigma^2)^{-1}d\sigma^2$$

$$\propto(2\pi c)^{-\frac{q_\gamma+1}{2}}|(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}|^{-\frac{1}{2}}\cdots$$

$$\cdots\int\left\{\int\exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma)^T(\mathbf{y}-\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma)+\frac{1}{c}(\boldsymbol{\beta}_\gamma-\tilde{\boldsymbol{\beta}}_\gamma)^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma(\boldsymbol{\beta}_\gamma-\tilde{\boldsymbol{\beta}}_\gamma))\right]d\boldsymbol{\beta}_\gamma\right\}\cdots$$

$$\ldots\ (\sigma^2)^{-\left(\frac{n+q_\gamma+1}{2}\right)-1}\, d\sigma^2$$

$$= (2\pi c)^{-\frac{q_\gamma+1}{2}} \,|(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}|^{-\frac{1}{2}} \ldots$$

$$\ldots \int\left\{\int \exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{y}^T\mathbf{y} + \boldsymbol{\beta}_\gamma^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma - 2\boldsymbol{\beta}_\gamma^T\mathbf{X}_\gamma^T\mathbf{y} + \frac{1}{c}\boldsymbol{\beta}_\gamma^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma \ldots\right.\right.\right.$$

$$\left.\left.\left.\ldots + \frac{1}{c}\tilde{\boldsymbol{\beta}}_\gamma^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma - \frac{2}{c}\boldsymbol{\beta}_\gamma^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma\right)\right] d\boldsymbol{\beta}_\gamma\right\} (\sigma^2)^{-\left(\frac{n+q_\gamma+1}{2}\right)-1}\, d\sigma^2$$

$$= (2\pi c)^{-\frac{q_\gamma+1}{2}} \,|(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}|^{-\frac{1}{2}} \ldots$$

$$\ldots \int\left\{\int \exp\left[-\frac{c+1}{2c\sigma^2}\left((\boldsymbol{\beta}_\gamma - \frac{c}{c+1}\hat{\boldsymbol{\beta}}_\gamma - \frac{1}{c+1}\tilde{\boldsymbol{\beta}}_\gamma)^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma(\boldsymbol{\beta}_\gamma - \frac{c}{c+1}\hat{\boldsymbol{\beta}}_\gamma - \frac{1}{c+1}\tilde{\boldsymbol{\beta}}_\gamma)\right)\right] d\boldsymbol{\beta}_\gamma\right\} \ldots$$

$$\ldots \exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{y}^T\mathbf{y} - \frac{c}{c+1}\mathbf{y}^T\mathbf{X}_\gamma(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma^T\mathbf{y} + \frac{1}{c+1}\tilde{\boldsymbol{\beta}}_\gamma^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma - \frac{2}{c+1}\mathbf{y}^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma\right)\right] \ldots$$

$$\ldots\ (\sigma^2)^{-\left(\frac{n+q_\gamma+1}{2}\right)-1}\, d\sigma^2$$

$$= (c+1)^{-\frac{q_\gamma+1}{2}} \int (\sigma^2)^{-\frac{n}{2}-1} \ldots$$

$$\ldots\ \exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{y}^T\mathbf{y} - \frac{c}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y} + \frac{1}{c+1}\tilde{\boldsymbol{\beta}}_\gamma^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma - \frac{2}{c+1}\mathbf{y}^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma\right)\right] d\sigma^2$$

$$\propto (c+1)^{-\frac{q_\gamma+1}{2}} \left(\mathbf{y}^T\mathbf{y} - \frac{c}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y} + \frac{1}{c+1}\tilde{\boldsymbol{\beta}}_\gamma^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma - \frac{2}{c+1}\mathbf{y}^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma\right)^{-n/2}$$

Joint posterior for β and σ^2:

$$f(\boldsymbol{\beta}_\gamma, \sigma^2 \mid \boldsymbol{\gamma}, \mathbf{y}, \mathbf{X}) = \mathbf{N}_{q_\gamma+1}\left(\frac{c}{c+1}\hat{\boldsymbol{\beta}}_\gamma + \frac{1}{c+1}\tilde{\boldsymbol{\beta}}_\gamma, \frac{c\sigma^2}{c+1}(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\right) \ldots$$

$$\mathbf{IG}\left(\frac{n}{2}, \frac{1}{2}\left[\mathbf{y}^T\mathbf{y} - \frac{1}{c+1}(c\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y} - \tilde{\boldsymbol{\beta}}_\gamma^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma - 2\mathbf{y}^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma)\right]\right)$$

Posterior Predictive distribution:

$f(\tilde{\mathbf{y}} \mid \tilde{\mathbf{X}}, \mathbf{X}, \mathbf{y}, \boldsymbol{\gamma}, c) = \int f(\tilde{\mathbf{y}} \mid \sigma^2, \tilde{\mathbf{X}}, \mathbf{X}, \mathbf{y}, \boldsymbol{\gamma}, c) f(\sigma^2 \mid \mathbf{y}, \mathbf{X}, \boldsymbol{\gamma}, c)\, d\sigma^2$ where

$f(\tilde{\mathbf{y}} \mid \sigma^2, \tilde{\mathbf{X}}, \mathbf{X}, \mathbf{y}, \boldsymbol{\gamma}, c) = \int f(\tilde{\mathbf{y}} \mid \sigma^2, \boldsymbol{\beta}, \tilde{\mathbf{X}}, \boldsymbol{\gamma}) f(\boldsymbol{\beta}_\gamma \mid \sigma^2, \mathbf{y}, \mathbf{X}, \boldsymbol{\gamma}, c)\, d\boldsymbol{\beta}_\gamma$ for the newly observed data:

$\tilde{\mathbf{y}} \,|\, \boldsymbol{\beta}_\gamma, \tilde{\mathbf{X}}_\gamma, \sigma^2 \sim N_m(\tilde{\mathbf{X}}_\gamma \boldsymbol{\beta}_\gamma, \sigma^2 I_m)$, hence $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}_\gamma \boldsymbol{\beta}_\gamma + \sigma\tilde{\varepsilon}$ where $\tilde{\varepsilon} \sim N_m(0, \mathbf{I}_m)$, and for the posterior for $\boldsymbol{\beta}$ we have $\boldsymbol{\beta}_\gamma \,|\, \sigma^2, \mathbf{y}, \mathbf{X}, \boldsymbol{\gamma}, c \sim N_{q_\gamma+1}\left(\frac{c}{c+1}\hat{\boldsymbol{\beta}}_\gamma + \frac{1}{c+1}\tilde{\boldsymbol{\beta}}_\gamma, \frac{c\sigma^2}{c+1}(\mathbf{X}_\gamma^T \mathbf{X}_\gamma)^{-1}\right)$ hence

$\boldsymbol{\beta}_\gamma = \left(\frac{c}{c+1}\hat{\boldsymbol{\beta}}_\gamma + \frac{1}{c+1}\tilde{\boldsymbol{\beta}}_\gamma\right) + \sigma\sqrt{\frac{c}{c+1}}\varepsilon$ where, $\varepsilon \sim N_{q_\gamma+1}(0, (\mathbf{X}_\gamma^T \mathbf{X}_\gamma)^{-1})$.

It follows: $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}_\gamma \left[\left(\frac{c}{c+1}\hat{\boldsymbol{\beta}}_\gamma + \frac{1}{c+1}\tilde{\boldsymbol{\beta}}_\gamma\right) + \sigma\sqrt{\frac{c}{c+1}}\varepsilon\right] + \sigma\tilde{\varepsilon} = \tilde{\mathbf{X}}_\gamma\left(\frac{c}{c+1}\hat{\boldsymbol{\beta}}_\gamma + \frac{1}{c}\tilde{\boldsymbol{\beta}}_\gamma\right) + \sigma\left(\sqrt{\frac{c}{c+1}}\tilde{\mathbf{X}}_\gamma \varepsilon + \tilde{\varepsilon}\right)$,

therefore: $f(\tilde{\mathbf{y}} \,|\, \sigma^2, \tilde{\mathbf{X}}, \mathbf{X}, \mathbf{y}, \boldsymbol{\gamma}, c) \sim N_m\left(\tilde{\mathbf{X}}_\gamma\left(\frac{c}{c+1}\hat{\boldsymbol{\beta}}_\gamma + \frac{1}{c+1}\tilde{\boldsymbol{\beta}}_\gamma\right), \sigma^2[\mathbf{I}_m + \frac{c}{c+1}\tilde{\mathbf{X}}_\gamma^T(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\tilde{\mathbf{X}}_\gamma]\right)$

$$f(\tilde{\mathbf{y}} \,|\, \tilde{\mathbf{X}}, \mathbf{X}, \mathbf{y}, \boldsymbol{\gamma}) = \int f(\tilde{\mathbf{y}} \,|\, \sigma^2, \tilde{\mathbf{X}}, \mathbf{X}, \mathbf{y}, \boldsymbol{\gamma}) f(\sigma^2 \,|\, \mathbf{y}, \mathbf{X}, \boldsymbol{\gamma})\, d\sigma^2$$

$$= \int N_m\left(\tilde{\mathbf{X}}_\gamma\left(\frac{c}{c+1}\hat{\boldsymbol{\beta}}_\gamma + \frac{1}{c+1}\tilde{\boldsymbol{\beta}}_\gamma\right), \sigma^2[\mathbf{I}_m + \frac{c}{c+1}\tilde{\mathbf{X}}_\gamma^T(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\tilde{\mathbf{X}}_\gamma]\right)\ldots$$

$$\ldots\; I.G\left(\frac{n}{2}, \frac{1}{2}[\mathbf{y}^T\mathbf{y} - \frac{c}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y} + \frac{1}{c+1}\tilde{\boldsymbol{\beta}}_\gamma^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma - \frac{2}{c+1}\mathbf{y}^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma]\right)d\sigma^2$$

$$\propto \int (\sigma^2)^{-m/2}\exp\left[-\frac{1}{2\sigma^2}\left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}_\gamma\boldsymbol{\mu}_\gamma\right)^T[\boldsymbol{\Sigma}_\gamma]^{-1}\left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}_\gamma\boldsymbol{\mu}_\gamma\right)\right]\ldots$$

$$\ldots(\sigma^2)^{-n/2-1}\exp\left[-\frac{1}{2\sigma^2}[\mathbf{y}^T\mathbf{y} - \frac{c}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y} + \frac{1}{c+1}\tilde{\boldsymbol{\beta}}_\gamma^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma - \frac{2}{c+1}\mathbf{y}^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma]\right]d\sigma^2$$

where: $\Sigma_\gamma = \left[\mathbf{I}_m + \frac{c}{c+1}\tilde{\mathbf{X}}_\gamma^T(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\tilde{\mathbf{X}}_\gamma\right]$, and $\boldsymbol{\mu}_\gamma = \tilde{\mathbf{X}}_\gamma\left(\frac{c}{c+1}\hat{\boldsymbol{\beta}}_\gamma + \frac{1}{c+1}\tilde{\boldsymbol{\beta}}_\gamma\right)$

$$\propto \left[(\tilde{\mathbf{y}} - \boldsymbol{\mu}_\gamma)^T\boldsymbol{\Sigma}_\gamma^{-1}(\tilde{\mathbf{y}} - \boldsymbol{\mu}_\gamma) + \mathbf{y}^T\mathbf{y} - \frac{c}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y} + \frac{1}{c+1}\tilde{\boldsymbol{\beta}}_\gamma^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma - \frac{2}{c+1}\mathbf{y}^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma\right]^{-(n+m)/2}$$

$$\propto \left[\frac{n(\tilde{\mathbf{y}} - \boldsymbol{\mu}_\gamma)^T\boldsymbol{\Sigma}_\gamma^{-1}(\tilde{\mathbf{y}} - \boldsymbol{\mu}_\gamma)}{n\left(\mathbf{y}^T\mathbf{y} - \frac{c}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y} + \frac{1}{c+1}\tilde{\boldsymbol{\beta}}_\gamma^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma - \frac{2}{c+1}\mathbf{y}^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma\right)} + 1\right]^{-(n+m)/2}$$

$$= T\left(n, \boldsymbol{\mu}_\gamma, \frac{\mathbf{y}^T\mathbf{y} - \frac{c}{c+1}\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y} + \frac{1}{c+1}\tilde{\boldsymbol{\beta}}_\gamma^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma - \frac{2}{c+1}\mathbf{y}^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma}{n}\boldsymbol{\Sigma}_\gamma\right)$$

B.3 Jeffreys Prior:

Derivation:

Jeffreys prior is obtained from the following relation:

$$f^J(\boldsymbol{\beta}_\gamma, \sigma^2) \propto \sqrt{|\mathbf{I}(\boldsymbol{\beta}_\gamma, \sigma^2)|} = \sqrt{-\mathrm{E}|\mathbf{H}(\boldsymbol{\beta}_\gamma, \sigma^2)|}$$

