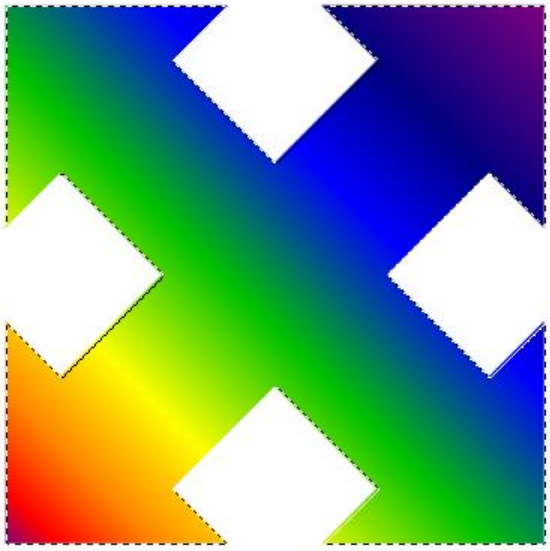
Institut für Angewandte Statistik
Johannes Kepler Universität Linz



Bayesian Variable Selection in Normal Regression Models

Masterarbeit zur Erlangung des akademischen Grades

"Master der Statistik" im Masterstudium Statistik

Gertraud Malsiner Walli

Betreuerin: Dr.^{in} Helga Wagner

November, 2010

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Masterarbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und alle den benutzten Quellen wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Linz, November 2010 Gertraud Malsiner Walli

I would like to thank my supervisor Dr. Helga Wagner for her stimulating suggestions and continuous support, she patiently answered all of my questions. I am also particularly grateful to my husband Johannes Walli for his loving understanding and encouragement for this work.

Abstract

An important task in building regression models is to decide which variables should be included into the model. In the Bayesian approach variable selection is usually accomplished by MCMC methods with spike and slab priors on the effects subject to selection. In this work different versions of spike and slab priors for variable selection in normal regression models are compared. Priors such as the Zellner's g-prior or as the fractional prior are considered, where the spike is a discrete point mass at zero and the slab a conjugate normal distribution. Variable selection under this type of prior requires to compute marginal likelihoods, available in closed form.

A second type of priors specifies both the spike and the slab as continuous distributions, e.g. as normal distributions (as in the SSVS approach) or as scale mixtures of normal distributions. These priors allow a simpler MCMC algorithm where no marginal likelihood has to be computed.

In a simulation study with different settings (independent or correlated regressors, different scales) the performance of spike and slab priors with respect to accuracy of coefficient estimation and variable selection is investigated with a particular focus on sampling efficiency of the different MCMC implementations.

Keywords: Bayesian variable selection; spike and slab priors; independence prior; Zellner's g-prior; fractional prior; normal mixture of inverse gamma distributions; stochastic search variable selection; inefficiency factor.

Contents

List of Figures

List of Tables

Introduction

Regression analysis is a widely applied statistical method to investigate the influence of regressors on a response variable. In the simplest case of a normal linear regression model, it is assumed that the mean of the response variable can be described as a linear function of influential regressors. Selection of the regressors is substantial. If more regressors are included in the model, a higher proportion of the response variability can be explained. On the other hand, overfitting, i.e. including regressors with zero effect worsens the predictive performance of the model and causes loss of efficiency.
Differentiating between variables which really have an effect on the response variable and those which have not can also have an impact on scientific result. For scientists, the regression model is not more than an instrument to represent the relationship between causes and effects of the reality which they want to detect and discover. The inclusion of non relevant quantities in the model or the exclusion of causal factors from the model yields wrong scientific conclusions and interpretations how things work.
So correct classification of these two types of regressors is a challenging goal for the statistical analysis. Thus methods for variable selection are needed to identify zero and non-zero effects.

In statistical tradition many methods have been proposed for variable selection. Commonly used methods are backward, forward and stepwise selection, where in every step regressors are added to the model or eliminated from the model according to a precisely defined testing schedule. Also information criteria like AIC and BIC are often used to assess the trade-off between model complexity and goodness-of-fit of the competing models. Recently also penalty approaches became popular, where coefficient estimation is accomplished by adding a penalty term to the likelihood function to shrink small effects to zero. Well known methods are the LASSO by Tibshirani (1996) and Ridge estimation.

Penalization approaches are very interesting from a Bayesian perspective since adding a penalty term to the likelihood corresponds to the assignment of an informative prior to the regression coefficients. A unifying overview of the relationship between Bayesian regression model analysis and frequentist penalization can be found in Fahrmeir et al. (2010).

In the Bayesian approach to variable selection prior distributions representing the subjective believes about parameters are assigned to the regressor coefficients. By applying Bayes' rule they are updated by the data and converted into the posterior distributions, on which all inference is based on. For the result of a Bayesian analysis the shape of the prior on the regression coefficients might be influential. If the prime interest of the analysis is coefficient estimation, the prior should be located over the a-priori guess value of the coefficient. If however the main interest is in distinguishing between large and small effects, a useful prior concentrates mass around zero and spreads the rest over the parameter space. Such a prior expresses the believe that there are coefficients close to zero on the one hand and larger coefficients on the other hand. These priors can be easily constructed as a mixture of two distributions, one with a "spike" at zero and the other with mass spread over a wide range of plausible values. This type of priors are called "spike" and "slab" priors. They are particularly useful for variable selection purposes, because they allow to classify the regression coefficients into two groups: one group consisting of large, important, influential regressors and the other group with small, negligible, probably noise effects. So Bayesian variable selection is performed by classifying regressor coefficients, rather than by shrinking small coefficient values to zero.

The aim of this master thesis is to analyze and compare five different spike-and-slab proposals with regard to variable selection. The first three, independence prior, Zellner's g-prior and fractional prior, are called "Dirac" spikes since the spike component consists of a discrete point mass on zero. The others, SSVS prior and NMIG prior, are mixtures of two continuous distributions with zero mean and different (a large and a small) variances. To address the two components of the prior, for each coefficient a latent indicator variable is introduced into the regression model. It indicates the classification of a coefficient to one of the two components: the indicator variable has the value 1, if the coefficient is assigned to the slab component of the prior, and 0 otherwise. To estimate the posterior probabilities of coefficients and indicator variables for all five priors a Gibbs sampling

scheme can be implemented. Variable selection is then based on the posterior distribution of the indicator variable which is estimated by the empirical frequency of the values 1 and 0, respectively. The higher the posterior mean of the indicator variable, the higher is evidence that the coefficient might be different from zero and therefore have an impact on the response variable.

The master thesis is organized as follows: In chapter 1 an introduction into the Bayesian analysis of normal regression models using conjugate priors is given. Chapter 2 describes Bayesian variable selection using stochastic search variables and implements a basic Gibbs sampling scheme to perform model selection and parameter estimation. In chapter 3 spike and slab priors for variable selection, independence prior, Zellner's g-prior and fractional prior, are introduced. In chapter 4 and 5 slab and spike priors with continuous spike component are studied: the stochastic search variable selection of George and McCulloch (1993), where the prior for a regression coefficient is a mixture of two normal distributions, and variable selection selection using normal mixtures of inverse gamma priors. Simulation studies in chapter 6 compare the presented approaches to variable selection with regard to accuracy of coefficient estimation, variable selection properties and efficiency for both independent and correlated regressors. Finally results and issues arisen during the simulations are discussed in chapter 7. The appendix summarizes derivations of formulas and R-codes.

Chapter 1

The Normal Linear Regression Model

In this section basic results of regression analysis are summarized and an introduction into Bayesian regression is given.

1.1 The standard normal linear regression model

In statistics regression analysis is a common tool to analyze the relationship between a dependent variable called the response and independent variables called covariates or regressors. It is assumed that the regressors have an effect on the response variable, and thus the researcher wants to quantify this influence. The simplest functional relationship between response variable and potentially influential variables is given by a linear regression model, in which the response can be described as a linear combination of the covariates with appropriate weights called regressor coefficients. More formally, given data as $(y_i, \mathbf{x_i})$, $i = 1, \ldots, N$, of N statistical units, the linear regression model is the following:

$$y_i = \mu + \alpha_1 x_{i1} + \alpha_2 x_{i2} + ... + \alpha_k x_{ik} + \epsilon_i \tag{1.1}$$

where y_i is the dependent variable and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})$ is the vector of potentially explanatory covariables. μ is the intercept of the model and $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_k)$ are the regression coefficients to be estimated. ϵ_i is the error term which should capture all other

unknown factors influencing the dependent variable y_i. In matrix notation model (1.1) is written as

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{y}$ is the N x 1 vector of the response variable, $\mathbf{X}$ is the N x k design matrix, $\boldsymbol{\beta}$ are the regression effects including the intercept, i.e. $\boldsymbol{\beta} = (\mu, \boldsymbol{\alpha})$, $\boldsymbol{\epsilon}$ is the n x 1 error vector and $\mathbf{X}_1$ denotes the design matrix $(\mathbf{1}, \mathbf{X})$. Without loss of generality we assume that the regressor columns $\mathbf{x_1}, ..., \mathbf{x_k}$ are centered.

Usually the unknown coefficient vector $\boldsymbol{\beta}$ is estimated by the ordinary-least-square method (OLS) where the sum of the squared distances between observed data y_i and the estimated data $\mathbf{x}_i\hat{\boldsymbol{\beta}}$ is minimized:

$$\sum_{i=1}^{N}(y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}})^2 \rightarrow min \tag{1.2}$$

Assuming that $\mathbf{X}_1'\mathbf{X}_1$ is of full rank, the solution is given by

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X_1}'\mathbf{X_1})^{-1}\mathbf{X_1}'\mathbf{y} \tag{1.3}$$

The estimator $\hat{\boldsymbol{\beta}}_{OLS}$ has many desirable properties, it is unbiased and the efficient estimator in the class of linear unbiased estimators, i.e. it has the so called BLUE-property (Gauß Markov theorem).

If the errors ϵ_i are assumed to be uncorrelated stochastic quantities following a Gaussian distribution with mean 0 and variance σ^2, the response $\mathbf{y}$ is also normally distributed:

$$\mathbf{y} \sim N(\mathbf{X}_1\boldsymbol{\beta}; \sigma^2\mathbf{I}) \tag{1.4}$$

In this case, the ML-estimator $\hat{\boldsymbol{\beta}}_{ML}$ of (1.4) coincides with the $\hat{\boldsymbol{\beta}}_{OLS}$ (1.3) and is normally distributed:

$$\hat{\boldsymbol{\beta}}_{ML} \sim N(\boldsymbol{\beta}; \sigma^2(\mathbf{X_1}'\mathbf{X_1})^{-1}) \tag{1.5}$$

Significance tests on the effects, e.g. whether an effect is significantly different from zero, are based on (1.5). Further information about maximum likelihood inference in normal regression models can be found in Fahrmeir et al. (1996).

1.2 The Bayesian normal linear regression model

In the Bayesian approach probability distributions are used to quantify uncertainty. Thus, in contrast to the frequentist approach, a joint stochastic model for response and parameters $(\boldsymbol{\beta}, \sigma^2)$ is specified. The distribution of the dependent variable $\mathbf{y}$ is specified *conditional* on the parameters $\boldsymbol{\beta}$ and σ^2:

$$\mathbf{y}|\boldsymbol{\beta}, \sigma^2 \sim N(\mathbf{X}_1\boldsymbol{\beta}; \sigma^2\mathbf{I}) \tag{1.6}$$

The analyst's certainty or uncertainty about the parameter *before* the data analysis is represented by the prior distribution for the parameters $(\boldsymbol{\beta}, \sigma^2)$. *After* observing the sample data $(y_i, \mathbf{x}_i)$, the prior distribution is updated by the empirical data applying Bayes' theorem,

$$p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}, \sigma^2)}{\int p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}, \sigma^2)d(\boldsymbol{\beta}, \sigma^2)} \tag{1.7}$$

yielding the so called posterior distribution $p(\boldsymbol{\beta}, \sigma^2|\mathbf{y})$ of the parameters $(\boldsymbol{\beta}, \sigma^2)$. Since the denominator of (1.7) acts as a normalizing constant and simply scales the posterior density, the posterior distribution is proportional to the product of likelihood function and prior. The posterior distribution usually represents less uncertainty than the prior distribution, since evidence of the data is taken into account. Bayesian inference is based only on the posterior distribution. Basic statistics like mean, mode, median, variance and quantiles are used to characterize the posterior distribution.

One of the most substantial aspects of a Bayesian analysis is the specification of appropriate prior distributions for the parameters. If the prior distribution for a parameter is chosen so that the posterior distribution follows the same distribution family as the prior, the prior distribution is said to be the conjugate prior of the likelihood. Conjugate priors ensure that the posterior distribution is a known distribution that can be easily derived.

The joint conjugate prior for $(\boldsymbol{\beta}, \sigma^2)$ has the structure

$$p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2) \tag{1.8}$$

where the conditional prior for the parameter vector $\boldsymbol{\beta}$ is the multivariate Gaussian distribution with mean $\mathbf{b}_0$ and covariance matrix $\sigma^2\mathbf{B}_0$:

$$p(\boldsymbol{\beta}|\sigma^2) = \mathcal{N}(\mathbf{b}_0, \mathbf{B}_0\sigma^2) \tag{1.9}$$

and the prior for σ^2 is the inverse gamma distribution with hyperparameter s_0 and S_0:

$$p(\sigma^2) = G^{-1}(s_0, S_0) \tag{1.10}$$

The posterior distribution is given by

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) &\propto p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2) \\ &\propto \frac{1}{(\sigma^2)^{N/2}} exp(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta})) \cdot \\ &\quad \frac{1}{(\sigma^2)^{(k+1)/2}} exp(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \mathbf{b}_0)'\mathbf{B}_0^{-1}(\boldsymbol{\beta} - \mathbf{b}_0)) \cdot \\ &\quad \frac{1}{(\sigma^2)^{(s_0+1)}} exp(-\frac{S_0}{\sigma^2}) \end{aligned} \tag{1.11}$$

This expression can be simplified, see Appendix A. It turns out that the joint posterior of $\boldsymbol{\beta}$ and σ^2 can be split into two factors being proportional to the product of a multivariate normal distribution and a inverse gamma distribution:

$$p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) \propto p_N(\boldsymbol{\beta}; b_N, \sigma^2\mathbf{B}_N)p_{IG}(s_N, S_N) \tag{1.12}$$

with parameters

$$\mathbf{B}_N = (\mathbf{X}_1'\mathbf{X}_1 + \mathbf{B}_0^{-1})^{-1} \tag{1.13}$$

$$\mathbf{b}_N = \mathbf{B}_N(\mathbf{X}_1'\mathbf{y} + \mathbf{B}_0^{-1}\mathbf{b}_0) \tag{1.14}$$

$$s_N = s_0 + N/2 \tag{1.15}$$

$$S_N = S_0 + \frac{1}{2}(\mathbf{yy} + \mathbf{b}_0\mathbf{B}_0^{-1}\mathbf{b}_0 - \mathbf{b}_N\mathbf{B}_N^{-1}\mathbf{b}_N) \tag{1.16}$$

If the error variance σ^2 is integrated out from the joint prior distribution of $\boldsymbol{\beta}$ and σ^2, the resulting unconditional prior for $\boldsymbol{\beta}$ is proportional to a multivariate Student distribution with $2s_0$ degrees of freedom, location parameter $\mathbf{b}_0$ and dispersion matrix $S_0/s_0\mathbf{B}_0$, see e.g. Fahrmeir et al. (2007):

$$p(\boldsymbol{\beta}) \propto t(2s_0, \mathbf{b}_0, S_0/s_0\mathbf{B}_0)$$

1.2.1 Specifying the coefficient prior parameters

Specifying the prior distribution of a single coefficient as

$$\beta_i|\sigma^2 \sim \mathcal{N}(b_{0i}, \sigma^2 B_{0ii})$$

especially the variance parameter B_{0ii} expresses the scientist's level of uncertainty about the parameter's location b_{0i}. If prior information is scarce, a large value for the variance parameter B_{0ii} should be chosen, so that the prior distribution is flat. In this case coefficient values far away from the mean b_{0i} are assigned a reasonable probability and the exact specification of b_{0i} is of minor significance. If at the extreme the variance becomes infinite every value on the parameter space has the same density, the analyst claims absolute ignorance about the coefficient's location. This type of prior is called "noninformative prior". On the other hand, if the analyst has considerable information about the coefficient β_i, he should choose a small value for the variance parameter B_{0ii}. If a high probability is assigned to values close to the mean b_{0i}, information in the data has to be very large to result in a posterior mean far away from b_{0i}.

Choice of the prior parameters $\mathbf{b}_0, \mathbf{B}_0$ of the prior distribution has an impact on the posterior mean

$$\mathbf{b}_N = \mathbf{B}_N(\mathbf{X}_1'\mathbf{y} + \mathbf{B}_0^{-1}\mathbf{b}_0)$$

and the posterior covariance matrix

$$\mathbf{B}_N = (\mathbf{X}_1'\mathbf{X}_1 + \mathbf{B}_0^{-1})^{-1}$$

If the prior information is vague, the prior covariance matrix $\mathbf{B}_0$ should be a matrix with large values representing the uncertainty about the location $\mathbf{b}_0$. The posterior covariance matrix $\mathbf{B}_N$ is then approximately $\sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}$ and the mean $\mathbf{b}_N \approx \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1\mathbf{y}$, which means that vague prior information leads to a posterior mean close to the OLS or ML estimator. If on the other hand the prior information about the coefficient vector is very strong, the prior covariance matrix $\mathbf{B}_0$ should contain small values. This yields the posterior covariance matrix $\boldsymbol{\beta}_N \approx \sigma^2\mathbf{B}_0$ and the mean $\mathbf{b}_N \approx \mathbf{b}_0$, and the Bayesian estimator is close to the prior mean.