Where **I** is the expected Fisher Information matrix and **H** is the Hessian matrix. The log likelihood is:

$$\ell(\boldsymbol{\beta}_\gamma, \sigma^2) = -(n/2)\ln(2\pi) - (n/2)\ln(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}_\gamma\boldsymbol{\beta}_\gamma)^T(\mathbf{y} - \mathbf{X}_\gamma\boldsymbol{\beta}_\gamma)$$

$$\ell(\boldsymbol{\beta}_\gamma, \sigma) = -(n/2)\ln(2\pi) - n\ln\sigma - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}_\gamma\boldsymbol{\beta}_\gamma)^T(\mathbf{y} - \mathbf{X}_\gamma\boldsymbol{\beta}_\gamma)$$

The grey boxes contain the equivalent derivation for σ instead of σ^2. We can derive the first partial derivative *w.r.t.* to $\boldsymbol{\beta}_\gamma$ simultaneously using matrix calculus:

$$\frac{\partial\ell(\boldsymbol{\beta}_\gamma, \sigma^2)}{\partial\boldsymbol{\beta}_\gamma} = -\frac{1}{2\sigma^2}(-2\mathbf{X}_\gamma^T\mathbf{y} + 2\mathbf{X}_\gamma^T\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma)$$

$$= \frac{1}{\sigma^2}(\mathbf{X}_\gamma^T\mathbf{y} - \mathbf{X}_\gamma^T\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma)$$

Now the second partial derivative:

$$\frac{\partial^2\ell(\boldsymbol{\beta}_\gamma, \sigma^2)}{\partial\boldsymbol{\beta}_\gamma^2} = \frac{1}{\sigma^2}(-\mathbf{X}_\gamma^T\mathbf{X}_\gamma)$$

$$= \frac{-\mathbf{X}_\gamma^T\mathbf{X}_\gamma}{\sigma^2}$$

$$-\mathrm{E}\left[\frac{\partial^2\ell(\boldsymbol{\beta}_\gamma, \sigma^2)}{\partial\boldsymbol{\beta}_\gamma}\right] = \frac{\mathbf{X}_\gamma^T\mathbf{X}_\gamma}{\sigma^2}$$

Now considering σ^2, we obtain the first partial derivative:

$$\frac{\partial\ell(\boldsymbol{\beta}_\gamma, \sigma^2)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(\mathbf{y} - \mathbf{X}_\gamma\boldsymbol{\beta}_\gamma)^T(\mathbf{y} - \mathbf{X}_\gamma\boldsymbol{\beta}_\gamma)$$

$$\frac{\partial\ell(\boldsymbol{\beta}_\gamma, \sigma)}{\partial\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}(\mathbf{y} - \mathbf{X}_\gamma\boldsymbol{\beta}_\gamma)^T(\mathbf{y} - \mathbf{X}_\gamma\boldsymbol{\beta}_\gamma)$$

Now the second partial derivative:

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{\beta}_\gamma, \sigma^2)}{\partial(\sigma^2)^2} &= \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}(\mathbf{y} - \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma)^T (\mathbf{y} - \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma) \\ &= \frac{n\sigma^2 - 2(\mathbf{y} - \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma)^T (\mathbf{y} - \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma)}{2(\sigma^2)^3}\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{\beta}_\gamma, \sigma)}{\partial(\sigma)^2} &= \frac{n}{\sigma^2} - \frac{3}{\sigma^4}(\mathbf{y} - \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma)^T (\mathbf{y} - \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma) \\ &= \frac{n\sigma^2 - 3(\mathbf{y} - \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma)^T (\mathbf{y} - \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma)}{\sigma^4}\end{aligned}$$

Now taking the negative expectation:

$$\begin{aligned}-\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\beta}_\gamma, \sigma^2)}{\partial(\sigma^2)^2}\right] &= -\left(\frac{n\sigma^2 - 2n\sigma^2}{2(\sigma^2)^3}\right) \\ &= \frac{n}{2(\sigma^2)^2}\end{aligned}$$

$$\begin{aligned}-\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\beta}_\gamma, \sigma)}{\partial(\sigma)^2}\right] &= -\left(\frac{n\sigma^2 - 3n\sigma^2}{\sigma^4}\right) \\ &= \frac{2n}{\sigma^2}\end{aligned}$$

Cross Products for β_j and σ^2:

$$\begin{aligned}-\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\beta}_\gamma, \sigma^2)}{\partial\beta_j \partial(\sigma^2)}\right] &= -\mathbb{E}\left[-\frac{1}{(\sigma^2)^2}(\mathbf{X}_\gamma^T \mathbf{y} - \mathbf{X}_\gamma^T \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma)\right] \\ &= \frac{1}{(\sigma^2)^2}\mathbf{X}_\gamma^T(\mathbf{X}_\gamma \boldsymbol{\beta}_\gamma - \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma) \\ &= 0\end{aligned}$$

$$\begin{aligned}-\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\beta}_\gamma, \sigma)}{\partial\beta_j \partial\sigma}\right] &= -\mathbb{E}\left[\frac{-2}{\sigma^3}(\mathbf{X}_\gamma^T \mathbf{y} - \mathbf{X}_\gamma^T \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma)\right] \\ &= \frac{2}{\sigma^3}\mathbf{X}_\gamma^T(\mathbf{X}_\gamma \boldsymbol{\beta}_\gamma - \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma) \\ &= 0\end{aligned}$$

Therefore:

$$\mathbf{I}(\boldsymbol{\beta}_\gamma,\sigma^2)=\frac{1}{\sigma^2}\begin{bmatrix}\mathbf{X}_\gamma^T\mathbf{X}_\gamma & 0\\ 0 & \dfrac{n}{2\sigma^2}\end{bmatrix}$$

$$\mathbf{I}(\boldsymbol{\beta}_\gamma,\sigma)=\frac{1}{\sigma^2}\begin{bmatrix}\mathbf{X}_\gamma^T\mathbf{X}_\gamma & 0\\ 0 & 2n\end{bmatrix}$$

and so:

$$\begin{aligned}|\mathbf{I}(\boldsymbol{\beta}_\gamma,\sigma^2)| &= (\sigma^2)^{-(q_\gamma+1)}\,|\mathbf{X}_\gamma^T\mathbf{X}_\gamma|\frac{n}{2}(\sigma^2)^{-2}\\ &=\frac{n}{2}|\mathbf{X}_\gamma^T\mathbf{X}_\gamma|(\sigma^2)^{-(q_\gamma+3)}\end{aligned},$$

and

$$\begin{aligned}|\mathbf{I}(\boldsymbol{\beta}_\gamma,\sigma)| &= (\sigma^2)^{-(q_\gamma+1)}\,|\mathbf{X}_\gamma^T\mathbf{X}_\gamma|\,2n(\sigma^2)^{-1}\\ &=2n|\mathbf{X}_\gamma^T\mathbf{X}_\gamma|(\sigma^2)^{-(q_\gamma+2)}\end{aligned}.$$

Hence,

$$\begin{aligned}f^J(\boldsymbol{\beta}_\gamma,\sigma^2) &\propto \sqrt{|\mathbf{I}(\boldsymbol{\beta}_\gamma,\sigma^2)|}\\ &\propto |\mathbf{X}_\gamma^T\mathbf{X}_\gamma|^{1/2}\,(\sigma^2)^{-(q_\gamma+3)/2}\end{aligned}$$

and

$$\begin{aligned}f^J(\boldsymbol{\beta}_\gamma,\sigma) &\propto \sqrt{|\mathbf{I}(\boldsymbol{\beta}_\gamma,\sigma)|}\\ &\propto |\mathbf{X}_\gamma^T\mathbf{X}_\gamma|^{1/2}\,(\sigma)^{-(q_\gamma+2)}\end{aligned}$$

Priors:

$$f(\boldsymbol{\beta}_\gamma,\sigma^2\,|\,\boldsymbol{\gamma},\mathbf{X})\propto(p)^{-\frac{q_\gamma}{2}}\,|\mathbf{X}_\gamma^T\mathbf{X}_\gamma|^{\frac{1}{2}}\,(\sigma^2)^{-\left(\frac{q_\gamma+1}{2}+1\right)},\text{ with } f(\boldsymbol{\gamma})\propto 1$$

Posterior for γ:

$$\begin{aligned}f(\boldsymbol{\gamma}\,|\,\mathbf{y},\mathbf{X}) &= \int\!\!\int f(\boldsymbol{\gamma},\boldsymbol{\beta}_\gamma,\sigma^2\,|\,\mathbf{y},\mathbf{X})\,d\boldsymbol{\beta}_\gamma d\sigma^2\\ &\propto \int\!\!\int f(\mathbf{y}\,|\,\boldsymbol{\gamma},\boldsymbol{\beta}_\gamma,\sigma^2,\mathbf{X})\,g(\boldsymbol{\beta}_\gamma,\sigma^2\,|\,\boldsymbol{\gamma},\mathbf{X})\,g(\boldsymbol{\gamma})\,d\boldsymbol{\beta}_\gamma d\sigma^2\\ &\propto (p)^{-\frac{q_\gamma}{2}}\,|\mathbf{X}_\gamma^T\mathbf{X}_\gamma|^{\frac{1}{2}}\int\left\{\int \mathbf{N}_n(\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma,\sigma^2\mathbf{I}_n)\,d\boldsymbol{\beta}_\gamma\right\}(\sigma^2)^{-\left(\frac{q_\gamma+1}{2}\right)-1}d\sigma^2\end{aligned}$$

$$\propto (p)^{-\frac{q_\gamma}{2}}\,|\mathbf{X}_\gamma^T\mathbf{X}_\gamma|^{\frac{1}{2}} \int\left\{\int \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma)^T(\mathbf{y}-\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma)\right]d\boldsymbol{\beta}_\gamma\right\}\ldots$$

$$\ldots(\sigma^2)^{-\left(\frac{n+q_\gamma+1}{2}\right)-1}\,d\sigma^2$$

$$=(p)^{-\frac{q_\gamma}{2}}\,|\mathbf{X}_\gamma^T\mathbf{X}_\gamma|^{\frac{1}{2}} \int\left\{\int \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{y}+\boldsymbol{\beta}_\gamma^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma-2\boldsymbol{\beta}_\gamma^T\mathbf{X}_\gamma^T\mathbf{y})\right]d\boldsymbol{\beta}_\gamma\right\}\ldots$$

$$\ldots(\sigma^2)^{-\left(\frac{n+q_\gamma+1}{2}\right)-1}\,d\sigma^2$$

$$=(p)^{-\frac{q_\gamma}{2}}\,|\mathbf{X}_\gamma^T\mathbf{X}_\gamma|^{\frac{1}{2}}\ldots$$

$$\ldots\int\left\{\int \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}+(\boldsymbol{\beta}_\gamma-\hat{\boldsymbol{\beta}}_\gamma)^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma(\boldsymbol{\beta}_\gamma-\hat{\boldsymbol{\beta}}_\gamma))\right]d\boldsymbol{\beta}_\gamma\right\}\ldots$$

$$\ldots(\sigma^2)^{-\left(\frac{n+q_\gamma+1}{2}\right)-1}\,d\sigma^2$$

$$\propto (p)^{-\frac{q_\gamma}{2}}\,|\mathbf{X}_\gamma^T\mathbf{X}_\gamma|^{\frac{1}{2}} \int\left\{\int \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{\beta}_\gamma-\hat{\boldsymbol{\beta}}_\gamma)^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma(\boldsymbol{\beta}_\gamma-\hat{\boldsymbol{\beta}}_\gamma)\right]d\boldsymbol{\beta}_\gamma\right\}\ldots$$

$$\ldots(\sigma^2)^{-\left(\frac{n+q_\gamma+1}{2}\right)-1}\exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})\right]d\sigma^2$$

$$\propto (p)^{-\frac{q_\gamma}{2}}\,|\mathbf{X}_\gamma^T\mathbf{X}_\gamma|^{\frac{1}{2}}\,(2\pi)^{\frac{q_\gamma}{2}}\,|(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}|^{\frac{1}{2}}\ldots$$

$$\ldots\int(\sigma^2)^{-\frac{n}{2}-1}\exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})\right]d\sigma^2$$

$$\propto\left(\frac{p}{2\pi}\right)^{-\frac{q_\gamma}{2}}(\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})^{-\frac{n}{2}}$$

Joint posterior for β and σ^2:

$$f(\boldsymbol{\beta}_\gamma,\sigma^2\,|\,\boldsymbol{\gamma},\mathbf{y},\mathbf{X}) \qquad = f(\boldsymbol{\beta}_\gamma\,|\,\sigma^2,\boldsymbol{\gamma},\mathbf{y},\mathbf{X})f(\sigma^2\,|\,\boldsymbol{\gamma},\mathbf{y},\mathbf{X})$$

$$= \mathbf{N}_{q_\gamma+1}(\hat{\boldsymbol{\beta}}_\gamma,\sigma^2(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1})\ \mathbf{IG}\left(\frac{n}{2},\frac{\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}}{2}\right)$$

Posterior Predictive distribution:

$f(\tilde{\mathbf{y}}\,|\,\tilde{\mathbf{X}},\mathbf{X},\mathbf{y},\boldsymbol{\gamma})=\int\ f(\tilde{\mathbf{y}}\,|\,\sigma^2,\tilde{\mathbf{X}},\mathbf{X},\mathbf{y},\boldsymbol{\gamma})f(\sigma^2\,|\,\mathbf{y},\mathbf{X},\boldsymbol{\gamma})\ d\sigma^2$ where

$f(\tilde{\mathbf{y}}\,|\,\sigma^2,\tilde{\mathbf{X}},\mathbf{X},\mathbf{y},\boldsymbol{\gamma})=\int\ f(\tilde{\mathbf{y}}\,|\,\sigma^2,\boldsymbol{\beta}_\gamma,\tilde{\mathbf{X}},\boldsymbol{\gamma})f(\boldsymbol{\beta}_\gamma\,|\,\sigma^2,\mathbf{y},\mathbf{X},\boldsymbol{\gamma})\ d\boldsymbol{\beta}_\gamma$ for the newly observed data:

$\tilde{\mathbf{y}} \,|\, \boldsymbol{\beta}_\gamma, \tilde{\mathbf{X}}_\gamma, \sigma^2 \sim N_m(\tilde{\mathbf{X}}_\gamma \boldsymbol{\beta}_\gamma, \sigma^2 I_m)$, hence $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}_\gamma \boldsymbol{\beta}_\gamma + \sigma\tilde{\varepsilon}$ where $\tilde{\varepsilon} \sim N_m(0, \mathbf{I}_m)$, and for the posterior for $\boldsymbol{\beta}$ we have $\boldsymbol{\beta}_\gamma \,|\, \sigma^2, \mathbf{y}, \mathbf{X}, \boldsymbol{\gamma} \sim N_{q_\gamma+1}(\hat{\boldsymbol{\beta}}_\gamma, \sigma^2(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1})$ hence $\boldsymbol{\beta}_\gamma = \hat{\boldsymbol{\beta}}_\gamma + \sigma\varepsilon$ where $\varepsilon \sim N_{q_\gamma+1}(0, (\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1})$.

It follows: $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}_\gamma(\hat{\boldsymbol{\beta}}_\gamma + \sigma\varepsilon) + \sigma\tilde{\varepsilon} = \tilde{\mathbf{X}}_\gamma\hat{\boldsymbol{\beta}}_\gamma + \sigma(\tilde{\mathbf{X}}_\gamma\varepsilon + \tilde{\varepsilon})$, therefore:

$$f(\tilde{\mathbf{y}} \,|\, \sigma^2, \tilde{\mathbf{X}}, \mathbf{X}, \mathbf{y}, \boldsymbol{\gamma}) \sim N_m(\tilde{\mathbf{X}}_\gamma\hat{\boldsymbol{\beta}}_\gamma, \sigma^2[\mathbf{I}_m + \tilde{\mathbf{X}}_\gamma^T(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\tilde{\mathbf{X}}_\gamma])$$

$$\begin{aligned}
f(\tilde{\mathbf{y}} \,|\, \tilde{\mathbf{X}}, \mathbf{X}, \mathbf{y}, \boldsymbol{\gamma}) \quad &= \int f(\tilde{\mathbf{y}} \,|\, \sigma^2, \tilde{\mathbf{X}}, \mathbf{X}, \mathbf{y}, \boldsymbol{\gamma}) f(\sigma^2 \,|\, \mathbf{y}, \mathbf{X}, \boldsymbol{\gamma})\, d\sigma^2 \\
&= \int N_m(\tilde{\mathbf{X}}_\gamma\hat{\boldsymbol{\beta}}_\gamma, \sigma^2[\mathbf{I}_m + \tilde{\mathbf{X}}_\gamma^T(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\tilde{\mathbf{X}}_\gamma])\; I.G\left(\frac{n}{2}, \frac{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}}{2}\right) d\sigma^2 \\
&\propto \int (\sigma^2)^{-m/2} \exp\left[-\frac{1}{2\sigma^2}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}_\gamma\hat{\boldsymbol{\beta}}_\gamma)^T[\mathbf{I}_m + \tilde{\mathbf{X}}_\gamma^T(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\tilde{\mathbf{X}}_\gamma]^{-1}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}_\gamma\hat{\boldsymbol{\beta}}_\gamma)\right] \ldots \\
&\qquad \ldots\ (\sigma^2)^{-n/2-1} \exp\left[-\frac{1}{2\sigma^2}\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}\right] d\sigma^2 \\
&\propto \left[(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}_\gamma\hat{\boldsymbol{\beta}}_\gamma)^T[\mathbf{I}_m + \tilde{\mathbf{X}}_\gamma^T(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\tilde{\mathbf{X}}_\gamma](\tilde{\mathbf{y}} - \tilde{\mathbf{X}}_\gamma\hat{\boldsymbol{\beta}}_\gamma) + \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y}\right]^{-(n+m)/2} \\
&\propto \left(\frac{n(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}_\gamma\hat{\boldsymbol{\beta}}_\gamma)^T[I_m + \tilde{\mathbf{X}}_\gamma^T(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\tilde{\mathbf{X}}_\gamma]^{-1}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}_\gamma\hat{\boldsymbol{\beta}}_\gamma)}{n(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})} + 1\right)^{-(n+m)/2} \\
&= T\left(n, \tilde{\mathbf{X}}_\gamma\hat{\boldsymbol{\beta}}_\gamma, \frac{(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_\gamma\mathbf{y})}{n}[\mathbf{I}_m + \tilde{\mathbf{X}}_\gamma^T(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\tilde{\mathbf{X}}_\gamma]\right)
\end{aligned}$$

APPENDIX C: Datasets

Ozone Data

Response:	Ozone concentration (ppm)
[k, n]:	[8, 80]
Size of Γ:	256
Description :	Ozone concentration at Upland, CA, USA.
Source:	R Software, Package {forward}

Predictors:

x^1: Temperature F (max for the day)

x^2: Inversion base height, feet

x^3: Daggett pressure gradient (mm Hg)

x^4: Visibility (miles)

x^5: Vandenburg 500 millibar height (m)

x^6: Humidity, percent

x^7: Inversion base temperature, degrees F

x^8: Wind speed, mph

Description/Comments:

This data set consist of the first 80 observations from a data set containing up to 360 observations from Breiman, L and Friedman, J. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation", *Journal of the American Statistical Association*, 80, 580-598. We use log(ozone) as **y**.

References:

Atkinson, A.C. and Riani, M. (2000), *Robust Diagnostic Regression Analysis*, First Edition. New York: Springer, Table A.7.

Physical Data

Response:	Mass (kg)
[k, n]:	[10, 22]
Size of Γ:	1024
Source:	http://www.statsci.org/data/oz/physical.txt

Predictors:

x^1: Forearm (Maximum circumference)

x^2: Bicep (Maximum circumference)

x^6: Calf (Maximum circumference of calf)

x^7: Height (Height from top to toe)

x^3: Neck (Distance around neck, approximately halfway up)

x^4: Chest (Distance around chest directly under the armpits)

x^5: Shoulder (Distance around shoulders, measured around the peak of the shoulder blades)

x^8: Waist (Distance around waist, approximately trouser line)

x^9: Head (Circumference of head at eye level)

x^{10}: Thigh (Circumference of thigh, measured halfway between the knee and the top of the leg)

Description/Comments:

The weight and various physical measurements for 22 male subjects aged 16 - 30 were recorded. Subjects were randomly chosen volunteers, all in reasonable good health. Subjects were requested to slightly tense each muscle being measured to ensure measurement consistency. All predictors are measurements in cm. There was no need to log transform the response variable the residuals are very well behaved.

References:

Larner, M. (1996). Mass and its Relationship to Physical Measurements. MS305 Data Project, Department of Mathematics, University of Queensland.

Bodyfat Data

Response:	Percentage bodyfat
[k, n]:	[13, 250]
Size of Γ:	8192
Source:	http://www.amstat.org/publications/jse/v4n1/datasets/fat.dat

Predictors:

x^1: Age (years)

x^2: Weight (pounds)

x^3: Height (inches)

x^4: Neck Circumference (cm)

x^5: Chest Circumference (cm)

x^6: Abdomen Circumference (cm)

x^7: Hip Circumference (cm)

x^8: Thigh Circumference (cm)

x^9: Knee Circumference (cm)

x^{10}: Ankle Circumference (cm)

x^{11}: Extended Biceps Circumference (cm)

x^{12}: Forearm Circumference (cm)

x^{13}: Wrist Circumference (cm)

Description:

Percentage of body fat estimated for 251 men using an underwater weighing technique. The logit transformation was used for **y** and the negative result truncated to zero was removed.

References:

Johnson, W. R. (1996) Fitting Percentage of Body Fat to Simple Body Measurements. Journal of Statistics Education Vol. 4 (1).

U.S. Crime Data

Response:	Rate of crime per head of population in US states
[k, n]:	[15, 47]
Size of $\boldsymbol{\Gamma}$:	32768
Source:	R Software, Package {MASS}

Predictors:

x^1: percentage of males aged 14-24

x^2: indicator variable for a southern state

x^3: mean years of schooling

x^4: police expenditure in 1960

x^5: police expenditure in 1959

x^6: labor force participation rate

x^7: number of males per 1000 females

x^8: state population

x^9: number of nonwhites per 1000 people

x^{10}: unemployment rate: urban males 14-24

x^{11}: unemployment rate: urban males 35-39

x^{12}: gross domestic product per head

x^{13}: income inequality

x^{14}: probability of imprisonment

x^{15}: average time served in state prisons

Description:

Data set records the crime rate in 47 U.S. states for various demographic predictors. All explanatory variables except x^2 were log transformed.

References:

The U.S. crime data has been analyzed often in the literature for variable selection such as Cripps *et al* (2006) and Liang *et al* (2008). It was first presented and analyzed in: Vandaele, W. (1978) “Participation in illegitimate activities; Ehrlich revisited,” *In:*

Blumstein, A., Cohen, J., Nagin, D. (Eds), Deterrence and Incapacitation. National Academy of Science Press, Washington, DC, 270 – 335.

Body Measurement Data

Response:	Weight (in kg)
[k, n]:	[24, 507]
Size of Γ:	16777216
Source:	http://www.sci.usq.edu.au/staff/dunn/Datasets/applications/biology/body.dat

Predictors:

x^1: Biacromial diameter (cm)
x^2: Biiliac diameter (pelvic breadth) (in cm)
x^3: Bitrochanteric diameter (cm)
x^4: Chest depth (cm)
x^5: Chest diameter (cm)
x^6: Elbow diameter (cm)
x^7: Wrist diameter (cm)
x^8: Knee diameter (cm)
x^9: Ankle diameter (cm)
x^{10}: Shoulder girth (cm)
x^{11}: Chest girth (cm)
x^{12}: Waist girth (cm)
x^{13}: Navel girth (cm)
x^{14}: Hip girth (cm)
x^{15}: Thigh girth (cm)
x^{16}: Bicep girth (cm)
x^{17}: Forearm girth (cm)
x^{18}: Knee girth (cm)
x^{19}: Calf girth (cm)
x^{20}: Ankle girth (cm)
x^{21}: Wrist girth (cm)
x^{22}: Age (years)
x^{23}: Height (cm)
x^{24}: Gender; 1 for males and 0 for females

Description:

The data give 21 body dimension measurements as well as age, weight, height, and gender on 507 individuals. The 247 men and 260 women were primarily individuals in their twenties and thirties, with a scattering of older men and women, all exercising several hours a week.

References:

Grete Heinz, Louis J. Peterson, Roger W. Johnson, and Carter J. Kerk. Exploring relationships in body dimensions. *Journal of Statistics Education*, Volume 11, Number 2.

Baseball Data

Response:	Salaries ($1000) of 337 Baseball players in 1992
[k, n]:	[27, 333]
Size of Γ:	134,217,728
Description:	The 27 variables collected are performance statistics from 1991, no baseball pitchers are included.
Source:	http://www.amstat.org/publications/jse/v6n2/datasets.watnik.html

Predictors:

x^{1}: batting average

x^{2}: on base percentage (obp)

x^{3}: runs scored (runs)

x^{4}: hits

x^{5}: doubles

x^{6}: triples

x^{7}: homeruns

x^{8}: runs batted in (rbi)

x^{9}: walks

x^{10}: strike outs (so)

x^{11}: stolen bases (sb)

x^{12}: errors

x^{13}: free-agent

x^{14}: arbitration

x^{15}: runs/sos

x^{16}: hits/sos

x^{17}: homeruns/sos

x^{18}: rbi/sos

x^{19}: walks/sos

x^{20}: obp/errors

x^{21}: runs/errors

x^{22}: hits/errors

x^{23}: homeruns/errors

x^{24}: sos*errors

x^{25}: sbs*obp

x^{26}: sbs*runs

x^{27}: sbs*hits

Description:

Performance statistics were collected for Major league players excluding pitchers along with the following years salary. Per analysis done in the reference below we remove the influential outliers observations: 205 268 284 322, and log transform the response variable.