1.2.2 Connection to regularisation

In contrast to the ML estimator the posterior mean estimator $\mathbf{b}_N$ is a biased estimator for $\boldsymbol{\beta}$. A criterion that allows to compare biased and unbiased estimators is the expected

quadratic loss (mean squared error, MSE):

$$\begin{aligned} MSE(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) &= E((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))') \\ &= Cov(\hat{\boldsymbol{\beta}}) + (E(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta})(E(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta})' \\ &= Cov(\hat{\boldsymbol{\beta}}) + Bias(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) Bias(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})' \end{aligned} \tag{1.17}$$

The ith of the diagonal elements of MSE is the mean squared error of the estimate for β_i:

$$E(\hat{\beta}_i - \beta_i)^2 = Var(\hat{\beta}_i) + (\hat{\beta}_i - \beta_i)^2 \tag{1.18}$$

As mean squared error is a function of both bias and variance, it seems reasonable to consider *biased* estimators which, however, considerably reduce variance.

A case where the variance of $\boldsymbol{\beta}$ can assume very large values is when columns of the data matrix are collinear. In this case the inverse of $\mathbf{X}_1'\mathbf{X}_1$ can have very large values, leading to high values of $\hat{\boldsymbol{\beta}}$ and $Var(\hat{\boldsymbol{\beta}})$. To regularise estimation, a penalty function penalizing large values of $\boldsymbol{\beta}$ can be included in the goal function. If the penalty is $\lambda\boldsymbol{\beta}'\boldsymbol{\beta}$, the so called "ridge estimator" results:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{ridge} &= argmin_\beta((\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{\beta}) \\ &= (\mathbf{X}_1'\mathbf{X}_1 + \lambda\mathbf{I})^{-1}\mathbf{X}_1\mathbf{y} \end{aligned} \tag{1.19}$$

See Appendix A for details. The ridge estimator is a biased estimator for $\boldsymbol{\beta}$, but its variance and also its MSE can be smaller than that of the OLS estimator, see Toutenburg (2003). The ridge estimator depends on a tuning parameter λ which controls the intensity of the penalty term. If λ=0, the common $\hat{\boldsymbol{\beta}}_{OLS}$ is obtained. With increasing λ, the influence of the penalty in the goal function grows: the fit on the data becomes weaker and the constraint on $\boldsymbol{\beta}$ dominates estimation.

Imposing a restriction on the parameter $\boldsymbol{\beta}$ in (1.19) also has a side effect: constraining the parameter vector $\boldsymbol{\beta}$ to lie around the origin causes a "shrinkage" of the parameter estimation. The size of shrinkage is controlled by λ: as λ goes to zero, $\boldsymbol{\beta}^{ridge}$ attains the $\boldsymbol{\beta}_{OLS}$, as λ increases $\boldsymbol{\beta}^{ridge}$ approaches 0. Shrinking is of interest for variable selection problems: ideally small true coefficients are shrunk to zero and the models obtained are

simpler including no regressors with small effect.

The ridge estimator can be interpreted as a Bayes estimator. If the prior for $\boldsymbol{\beta}$ is specified as

$$p(\boldsymbol{\beta}|\sigma^2) = \mathcal{N}(\mathbf{0}, c\mathbf{I}\sigma^2)$$

the posterior mean of $\boldsymbol{\beta}$ is given by

$$\begin{aligned} \mathbf{b}_N &= (\mathbf{X}_1'\mathbf{X}_1 + \mathbf{B}_0^{-1})^{-1}(\mathbf{X}_1'\mathbf{y} + \mathbf{B}_0^{-1}\mathbf{b}_0) \\ &= (\mathbf{X}_1'\mathbf{X}_1 + 1/c\mathbf{I})^{-1}\mathbf{X}_1'\mathbf{y} \end{aligned}$$

which is exactly the ridge estimator from (1.19) with $\lambda = 1/c$. This means that choosing a prior for $\boldsymbol{\beta}$ causes regularization and shrinkage of the estimation of $\boldsymbol{\beta}$. The tuning parameter $c = 1/\lambda$ controls the size of coefficients and the amount of regularisation. However, the variance of this estimator is $\mathbf{B}_N = (\mathbf{X}_1'\mathbf{X}_1 + 1/c\mathbf{I})^{-1}$ in a Bayes interpretation and $(\mathbf{X}_1'\mathbf{X}_1 + 1/c\mathbf{I})^{-1}\mathbf{X}_1'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1 + 1/c\mathbf{I})\sigma^2$ in a classical interpretation. For more details on the relationship between regularisation and Bayesian analysis see Fahrmeir et al. (2010).

1.2.3 Implementation of a Gibbs sampling scheme to simulate the posterior distribution

Although under a conjugate prior for $(\boldsymbol{\beta}, \sigma^2)$ the posterior $p(\boldsymbol{\beta}, \sigma^2|\mathbf{y})$ is available in closed form, we will describe a Gibbs sampling scheme to sample from the posterior distribution. This Gibbs sampling scheme is the basic algorithm which will be extented later to allow variable selection.

A Gibbs sampler is an MCMC (Markov Chain Monte Carlo) method to generate a sequence of samples from the joint posterior distribution by breaking it down into more manageable univariate or multivariate distributions. To implement a Gibbs sampler for the posterior distribution $p(\boldsymbol{\beta}, \sigma^2|\mathbf{y})$ the parameter vector is split into two blocks $\boldsymbol{\beta}$ and σ^2. Then random values are drawn from the conditional distributions $p(\boldsymbol{\beta}|\sigma^2, \mathbf{y})$ and $p(\sigma^2|\boldsymbol{\beta}, \mathbf{y})$ alternately, where in each step a draw from the conditional posterior conditioning on the current value of the other parameter is produced. Due to Markov Chain Theory, after a burnin period the sampled values can be regarded as realisations from the

marginal distributions $p(\boldsymbol{\beta}|\mathbf{y})$ and $p(\sigma^2|\mathbf{y})$. The conditional distribution $p(\boldsymbol{\beta}|\sigma^2, \mathbf{y})$ is the normal distribution $\mathcal{N}(\mathbf{b}_N, \mathbf{B}_N\sigma^2)$. For the conditional distribution of σ^2 given $\boldsymbol{\beta}$ and $\mathbf{y}$ we obtain (see appendix A for details)

$$p(\sigma^2|\boldsymbol{\beta}, \mathbf{y}) = \mathcal{G}^{-1}(s_N^*, S_N^*) \tag{1.20}$$

with

$$\begin{aligned} s_N^\star &= s_0 + N/2 + (k+1)/2 & (1.21)\\ S_N^\star &= S_0 + \frac{1}{2}(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}) + \frac{1}{2}(\boldsymbol{\beta} - \mathbf{b}_0)'\mathbf{B}_0^{-1}(\boldsymbol{\beta} - \mathbf{b}_0) & (1.22) \end{aligned}$$

Sampling from the posterior is feasible by a two-block Gibbs sampler. After assigning starting values to the parameters, the following steps are repeated:

(1) sample σ^2 from $\mathcal{G}^{-1}(s_N, S_N)$, parameters s_N, S_N are given in (1.21) and (1.22)

(2) sample $\boldsymbol{\beta}$ from $\mathcal{N}(\mathbf{b}_N, \mathbf{B}_N\sigma^2)$, parameters b_N, B_N are given in (1.13) and (1.14)

We could sample from the posterior alternatively using a one-block Gibbs sampler, where σ^2 and $\boldsymbol{\beta}$ are sampled in one step:

(1a) sample σ^2 from $G^{-1}(s_N, S_N)$ with parameters s_N, S_N given in (1.15) and (1.16)

(1b) sample $\boldsymbol{\beta}|\sigma^2$ from $N(\mathbf{b}_N, \mathbf{B}_N\sigma^2)$ with parameters $\mathbf{b}_N, \mathbf{B}_0$ given in (1.13) and (1.14).

Chapter 2

Bayesian Variable Selection

The goal of variable selection in regression models is to decide which variables should be included in the final model. In practical applications often a large number of regressors could be used to model the response variable. However, usually only a small subset of these potential regressors actually have an influence on the response variable, whereas the effect of most covariates is very small or even zero. Variable selection methods try to identify regressors with nonzero and zero effects. This is a challenging task, see e.g. O'Hara and Sillanpää (2009) who summarize the variable selection problem as follows: "In any real data set, it is unlikely, that the true regressor coefficients are either zero or large; the sizes are more likely to be tapered towards zero. Hence, the problem is not one of finding the zero coefficients, but of finding those that are small enough to be insignificant, and shrinking them towards zero". The stochastic search variable approach to differentiate between regressors with (almost) zero and nonzero effects proceeds by introducing auxiliary indicator variables: for each regressor a binary variable is defined indicating whether the effect of that regressor is zero or nonzero. A Gibbs sampling scheme can be implemented to generate samples from the joint posterior distribution of indicator variables and regression coefficients. The posterior probability that an indicator variable takes 1 can be interpreted as the posterior inclusion probability of the corresponding regressor, i.e. probability that this regressor should be included in the final model.

2.1 Bayesian Model Selection

Variable selection is a special case of model selection. Model selection means to choose the "best" model for the data $\mathbf{y}$ from a set of candidate models $M_1, M_2, M_3....$ In the Bayesian approach a prior probability $p(M_l)$ is assigned to each candidate model M_l and the posterior probability of the model M_l is obtained by applying Bayes' rule:

$$p(M_l|y) \propto p(y|M_l)p(M_l)$$

The probability $p(\mathbf{y}|M_l)$ is called the marginal likelihood of the model M_l and it is the key quantity in the equation above. It is the density of the sample given only the model structure without any information on specific parameters. It is computed by integrating out the parameters of the model:

$$p(\mathbf{y}|M_l) = \int_{\theta_l} \underbrace{p(\mathbf{y}|\theta_l, M_l)}_{likelihood} \underbrace{p(\theta_l|M_l)}_{prior} d\theta_l$$

For every candidate model M_l, the posterior model probability $p(M_l|\mathbf{y})$ has to be computed and the model with the highest posterior probability is the favorite model. For the normal linear regression model under conjugate priors this integration can be solved analytically, and the marginal likelihood is given as

$$\begin{aligned} p(\mathbf{y}|M) &= \iint p(\mathbf{y}|\boldsymbol{\delta}, \boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{\delta})d\boldsymbol{\beta}d\sigma^2 \\ &= \frac{1}{(2\pi)^{N/2}} \frac{|\mathbf{B}_N|^{1/2}}{|\mathbf{B}_0|^{1/2}} \frac{\Gamma(s_N)S_0^{s_0}}{\Gamma(s_0)(S_N)^{s_N}} \end{aligned} \tag{2.1}$$

where the parameters s_N, S_N and B_N are given in (1.13) to (1.16). If selection among competing regression models is considered choosing a model M_l means to choose a set of regressor variables, usually a subset of $\mathbf{x}_1, \ldots, \mathbf{x}_k$, which corresponds to M_l. However, in a regression model with k regressors considering all models with all possible subsets of regressors is a computational challenge, even if k is moderate, e.g. for k=30 $1.1 \cdot 10^{12}$ marginal likelihoods would be required. Therefore, a different approach is to perform a stochastic search for models with high posterior probability using MCMC methods. This approach is presented in the following section.

2.2 Model selection via stochastic search variables

To perform a stochastic search for models with high posterior probability a new indicator variable δ_j for each regressor coefficient α_j is defined, where $\delta_j = 0$ or $\delta_j = 1$ represents inclusion or exclusion of the regressor in the model:

$$\delta_j = \begin{cases} 0 & \text{if } \alpha_j = 0 \\ 1 & \text{otherwise} \end{cases}$$

Regression model (1.1) can be rewritten as

$$y_i = \mu + \delta_1\alpha_1 x_{i1} + \delta_2\alpha_2 x_{i2} + ... + \delta_k\alpha_k x_{ik} + \epsilon_i \tag{2.2}$$

The vector $\boldsymbol{\delta} = (\delta_1, ..., \delta_k)$ determines the model, i.e. the vector contains the information which elements of $\boldsymbol{\alpha}$ are included in the model or restricted to be zero. Also each of the 2^k candidate models is represented by a realisation of $\boldsymbol{\delta}$. Model selection means to choose a value of $\boldsymbol{\delta}$. Once a value $\boldsymbol{\delta}$ is chosen, the following reduced normal regression model is obtained (in matrix notation, with $\boldsymbol{\beta} = (\mu, \boldsymbol{\alpha})$):

$$\mathbf{y} = \mathbf{X}_1^{\delta}\boldsymbol{\beta}^{\delta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathcal{I}),$$

where $\boldsymbol{\beta}^{\boldsymbol{\delta}}$ contains only the nonzero elements of $\boldsymbol{\beta}$ and the design matrix $\mathbf{X}_1^{\delta}$ only the columns of $\mathbf{X}_1$ corresponding to nonzero effects. The intercept does not have an indicator variable as μ is included in the model in any case.

2.2.1 Prior specification

In representation (2.2) model parameters are the model indicator $\boldsymbol{\delta}$, the regressor effects $\boldsymbol{\beta}$ and the error variance σ^2. In the Bayesian framework the goal is to derive the posterior density of all the parameters $\boldsymbol{\delta}$, $\boldsymbol{\beta}$ and σ^2:

$$p(\boldsymbol{\delta}, \boldsymbol{\beta}, \sigma^2|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\delta}, \boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\delta}, \boldsymbol{\beta}, \sigma^2)$$

To specify a prior distribution, we assume that it has the structure

$$p(\boldsymbol{\delta}, \sigma^2, \boldsymbol{\beta}) = p(\sigma^2, \boldsymbol{\beta}|\boldsymbol{\delta}) \prod_j p(\delta_j)$$

As δ_j is a binary variable, a straightforward choice of a prior for δ_j is

$$p(\delta_j = 1) = \pi, \quad j = 1, \ldots, k$$

where π is a fixed inclusion probability between 0 and 1. For σ^2 and those effects of $\boldsymbol{\beta}$ which are not restricted to be zero, $\boldsymbol{\beta}^{\delta}$, we use conjugate priors:

$$\begin{aligned} \sigma^2 &\sim \mathcal{G}^{-1}(s_0, S_0) \\ \boldsymbol{\beta}^{\boldsymbol{\delta}} | \sigma^2, \boldsymbol{\delta} &\sim N(\mathbf{b_0}^{\delta}, \mathbf{B_0}^{\delta} \sigma^2) \end{aligned}$$

2.2.2 Gibbs sampling

A naive approach for a Gibbs sampler to draw from the joint posterior $p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}|\mathbf{y})$ would be to add a further step to the algorithm described in section 1.2.3:

(1) sample $\boldsymbol{\delta}$ from $p(\boldsymbol{\delta}|\boldsymbol{\beta}, \sigma^2, \mathbf{y})$

(2) sample $\boldsymbol{\beta}, \sigma^2$ from $p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{\delta}, \mathbf{y})$

However, this scheme does not define an irreducible Markov chain: whenever $\delta_j = 0$ also $\alpha_j = 0$ and hence the chain has absorbing states. This problem can be avoided by sampling $\boldsymbol{\delta}$ from the marginal posterior distribution $p(\boldsymbol{\delta}|\mathbf{y})$. Formally, the marginal posterior distribution is given as

$$p(\boldsymbol{\delta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\delta}) p(\boldsymbol{\delta})$$

where $p(\mathbf{y}|\boldsymbol{\delta})$ is the marginal likelihood from the likelihood of the regression model with regressors $\mathbf{X}_1^{\delta}$, where the effects $\boldsymbol{\beta}^{\delta}$ and error variance σ^2 are integrated out:

$$p(\mathbf{y}|\boldsymbol{\delta}) = \iint p(\mathbf{y}|\boldsymbol{\delta}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{\delta}) d\boldsymbol{\beta} d\sigma^2$$

This integral is available in a closed form and given as

$$p(\mathbf{y}|\boldsymbol{\delta}) = \frac{1}{(2\pi)^{N/2}} \frac{|\mathbf{B}_N^{\delta}|^{1/2}}{|\mathbf{B}_0^{\delta}|^{1/2}} \frac{\Gamma(s_N) S_0^{s_0}}{\Gamma(s_0)(S_N^{\delta})^{s_N}}$$

with posterior moments

$$B_N^{\delta} = \left((\mathbf{X}_1^{\delta})' \mathbf{X}_1^{\delta} + (\mathbf{B}_0^{\delta})^{-1} \right)^{-1} \tag{2.3}$$

$$b_N^{\delta} = \mathbf{B}_N^{\delta} \left((\mathbf{X}_1^{\delta})' \mathbf{y} + (\mathbf{B}_0^{\delta})^{-1} \mathbf{b}_0^{\delta} \right) \tag{2.4}$$

$$s_N = s_0 + N/2 \tag{2.5}$$

$$S_N^{\delta} = S_0 + \frac{1}{2} \left(\mathbf{y}'\mathbf{y} + (\mathbf{b}_0^{\delta})'(\mathbf{B}_0^{\delta})^{-1} \mathbf{b}_0^{\delta} - (\mathbf{b}_N)'(\mathbf{B}_N^{\delta})^{-1} \mathbf{b}_N^{\delta} \right) \tag{2.6}$$

To sample from the posterior $p(\boldsymbol{\delta}, \boldsymbol{\beta}, \sigma^2|\mathbf{y})$ a one block Gibbs sampler can be implemented, which involves the following steps:

(1) Sample $\boldsymbol{\delta}$ marginally from $p(\boldsymbol{\delta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\delta})p(\boldsymbol{\delta})$.