References:

M.R. Watnik (1998), "Pay for Play: Are Baseball Salaries Based on Performance", *Journal of Statistics Education,* Volume 6, number 2

APPENDIX D: Matlab Code

We now list the main functions used in this research, we do not include code used in the summary and production of figures from the output results of these functions. For full details on input and output arguments see the relevant code. All main functions (listed below in alphabetical order) have been made stand alone, thus no call structure for the required sub-routines is required. This means all required sub functions that are not standard Matlab functions are contained therein.

The functions in alphabetical order are:

1. cRejection1.m
2. cRejection2.m
3. GibbsOSampler.m
4. GibbsPerfect.m
5. GibbsSampler.m
6. gRejection.m
7. Jeffrey.m
8. ModelCheck.m
9. Zellner.m

```
function [sampc m par] = cRejection1(y,X,g,a,N)

%%%%%%%%%%%%%%Jason Bentley (2008) University of Canterbury%%%%%%%%%%%%%
%Rejection sampler for the conditional distribution of c using Zellner’s  %
%prior. The prior for c is the Hyper-G-n.                                 %
%INPUT: y is the response vector                                          %
%       X is the predictor matrix                                         %
%       g is the chosen model                                             %
%       a is the hyper-hyper-parameter for the hyper-G-n prior            %
%       N is the number of samples to be generated                        %
%OUTPUT:samp is N i.i.d samples from the required conditional posterior   %
%       for c.                                                            %
%       m is the N waiting times to generate each sample point            %
%       par is the a and b parameters of an I.G. distribution that        %
%         approximates the conditional posterior of c                     %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

n =length(y); v = 2/(a-2); count = 1; C1 = -(sum(g)/2); %constants
C2 = y'*y; Xg = X(:,g==1); C3 = y'*Xg*inv(Xg'*Xg)*Xg'*y; C4 = -n/2; %constants
cm = -(sum(g)*C2-n*C3)/(sum(g)*(C2-C3)); %maximum
bound = ((cm+1)^C1)*((C2-(cm/(cm+1))*C3)^C4); %optimal bound
while count <= N
    steps = 1; %relative to the previous accepted point
    check = 0;
    while check == 0;
```

```
        vv = rand;
        prp = -n*(((vv-1)^v)-1)/((vv-1)^v); %propose a new value
        prmove = (C1*log(prp+1)+C4*log(C2-(prp/(prp+1))*C3))-log(bound);
%acceptance probability
        if rand <= exp(prmove)
            sampc(count) = prp; %store accepted value
            m(count) = steps; %store length of run for accepted value
            check = 1; %sample point obtained
        end
        steps = steps + 1;
    end
    count = count + 1; %update sample count
end
par = gamfit(1./sampc); %calculate Inverse Gamma Approximation
par = [par(1) 1/par(2)];
```

function [sampc m] = cRejection2(y,X,g,a,r,s,N)

```
%%%%%%%%%%%%%%%Jason Bentley (2008) University of Canterbury%%%%%%%%%%%%%%%
%Rejection sampler for the conditional distribution of c using Zellner’s  %
%prior. The prior for c is the Hyper-G-n.                                 %
%INPUT: y is the response vector                                          %
%       X is the predictor matrix                                         %
%       g is the chosen model                                             %
%       a is the hyper-hyper-parameter for the hyper-G-n prior            %
%       r and s are the required parameters for an I.G. approximation to  %
%       the conditional posterior of c                                    %
%       N is the number of samples to be generated                        %
%OUTPUT:samp is N i.i.d samples from the required conditional posterior   %
%       for c.                                                            %
%       m is the N waiting times to generate each sample point            %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

n =length(y); C1 = -(sum(g)/2); C2 = y'*y; Xg = X(:,g==1); count = 1;
C3 = y'*Xg*inv(Xg'*Xg)*Xg'*y; C4 = -n/2; v = 2/(a-2);
[cmv bb] = solvepolycm(C2,Xg,C3,n,g,a,r,s);
cm = cmv(bb==max(bb))
bound = max(bb) %optimal bound
while count <= N
    steps = 1; %relative to the previous accepted point
    check = 0;
    while check == 0;
        vv = rand;
        prn = 1/gamrnd(r,1/s); %propose a new value
        prp = (((prn+1)^(C1))*((C2-(prn/(prn+1))*C3)^(C4))*...
              ((1+(prn/n))^(-a/2)))/((prn^-(r+1))*exp(-s/prn));
        if rand <= (prp/bound)
            sampc(count) = prn; %store accepted value
            m(count) = steps; %store length of run for accepted value
            check = 1; %sample point obtained
        end
        steps = steps + 1;
    end
    count = count + 1; %update sample count
end
```

function [cmv bb] = solvepolycm(A,Xg,B,n,g,a,v,w)

```
%the w must be as the parameter for the IG
q = sum(g)-1;
%these are the polynomial coefficients
r1=-A+B+q*A-q*B+a*A-a*B-2*v*A+2*v*B;
r2=q*A*n+q*A-A*n+2*a*A-a*B-4*v*A+2*v*B+2*w*A-2*w*B+2*B-3*A-q*B*n-...
    2*v*A*n+2*v*B*n;
```

```
r3=q*A*n-n*n*B-3*A*n+a*A-2*v*A+4*w*A-2*w*B+2*n*B-2*A-4*v*A*n+2*v*B*...
    n+2*w*A*n-2*w*n*B;
r4=-2*A*n-2*v*A*n+2*w*A-2*w*n*B+4*w*A*n;
r5=2*w*A*n;
cof = [r1 r2 r3 r4 r5]; cm = roots(cof);
for j = 1:length(cm)
    cmr(j) = isreal(cm(j));
end
cmv1 = cm(cmr==1); cmv = cmv1(cmv1>0);
for i = 1:length(cmv);
    bb(i) = (((cmv(i)+1)^(-(q+1)/2))*((A-(cmv(i)/(cmv(i)+1))*B)^(-...
n/2))*((1+(cmv(i)/n))^(-a/2)))/((cmv(i)^-(v+1))*exp(-w/cmv(i)));
end

function [dec runtime] = GibbsOSampler(y,X,gstart,tau,penalty,shrink,N)

%%%%%%%%%%%%%%%Jason Bentley (2008) University of Canterbury%%%%%%%%%%%%%%%
%Perfect Gibbs sampler for Bayesian variable selection with an orthogonal %
%design matrix W in linear regression using Zellner's prior or Jeffreys   %
%prior with the binomial prior for gamma                                  %
%INPUT: y is the n.1 response vector                                      %
%       X is n.(k+1) design matrix                                        %
%       gstart is the specified starting value for the Gibbs sampler, it  %
%           must contain a 1 in the first position and be of length k+1,  %
%           it may be obtained using the perfect sampler                  %
%       tau is the choice of hyper-parameter in the constant Bernoulli    %
%           prior (using tau = 0.5 corresponds to a uniform prior)        %
%       penalty and shrink specify whether Jeffreys prior or Zellner's    %
%           prior is used. For Jeffreys penalty corresponds to p = 2*pi*  %
%           penalty with shrink = 1. For Zellner's prior penalty = (c+1)  %
%           and shrink should be set to c/(c+1).                          %
%       N is the number of samples to be generated                        %
%OUTPUT:dec is an N.1 vector of samples represented in decimal form       %
%       runtime is the required cputime to generate the N samples, memory %
%           for output storage                                            %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

t = cputime;
[n k] = size(X); i = 1; samp = zeros(N,k);
for l = 1:k; hats(l) = y'*X(:,l)*inv(X(:,l)'*X(:,l))*X(:,l)'*y; end
A = [y'*y sqrt(penalty)*((1-tau)/tau) n/2 shrink];
samp(1,:) = ExactStart(hats,A,k);
while i < N
    i = i + 1;
    g = samp(i-1,:); %take previous value
    for j = 2:k %Gibbs sampler
        [P1] = UpdateOrth(g,j,hats,A); %calculate the P(gi=1)
        g(j) = rand <= P1;
    end % for j = 2:k+1
    samp(i,:) = g; %record updated vector as the next g vector
end
runtime = (cputime - t);
dec = bin2dec(num2str(samp(:,2:k),'%1.f')); %turn vectors into dec values, can
do a max of 52 values!

function P1 = UpdateOrth(g,j,hats,A)

%%%%%%%%%%%%%%%Jason Bentley (2008) University of Canterbury%%%%%%%%%%%%%%%
%Update function for the conditional distribution of the posterior for    %
%model probabilities                                                      %
%INPUT: g is current binary vector                                        %
%       j is the current component being updated                          %
%       hats is the of individual Sums of Squares for each predictor      %
%       A is various constants as above n the binomial                    %
%OUTPUT:P1 is the probability the component is = 1.                       %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

g0 = g; g0(j) = 0; g0hat = sum(hats(g0==1));
P1 = 1/(1+exp(log(A(2)) + A(3)*(log(1-((A(4)*hats(j))/(A(1)-(A(4)*g0hat)))))));
```

function [dec bct runtime] = GibbsPerfect(y,X,tau,penalty,shrink,N)

```
%%%%%%%%%%%%%%%Jason Bentley (2008) University of Canterbury%%%%%%%%%%%%%%%
%Perfect Gibbs sampler for Bayesian variable selection with an orthogonal %
%design matrix W in linear regression using Zellner's prior or Jeffreys   %
%prior with the binomial prior for gamma                                  %
%INPUT: y is the n.1 response vector                                      %
%       X is n.(k+1) design matrix                                        %
%       tau is the choice of hyper-parameter in the constant Bernoulli    %
%           prior (using tau = 0.5 corresponds to a uniform prior)        %
%       penalty and shrink specify whether Jeffreys prior or Zellner's    %
%           prior is used. For Jeffreys penalty corresponds to p = 2*pi*  %
%           penalty with shrink = 1. For Zellner's prior penalty = (c+1)  %
%           and shrink should be set to c/(c+1).                          %
%       N is the number of samples to be generated                        %
%OUTPUT:dec is an N.1 vector of samples represented in decimal form       %
%       runtime is the required cputime to generate the N samples, memory %
%           for output storage                                            %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

t = cputime; [n k] = size(X); v = 0; samp = zeros(N,k); bct = zeros(N,1);
for l = 1:k; hats(l) = y'*X(:,l)*inv(X(:,l)'*X(:,l))*X(:,l)'*y; end
A = [y'*y sqrt(penalty)*((1-tau)/tau) n/2 shrink];
while v <= N
v = v + 1;
m = 1;
k1 = k - 1;
g1 = ones(1,k);
g0 = [1 zeros(1,k1)];
u = rand(k1,m);
while ~isequal(g0,g1)
    u = [rand(k1,m) u]; m = 2*m; g1 = ones(1,k); g0 = [1 zeros(1,k1)];
    for i = 1:m % iterate from the past
       for j = 2:k % Gibbs sampler
            P10 = UpdateOrth(g0,j,hats,A);
            g0(j) = u(j-1,i) < P10;
            if g0(j) == 1;
                g1(j) = 1; % by monotonicity: if g0(j) is 1, so must g1(j)
            else P11 = UpdateOrth(g1,j,hats,A);
                g1(j) = u(j-1,i) < P11;
            end
       end % for j = 2:k
    end % for i = 1:m
end % while ~isequal(g0,g1)
samp(v,:) = g0; bct(v) = m;
end %runtime loop
runtime = (cputime - t);
dec = bin2dec(num2str(samp(:,2:k),'%1.f')); %turn vectors into dec values, can
do a max of 52 values!
```