(2) Sample $\sigma^2|\boldsymbol{\delta}$ from $p(\sigma^2|\mathbf{y}, \boldsymbol{\delta})$, which is the gamma inverse distribution $\mathcal{G}^{-1}(s_N^\delta, S_N^\delta)$.

(3) Sample $\boldsymbol{\beta}^\delta|\sigma^2$ in one block from $p(\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\delta}, \sigma^2)$. This conditional posterior is the normal distribution $N(\mathbf{b_N}^\delta, \mathbf{B}_N^\delta\sigma^2)$.

In this scheme model all parameters are sampled jointly. However, model selection could be performed without parameter estimation by iterating only step 1.

If the number of regressors k is large, the sampling scheme above is computationally too expensive because calculation of the marginal likelihood of all 2^k models is required in each iteration. A faster sampling scheme can be implemented, which updates each component δ_j conditional on the actual value $\delta_{\backslash j}$ of the other elements of $\boldsymbol{\delta}$. Thus for each component only the computation of two marginal likelihoods (for $\delta_j = 0$ and $\delta_j = 1$) and therefore in one iteration step only the evaluation of 2k marginal likelihoods is required. This leads to the following sampling scheme:

(1) sample each element δ_j of the indicator vector $\boldsymbol{\delta}$ seperately conditional $\delta_{\backslash j}$. This step is carried out by first choosing a random permutation of the column numbers 1,...,k and updating the elements of $\boldsymbol{\delta}$ in this order.

(2) sample the error variance σ^2 conditional on $\boldsymbol{\delta}$ from the $\mathcal{G}^{-1}(s_N^\delta, S_N^\delta)$ distribution.

(3) sample the non-zero elements $\boldsymbol{\beta}^\delta$ in one block conditional on σ^2 from $N(\mathbf{b_N}^\delta, \mathbf{B}_N^\delta\sigma^2)$.

2.2.3 Improper priors for intercept and error term

We have assumed that the columns of the design matrix $\mathbf{X}$ are centered and therefore orthogonal to the $\mathbf{1}$ vector. In this case the intercept is identical for all models. Whereas model selection cannot be performed using improper priors on the effects subject to selection, Kass and Raftery (1995) showed that model comparison can be performed properly when improper priors are put on parameters common to all models. Therefore, a flat prior for the intercept μ and a improper prior for the error term σ^2 can be assumed:

$$p(\mu) \propto 1 \tag{2.7}$$

$$p(\sigma^2) = \frac{1}{\sigma^2} \sim \mathcal{G}^{-1}(0,0) \tag{2.8}$$

An advantage of the improper prior on the intercept is that in case of centered covariates the mean μ can be sampled independently of the other regressor effects. If the flat prior on the intercept is combined with a proper prior $N({\mathbf{a_0}}^{\delta}, \mathbf{A}_0^{\delta}\sigma^2)$ on the other unrestricted elements of $\boldsymbol{\alpha}^{\delta}$, a so called partially proper prior is obtained. In chapter 3 different proposals how to select the prior parameters $\mathbf{a}_0, \mathbf{A}_0$ are described.

2.3 Posterior analysis from the MCMC output

After having generated samples from the posterior distributions model parameters are estimated by the sample means of their posterior probabilities respectively:

$$\hat{\mu} = \frac{1}{M}\sum_{m=1}^{M}(\mu)^{(m)} \tag{2.9}$$

$$\hat{\boldsymbol{\beta}} = \frac{1}{M}\sum_{m=1}^{M}\boldsymbol{\beta}^{(m)} \tag{2.10}$$

$$\hat{\sigma^2} = \frac{1}{M}\sum_{m=1}^{M}(\sigma^2)^{(m)} \tag{2.11}$$

$$\hat{\delta_j} = \frac{1}{M}\sum_{m=1}^{M}\delta_j^{(m)}, \quad j = 1 \ldots k \tag{2.12}$$

It should be remarked that the estimates of $\mu, \boldsymbol{\alpha}$ and σ^2 are averages over different regression models as during the MCMC procedure they are drawn conditional on different model indicator variables $\boldsymbol{\delta}$. This is known as "Bayesian model averaging".
The posterior probability $p(\delta_j = 1|\mathbf{y})$ for the regressor x_j to be included in the model can be estimated by the mean $\hat{\delta_j}$ or alternatively by the mean of $p(\delta_j^{(m)} = 1|\mathbf{y})$. To select the final model one of the following choices is possible:

- Selection of the median probability model: those regressors are included in the final model which have a posterior probability greater than 0.5.

- Selection of the highest probability model: the highest probability model is indicated by that $\boldsymbol{\delta}$ which has occured most often during the MCMC iterations.

It has been shown by Barbieri and Berger (2004) that the median probability model outperforms the highest probability model in regard to predictive optimality.

Chapter 3

Variable selection with Dirac Spikes and Slab Priors

If indicator variables for inclusion of regression coefficients are introduced as in chapter 2, the resulting prior on the regression coefficients is a mixture of a point mass (Dirac mass) at zero and an absolutely continuous component. In this chapter different normal distributions for the absolutely continuous component are considered and the MCMC scheme for all model parameters is described.

3.1 Spike and slab priors with a Dirac spike and a normal slab

By introducing the indicator variable δ_j in the normal linear regression model, the resulting prior for the the regression coefficients is an example of a spike and slab prior. A spike and slab prior is a mixture of two distributions, a "spike" and a "slab" distribution respectively, where the "spike" is a distribution with its mass concentrated around zero and the "slab" a flat distribution spread over the parameter space. Mitchell and Beauchamp (1988) introduced this type of prior to facilitate variable selection by constraining regressor coefficients to be zero or not.
More formally the prior can be written as

$$p(\alpha_j|\delta_j) = \delta_j p_{slab}(\alpha_j) + (1-\delta_j)p_{spike}(\alpha_j)$$

Here we consider a spike and a point mass at zero (Dirac spike) and normal slabs:

$$\begin{aligned} p_{slab}(\boldsymbol{\alpha}) &= N(\mathbf{a}_0, \mathbf{A}_0\sigma^2) \\ p_{spike}(\alpha_j) &= I_{\{0\}}(\alpha_j) \end{aligned}$$

Depending on the value of δ_j, the coefficient is assumed to belong to either the spike distribution or the slab distribution. Spike and slab priors allow classification of the regressor coefficients as "zero" and "non-zero" effects by analyzing the posterior inclusion probabilities estimated by mean of the indicator variables. Also the priors presented in chapter 4 and 5 are spike and slab priors, where the Dirac spike will be replaced by an absolutely continuous density with mean zero and small variance.

We consider specifications with different prior parameters $\mathbf{a}_0$ and $\mathbf{A}_0$ of the $N(\mathbf{a_0}^\delta, \mathbf{A}_0^\delta\sigma^2)$-slab:

- independence prior: $\mathbf{a}_0^\delta = 0, \mathbf{A}_0^\delta = c\mathbf{I}$
- g-prior: $\mathbf{a}_0^\delta = 0, \mathbf{A}_0^\delta = g((\mathbf{X}^\delta)'(\mathbf{X}^\delta))^{-1}$
- fractional prior: $\mathbf{a}_0^\delta = ((\mathbf{X}^\delta)'(\mathbf{X}^\delta))^{-1}(\mathbf{X}^\delta)'\mathbf{y}_c, \mathbf{A}_0^\delta = \frac{1}{b}((\mathbf{X}^\delta)'(\mathbf{X}^\delta))^{-1}$

where $c, g, 1/b$ are constants and $\mathbf{y}_c$ is the centered response. All priors considered here are still conjugate priors and allow straightforward computation of the posterior model probability.

Hierarchical prior for the inclusion probability

To achieve more flexibility than using a fixed prior inclusion probability $p(\delta_j = 1) = \pi$, we use a hierarchical prior where the inclusion probability follows a Beta distribution with parameters c_0 and d_0:

$$\pi \sim B(c_0, d_0)$$

Then the induced prior for the indicator vector $\boldsymbol{\delta}$ equals the following Beta distribution:

$$p(\boldsymbol{\delta}) = B(p_\delta + c_0, k - p_\delta + d_0)$$

where p_δ is the number of non zero elements in $\boldsymbol{\delta}$ and k is the number of covariates. If c_0 and d_0 equals 1, the prior is uninformative, but also an informative prior could be used to model prior information. Ley and Steel (2007) have shown that the hierarchical prior outperforms the prior with fixed inclusion probabilities.

3.1.1 Independence slab

In the simplest case of a non informative prior for the coefficients each regressor effect is assumed to follow the same distribution and to be independent of the other regressors:

$$p(\boldsymbol{\alpha}^{\delta}|\sigma^2) = N(\mathbf{a}_0^{\delta}, \mathbf{A}_0^{\delta}\sigma^2) = N(\mathbf{0}, c\mathbf{I}\sigma^2) \tag{3.1}$$

where c is an appropriate constant.

Since

$$p(\alpha_j) = p(\alpha_j|\delta_j = 1)p(\delta_j = 1) + p(\alpha_j|\delta_j = 0)p(\delta_j = 0) \tag{3.2}$$

the resulting prior for a regression coefficient is mixture of a (flat) normal distribution (for $\delta_j = 1$) and a (spike) point mass at point zero (for $\delta_j = 0$). In figure (3.1) the contour plot of the independence prior for 2 regressors is plotted, the blue point at 0 marks the discrete point mass.

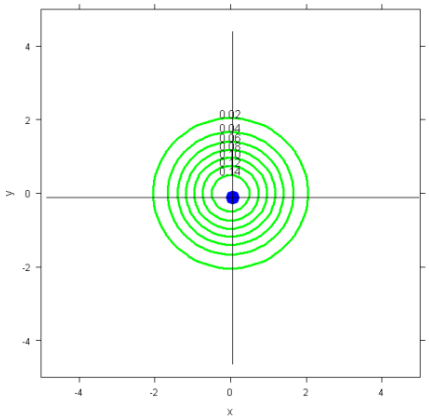


Figure 3.1: Contour plot of independence prior for 2 regressors, c=10, $\sigma^2 = 1$, and $\boldsymbol{\delta} = (1, 1)$ (green) and $\boldsymbol{\delta} = (0, 0)$ (blue).

3.1.2 Zellner's g-prior slab

The g-prior introduced by Zellner (1986) is the most popular prior slab used for model estimation and selection in regression models, see e.g. in Liang et al. (2008) and Maruyama and George (2008). Like the independence prior the g-prior assumes that the effects are a priori centered at zero, but the covariance matrix $\mathbf{A}_0$ is a scalar multiple of the Fisher information matrix, thus taking the dependence structure of the regressors into account:

$$p(\boldsymbol{\alpha}^{\delta}|\sigma^2) = N\left(\mathbf{0}, g\left((\mathbf{X}^{\delta})'(\mathbf{X}^{\delta})\right)^{-1}\sigma^2\right)$$

Commonly g is chosen as n or k^2. However, there are various options in specifying the constant g, see Liang et al. (2008), section 2.4, for an overview. Figure (3.2) shows the g-prior for 2 correlated regressors with ρ=0.8, g=400 and 40 observations.

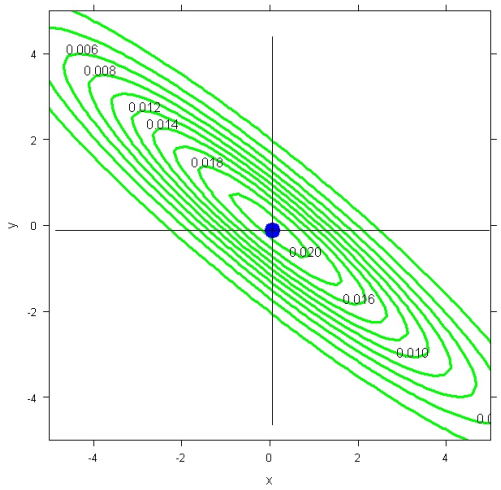


Figure 3.2: Contour plot of Zellner's g-prior for 2 correlated regressors with ρ=0.8, g=400, N=40 observations and $\boldsymbol{\delta} = (1,1)$ (green) and $\boldsymbol{\delta} = (0,0)$ (blue).

Popularity of the g-prior is at least partly due to the fact that it leads to simplifications, e.g. in evaluating the marginal likelihood, see Liang et al. (2008). The marginal likelihood can be expressed as a function of the coefficient of determination in the regression model, evaluation of determinants of prior and posterior covariances for the regressor effects is not necessary. Marginal likelihood and posterior moments are given in section 3.2.1.

Recently shrinkage properties of the g-prior were discussed, e.g. by George and Maruyama (2010), who critized that "the variance is put in wrong place" and proposed a modified version of the g-prior. Following Brown et al. (2002) who compare the shrinkage properties of g-prior and independence prior, we consider the singular value decompensation of $\mathbf{X}$, $\mathbf{X} = \mathbf{T}\boldsymbol{\Lambda}^{1/2}\mathbf{V}'$. $\mathbf{T}$ and $\mathbf{V}$ are orthonormal matrices and $\mathbf{X}'\mathbf{X}$ is assumed to have full rank.

The normal regression model

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

can be transformed to

$$\mathbf{u} = \mathbf{T}'\mathbf{X}\mathbf{V}\mathbf{V}'\boldsymbol{\beta} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

where $\mathbf{u} = \mathbf{T}'\mathbf{y}$ and $\boldsymbol{\theta} = \mathbf{V}'\boldsymbol{\beta}$.
The independence prior $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, c\sigma^2\mathbf{I})$ induces $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, c\sigma^2\mathbf{I})$. As

$$\mathbf{V}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{V} = \mathbf{V}'(\mathbf{V}\boldsymbol{\Lambda}^{1/2}\mathbf{T}'\mathbf{T}\boldsymbol{\Lambda}^{1/2}\mathbf{V}')^{-1} = \boldsymbol{\Lambda}^{-1}$$

the g-prior $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, g\sigma^2(\mathbf{X}'\mathbf{X})^{-1})$ induces $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, g\sigma^2\boldsymbol{\Lambda}^{-1})$. Under a normal prior $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_0\sigma^2)$, the posterior mean of $\boldsymbol{\theta}$ is given as (see (1.13) and (1.14)):

$$\begin{aligned} E(\boldsymbol{\theta}|\mathbf{u}) &= (\boldsymbol{\Lambda}^{1/2'}\boldsymbol{\Lambda}^{1/2} + \mathbf{D}_0^{-1})^{-1}\boldsymbol{\Lambda}^{1/2'}\mathbf{u} \\ &= (\boldsymbol{\Lambda} + \mathbf{D}_0^{-1})^{-1}\boldsymbol{\Lambda}^{1/2}\mathbf{u} \\ &= (\mathbf{I} + \boldsymbol{\Lambda}^{-1}\mathbf{D}_0^{-1})^{-1}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}^{1/2}\mathbf{u} \\ &= (\mathbf{I} + \boldsymbol{\Lambda}^{-1}\mathbf{D}_0^{-1})^{-1}\hat{\boldsymbol{\theta}}_{OLS} \end{aligned}$$

since $\hat{\boldsymbol{\theta}}_{OLS} = (\boldsymbol{\Lambda}^{1/2'}\boldsymbol{\Lambda}^{1/2})^{-1}\boldsymbol{\Lambda}^{1/2'}\mathbf{u}$. Therefore under the independence prior the mean of θ_i is given as

$$E(\theta_i|\mathbf{u}) = \frac{\lambda_i}{\lambda_i + 1/c}\hat{\theta}_i \tag{3.3}$$

whereas under the g-prior the mean of θ_i is given as

$$E(\theta_i|\mathbf{u}) = \frac{g}{g+1}\hat{\theta}_i \tag{3.4}$$

Comparing the factors of proportionality in (3.3) and (3.4), it can be seen that under the independence prior the amount of shrinkage depends on the eigenvalue λ_i: increasing the eigenvalue the shrinkage factor increases to 1, meaning that shrinkage will disappear and the posterior mean will approximate the ML estimate. On the other hand, in direction of small eigenvalues the ML estimate are shrunk towards zero. On the contrary, the posterior mean under the g-prior (3.4) is shrunk equally in the directions of all eigenvalues, which is an undesirable effect. In figure (3.3) the posterior mean under independence prior and g-prior is plotted for different amounts of eigenvalues.