function P1 = UpdateOrth(g,j,hats,A)

```
%%%%%%%%%%%%%%%Jason Bentley (2008) University of Canterbury%%%%%%%%%%%%%%%
%Update function for the conditional distribution of the posterior for    %
%model probabilities                                                      %
```

```
%INPUT: g is current binary vector                                        %
%       j is the current component being updated                          %
%       hats is the of individual Sums of Squares for each predictor      %
%       A is various constants as above n the binomial                    %
%OUTPUT:P1 is the probability the component is = 1.                       %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

g0 = g; g0(j) = 0; g0hat = sum(hats(g0==1));
P1 = 1/(1+exp(log(A(2)) + A(3)*(log(1-((A(4)*hats(j))/(A(1)-(A(4)*g0hat)))))));
```

```
function [dec runtime] = GibbsSampler(y,X,gstart,tau,penalty,shrink,N)

%%%%%%%%%%%%%%%Jason Bentley (2008) University of Canterbury%%%%%%%%%%%%%%%
%Gibbs sampler for Bayesian variable selection in linear regression using %
%Zellner's prior or Jeffreys prior with the binomial prior for gamma.     %
%INPUT: y is the n.1 response vector                                      %
%       X is n.(k+1) design matrix                                        %
%       gstart is the specified starting value for the Gibbs sampler, it  %
%           must contain a 1 in the first position and be of length k+1   %
%       tau is the choice of hyper-parameter in the constant Bernoulli    %
%           prior (using tau = 0.5 corresponds to a uniform prior)        %
%       penalty and shrink specify whether Jeffreys prior or Zellner's    %
%           prior is used. For Jeffreys penalty corresponds to p = 2*pi*  %
%           penalty with shrink = 1. For Zellner's prior penalty = (c+1)  %
%           and shrink should be set to c/(c+1).                          %
%       N is the number of samples to be generated                        %
%OUTPUT:dec is an N.1 vector of samples represented in decimal form       %
%       runtime is the required cputime to generate the N samples, memory %
%           for output storage                                            %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

t = cputime; %start recording the time for the Gibbs sampler
[n k] = size(X); %k includes the intercept
samp = zeros(N,k); %storage
samp(1,:) = gstart; %first row of sample is starting gamma vector
vv = X'*y; V = X'*X; A = [y'*y n/2 sqrt(penalty)*((1-tau)/tau) shrink];
%constants
i = 1; %track sample size
while i < N %running time
    i = i + 1;
    g = samp(i-1,:); %take previous value
    for j = 2:k %Gibbs sampler
        [P1] = Update(V,g,j,A,vv); %calculate the P(gi=1)
        g(j) = rand <= P1;
    end % for j = 2:k+1
    samp(i,:) = g; %record updated vector as the next g vector
end
runtime = (cputime - t); %record cputime not including conversion of g to dec
dec = bin2dec(num2str(samp(:,2:k),'%1.f'));
%turn vectors into dec values can do a max of 52 values!
```

```
function [P1] = Update(V,g,j,A,vv)

%%%%%%%%%%%%%%%Jason Bentley (2008) University of Canterbury%%%%%%%%%%%%%%%
%Update function for the conditional distribution of the posterior for    %
%model probabilities                                                      %
%INPUT: g is current binary vector                                        %
%       j is the current component being updated                          %
%       hats is the of individual Sums of Squares for each predictor      %
%       A is various constants as above n the binomial                    %
%       vv is the covariance matrix for the full model                    %
%OUTPUT:P1 is the probability the component is = 1.                       %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
g1 = g; g1(j) = 1; g0 = g; g0(j) = 0;
Vg1 =inv(V(g1==1,g1==1)); Vg0 = inv(V(g0==1,g0==1));
P1 = 1/(1+exp(log(A(3)) + A(2)*(log(A(1)-A(4)*vv(g1==1)'*Vg1*vv(g1==1))-...
    log(A(1)-A(4)*vv(g0==1)'*Vg0*vv(g0==1)))));
```

function [GPC Low GS GS1] = OrthDesign(Xo,y)

```
%%%%%%%%%%%%%%%Jason Bentley (2008) University of Canterbury%%%%%%%%%%%%%%%
%Gibbs sampler for Bayesian variable selection in linear regression using %
%Zellner's prior or Jeffreys prior with the binomial prior for gamma.     %
%INPUT: y is the n.1 response vector                                      %
%       X is n.(k+1) design matrix                                        %
%       gstart is the specified starting value for the Gibbs sampler, it  %
%           must contain a 1 in the first position and be of length k+1   %
%       tau is the choice of hyper-parameter in the constant Bernoulli    %
%           prior (using tau = 0.5 corresponds to a uniform prior)        %
%       penalty and shrink specify whether Jeffreys prior or Zellner's    %
%           prior is used. For Jeffreys penalty corresponds to p = 2*pi*  %
%           penalty with shrink = 1. For Zellner's prior penalty = (c+1)  %
%           and shrink should be set to c/(c+1).                          %
%       N is the number of samples to be generated                        %
%OUTPUT:dec is an N.1 vector of samples represented in decimal form       %
%       runtime is the required cputime to generate the N samples, memory %
%           for output storage                                            %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

[n k] = size(Xo); X = Xo(:,2:k); Int = ones(n,1)./sqrt(n);
for i = 1:k-1; X(:,i) = X(:,i)-mean(X(:,i)); end %centre every predictor

%%%%%Principal Components%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

D = diag(diag(X'*X).^0.5); A = D*(X'*X)*D; [U V] = eig(A); PC1 = X*(D*U);
%U is eigen vectors, V is eigen values.
for i = 1:(k-1); PC2(:,i)=PC1(:,i)./norm(PC1(:,i)); end; GPC = [Int PC2];

%%%%%SVD to Lowdin%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

[M N K] = svd(X,0); SV = [Int M]; %economy size decomposition
ASV = (N*K'); Low = M*K'; Low = [Int Low];

%%%%%Gram-Schmidt%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

[GS1] = cgrscho1(y,X); GS1 = [Int GS1]; %method 1
[GS2] = cgrscho2(y,X); GS2 = [Int GS2]; %method 2
```

function [A] = cgrscho1(y,X)

```
% Created by A. Trujillo-Ortiz, R. Hernandez-Walls, A. Castro-Perez
%            and K. Barba-Rojo
%            Facultad de Ciencias Marinas
%            Universidad Autonoma de Baja California
%            Apdo. Postal 453
%            Ensenada, Baja California
%            Mexico.
%            atrujo@uabc.mx
% Copyright. September 28, 2006.

[mag pos] = sort(abs(corr(y,X)),'descend');
X = X(:,pos); %re-ordering of X based on correlations
A = X; [m n]=size(A);
for j= 1:n
    R(1:j-1,j)=A(:,1:j-1)'*A(:,j);
    A(:,j)=A(:,j)-A(:,1:j-1)*R(1:j-1,j);
    R(j,j)=norm(A(:,j));
    A(:,j)=A(:,j)/R(j,j);
end
return,
```

function [A] = cgrscho2(y,X)

```
% Created by A. Trujillo-Ortiz, R. Hernandez-Walls, A. Castro-Perez
%            and K. Barba-Rojo
%            Facultad de Ciencias Marinas
%            Universidad Autonoma de Baja California
%            Apdo. Postal 453
%            Ensenada, Baja California
%            Mexico.
%            atrujo@uabc.mx
% Copyright. September 28, 2006.
% This copyright does not include the sub-routine orderyX
[X order] = orderyX(y,X); X = X(:,order); %re-order X based on correlation
%with y and X
A = X; [m n]=size(A);
for j= 1:n
    R(1:j-1,j)=A(:,1:j-1)'*A(:,j);
    A(:,j)=A(:,j)-A(:,1:j-1)*R(1:j-1,j);
    R(j,j)=norm(A(:,j));
    A(:,j)=A(:,j)/R(j,j);
end
return,
```

function [Xnew order] = orderyX(y,X)

```
[n k] = size(X); [mag pos] = sort(abs(corr(y,X)),'descend');
Xnew(:,1) = X(:,pos(1)); %the most correlated variable with y
order(1) = pos(1); mag = mag(2:k); pos = pos(2:k); %remove the first predictor
chosen
for i = 2:(k-1)
    stage2 = abs(corr(Xnew(:,i-1),X(:,pos))); %the correlation between the most
correlated variable with y and repeat this in a loop for k.
    d = sqrt((stage2.^2)+((1-mag).^2)); %vector operations for distances
    ind = find(d==min(d));
    Xnew(:,i) = X(:,pos(ind)); %make the next predictor that which minimizes
the distance to 0,1 or min corr with previous Xnew and max corr with y
    order(i) = pos(ind);
    pos = setdiff(pos,pos(ind)); %need to update pos vector by removing the
most recent added variable
    mag = setdiff(mag,mag(ind));
end
Xnew(:,k) = X(:,pos); %whats left not working!!!!!!
order(:,k) = pos;
```

function [samp count] = gRejection(y,X,penalty,shrink,tau,M)

```
%%%%%%%%%%%%%%%Jason Bentley (2008) University of Canterbury%%%%%%%%%%%%%%%
%Rejection sampler for the conditional distribution of c using Zellner’s  %
%prior. The prior for c is the Hyper-G-n.                                 %
%INPUT: y is the response vector                                          %
%       X is the predictor matrix                                         %
%       tau is the choice of hyper-parameter in the constant Bernoulli    %
%           prior (using tau = 0.5 corresponds to a uniform prior)        %
%       penalty and shrink specify whether Jeffreys prior or Zellner's    %
%           prior is used. For Jeffreys penalty corresponds to p = 2*pi*  %
%           penalty with shrink = 1. For Zellner's prior penalty = (c+1)  %
%           and shrink should be set to c/(c+1).                          %
%       M is the number of samples to be generated                         %
%OUTPUT:samp is N i.i.d samples from the required conditional posterior   %
%       for c.                                                             %
%       count is the N waiting times to generate each sample point        %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

format long g;
[n k1] = size(X); k = k1-1; d = 2^k; A = y'*y; XX = X'*X; BX = (X'*y); N = n/2;
%constants
for i = 1:k1
Q(i)=(nchoosek(k,(i-1))*((penalty)^-(i/2)))*(tau^(i-1))*((1-tau)^(k-(i-1)));
end; Q = Q./sum(Q); %calculate proposal density
B = ((y'*y-shrink*y'*X*inv(X'*X)*X'*y)^-N); %compute bound
cn = 0; v = 0;
while cn < M
    g = zeros(1,k); q = randsample(0:k,1,'true',Q);
    if q > 0; g(randsample(1:k,q))=1; g = [1 g]; else; g = [1 zeros(1,k)]; end
    %generate a proposal gamma
    RSS = (BX(g==1)'*inv(XX(g==1,g==1))*BX(g==1)); %hat matrix
    P = ((y'*y-shrink*RSS)^-N)/B;
    if rand <= P
        cn = cn + 1
        samp(cn) = bin2dec(num2str(g(2:k1))); %store decimal
        %samp(cn,:) = g; %can choose to store g requires more memory
        count(cn) = v;
        v = 0;
    end
    v = v + 1;
end
```

function [P Pqg Eqg Map Med Marg yBMA DIC] = Jeffrey(y,X,penalty,tau)

```
%%%%%%%%%%%%%%%Jason Bentley (2008) University of Canterbury%%%%%%%%%%%%%%%%
%Posterior for gamma for Jeffreys prior with the binomial prior for qg.   %
%INPUT: y is the n.1 response vector                                       %
%       X is n.(k+1) design matrix                                         %
%       tau is the choice of hyper-parameter in the constant Bernoulli    %
%           prior (using tau = 0.5 corresponds to a uniform prior)        %
%       penalty should be set p = 2*pi*(c+1) to mimic the penalty of      %
%           Zellner's prior.                                               %
%           and shrink should be set to c/(c+1).                           %
%OUTPUT:P is the normalized posterior probability for eacg gamma          %
%       Pqg is the posterior probability of the model sizes 0:k           %
%       Eqg is the expected model size                                     %
%       Map is maximum aposteriori estimate model                          %
%       Med is the median probability model which incldues all predictors %
%           with MIP > 0.5                                                 %
%       Marg are the MIP                                                   %
%       yBMA is the model averaged fitted response                         %
%       DIC is the model averaged DIC                                      %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

format long g; [n k1] = size(X); %constants
k = k1-1; d = 2^k; C2 = log(penalty/(2*pi)); AA = y'*y; %constants
P = zeros(1,d); S = P; Q = P; D = P; AIC = P; VV = P; %storage
for i = 1:d %loop through models
    g = [1 str2num(dec2bin(i-1,k)')']; %generate a single gamma vector to be
evaluated
    Xg = X(:,g==1);
    S(i) = sum(g)-1; %the sum of gamma
    bhat = inv(Xg'*Xg)*Xg'*y;
    Q(i) = log(AA-y'*Xg*bhat); %the log of the quadratic term
    sighat = exp(Q(i))/(n-2); %the posterior expectation
    DIC(i) = -2*(sum(log(normpdf(y,Xg*bhat,sqrt(sighat))))) + 2*S(i) + 4; %This
is also DIC
end %end loop which has generated two vectors of values one for each quadratic
term in
%the posterior and the second for the sum of the gamma vector.
for i = 1:d; B = ((((S(i)-S(j~=i))/2)*C2)+N*(Q(i)-Q(j~=i)));
    P(i) = ((1 + sum(exp(B)))^-1)*((tau^S(i))*((1-tau)^(k-S(i))));
end; clear Q;
P = P./sum(P); %renormalize after adding the proportional prior
for i = 1:d; yhat(:,i) = P(i).*yhat(:,i); end %multiply each models predicted
values
%by the posterior probability
dec = find(P==max(P)); Map = [1 str2num(dec2bin(dec-1,k)')']; %find the MAP
Marg = Margprob(P,k); %Marginal inclusion probabilities this includes 1 for the
intercept
Med = [1 Marg>=0.5]; %calculate the median model
for i = 1:k+1; Pqg(i) = sum(P(S==(i-1))); end %Posterior for model size
yBMA = sum(yhat,2); rBMA = y - yBMA; %the sum across rows and model averaged
residuals
Eqg = sum(P.*S); clear S
for i = 1:k+1; Pqg(i) = sum(P(S==(i-1))); end
DIC = sum(P.*DIC);
```

function M = Margprob(P,k)

```
Mat = zeros(2^k,k); for i = 1:2^k;
Mat(i,:) = P(i)*[str2num(dec2bin(i-1,k)')']; end; M = sum(Mat);
```

function [tails PC tailstats] = ModelCheck(y,X,P,shrink,N,a)

```
%%%%%%%%%%%%%Jason Bentley (2008) University of Canterbury%%%%%%%%%%%%%%%%%
%Function for checking model adequacy in linear regression using tail     %
%probabilities for the observations and statistics of y (min, max, median,%
%std. dev.) and the predictive coverage.                                  %
%INPUT: y is the n.1 response vector                                      %
%       X is n.(k+1) design matrix                                        %
%       P is the (2^k).1 vector of posterior probabilities.               %
%       shrink specify whether Jeffreys prior or Zellner's prior is used. %
%           For Jeffreys shrink = 1. For Zellner's prior shrink = c/(c+1).%
%       N is the number of samples to be generated for each model from the%
%           PPD to estimate tail prob for statistics of y.                %
%       a is the tail probability for the (a/2), 1-(a/2) interval for     %
%           assessing predictive coverage.
%OUTPUT:tails is an N.1 vector of tail probabilities model averaged for   %
%           each observation.                                             %
%       PC is the modelaaveraged predictive coverage under the PPD.       %
%       tailtstats are the tail probabilities for the min, max, median and%
%           std. dev. of y.                                               %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

[n k1] = size(X); k = k1 -1;
for i = 1:length(P)
    g = [1 str2num(dec2bin(i-1,k)')'];
    [tprob PCi(i) pval] = PPD(y,X,g,N,shrink,a);
    tprob1(:,i) = P(i).*tprob;
    pval1(i,:) = P(i).*pval;
end
tails = sum(tprob1,2);
tailstats = sum(pval1,1);
PC = sum(PCi.*P)*100;
```

function [tprob PCi pval] = PPD(y,X,g,N,shrink,a)

```
format long g; n = length(y); Ig = eye(n); Xg = X(:,g==1);
Hg = shrink*Xg*inv(Xg'*Xg)*Xg'; sig = ((y'*(Ig-Hg)*y)/n)*(Ig + Hg); mu = Hg*y;
ytrans = (y - mu)./(sqrt(diag(sig))); %transform to a standard t r.v.
PCi = sum(abs(ytrans) <= tinv(1-(a/2),n))/n;
tprob = tcdf(ytrans,n); %from -inf to x so the left hand tail
tprob(find(tprob>=0.5)) = 1-tprob(find(tprob>=0.5));
Y = MVTpRnd(n,mu,sig,N); %simulate from the PPD
sumin = min(Y); A = min(y); pval(1) = min(sum(sumin<=A)/N,sum(sumin>=A)/N);
sumax = max(Y); B = max(y); pval(2) = min(sum(sumax<=B)/N,sum(sumax>=B)/N);
sumed = median(Y); C = median(y); pval(3) =
min(sum(sumed<=C)/N,sum(sumed>=C)/N);
sumstd = std(Y); D = std(y); pval(4) = min(sum(sumstd<=D)/N,sum(sumstd>=D)/N);
```

function Y = MVTpRnd(v,mu,sig,N)

```
p = length(mu); Y = zeros(p,N); sig1 = zeros(p,p); sig1 = diag(diag(sig));
X = csmvrnd(zeros(p,1),sig1,N); s = sqrt(chi2rnd(v,N)./v);
for i = 1:N; Y(:,i) = (X(i,:)./s(i))'+mu; end
```

function [P Pqg Eqg Map Med Marg yBMA DIC] = Zellner(y,X,c,tau,dic,tol,v)

```
%%%%%%%%%%%%%%%Jason Bentley (2008) University of Canterbury%%%%%%%%%%%%%%%
%Posterior for gamma for Zellner's prior with the binomial prior for qg.  %
%INPUT: y is the n.1 response vector                                      %
%       X is n.(k+1) design matrix                                        %
%       tau is the choice of hyper-parameter in the constant Bernoulli    %
%           prior (using tau = 0.5 corresponds to a uniform prior)        %
%       c choice of c in Zellner's prior                                  %
%       dic is a ~=1, 1 option 1 = compute dic, ~=1 = do not compute DIC  %
%       tol minimum error (standard deviation) for the simulation estimate%
%           of deviance and DIC                                           %
%       v is the numer of samples to generate in simulation of deviance   %
%           between each check of the simulation error                    %
%OUTPUT:P is the normalized posterior probability for eacg gamma          %
%       Pqg is the posterior probability of the model sizes 0:k           %
%       Eqg is the expected model size                                    %
%       Map is maximum aposteriori estimate model                         %
%       Med is the median probability model which incldues all predictors %
%           with MIP > 0.5                                                %
%       Marg are the MIP                                                  %
%       yBMA is the model averaged fitted response                        %
%       DIC is the model averaged DIC                                     %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

format long g; [n k1] = size(X); k = k1-1; d = 2^k; C1 = c/(c+1); %constants
C2 = log(c+1); N = n/2; I = eye(n); j = 1:d; %constants
P = zeros(1,d); S = P; Q = P; DIC = P; BIC = P; yhat = zeros(n,d); %storage
for i = 1:d %loop through models
    g = [1 str2num(dec2bin(i-1,k)')']; %generate a single gamma vector
    Xg = X(:,g==1); %predictor matrix adjusted by gamma
    Hg = Xg*inv(Xg'*Xg)*Xg'; %hat matrix
    bhat = C1*inv(Xg'*Xg)*Xg'*y;
    yhat(:,i) = Xg*bhat;
    S(i) = sum(g)-1; %the sum of gamma
    Q(i) = log(y'*(I-(C1*Hg))*y); %the log of the quadratic term
    sighat = exp(Q(i))/(n-2); %the posterior expectation
    VV(i) = -2*(sum(log(normpdf(y,Xg*bhat,sqrt(sighat)))));
end %end loop which has generated two vectors of values one for each quadratic
term in
%the posterior and the second for the sum of the gamma vector.
for i = 1:d; B = ((((S(i)-S(j~=i))/2)*C2)+N*(Q(i)-Q(j~=i)));
    P(i) = ((1 + sum(exp(B)))^-1)*((tau^S(i))*((1-tau)^(k-S(i))));
```

```
end; clear Q;
P = P./sum(P); %renormalize after adding the proportional prior
for i = 1:d; yhat(:,i) = P(i).*yhat(:,i); end %multiply each models predicted
values by the posterior probability
dec = find(P==max(P)); Map = [1 str2num(dec2bin(dec-1,k)')']; %find the MAP
Marg = Margprob(P,k); %Marginal inclusion probabilities this includes 1 for the
intercept
Med = [1 Marg>=0.5]; %calculate the median model
for i = 1:k+1; Pqg(i) = sum(P(S==(i-1))); end %Posterior for model size
yBMA = sum(yhat,2); %the sum across rows and model averaged residuals
Eqg = sum(S.*P); clear S
if dic == 1;[Dev] = PostExpDevZell(y,X,P,v,c,tol); else; dic = 'NA'; end
%calculate Deviance
pd = dev-sum(VV.*P); DIC = dev + pd;
```

function M = Margprob(P,k)

```
Mat = zeros(2^k,k); for i = 1:2^k;
Mat(i,:) = P(i)*[str2num(dec2bin(i-1,k)')']; end; M = sum(Mat);
```

function [Dev] = PostExpDevZell(y,X,P,n,c,tol)

```
sd = 10;
[D] = DevSimZell(y,X,P,c);
while sd > tol
for i = 1:n
[Dev(i)] = DevSimZell(y,X,P,c);
end
D = [D Dev]; m = length(D);
sd = sqrt((1/m)*((1/m)*sum(D.^2-mean(D)^2)));
end
Deviance = mean(D);
```

function [Dev] = DevSimZell(y,X,P,c)

```
[n k1] = size(X); k = k1 - 1; CDF = cumsum(P);
g = PostGamSimX(CDF,k);
Xg = X(:,g==1); bhat = inv(Xg'*Xg)*Xg'*y;
B = ((y'*y)/2)-((c/(2*(c+1)))*y'*Xg*inv(Xg'*Xg)*Xg'*y); %beta parameter
sig2sim = 1/gamrnd(n/2,B^-1,1,1); %simulate sigma2
bhatsim = csmvrnd((c/(c+1))*bhat,((sig2sim*c)/(c+1))*inv(Xg'*Xg),1);
Dev = -2*(sum(log(normpdf(y,Xg*bhatsim',sqrt(sig2sim)))));
```

APPENDIX E: Zellner’s Prior Additional DIC Figures

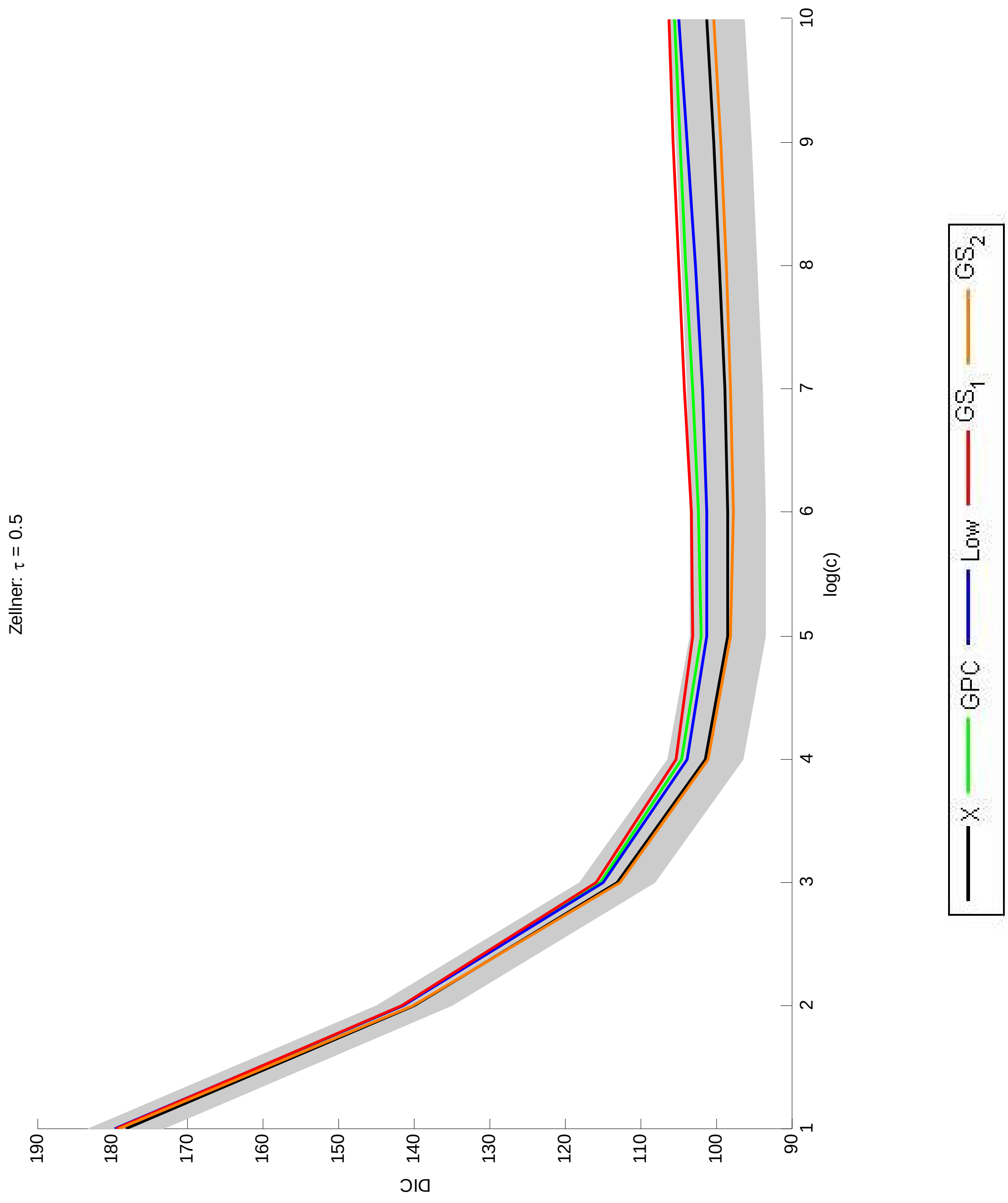


Figure E.1 Enlargement of DIC plot over the choice of penalty for Zellner’s prior and the ozone data.