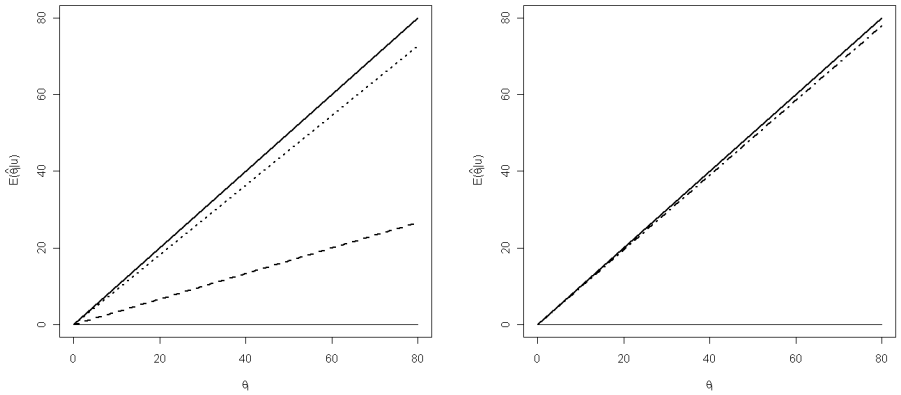


Figure 3.3: Left: Independence prior: posterior mean of coefficient estimation in direction of different eigenvalues, $\lambda_i = 0.5$ (dashed line) and $\lambda_i = 10$ (dotted line). Parameter c=1
Right: G-prior: posterior mean of coefficient estimation shrinks the ML-estimation equally in all direction of eigenvalues. Parameter g=40.

3.1.3 Fractional prior slab

The basic idea of the fractional prior introduced by O'Hagan (1995) is to use a fraction b of the likelihood of the centered data $\mathbf{y}_c$ to construct a proper prior under the improper prior $p(\boldsymbol{\alpha}^\delta) \propto 1$. So the following proper prior on the unrestricted elements $\boldsymbol{\alpha}_j$ is obtained:

$$p(\boldsymbol{\alpha}^\delta|\sigma^2) = N\left(\left((\mathbf{X}^\delta)'(\mathbf{X}^\delta)\right)^{-1}(\mathbf{X}^\delta)'\mathbf{y}_c, 1/b\left((\mathbf{X}^\delta)'(\mathbf{X}^\delta)\right)^{-1}\sigma^2\right)$$

The fractional prior is centered at the value of the ML (OLS) estimate, with the variance-covariance matrix multiplied by the factor $1/b$. For $0 < b < 1$, usually $b \ll 1$, the prior is considerably more spread than the sampling distribution. Since information used for constructing the prior should not reappear in the likelihood, the fractional prior is combined with the remaining part of the likelihood and yields the posterior (Frühwirth-Schnatter and Tüchler (2008)):

$$p(\boldsymbol{\alpha}^\delta|\sigma^2, \mathbf{y}) = N(\mathbf{a}_N^\delta, \mathbf{A}_N^\delta\sigma^2)$$

with parameters

$$\mathbf{A}_N^\delta = \left((\mathbf{X}^\delta)'(\mathbf{X}^\delta)\right)^{-1} \tag{3.5}$$

$$\mathbf{a}_N^\delta = \mathbf{A}_N(\mathbf{X}^\delta)'\mathbf{y}_c \tag{3.6}$$

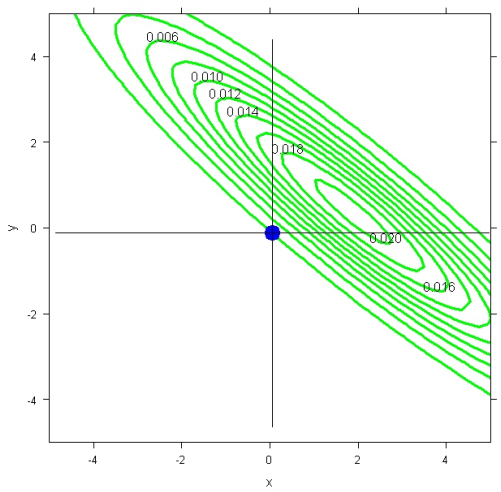


Figure 3.4: Contour plot of fractional prior for 2 correlated regressors with ρ=0.8, b=1/400, N=40 observations, and $\boldsymbol{\delta} = (1, 1)$ (green) and $\boldsymbol{\delta} = (0, 0)$ (blue).

In figure (3.4) the contour fractional prior slab is plotted for two correlated regressors with ρ=0.8, f=1/400, N=40 observations.

3.2 MCMC scheme for Dirac spikes

Under the improper prior for intercept and error variance defined in (2.8) and (2.7), the following MCMC scheme allows to sample the model parameters $(\boldsymbol{\delta}, \mu, \boldsymbol{\alpha}, \sigma^2)$:

(1) sample each element δ_j of $\boldsymbol{\delta}$ separately from $p(\delta_j|\boldsymbol{\delta}_{\setminus j}, \mathbf{y}) \propto p(\mathbf{y}|\delta_j, \delta_{\setminus j})p(\delta_j, \boldsymbol{\delta}_{\setminus j})$, where $\delta_{\setminus j}$ denotes the vector $\boldsymbol{\delta}$ without element δ_j.

(2) sample $\sigma^2|\boldsymbol{\delta}$ from $\mathcal{G}^{-1}(s_N^{\delta}, S_N^{\delta})$

(3) sample the intercept μ from $N(\bar{y}, \sigma^2/N)$

(4) sample the nonzero elements $\boldsymbol{\alpha}^{\delta}|\sigma^2$ in one block from $N(\mathbf{a_N}^{\delta}, \mathbf{A}_N^{\delta}\sigma^2)$

The marginal likelihood of the data conditioning only on the indicator variables is given as

$$p(\mathbf{y}|\boldsymbol{\delta}) = \frac{N^{-1/2}}{(2\pi)^{(N-1)/2}} \frac{|\mathbf{A}_N^\delta|^{1/2}}{|\mathbf{A}_0^\delta|^{1/2}} \frac{\Gamma(s_N) S_0^{s_0}}{\Gamma(s_0)(S_N)^{s_N}}$$

with posterior moments

$$\begin{aligned}
\mathbf{A}_N^\delta &= \left((\mathbf{X}^\delta)'\mathbf{X}^\delta + (\mathbf{A}_0^\delta)^{-1}\right)^{-1} && (3.7)\\
\mathbf{a}_N^\delta &= \mathbf{A}_N^\delta \left((\mathbf{X}^\delta)'\mathbf{y} + (\mathbf{A}_0^\delta)^{-1}\mathbf{a}_0^\delta\right) && (3.8)\\
s_N &= s_0 + (N-1)/2 && (3.9)\\
S_N &= S_0 + \frac{1}{2}\left(\mathbf{y_c'y_c} + (\mathbf{a}_0^\delta)'(\mathbf{A}_0^\delta)^{-1}\mathbf{a}_0^\delta - (\mathbf{a}_N)'(\mathbf{A}_N^\delta)^{-1}\mathbf{a}_N^\delta\right) && (3.10)
\end{aligned}$$

3.2.1 Marginal likelihood and posterior moments for a g-prior slab

Under the g-prior the marginal likelihood is given as

$$p(\mathbf{y}|\boldsymbol{\delta}) = \frac{(1+g)^{-q_\delta/2}}{\sqrt{N}(\pi)^{(N-1)/2}} \frac{\Gamma((N-1)/2)}{S(\mathbf{X}^\delta)^{(N-1)/2}} \tag{3.11}$$

where

$$S(\mathbf{X}^\delta) = \frac{\|\mathbf{y}_c\|^2}{1+g}(1 + g(1 - R(\mathbf{X}^\delta)^2)) \tag{3.12}$$

where q_δ is the number of nonzero elements in $\boldsymbol{\delta}$ and $R(\mathbf{X}^\delta)$ is the coefficient of determination $\mathbf{y}_c'\mathbf{X}^\delta(\mathbf{X}^{\delta'}\mathbf{X}^\delta)^{-1}\mathbf{X}^{\delta'}\mathbf{y}_c$.

The posterior moments are given as:

$$\begin{aligned}
\mathbf{A}_N^\delta &= \frac{g}{1+g}(\mathbf{X}^{\delta'}\mathbf{X}^\delta)^{-1} && (3.13)\\
\mathbf{a}_N^\delta &= \frac{g}{1+g}(\mathbf{X}^{\delta'}\mathbf{X}^\delta)^{-1}\mathbf{X}^{\delta'}\mathbf{y}_c && (3.14)\\
s_N &= s_0 + (N-1)/2 && (3.15)\\
S_N &= S_0 + \frac{1}{2}S(\mathbf{X}^\delta) && (3.16)
\end{aligned}$$

3.2.2 Marginal likelihood and posterior moments for the fractional prior slab

For the fractional prior the marginal likelihood can be expressed as:

$$p(\mathbf{y}|\boldsymbol{\delta}) = \frac{b^{q_\delta/2}\Gamma(s_N)S_0^{s_0}}{(2\pi)^{(N-1)(1-b)/2}\Gamma(s_0)(S_N)^{s_N}}$$

where q_δ is the number of nonzero elements of $\boldsymbol{\delta}$, while $\mathbf{a}_N^\delta$ and $\mathbf{A}_N^\delta$ are given in (3.5) and (3.6) and

$$s_N = s_0 + (1-b)(N-1)/2 \tag{3.17}$$

$$S_N = S_0 + \frac{(1-b)}{2}(\mathbf{y}_c'\mathbf{y}_c - ((\mathbf{a}_N^\delta)'(\mathbf{A}_N^\delta)^{-1})\mathbf{a}_N^\delta) \tag{3.18}$$

Chapter 4

Stochastic Search Variable Selection (SSVS)

4.1 The SSVS prior

The "traditional" Bayesian approach presented in chapter 3 is related to model-selection-criteria: searching the best model is equivalent to maximizing the marginal likelihood. The following model selection approach is more in the spirit of significance-testing: in the full model one has to decide whether a regression coefficient is close to zero or far apart. Stochastic search variable selection was introduced by George and McCulloch (1993) and has the following basic idea: Each regressor coefficient is modeled as coming from a mixture of two normal distributions with different variances: one with density concentrated around zero, the other with density spread out over large plausible values. For every coefficient α_j a Bernoulli variable ν_j is defined taking values 1 and c ($\ll 1$) with probability $p(\nu_j = 1) = \omega$ and $p(\nu_j = c) = 1 - \omega$. ν_j acts as an indicator variable to address the two mixing components. If $\nu_j = 1$, α_j is sampled from the flat distribution implicating that a nonzero estimate of α_j should be included in the final model. Otherwise, if $\nu_j = c$, the value of α_j is sampled from the density with mass concentrated close to zero and the regressor x_j has negligible effect.

Formally the prior construction is the following:

- $\alpha_j|\nu_j \sim \nu_j\mathcal{N}(0, \psi^2) + (1 - \nu_j)\mathcal{N}(0, c\psi^2)$

- $$\nu_j = \begin{cases} c & \text{c very small } (e.g.\ 0.001) \quad p(c) = 1 - \omega \\ 1 & \qquad\qquad\qquad\qquad\quad p(1) = \omega \end{cases}$$
- $\omega \sim Beta(c_0, d_0)$

where ψ^2 is a fixed value chosen large enough to cover all reasonable values. In Swartz et al. (2008) ψ^2 and c are set 1000 and 1/1000 respectively. Figure (4.1) shows the plot of the two normal distributions.

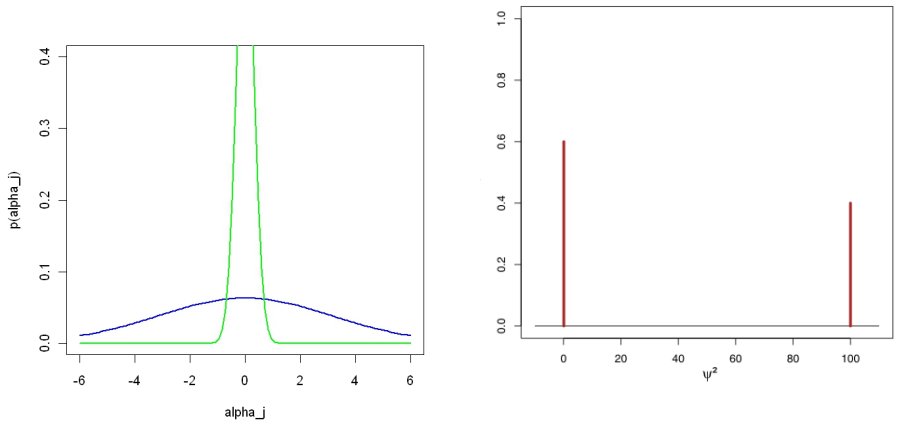


Figure 4.1: Left: SSVS prior for a single regressor: mixture of a slab (blue) and a spike (green) normal distribution. Right: The variances of the slab-and-spike components assume two values, with mass ω and $1 - \omega$.

It should be remarked that the prior for α_j is a spike and slab prior. In contrast to the priors presented in chapter 3 where the "spike" distribution was a discrete point mass, the prior here is a mixture of two continuous distributions meaning that also the "spike" distribution is continuous. This makes it easier to sample the indicator variable ν_j as for a continuous spike α_j is not exactly zero when $\nu_j = c$. The indicator can be drawn conditionally on the regressor coefficient α_j and computation of the marginal likelihood is not required. However, in each iteration the full model is fitted. Nonrelevant regressors remain in the model instead of being removed as under a Dirac spike. So model complexity cannot be reduced during the MCMC steps. This can be quite cumbersome for data sets with many covariables.

4.2 MCMC scheme for the SSVS prior

For MCMC estimation of the parameter $(\mu, \boldsymbol{\nu}, \omega, \boldsymbol{\alpha}, \sigma^2)$ the following Gibbs sampling scheme can be implemented:

(1) sample μ from $\mathcal{N}(\bar{y}, \sigma^2/N)$

(2) sample each ν_j, $j = 1 \ldots j$, from $p(\nu_j|\alpha_j, \omega) = (1-\omega) f_N(\alpha_j; 0, c\psi^2) I_{\{\nu_j=c\}} + \omega f_N(\alpha_j; 0, \psi^2) I_{\{\nu_j=1\}}$

(3) sample ω from $\mathcal{B}(c_0 + n_1, d_0 + k - n_1)$, where $n_1 = \sum_j I_{\{\nu_j=1\}}$

(4) sample $\boldsymbol{\alpha}$ from $N(\mathbf{a}_N, \mathbf{A}_N)$, where $\mathbf{A}_N^{-1} = \mathbf{X}'\mathbf{X}/\sigma^2 + \mathbf{D}^{-1}$, $a_N = \mathbf{A_N}\mathbf{X}'\mathbf{y}_c/\sigma^2$ and $\mathbf{D} = diag(\psi^2\nu_j)$

(5) sample σ^2 from $\mathcal{G}^{-1}(s_N, S_N)$,where $s_N = s_0 + (N-1)/2$ and $S_N = \frac{1}{2}((\mathbf{y}_c - \mathbf{X}\boldsymbol{\alpha})'(\mathbf{y}_c - \mathbf{X}\boldsymbol{\alpha}))$.

Chapter 5

Variable selection Using Normal Mixtures of Inverse Gamma Priors (NMIG)

5.1 The NMIG prior

An extension of the SSVS presented in the previous chapter was proposed by Ishwaran and Rao (2003). To avoid an arbitrary choice of the variance parameter ψ^2 as in the SSVS prior, a hierarchical formulation is proposed: the variance ψ^2 itself is assumed to be random and to follow a Gamma inverse distribution. The marginal prior for an effect α_j is a mixture of two Student distributions with mean zero, one with a very small and the other with a larger variance. As in SSVS an effect will drop out of the model, if the posterior probability that it belongs to the component with small variance is high.
The prior construction for the regressor coefficients is the following:

- $\alpha_j|\nu_j \sim \nu_j\mathcal{N}(0,\psi_j^2) + (1-\nu_j)\mathcal{N}(0,c\psi_j^2)$
- $\psi_j^2 \sim \mathcal{G}^{-1}(a_{\psi 0}, b_{\psi 0})$
- $$\nu_j = \begin{cases} c(=0.000025) & p(c) = 1-\omega \\ 1 & p(1) = \omega \end{cases}$$
- $\omega \sim Beta(c_0, d_0)$

As in SSVS c is a fixed value close to zero, ν_j=1 indicates the slab component and ν_j=c the spike component. The resulting prior for the variance parameter $\phi_j^2 = \nu_j \psi_j^2$ is a mixture of scaled inverse Gamma distributions:

$$p(\phi_j^2) = (1-\omega)\mathcal{G}^{-1}(\phi_j^2 | a_{\psi 0}, s_0 b_{\psi 0}) + \omega \mathcal{G}^{-1}(\phi_j^2 | a_{\psi 0}, s_1 b_{\psi 0})$$

It can be shown that the marginal distribution for the components of α_j is a mixture of scaled t-distributions, (see Konrath et al. (2008) for more detail):

$$p(\alpha_j | s_0, s_1, a_{\psi 0}, b_{\psi 0}) = 0.5 t_{df}(\alpha_j; 0, \tau_0) + 0.5 t_{df}(\alpha_j; 0, \tau_1)$$

with df=$2a_{\psi 0}$ degrees od freedom and scale parameter $\tau_i = \sqrt{\frac{s_i b_{\psi 0}}{a_{\psi 0}}}, i = 0, 1$. In figure (5.1) the two t-distributions are plotted with $a_{\psi 0} = 5, b_{\psi 0} = 50, s_0 = 0.000025, s_1 = 1$.