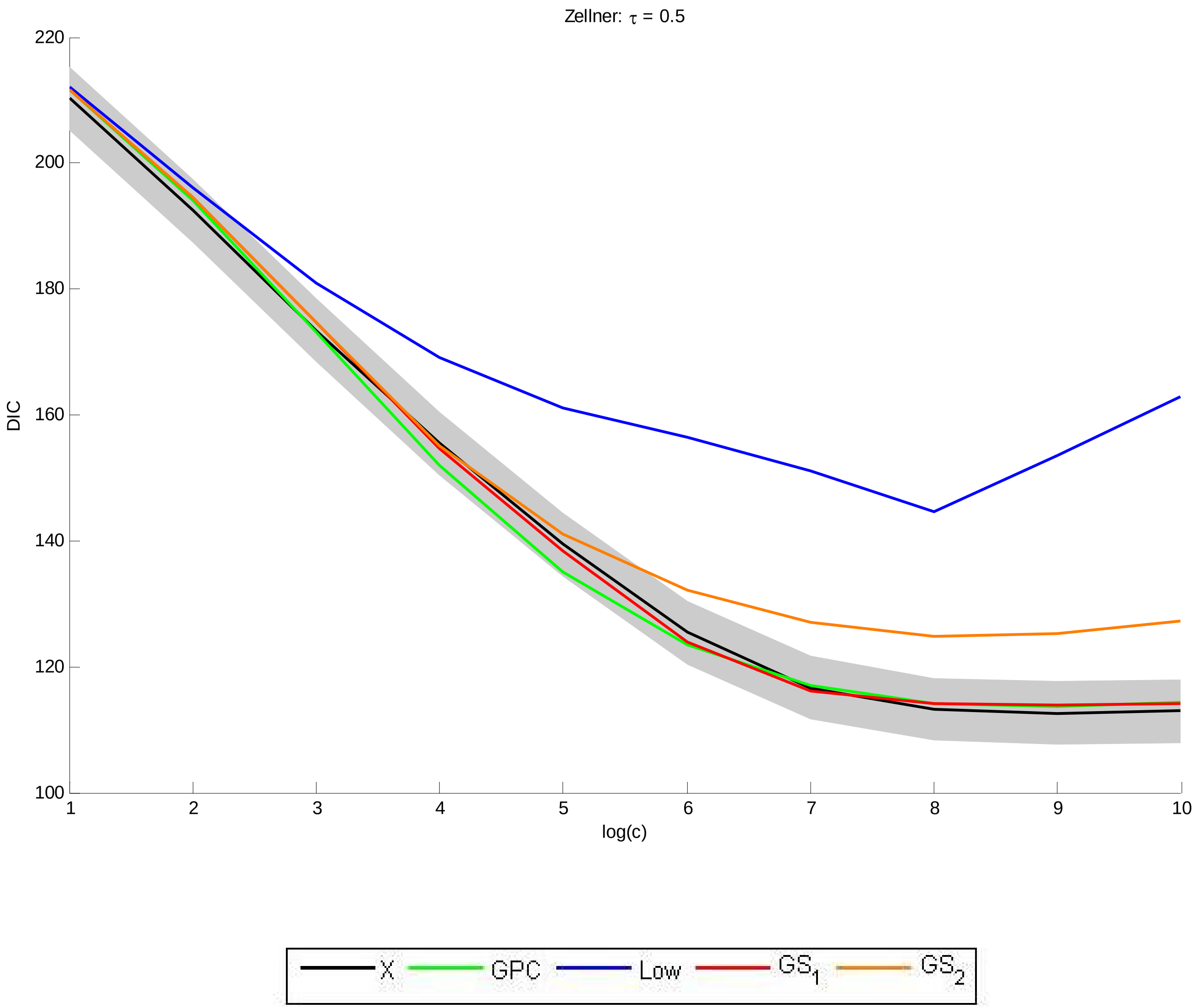


Figure E.2 Enlargement of DIC plot over the choice of penalty for Zellner's prior and the physical data.

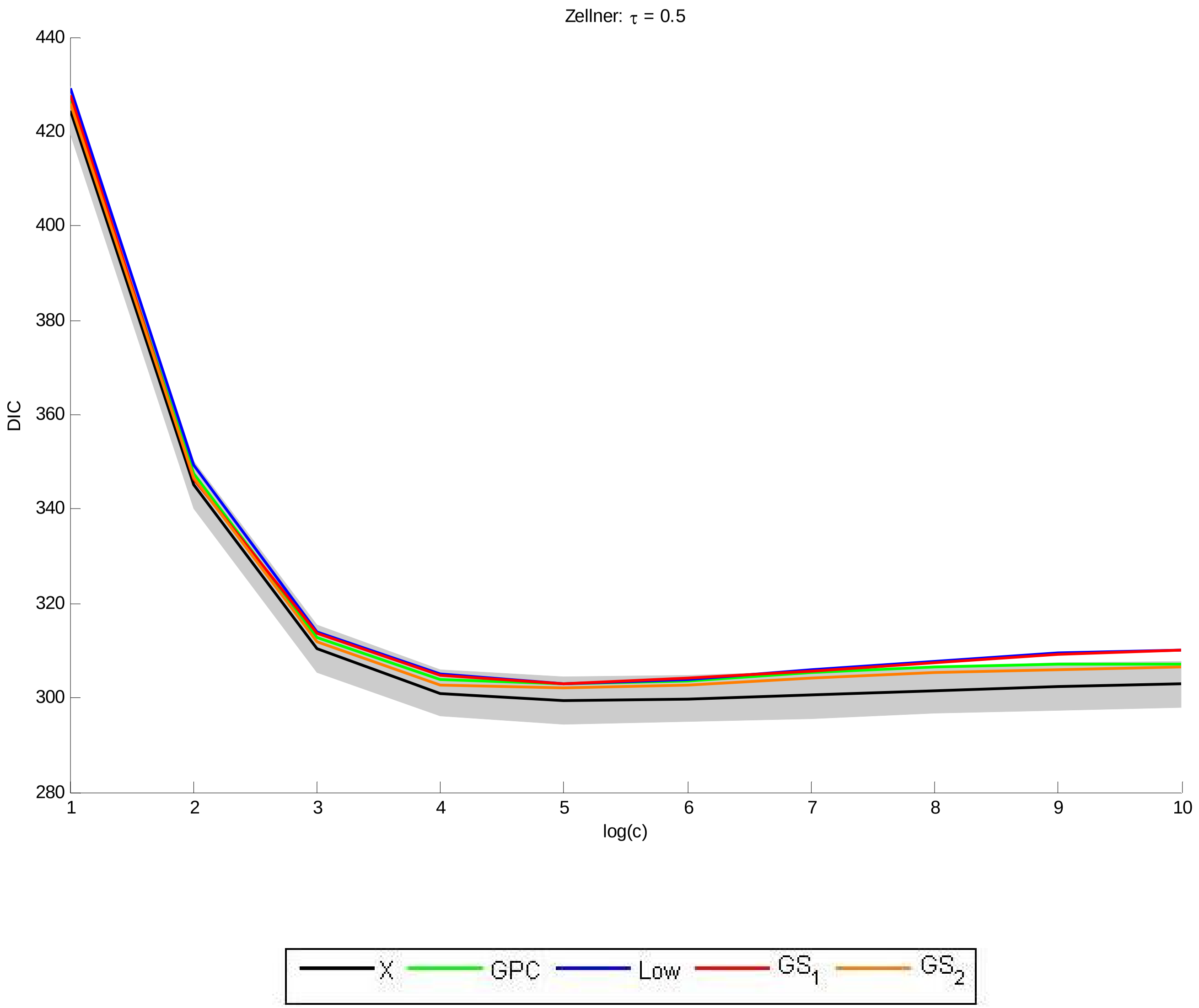


Figure E.3 Enlargement of DIC plot over the choice of penalty for Zellner's prior and the bodyfat data.

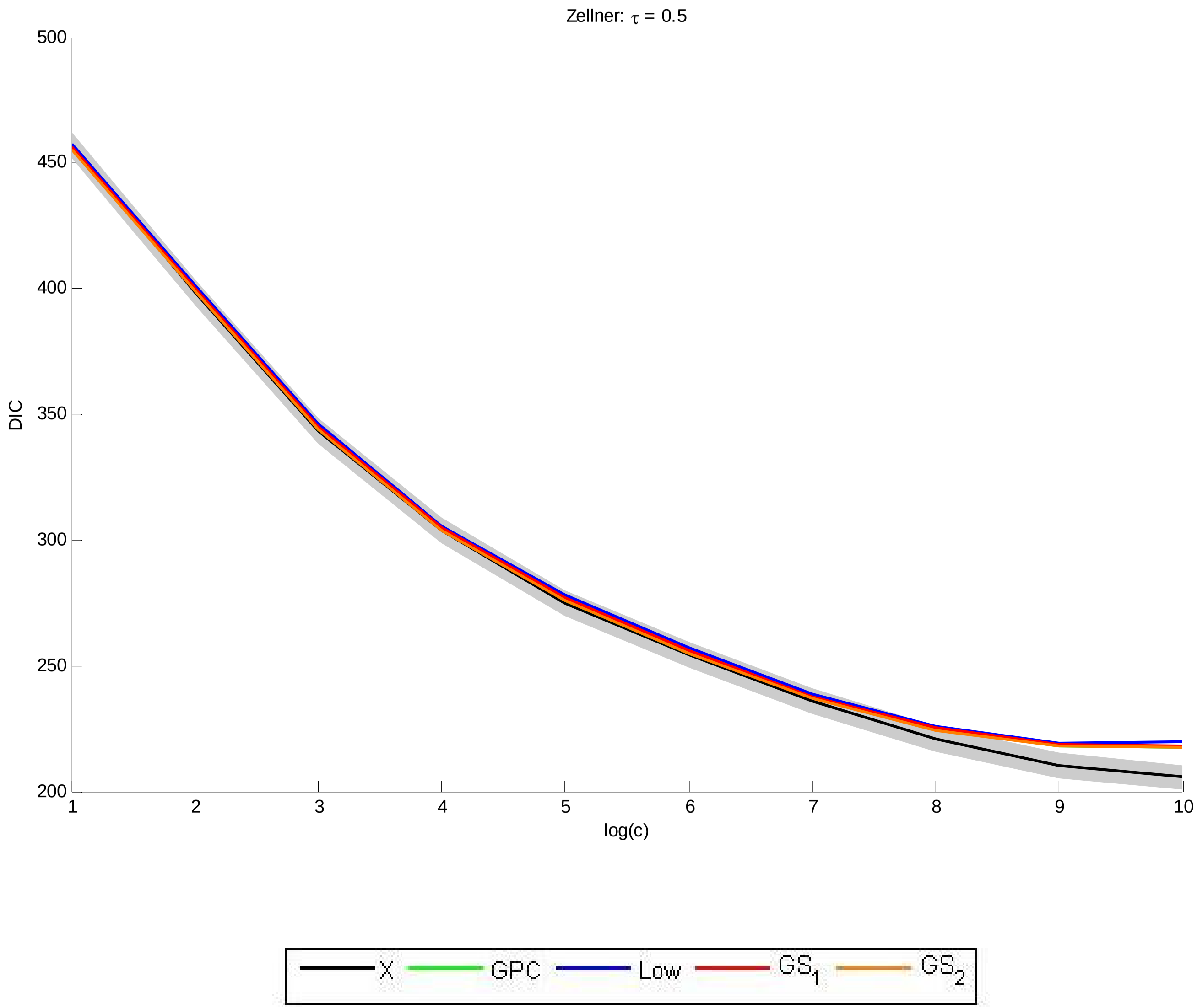


Figure E.4 Enlargement of DIC plot over the choice of penalty for Zellner's prior and the crime data.

REFERENCES

1. Athreya, K.B., and Ney, P. (1978). A new approach to the limit theory of recurrent Markov chains. *Transactions of the American Mathematical Society*, 245, 493-501.
2. Ambler, G.K., and Silverman, B.W. (2004). Perfect simulation for Bayesian wavelet thresh-holding with correlated coefficients. Preprint from http://dbwilson.com/exact.
3. Bartlett, M. (1957). A comment on D.V. Lindley's statistical paradox. *Biometrika*, 44, 533-534.
4. Beaver, S. (2007). Lowdin orthogonalization – A natural supplement to Gram-Schmidt. Preprint from http://www.wou.edu/~beavers/Talks/TalksPage.html.
5. Berger, J.O., and Pericchi, L. (2001). Objective Bayesian methods for model selection: Introduction and comparison. In *Model Selection*, IMS monograph 38, Lahiri, P. (ed). Institute of Mathematical Statistics: Beachwood, OH, 135-193.
6. Breyer, L.A., and Roberts, G.O. (2000). Catalytic perfect simulation. *Methodology and Computing in Applied Probability*, 3(2), 161-171.
7. Brooks, S.P., Yanan, F., Rosenthal, J.S. (2006). Perfect forward simulation via simulated tempering. *Communications in Statistics – Simulation and Computation*, 35, 683-713.
8. Brown, P.J., Vannucci, M., and Fearn, T. (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society Series B*, 64(3), 519-536.
9. Burdzy, K., and Kendall, W.S. (2000). Efficient Markovian couplings: examples and counter examples. *The annals of applied probability*, 10(2), 362-409.
10. Cai, Y., and Kendall, W.S. (2002). Perfect simulation for correlated Poisson random variables conditioned to be positive. *Statistics and Computing*, 12, 229-243.
11. Cai, Y. (2005). A non-monotone CFTP perfect simulation method. *Statistica Sinica*, 15, 927-943.

12. Carlin, B., and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo. *Journal of the Royal Statistical Society Series B*, 57, 473-484.

13. Carvalho, M.S., and Corcoran, J.N. (2005). Finding the stationary distribution of ARCH models through perfect simulation. Preprint from http://ww.amath.colorado.edu/~corcoran/Papers/arch.pdf.

14. Casella, G., Mengersen, K.L., Robert, C.P., and Titterington, D.M. (2002). Perfect slice samplers for mixtures of distributions. *Journal of the Royal Statistical Society Series B*, 64(4), 777-790.

15. Celeux, G., Marin, J.-M., and Robert, C.P. (2007). Sélection bayésienne de variables en régression linéaire. *Journal de la Société Française de Statistique*, 147, 59-79.

16. Chipman, H., Georege, E., and McCulloch, R. (2001). The practical implementation of Bayesian model selection. In *Model Selection*, IMS monograph 38, Lahiri, P. (ed). Institute of Mathematical Statistics: Beachwood, OH, 67-116.

17. Clyde, M., Desimone, H., and Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association*, 91, 1197-1208.

18. Clyde, M., and George, E.I. (2004). Model Uncertainty. *Statistica Scinica*, 19, 81-94.

19. Connor, S.B., and Kendall, W.S. (2007). Perfect simulation for a class of positive recurrent Markov chains. *Annals of Applied Probability*, 17(3), 781-808.

20. Corcoran, J.N., and Tweedie, R.L. (2000). Perfect sampling from independent Metropolis-Hastings chains. *Journal of statistical planning and inference*, 104, 207-314.

21. Cripps, E., Kohn, R., and Nott, D. (2006). Bayesian subset selection and model averaging using a centered and dispersed prior for the error variance. *Australian and New Zealand Journal of Statistics*, 48(2), 237-252.

22. Czado, C., and Raftery, A.E. (2006). Choosing the link function and accounting for link uncertainty in generalized linear models using Bayes factors. *Statistical Papers*, 47, 419-442.

23. Dellaportas, P., Forster, J., and Ntzoufras, I. (2002). On Bayesian model and

variable selection using MCMC. *Statistics and Computing*, 12, 27-36.

24. Dimakos, X.K. (2001). A guide to exact simulation. *International Statistics review*, 69(1), 27-48.

25. Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society Series B*, 57, 45-97.

26. Draper, N.R., and Smith, H. (1998) *Applied Regression Analysis*. 3rd Ed. John Wiley and Sons Inc.

27. Eklund, J., and Karlsson, S. (2007). Computational efficiency in Bayesian model and variable selection. Preprint from *econpapers.repec.org/paper/hhsoruesi/2007_5f004.html.*

28. Elsalloukh, H., Yound, D., and Guardiola, J. (2005). The epsilon-skew exponential power distribution family. *Far East Journal of Theoretical Statistics*, 16, 97-112.

29. Ferrari, P.A., Fernandez, R., and Garcia, N.L. (2002). Perfect simulation for interacting point processes, loss networks and Ising models. *Stochastic Processes and their Applications*, 1, 63-88.

30. Fill, J.A. (1998). An interruptible algorithm for perfect sampling via Markov chains. *Annals of Applied Probability*, 8, 131-162.

31. Foss, S.G., and Tweedie, R.L. (1998). Perfect simulation and backward coupling. *Stochastic Models*, 14, 187-203.

32. Foster, D.P., and George, E.I. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics*, 22, 1947-1975.

33. Frigessi, A., Gasemyr, J., and Rue, H. (2000). Antithetic coupling of two Gibbs sampler chains. *The Annals of Statistics*, 28(4), 1128–1149.

34. Gamerman, D., and Lopes, H.F. (2006). *Markov chain Monte Carlo*. Chapman and Hall.

35. Gelfand, A., and Dey, D. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society*, *Series B*, 56(3), 501-514.

36. George, E.I., and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88, 881-891.

37. George, E.I., and McCulloch, R.E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7, 339-374.

38. George, E.I., and Foster, D.P. (2000). Calibration and empirical Bayes variable selection. *Biometrika,* 87, 731-747.

39. Givens, G.H., and Hoeting, J.A. (2005). *Computational Statistics*. Wiley and Sons.

40. Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, ***82***, 711-732.

41. Green, P., and Murdoch, D. (1999). Exact sampling for Bayesian inference: towards general purpose algorithms. In J. Murdoch, J. Bernardo, A. Dawid, D. Linley, and A. Smith (Eds.) *Bayesian Statistics 6*, 302-321. Oxford University Press.

42. Haggstrom, O., and Nelander, K. (1998). Exact sampling from anti-monotone systems. *Statistica Neerlandica*, 52(3), 360 - 380.

43. Haggstrom, O., van Lieshout, M.N.M., and Møller, J. (1999). Characterization results and Markov chain Monte Carlo algorithms including exact simulation for some spatial point processes. *Bernoulli*, 5, 641 - 658.

44. Haggstrom, O. (2002). *Finite Markov chains and algorithm applications*, Vol 52 of London Mathematical society student texts. Cambridge University Press, Cambridge.

45. Hansen, M.H., and Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96, 746-774.

46. Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97-109.

47. Hobert, J.P., Robert, C.P., and Titterington, D.M. (1999). On perfect simulation for some mixtures of distributions. *Statistics and Computing*, 9, 287-298.

48. Hobert, J.P., and Robert, C.P. (2004). A mixture representation of π with applications in Markov chain Monte Carlo and perfect sampling. *The Annals of Applied probability*, 14(3), 1295-1305.

49. Hoeting, J.A., Madigan, D., Raferty, A.E., and Volinsky, C.T. (1999). Bayesian model averaging: A Tutorial. *Statistical Science*, 14(4), 382-417.

50. Hoeting, J.A. (2002). Methodology for Bayesian model averaging: an update. Preprint from http://www.state.colostate.edu/~nsu/starmap/jah.ibcbma.pdf.

51. Holmes, C.C., and Mallick, B.K. (1998). Perfect simulation for orthogonal model

mixing. Preprint from http://dbwilson.com/exact.

52. Holmes, C.C., and Denison, D.G.T. (2002). Perfect sampling for the wavelet reconstruction of signals. *IEEE Transactions on Signal Processing*, 50(2), 337-344.

53. Holmes, C.C., and Mallick, B.K. (2003). Perfect simulation for Bayesian curve and surface fitting. Preprint from *www.stat.tamu.edu/~bmallick/papers/perf.ps.*

54. Huang, Y., and Djuric, P.M. (2002). Variable selection by perfect sampling. *EURASIP Journal on Applied Signal Processing*, 1, 38-45.

55. Huber, M. (1998). Efficient exact sampling from the Ising model using Swendsen-Wang. *Tenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 921-922.

56. Huber, M. (2003). A bounding chain for the Swendsen-Wang. *Random Structures and Algorithms*, 22, 43-59.

57. Huber, M. (2004). Perfect sampling using bounding chains. *The Annals of Applied Probability*, 14(2), 734-753.

58. Johnson, V.E. (1998). A coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms. *Journal of the American Statistical Association*, 93, 238-248.

59. Kass, R.E., and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses with large samples. *Journal of the American Statistical Society*, 90, 928-934.

60. Kendall, W.S. (1997). Perfect simulation for spatial point processes. In *Proceedings of the 51st Session of the ISI*, Istanbul.

61. Kendall, W.S. (1998). Perfect simulation for the area-interaction point process. In L. Accardi and C.C. Heyde, editors, *Probability Towards 2000*, Springer-Verlag, New York, 218–234.

62. Kendall, W.S., and Thonnes, E. (1999). Perfect simulation in stochastic geometry. *Pattern recognition*, 32(9), 1569 - 1586.

63. Kendall, W.S., and Møller, J. (2000). Perfect simulation using dominating processes on ordered state spaces, with application to locally stable point processes. *Advanced in Applied Probability,* 32, 844-865.

64. Kendall, W.S. (2004). Geometric ergodicity and perfect simulation. *Electronic*

communications in Probability, 9, 140-151.

65. Kendall, W.S. (2005). Notes on perfect simulation. Preprint from http://dbwilson.com/exact/

66. Kuo, L., and Mallick, B. (1998). Variable selection and regression models. *Sankhya B*, 60, 65-81.

67. Lawson, A., and Clark, A. (2002) Spatial mixture relative risk models applied to disease mapping. *Statistics in Medicine*, 21, 359-370.

68. Lee, D.S., Iyengar, S., Czanner, G., and Roscoe, J. (2005). *Bayesian Wavelet representation of neuron signals via perfect sampling*. 2nd IMS-ISBA Joint meeting, Bormio, Italy.

69. Lee, D.S. (2008). Exact Markov Chain Monte Carlo algorithms and their applications in probabilistic data analysis and inference. In H.-F. Wang (Ed.), *Intelligent Data Analysis*: Developing New Methodologies Through Pattern Discovery and Recovery. Information Science Reference, IGI Publishing, 161-183.

70. Liang, F., Truong, Y.K., and Wong, W.H. (2001). Automatic Bayesian model averaging for linear regression and application in Bayesian curve fitting. *Statistica Sinica*, 11, 1005-1029.

71. Liang, F., Paulo, R., Molina, G., Clyde, M.A., and Berger, J.O. (2008). Mixtures of *g*-priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103, 410-423.

72. Lindvall, T. (2002). *Lectures on the Coupling method*. Dover Publications Inc., Mineola, NY.

73. Liu, J.S. (1996). Metropolized independent sampling. *Statistics and Computing*, 3, 113-119.

74. Mira, A., Møller, J., and Roberts, G.O. (2001). Perfect slice samplers. *Journal Royal Statistical Society*. 63(3), 251-264.

75. Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1087-1092.

76. Murdoch, D., J. and Meng, X.-L. (2001). Towards perfect sampling for Bayesian mixture priors. In *Bayesian Methods, with Applications to Science, Policy and*

Official Statistics, E. I. George ed., (proceedings of the ISBA 2000 conference, Hersonnissos, Crete), Office for Official Publications of the European Communities, Luxembourg, 381-390.

77. Møller, J. (1999). Perfect simulation of conditionally specified models. *Journal of the Royal Statistical Society Series B*, 61(1), 251-264.

78. Møller, J., and Nicholls, G.K. (1999). Perfect simulation for sample based inference. Preprint from http://www.math.auckland.ac.nz/~nicholls/linkfiles/ papers.html

79. Montgomery, J., and Nyhan, B. (2008). Bayesian model averaging: Theoretical developments and practical applications. Preprint from http://www.duke.edu/~bjn3/montgomery-nyhan-bma.pdf.

80. Murdoch, D.J., and Green, P.J. (1998). Exact sampling for a continuous state space. *Scandinavian Journal of Statistics*, 25, 483-502.

81. Murdoch, D.J. (2000). Exact Sampling for Bayesian Inference. Unbounded state spaces. *Monte Carlo methods - Fields Inst. Communications*. 26, 111-121.

82. Murray, I. (2004). Note on rejection sampling and exact sampling with the metropolized independence sampler. Preprint from http://www.cs.toronto.edu/~murray/pub/04rejection_cftp/rejection_cftp.pdf.

83. Nott, D.J., and Green, P.J. (2004). Bayesian variable selection and the Swendsen-Wang algorithm. *Journal of Computational and Graphical Statistics*, 13, 141-157.

84. Nott, D.J., and Kohn, R. (2005). Adaptive sampling for Bayesian variable selection. *Biometrika*, 92, 747-763.

85. Nummelin, E. (1984). *General irreducible Markov chains and non-negative operators*. Cambridge University Press.

86. Propp, J.G., and Wilson, D.B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9, 223-252.

87. Propp, J.G. and Wilson, D.B. (1999). Coupling from the past: A user's guide. Preprint from *research.microsoft.com/en-us/um/people/dbwilson/exact/user.ps.gz.*

88. Raftery, A.E., Madigan, D., and Hoeting, J. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92, 179-

191.

89. Roberts, C.P. (2001). *The Bayesian choice*. Springer.

90. Roberts, C.P., and Casella, G. (2004). *Monte Carlo statistical methods*. Springer.

91. Savage, C. (1997). A Survey of Combinatorial Gray Codes. *Society of Industrial and Applied Mathematics Review*, 39, 605–629.

92. Schneider, U., and Corcoran, J.N. (2004). Perfect sampling for Bayesian Variable selection in a linear regression model. *Journal of Statistical Planning and Inference*, 126, 153-171.

93. Smith, M., and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75, 317-343.

94. Smith, M., Sheather, S., and Kohn, R. (1996). Finite sample performance of robust Bayesian regression. *Journal of computational statistics*, 11(3), 260-301.

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

96. Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *Annals of Statistics,* 28, 40-74.

97. Thonnes, E. (1999). Perfect Simulation of some point processes for the impatient user. *Annals of Applied Probability*, 41(1), 69-87.

98. Thonnes, E. (2000). A primer on perfect simulation. In K.R. Mecke & D. Stoyan (Eds.), *Statistical Physics and Spatial Statistics*, pp. 349-378. Springer.

99. Thorrison, H. (2000). *Coupling, Stationarity, and regeneration*. Springer-Verlag, New York.

100. Tweedie, R.L., and Corcoran, J.N. (2001). Perfect Sampling and queuing models. *Proceedings of the 38th Annual Allerton Conference on Communication, Control and Computing.*

101. Wasserman, L. (2000). Bayesian Model Selection and Model Averaging. *Journal of Mathematical Psychology* 44, 92-107.

102. Wilson, D.B. (2000). How to couple from the past using a read once source of randomness. *Random Structures and Algorithms*, 16(1), 85-113.

103. Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia* (Spain), (eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 585-603. Valencia: University Press.

104. Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (Geol, P. K. and Zellner, A., eds) pp. 233-243. North-Holland, Amsterdam.