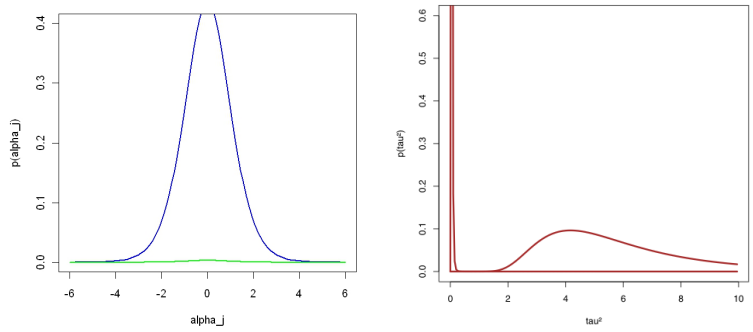


Figure 5.1: Left: NMIG-prior for a single regressor: mixture of two scaled t-distributions with $a_{\psi 0} = 5, b_{\psi 0} = 50, s_0 = 0.000025$ (blue line), $s_1 = 1$ (green line). Right: the induced prior for the variance follows a mixture of two inverse Gamma distributions.

5.2 MCMC scheme for the NMIG prior

The Gibbs sampling scheme for estimating the posterior distributions of the parameters $(\boldsymbol{\nu}, \boldsymbol{\psi}, \omega, \boldsymbol{\alpha}, \mu, \sigma^2)$ involves the following steps:

(1) sample the common mean μ from $\mathcal{N}(\bar{y}, \sigma^2/N)$

(2) for each j=1,. . . ,k sample ν_j from $p(\nu_j|\alpha_j, \psi_j^2, \omega, \mathbf{y}) = (1-\omega) f_N(\alpha_j; 0, c\psi_j^2) I_{\{\nu_j=c\}} + \omega f_N(\alpha_j; 0, \psi_j^2) I_{\{\nu_j=1\}}$

(3) for each j=1,...,k sample ψ_j^2 from $p(\psi_j^2|\alpha_j, \nu_j) = \mathcal{G}^{-1}(a_{\psi 0} + 1/2; b_{\psi 0} + 0.5\alpha_j^2/\nu_j)$

(4) sample ω from $p(\omega|\boldsymbol{\nu}) = B(c_0 + n_1, d_0 + k - n_1)$ where $n_1 = \sum_j I_{\{\nu_j=1\}}$

(5) sample the regression coefficients $\boldsymbol{\alpha}$ from $p(\boldsymbol{\alpha}|\boldsymbol{\nu}, \psi^2, \sigma^2, \mathbf{y}) = \mathcal{N}(\mathbf{a_N}, \mathbf{A}_N)$ where $\mathbf{A}_N^{-1} = \mathbf{X}'\mathbf{X}/\sigma^2 + \mathbf{D}^{-1}$, $\mathbf{a}_N = \mathbf{A}_N\mathbf{X}'\mathbf{y}_c/\sigma^2$, $\mathbf{D} = diag(\psi_j^2\nu_j)$

(6) sample the error variance σ^2 from $\mathcal{G}^{-1}(s_N, S_N)$ with parameters $s_N = s_0 + (N-1)/2$, $S_N = \frac{1}{2}((\mathbf{y}_c - \mathbf{X}\boldsymbol{\alpha})'(\mathbf{y}_c - \mathbf{X}\boldsymbol{\alpha}))$.

Chapter 6

Simulation study

To compare the performance of the different spike and slab priors described in chapters 3-5 a simulation study was conducted. Estimation, variable selection and efficiency of the MCMC-draws are investigated. To simplify notation the following abbreviations are used for the different priors:

- c ... (classical) ML estimation
- p ... independence prior: $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, c\mathbf{I})$
- g ... g-prior: $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, g(\mathbf{X}'\mathbf{X})^{-1}\sigma^2)$
- f ... fractional prior: $\boldsymbol{\beta} \sim \mathcal{N}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \frac{1}{b}(\mathbf{X}'\mathbf{X})^{-1}\sigma^2)$
- n ... NMIG prior: $\beta_j \sim \mathcal{N}(0, \nu_j\psi_j^2), \psi_j^2 \sim \mathcal{G}^{-1}(a_{\psi,0}, b_{\psi,0})$
- s ... SSVS prior: $\beta_j \sim \nu_j\mathcal{N}(0, \tau^2) + (1-\nu_j)\mathcal{N}(0, c\tau^2)$

For the independence prior 'p' as abbreviation is used instead of 'i', because the later is reserved for indices.
For error term σ^2 and intercept μ the improper priors given in (2.8) and (2.7) are used.
Concerning the priors for the regression coefficients the tuning of the prior variances is substantial. Considering the effect of penalization which depends on the size of the prior variance discussed in section 1.2.2, the magnitude of the prior variance influences estimation results. Therefore, to make the different priors for the regressor coefficients comparable, the prior variances are specified in order to obtain approximately covariance matrices of the same size, i.e. the constants $c, g, b, b_{\psi,0}a_{\psi,0}, \tau$ are chosen so that

prior	variance parameter	scaling groups		
		1	2	3
p	c	100	1	0.25
g	g	4000	40	10
f	b	$\frac{1}{4000}$	$\frac{1}{40}$	$\frac{1}{10}$
n	$(a_{\psi 0}, b_{\psi 0})$	(5,500)	(5,5)	(5,1.25)
s	τ^2	100	1	0.25

Table 6.1: Table of prior variance scaling groups

$$c\mathbf{I} \approx g(X'X)^{-1} \approx 1/b(X'X)^{-1} \text{ and } c \approx variance(\psi) \approx b_{\psi,0}/a_{\psi,0} \approx \tau^2$$

Table (6.1) shows 3 different prior variance groups used for simulations. Results are compared within the group, prior variance groups are denoted by the value of c.
For each scaling group 100 data sets consisting of N=40 responses with 9 covariates are generated, according to the model

$$y_i = \beta_0 + \boldsymbol{\beta_i}\mathbf{x_i} + \epsilon_i$$

For all data sets the intercept is set to 1 and the error term is drawn from the $N(0,1)$ distribution. To obtain independent regressors the covariates $\mathbf{x}_i = (x_{i1}, \ldots, x_{i9})$ are drawn from a multivariate Gaussian distribution with covariance matrix equal to the identity matrix. To generate correlated regressors the configuration of Tibshirani (1996) is used where the covariance matrix $\boldsymbol{\Sigma}$ is set as $\Sigma_{ij} = corr(x_i, x_j) = \rho^{|i-j|}$ with $\rho = 0.8$.
To study the behavior of selection of both 'strong' and 'weak' regressors the coefficient vector is set to

$$\boldsymbol{\beta} = (2, 2, 2, 0.2, 0.2, 0.2, 0, 0, 0) \tag{6.1}$$

where an effect of "2" is strong and an effect of "0.2" is weak. For the simulations with highly correlated regressors the parameter vector is set to:

$$\boldsymbol{\beta} = (2, 2, 0, 2, 0.2, 0, 0, 0.2, 0.2) \tag{6.2}$$

Correlation between regressors are highest between "neighbouring" regressors. This setting allows to study different scenarios, e.g. zero effects highly correlated with strong or

weak effects etc.

For each data set coefficient estimation and variable selection is performed jointly. MCMC is run for M=1000 iterations without burn in. Additionally, ML estimates of the full model are computed.

6.1 Results for independent regressors

6.1.1 Estimation accuracy

In a first step accuracy of coefficient estimation under the different priors measured by the squared error (SE)

$$SE(\alpha_i) = (\alpha_i - \hat{\alpha})^2$$

is compared. Estimated coefficients and squared errors are displayed in box plots in figures (6.1) to (6.6).

Starting with the prior variance group $c = 100$ for the independence prior (for the prior variance parameters of the other priors see table (6.1)), the results of coefficient estimations can be seen in figure (6.1). In the first row box plots of the estimated coefficients of "strong" regressors ($\alpha_i = 2$) are displayed, in the second one those of weak regressors ($\alpha_i = 0.2$), and in the third row those of the zero effects ($\alpha_i = 0$). The red lines mark the true values.

The mean of the estimates is close to the true coefficient values for both strong and zero effects, but smaller for weak effects, where shrinkage to zero occurs. Considering the discussion on the shrinkage property in section 1.2.2, it can be concluded that a large prior variance causes negligible shrinkage and Bayes estimates approximately coincide with ML-estimates. Bayes estimation with spike and slab priors implies model averaging as the posterior mean of a coefficient is an average over different models. Inclusion probabilities displayed in figure (6.8) show that strong regressors are included in almost all data sets; weak and zero regressors however have lower inclusion probabilities which means they are either set to zero for a Dirac spike or shrunk close to zero for a continuous spike. Since the posterior mean is a weighted average of the estimate under the slab prior (approximately equal to the OLS estimator) and the heavily shrunk estimate under the

spike prior, weak regressors are underestimated.

If the prior variance is smaller with $c = 1$ or $c = 0.25$, the shrinkage effect of a small prior variance discussed in section 1.2.2 can be seen in figure (6.5). Although the inclusion probability of strong regressors is still close to 1 (see figure (6.10)), the estimated coefficients are smaller than the true value. This can be observed in particular for the fractional prior, g-prior and SSVS prior. Also coefficients of weak regressors are shrunk, but due to the increased inclusion probability (see figure (6.10)) implying a larger weight on the almost unshrunk estimates under the slab prior, estimates are higher compared to a prior variance of $c = 100$. Also the squared error of weak regressors is reduced and of comparable size as the squared error of the MLE. Zero effects have an increased inclusion probability too, but their estimates are still zero or close to zero. Again the squared error of the zero effects is smaller for Bayes estimator than for the ML estimator.

The conclusions from the simulation study are:

- For estimation the shrinkage of the slab is not so pronounced. Due to model averaging it is relevant how often a coefficient is sampled from the spike component leading to a high shrinkage to zero. The inclusion probability depends on the variance of the slab component. This leads to following recommendations:
  - To estimate the effect of strong regressors a slab with large prior variance should be chosen.
  - To estimate (and detect) weak regressors a slab with small prior variance should be chosen.
  - To exclude zero effects from a final model a slab with large prior variance should be chosen.
- In figure (6.7) the sum of squared errors over all coefficients and for all data sets is shown by box plots. For c=100 and c=1 the Bayes estimates under spike and slab priors are smaller than those of the ML estimator.
- All priors perform rather similar.

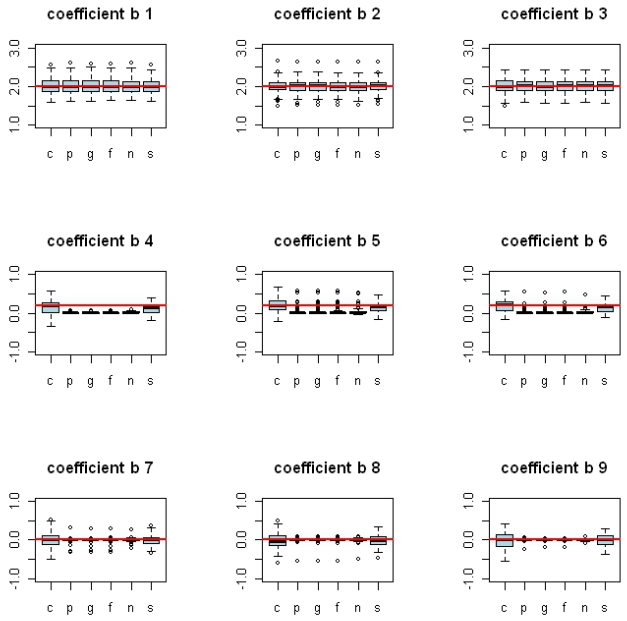


Figure 6.1: Box plots of coefficient estimates, the red line marks the true value. Prior variance group c=100.

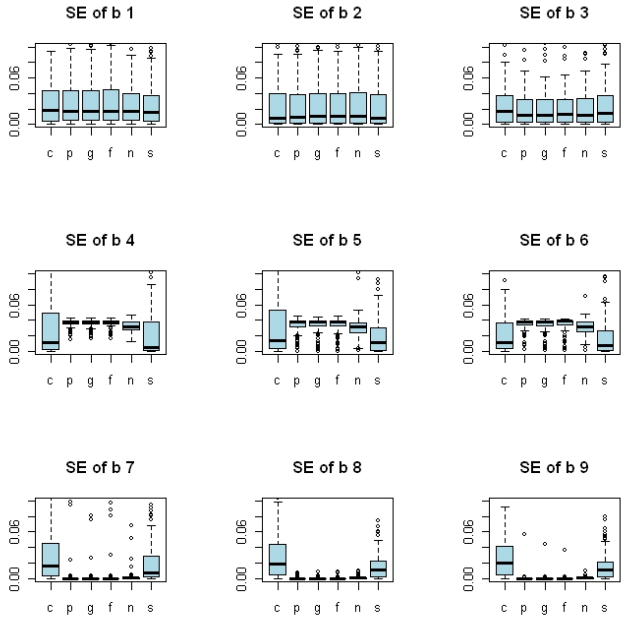


Figure 6.2: Box plots of SE of coefficient estimates. Prior variance group c=100.

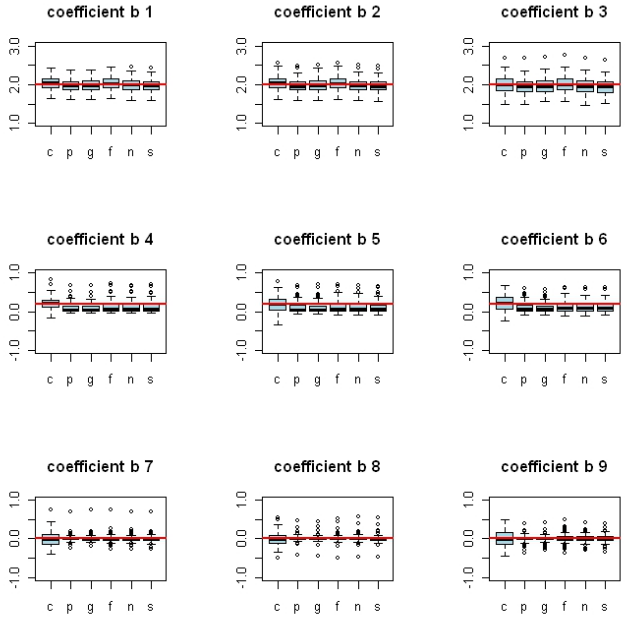


Figure 6.3: Box plots of coefficient estimates, the red line marks the true value. Prior parameter group c=1.

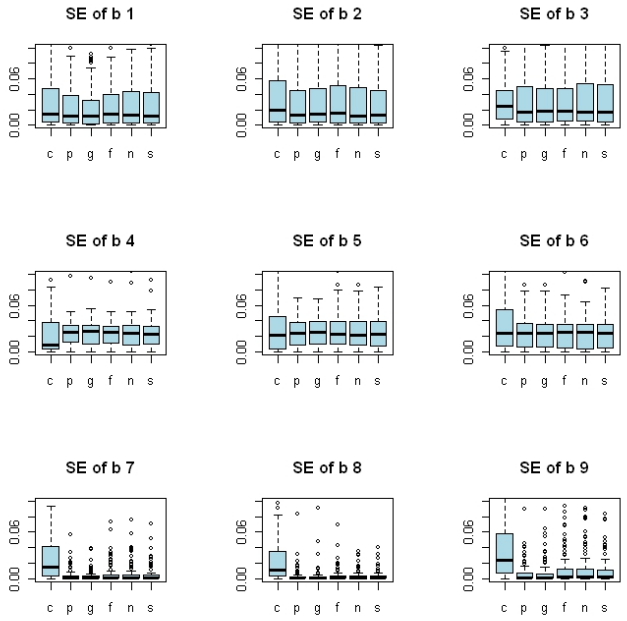


Figure 6.4: Box plots of SE of coefficient estimates. Prior variance group c=1.

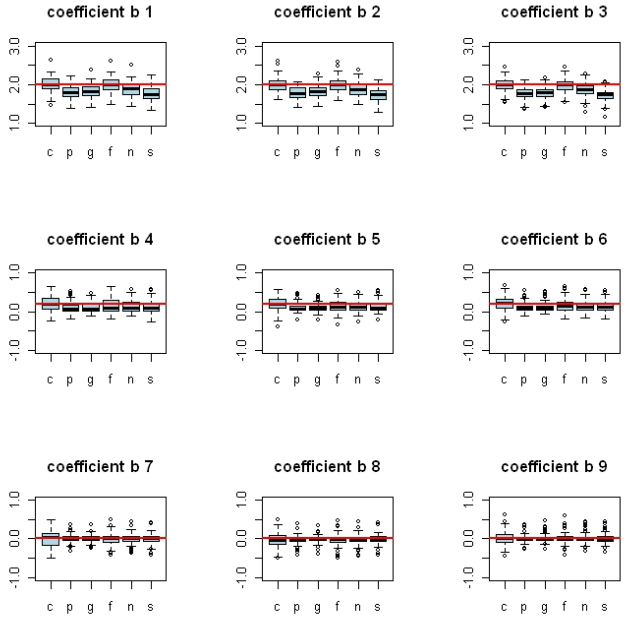


Figure 6.5: Box plots of coefficient estimates, the red line marks the true value. Prior variance group c=0.25.

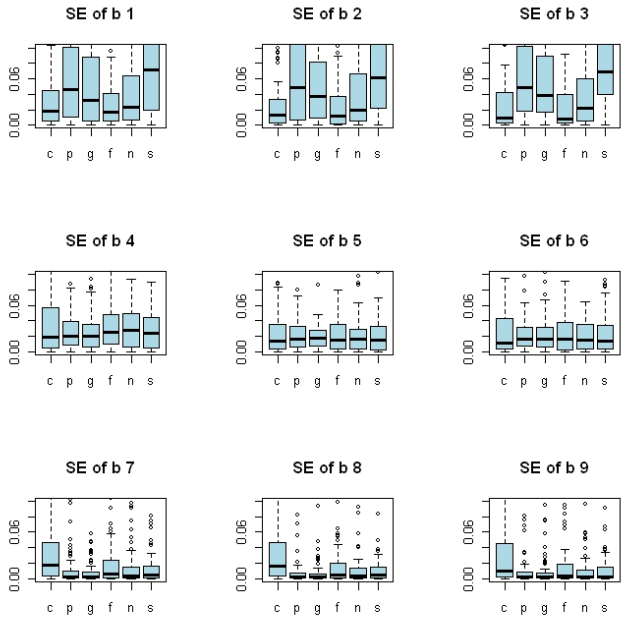
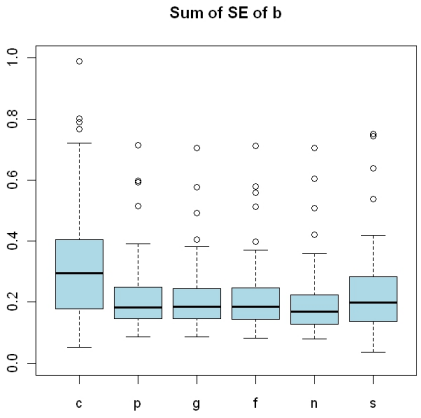
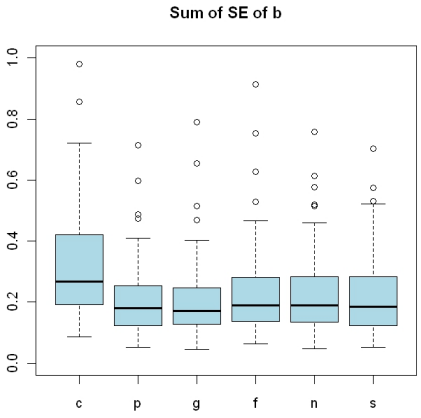


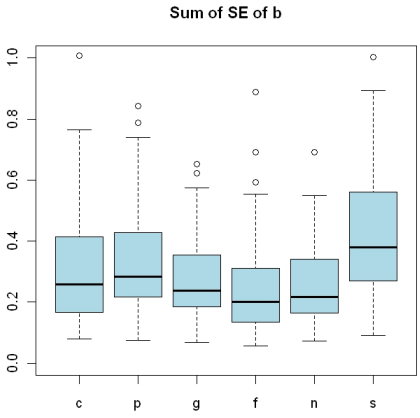
Figure 6.6: Box plots of SE of coefficient estimates. Prior variance group c=0.25.



(a) Sum of SE, c=100



(b) Sum of SE, c=1



(c) Sum of SE, c=0.25

Figure 6.7: Sum of SE of coefficient estimates for different prior variances

6.1.2 Variable selection

Variable selection means to decide for each regressor individually whether it should be included in the final model or not. Following Barbieri and Berger (2004), in a Bayesian framework the final model should be the median probability model consisting of those variables whose posterior inclusion probability $p(\delta_j = 1|\mathbf{y})$ is at least 0.5. The inclusion probability of a regressor is estimated by the posterior mean of the inclusion probability. The mean corresponds to the proportion of draws of a coefficient from the slab component of the prior. A larger posterior inclusion probability indicates that the corresponding regressor x_j has an effect which is not close to zero.
In figures (6.8), (6.9) and (6.10), the inclusion probabilities for each regressor are displayed in box plots for different prior variance settings. The first row of the plots shows the inclusion probabilities of "strong" regressors ($\beta_i = 2$), the second row those of "weak" regressors ($\beta_i = 0.2$) and the third row those of the zero effects ($\beta_i = 0$). For ML estimates the relative frequency of inclusion of a regressor based on a significance test with significance level $\alpha = 0.05$ is shown.

For strong regressors the inclusion probability is equal to one for all prior variances and all priors. That means, that strong coefficients are sampled only from the slab component of the priors, no matter what size of prior variance was chosen. For weak regressors, however, the inclusion probability depends on the prior variance. If the prior variance is large, the inclusion probability of weak regressors is low. The smaller the variance of the slab distribution, the higher the inclusion probability. Inclusion probabilities of zero effects show a similar behavior as those of weak regressors. For large variances the effect is assigned to the spike component, for smaller slab variances posterior inclusion probabilities increase as the effects are occasionally assigned to the slab component.

In the next step the influence of the size of prior variance on the number of misclassified regressors is examined. For this purpose, the false-discovery-rate (FDR) and the non-discovery-rate (NDR) defined as:

$$FDR = \frac{h(\delta_i = 1|\alpha_i = 0)}{h(\alpha_i = 0)}$$

$$NDR = \frac{h(\delta_i = 0 | \alpha_i \neq 1)}{h(\alpha_i \neq 1)}$$

are calculated, where h denotes the absolute frequency. Figure (6.11) shows how FDR and NDR change by varying the prior variance. As a benchmark line the FDR and NDR of the classical approach are plotted, which are 0.05 for the FDR (α-error of coefficient testing) and 0.40 for NDR (β-error of coefficient testing) respectively. Under all priors FDR and NDR show a similar behavior: if the variance of the slab component becomes smaller, the NDR decreases and the FDR increases. Therefore, looking at the total sum of misclassified regressors defined as

$$MISS = \frac{1}{k} \sum_{i=1}^{k} (1_{\{\delta_i = 1, \alpha_i = 0\}} + 1_{\{\delta_i = 0, \alpha_i \neq 0\}})$$

displayed in figure (6.11), it can be seen that the total sum of the misclassified variables remains roughly constant by varying the prior variance scaling.
Conclusions are:

- The inclusion probability depends on the size of the variance of the slab component. By increasing the variance, the inclusion probability of weak and zero effects decreases. However, the inclusion probability of strong regressors remains close to one for all variance settings.

- It is almost impossible to distinguish between small and zero effects: either both small and zero effects are included in the model or both are excluded simultaneously. The number of misclassified regressors remains roughly constant if the slab variance is varied.

- The different priors yield similar results for large variances. For a small prior variance ($c = 0.25$) the inclusion probability of weak and zero effects is smaller under the independence prior and g-prior compared to the other priors.

To identify zero and nonzero effects the following recommendations can be given:

- Strong regressors are detected under each prior in each prior variance parameter setting. Weak regressors are hard to detect in our simulation setup.

- For reasonable prior variances ($c = 100$ or $c = 1$) also zero effects are correctly identified under each prior.

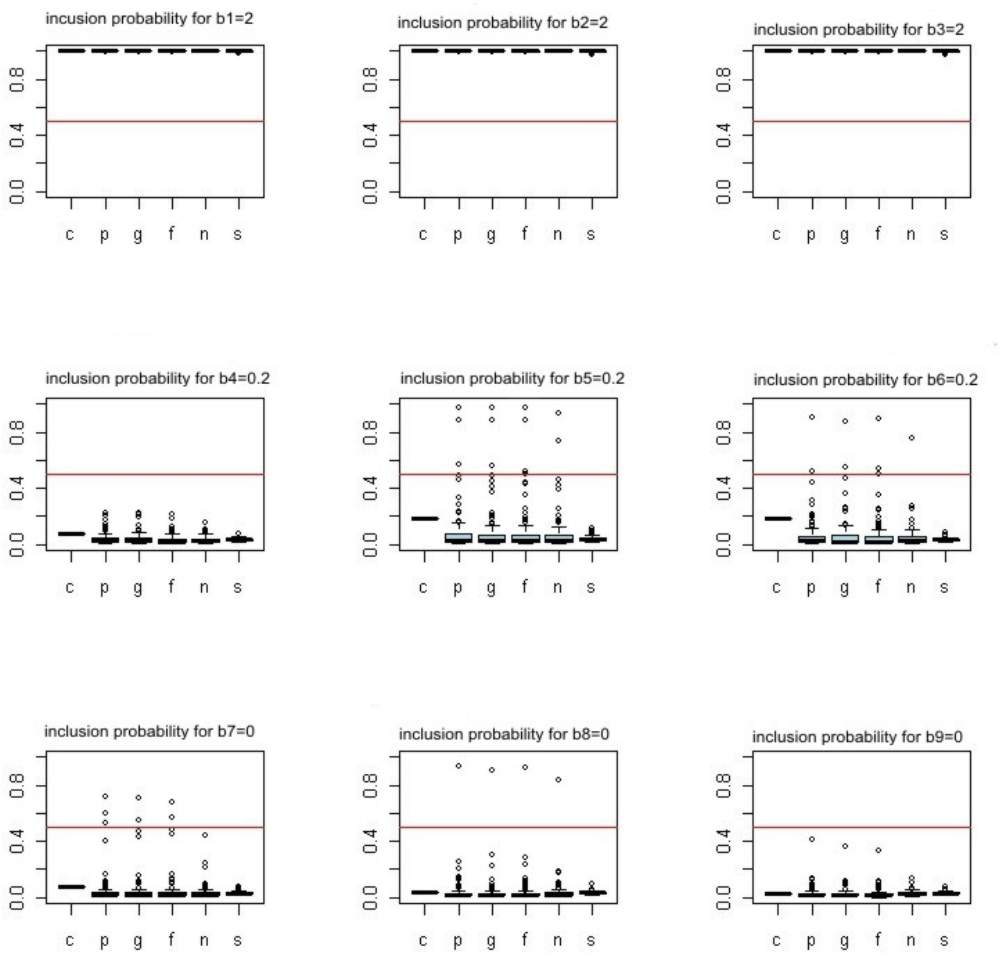


Figure 6.8: Box plots of the posterior inclusion probabilities $p(\delta_j = 1|\mathbf{y})$. Prior variance group c=100.

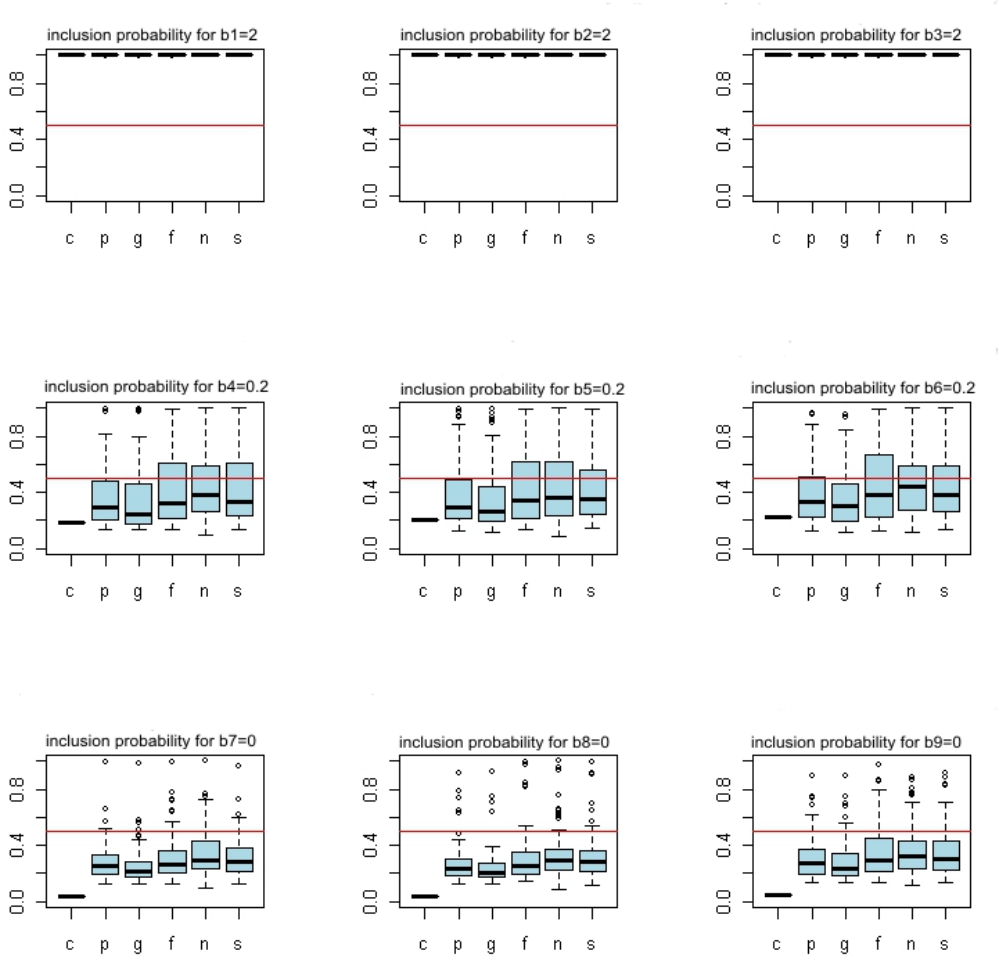


Figure 6.9: Box plots of the posterior inclusion probabilities $p(\delta_j = 1|\mathbf{y})$. Prior variance group c=1.

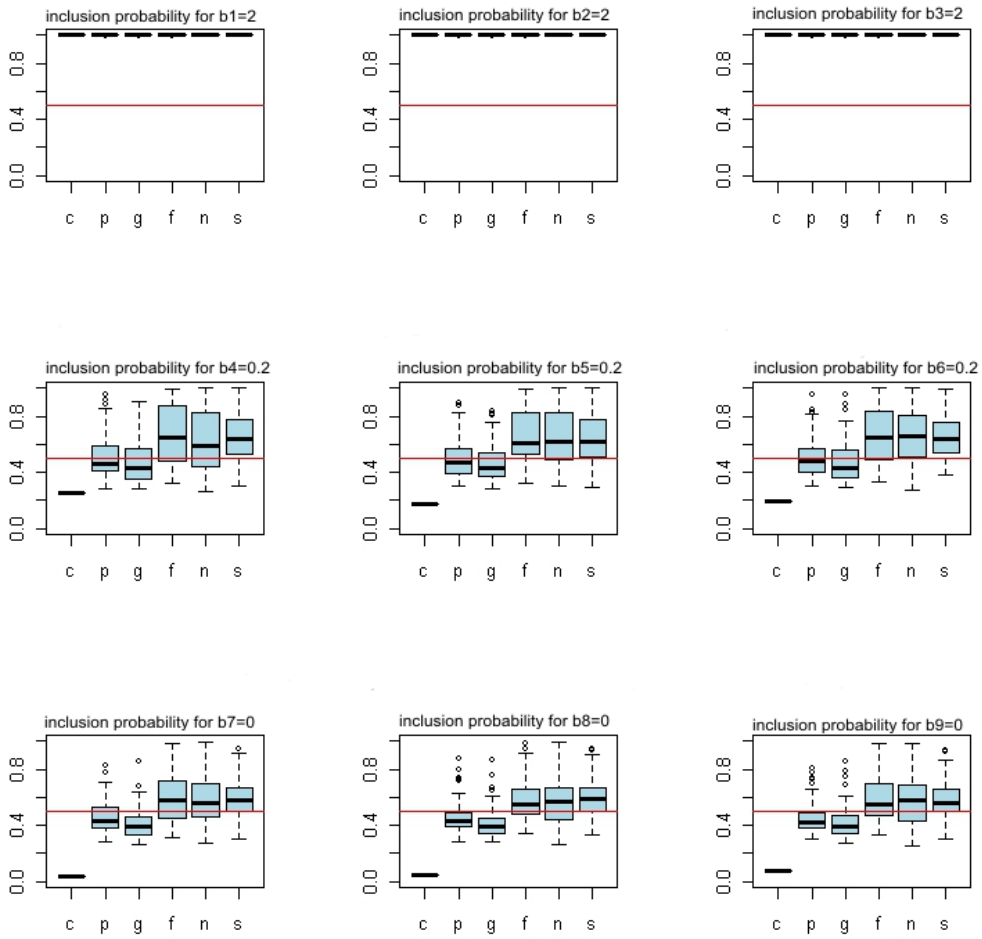
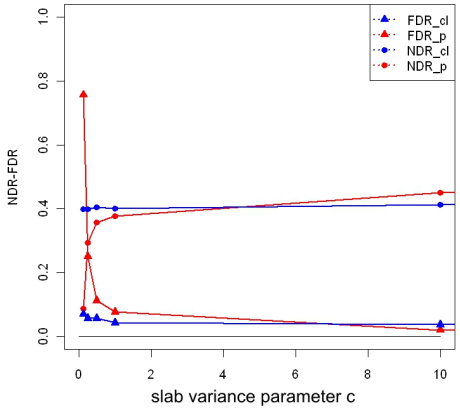
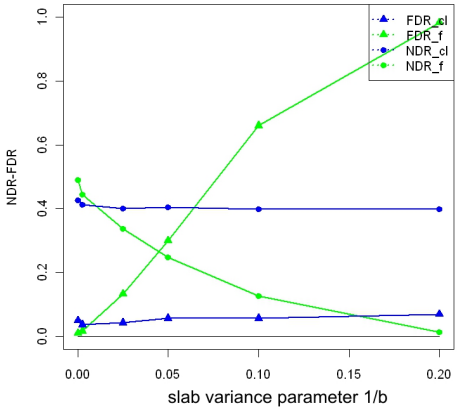


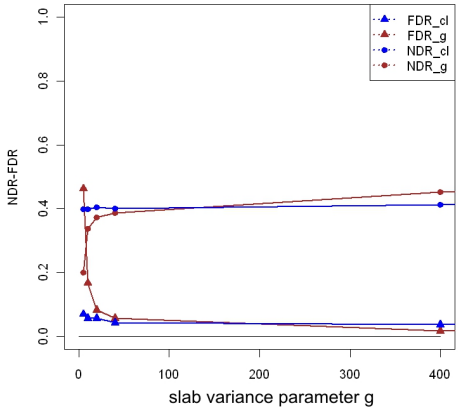
Figure 6.10: Box plots of the posterior inclusion probabilities $p(\delta_j = 1|\mathbf{y})$. Prior variance group c=0.25.



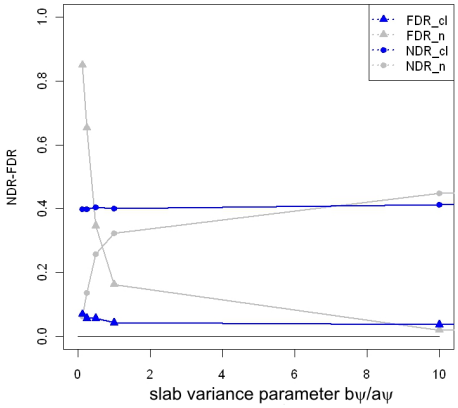
(a) independence prior



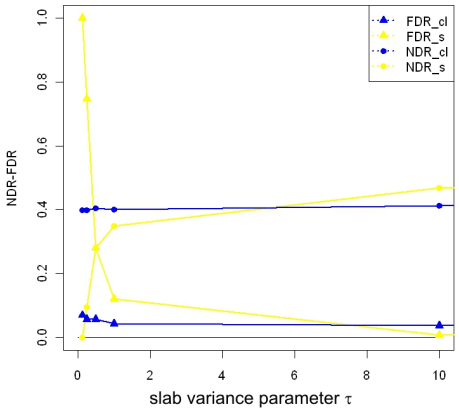
(b) fractional prior



(c) g prior



(d) NMIG prior



(e) SSVS prior

Figure 6.11: NDR and FDR for different priors as a function of the prior variance parameters.

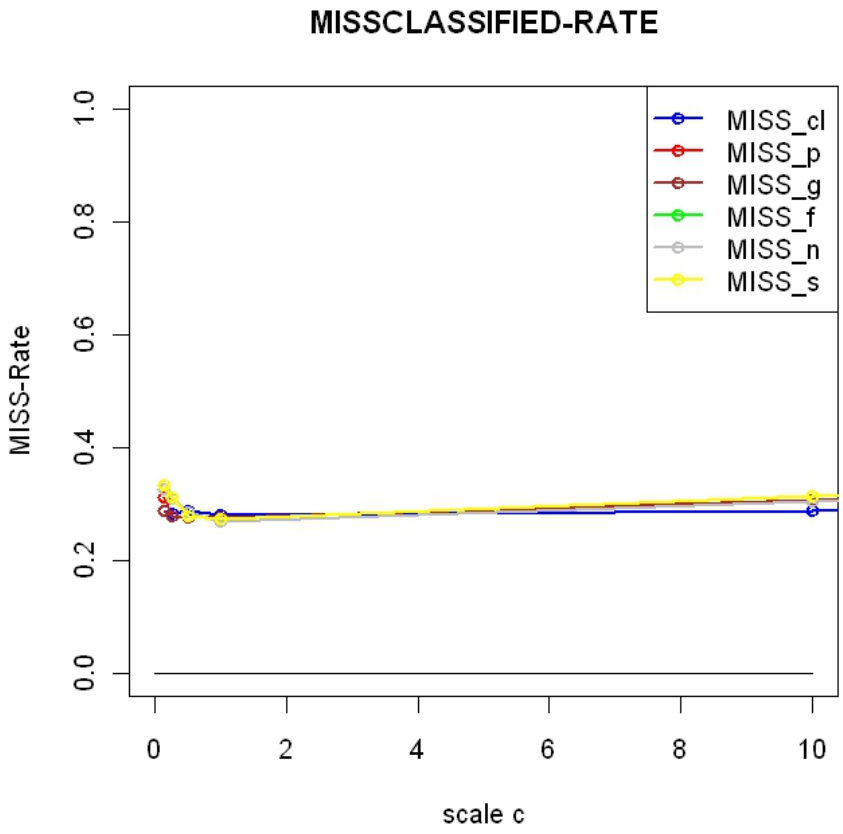


Figure 6.12: Proportion of misclassified effects as a function of the prior variance scale

6.1.3 Efficiency of MCMC

In this chapter we compare computational effort and MCMC efficiency under different priors. Independence prior, fractional prior and g-prior require the time-consuming calculation of 2k marginal likelihoods in every step of iteration. Computation of the marginal likelihood is not necessary for the NMIG-prior and the SSVS-prior, but model complexity is not reduced during MCMC, as no effects are shrunk exactly to zero. Draws from an MCMC implementation are not independent but in general correlated. To measure the loss of information of a dependent sample compared to a independent sample, the inefficiency factor f defined as

$$f = 1 + \sum_{s=1}^{\infty} \rho_s$$

is calculated where ρ_s is the autocorrelation at lag s. If the number of draws is divided by the inefficiency factor f, the effective sample size of the sample (ESS) is obtained :

$$ESS = \frac{M}{f}$$

The effective sample size is the number of independent draws that correspond to the dependent draws. For practical computation of the inefficiency factor autocorrelations are summed only up to lag s. To determine s Geyer (1992) proposed to calculate the function $\Gamma_m = \rho_{2m} + \rho_{2m+1}$, which is the sum of two adjacent pairs of autocorrelations. He showed that for an irreducible, reversible Markov chain Γ_m is a strictly positive, strictly decreasing and strictly convex function. s is determined as the lag where the conditions are violated for the first time. Geyer (1992) showed also that with the obtained efficiency factor the true variance is overestimated. If the ESS is divided by the computation time needed, the number of effective iterations per second is obtained, which can be compared for different priors.

To study the efficiency of MCMC under the presented priors, one data set with 40 observations is generated as described on page 41. MCMC is performed under each prior with the variance parameters $c = 10$ for independence prior, $g = 400$ for g-prior, $b = 1/400$ for fractional prior, $(a_{\psi 0}, b_{\psi 0}) = (5, 50)$ for NMIG prior and $\tau^2 = 10$ for SSVS prior. The MCMC algorithms were run for 10000 iterations and the posterior inclusion probability

$p(\delta_j^{(m)} = 1|\mathbf{y})$ in each iteration is saved for each regressor. Autocorrelations, inefficiency factors and ESS of the draws $p(\delta_j^{(m)} = 1|\mathbf{y})$ are computed.

Autocorrelations of the posterior inclusion probabilities under the independence prior are shown in figure (6.13). The autocorrelations decay quickly, and this also is true for the g-, fractional and SSVS prior. However, autocorrelations are high for the NMIG prior, see figure (6.14). Efficiency factors summarized in table (6.2) vary between 1 and 2 under Dirac spikes, between 1 and 5.3 for SSVS prior and between 1 and 64 for the NMIG prior. In table (6.3) implementations for the different priors are compared taking into account computation time: ESS per second for the independence prior, g-prior and fractional prior are between 10 and 20, for the SSVS prior between 62 and 335, and for the NMIG prior between 2 and 284. Performance is best for the SSVS as draws are nearly autocorrelated and sampling is fast.
To investigate whether models are sampled according to their model probability, the observed frequencies of the drawn models are compared to the posterior model probabilities. For the independence prior, g-prior and fractional prior the results are summarized in tables (6.6), (6.7) and (6.8). The model frequencies from MCMC closely correspond to the posterior model probabilities: this means that the models are actually sampled according to their posterior probability.
To determine the sample size necessary to achieve a good approximation to the posterior model probability, MCMC is run for 100, 1000, 10000 and 20000 iterations under the independence prior. The results shown in table (6.4) indicate that only 1000 draws yield good approximation. However, posterior model probabilities can differ considerably under different priors (see table (6.6), (6.7) and (6.8)).

We arrive at the following conclusions:

- Autocorrelations: MCMC draws of the inclusion probabilities have small autocorrelations for the independence prior, g-prior, fractional prior and SSVS prior, but can be highly correlated for the NMIG prior.

- Computational time: MCMC for Dirac spikes is time consuming, CPU time is 10 times higher then for NMIG and SSVS prior.

- ESS: MCMC for the SSVS prior performs best. MCMC under the NMIG prior can outperform but also perform worse than MCMC for Dirac spikes.
- Under the independence prior, g-prior and fractional prior models are sampled according to their model probability.
- Convergence: For the independence prior 1000 iterations are enough to achieve a reasonable approximation to the posterior model probability.

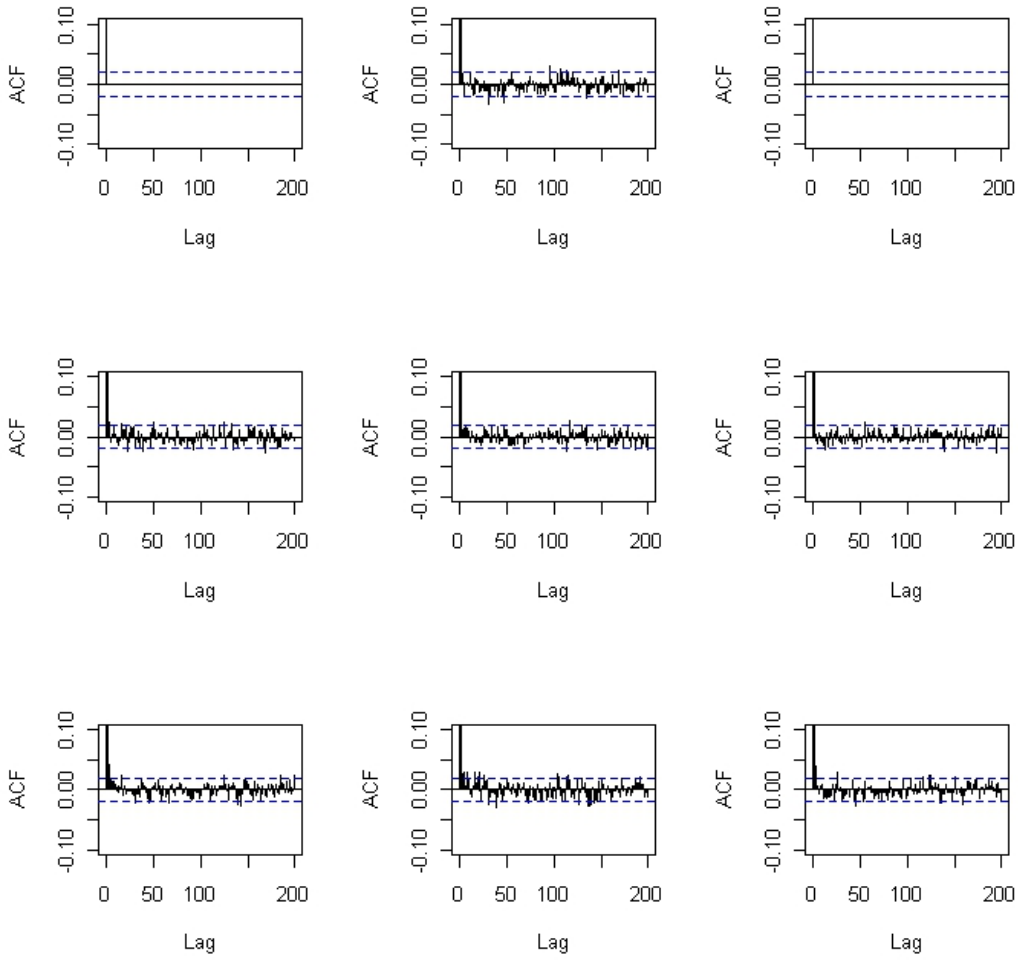


Figure 6.13: ACF of the posterior inclusion probabilities $p(\delta = 1|y)$ under the independence prior, prior variance parameter c=10, M=10000.

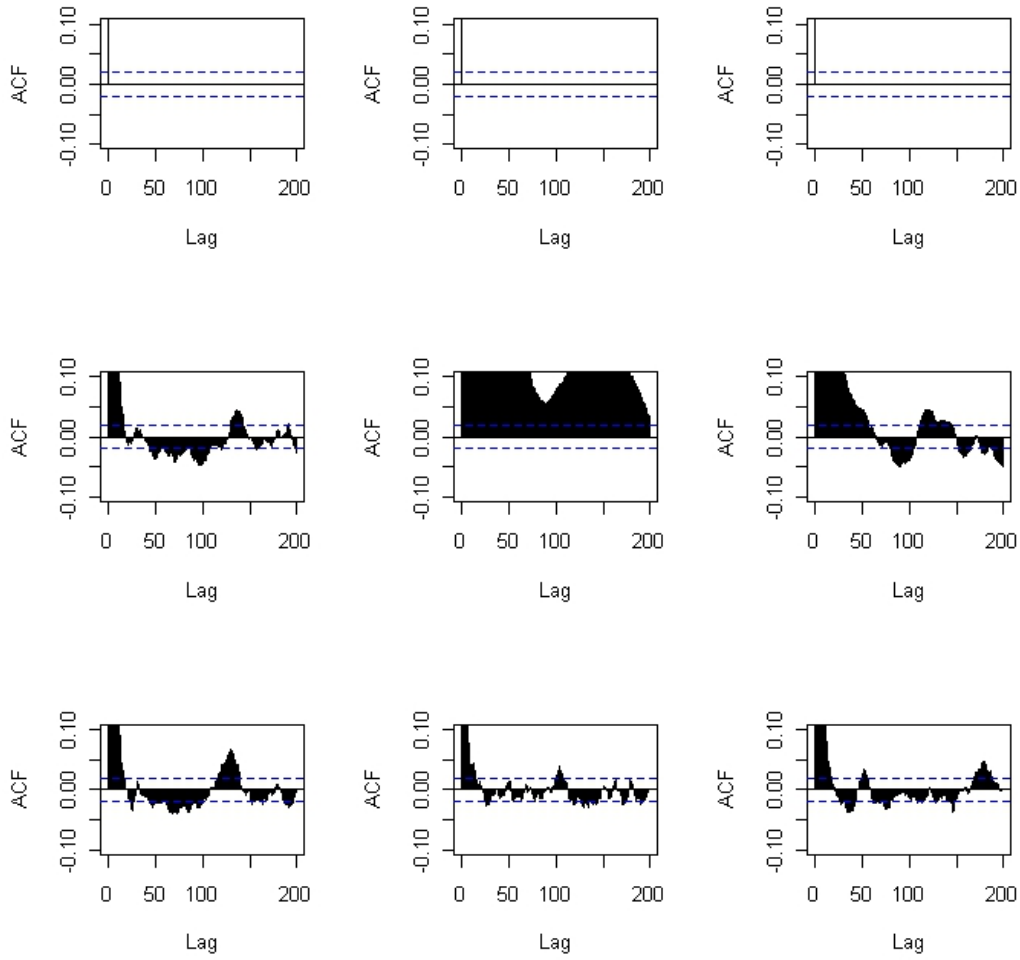


Figure 6.14: ACF of the posterior inclusion probabilities $p(\delta = 1|y)$ under the NMIG prior, prior variance parameter $(a_{\psi 0}, b_{\psi 0})$=(5,50), M=10000.

	p		g		f		n		s	
	f	ms	f	ms	f	ms	f	ms	f	ms
x1	1.00	1	1.00	1	1.00	1	1.00	1	1.00	1
x2	1.55	3	1.00	1	1.52	5	1.00	1	1.00	1
x3	1.00	1	1.43	1	1.00	1	1.00	1	1.00	1
x4	1.63	3	2.00	7	2.04	5	9.70	17	1.98	5
x5	1.47	1	1.78	3	1.63	3	64.54	87	5.30	9
x6	1.47	1	1.84	3	2.03	9	23.53	57	1.90	9
x7	1.70	3	2.02	3	2.09	5	9.48	19	1.57	7
x8	1.59	1	2.01	3	2.11	9	5.42	9	1.57	3
x9	1.71	5	2.04	3	1.98	3	8.68	19	1.61	3

Table 6.2: Posterior indicator probability $p(\delta_j = 1|\mathbf{y})$: inefficiency factor (f) and number of the autocorrelations (ms) summed up for computation of the inefficiency factor

	p		g		f		n		s	
	ESS	iter	ESS	iter	ESS	iter	ES	iter	ESS	iter
x1	9999	21	10000	22	10000	22	10000	284	10000	333
x2	6465	14	10000	22	6579	15	10000	284	10000	333
x3	10000	21	6995	15	10000	22	10000	284	10000	333
x4	6144	13	5009	11	4900	11	1031	29	5039	167
x5	6823	14	5623	12	6130	14	154	4	1886	62
x6	6801	14	5445	12	4925	11	424	12	5273	175
x7	5887	12	4944	11	4784	10	1055	30	6360	212
x8	6270	13	4978	11	4729	10	1845	52	6380	212
x9	5860	12	4905	11	5047	11	1152	32	6207	206

Table 6.3: Posterior inclusion probability $p(\delta_j = 1|\mathbf{y})$: effective sample size (ESS) and ESS per second (iter)

model	x1	x2	x3	x4	x5	x6	x7	x8	x9	h_100	h_1000	h_10000	h_20000	p_mM
1	1	1	1	0	1	0	0	0	0	0.30	0.36	0.38	0.38	0.38
2	1	1	1	0	0	0	0	0	0	0.45	0.38	0.36	0.35	0.35
3	1	1	1	0	0	1	0	0	0	0.04	0.04	0.04	0.04	0.04
4	1	1	1	1	1	0	0	0	0	0.02	0.04	0.04	0.04	0.04
5	1	1	1	0	1	1	0	0	0	0.01	0.02	0.02	0.03	0.02
6	1	1	1	0	1	0	0	0	1	0.01	0.03	0.02	0.02	0.02
7	1	1	1	0	1	0	1	0	0	0.02	0.03	0.02	0.02	0.02
8	1	1	1	0	1	0	0	1	0	0.01	0.02	0.02	0.02	0.02
9	1	1	1	0	0	0	1	0	0	0.03	0.01	0.02	0.02	0.02
10	1	1	1	1	0	0	0	0	0	0.00	0.02	0.02	0.02	0.02

Table 6.4: Observed frequencies of the models under independence prior; number of MCMC iterations m=100, 1000,10000, 20000. p_mM denotes the model probability under the independence prior.

model	x1	x2	x3	x4	x5	x6	x7	x8	x9	h_p	h_g	h_f	h_n	h_s	p_mM
1	1	1	1	0	1	0	0	0	0	0.369	0.239	0.265	0.362	0.189	0.380
2	1	1	1	0	0	0	0	0	0	0.354	0.158	0.080	0.395	0.565	0.351
3	1	1	1	0	0	1	0	0	0	0.037	0.044	0.027	0.037	0.038	0.037
4	1	1	1	1	1	0	0	0	0	0.036	0.054	0.074	0.028	0.018	0.036
5	1	1	1	0	1	1	0	0	0	0.026	0.043	0.054	0.022	0.011	0.024
6	1	1	1	0	1	0	0	0	1	0.024	0.037	0.044	0.019	0.011	0.024
7	1	1	1	0	1	0	1	0	0	0.021	0.038	0.050	0.017	0.015	0.021
8	1	1	1	0	1	0	0	1	0	0.021	0.037	0.043	0.017	0.010	0.018
9	1	1	1	0	0	0	1	0	0	0.017	0.022	0.012	0.018	0.030	0.017
10	1	1	1	1	0	0	0	0	0	0.018	0.018	0.011	0.015	0.028	0.017

Table 6.5: Frequencies of the models for different priors. (p_mM)is the model probability under the independence prior.

model	x1	x2	x3	x4	x5	x6	x7	x8	x9	h_10000_p	p_mM_p
1	1	1	1	0	1	0	0	0	0	0.369	0.380
2	1	1	1	0	0	0	0	0	0	0.354	0.351
3	1	1	1	0	0	1	0	0	0	0.037	0.037
4	1	1	1	1	1	0	0	0	0	0.036	0.036
5	1	1	1	0	1	1	0	0	0	0.026	0.024
6	1	1	1	0	1	0	0	0	1	0.024	0.024
7	1	1	1	0	1	0	1	0	0	0.021	0.021
8	1	1	1	0	1	0	0	1	0	0.021	0.018
9	1	1	1	0	0	0	1	0	0	0.017	0.017
10	1	1	1	1	0	0	0	0	0	0.018	0.017

Table 6.6: Independence prior: observed frequencies (h_10000_p) and probability (p_mM_p) of different models.

model	x1	x2	x3	x4	x5	x6	x7	x8	x9	h_10000_g	p_mM_g
1	1	1	1	0	1	0	0	0	0	0.239	0.231
2	1	1	1	0	0	0	0	0	0	0.158	0.165
3	1	1	1	1	1	0	0	0	0	0.054	0.054
4	1	1	1	0	0	1	0	0	0	0.044	0.043
5	1	1	1	0	1	1	0	0	0	0.043	0.043
6	1	1	1	0	1	0	1	0	0	0.038	0.039
7	1	1	1	0	1	0	0	0	1	0.037	0.039
8	1	1	1	0	1	0	0	1	0	0.037	0.036
9	1	1	1	0	0	0	1	0	0	0.022	0.023
10	1	1	1	1	0	0	0	0	0	0.018	0.021

Table 6.7: G-prior: observed frequencies (h_10000_g) and probability (p_mM_g) of different models.

model	x1	x2	x3	x4	x5	x6	x7	x8	x9	h_10000_f	p_mM_f
1	1	1	1	0	1	0	0	0	0	0.265	0.268
2	1	1	1	0	0	0	0	0	0	0.080	0.080
3	1	1	1	1	1	0	0	0	0	0.074	0.073
4	1	1	1	0	1	1	0	0	0	0.054	0.053
5	1	1	1	0	1	0	1	0	0	0.050	0.047
6	1	1	1	0	1	0	0	0	1	0.044	0.047
7	1	1	1	0	1	0	0	1	0	0.043	0.043
8	1	1	1	0	0	1	0	0	0	0.027	0.028
9	1	1	1	1	1	0	0	0	1	0.023	0.022
10	1	1	1	1	1	1	0	0	0	0.021	0.021

Table 6.8: Fractional prior: observed frequencies (h_10000_f) and probability (p_mM_f) of different models

6.2 Results for correlated regressors

To study performance of variable selection under different priors another simulation study with highly correlated regressors is carried out. Highly correlated data sets as described on page 41 are generated and the same simulations as in the uncorrelated case are performed.

6.2.1 Estimation accuracy

As for the uncorrelated regressors Bayes estimates are investegated under 3 different slab variance parameters. Box plots of estimates are given in figures (6.15), (6.17) and (6.19), and those of the squared error in figures (6.16), (6.18), (6.20). Strong regressors ($\beta = 2$) are b1, b2 and b4, weak regressors ($\beta = 0.2$) are b5, b8 and b9, and zero-effects are b3, b6 and b7. The result is similar to that for independent regressors: the mean of the Bayes estimates for strong and zero effects is close to the true value, whereas weak effects are underestimated again under all priors. Compared to independent regressors estimation error is increased, see figures (6.2) and (6.6). This is expected, since in the presence of correlated regressors the matrix $\mathbf{X}'\mathbf{X}$ tends to be ill-conditioned yielding unstable coefficient estimations and large estimation errors. The increase of the estimation error is higher for the ML estimates than for the Bayes estimates. In detail, for weak and zero effects the SE of the Bayes estimates is much smaller than for the ML estimates, whereas for strong effects Bayes estimates perform similar as the ML estimates. Comparing figures (6.7) and (6.21), we see that the total sum of SE for ML estimation is approximately three times as high as in the independent case, whereas it is only doubled for the Bayes estimates. This illustrates the regularisation property of Bayes estimates. With regard to SE Bayes estimates clearly outperform ML estimates.

As adjacent regressors are highly correlated, the influence of correlation can be seen by looking at the position of a regressor within the coefficient vector. It seems that estimation accuracy is not much affected by the correlation among regressors. For example, the squared estimation error of a strong regressor differs not much in cases where its neighbour is another strong or a weak regressor.

If the prior variances get smaller, the results remains essentially the same, apart from the fact that the priors begin to act differently. As shown in figure (6.17) the smallest SE

is achieved by the Bayes estimate under the g- prior. Generally, all Bayes estimates still outperform the ML estimation, see figure (6.21).

Conclusions are:

- For correlated regressors the SE of effect estimates are higher than for independent regressors. The increase is smaller for Bayes estimates than for MLE.
- For large slab variance Bayes estimates under all priors perform similar, for very small slab variance the g-prior perform best.

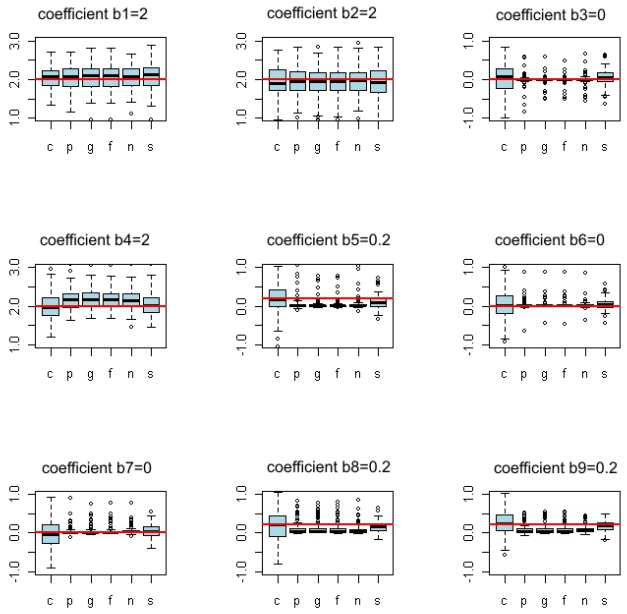


Figure 6.15: Correlated regressors: Box plot of coefficient estimates, the red line marks the true value. Prior variance group c=100.

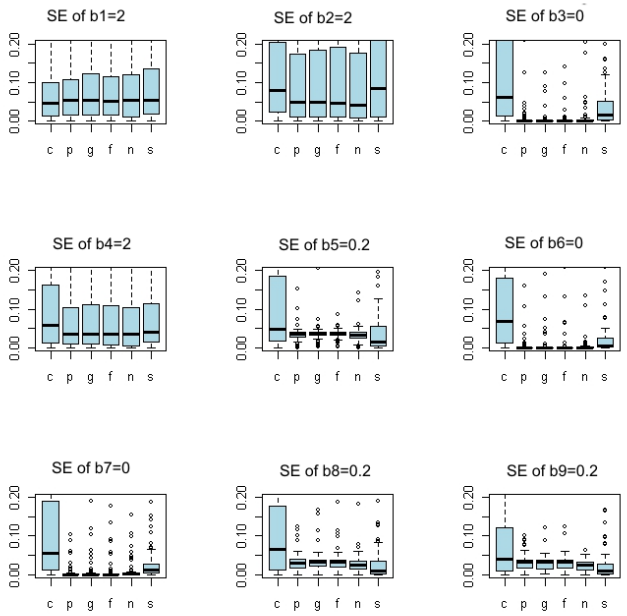


Figure 6.16: Correlated regressors: Box plot of SE of coefficient estimates. Prior variance group c=100.

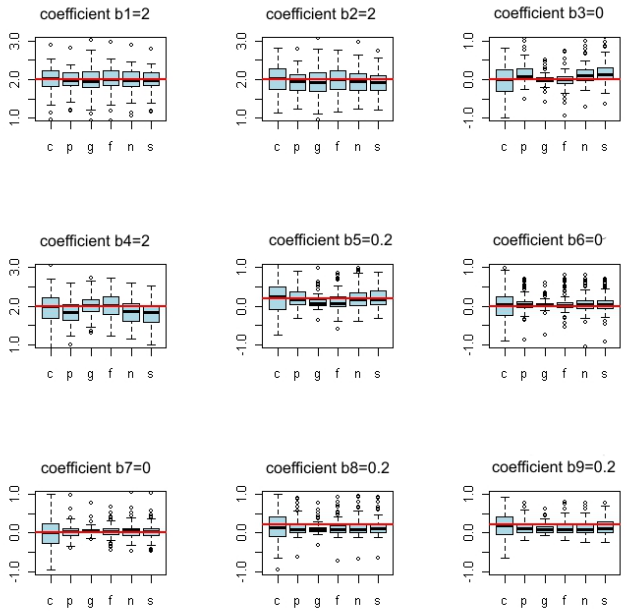


Figure 6.17: Correlated regressors: Box plot of coefficient estimates, the red line marks the true value. Prior variance group c=1.

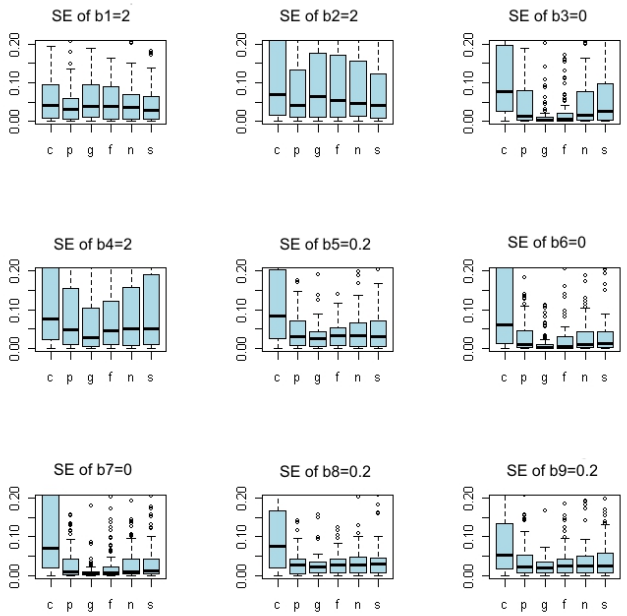


Figure 6.18: Correlated regressors: Box plot of SE of coefficient estimates. Prior variance group c=1.

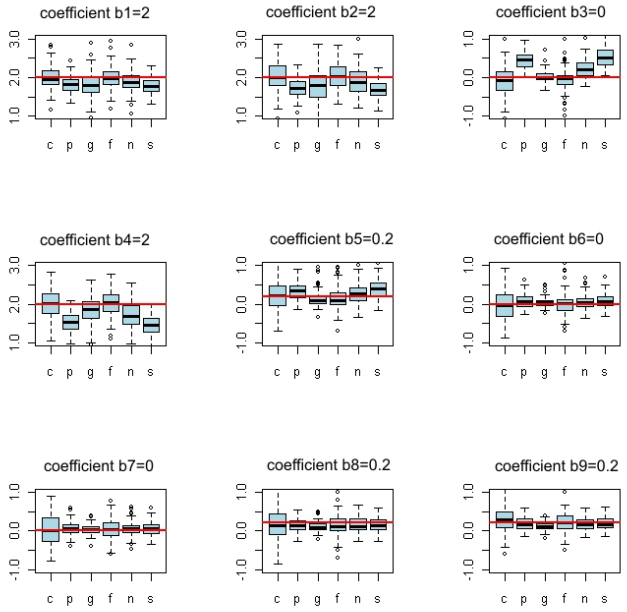


Figure 6.19: Correlated regressors: Box plot of coefficient estimates, the red line marks the true value. Prior variance group c=0.25.

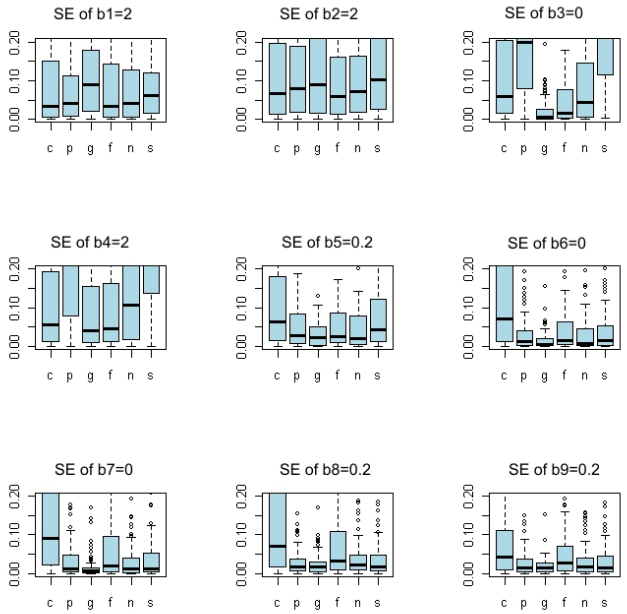
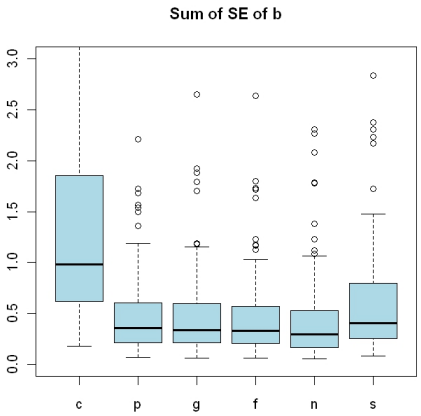
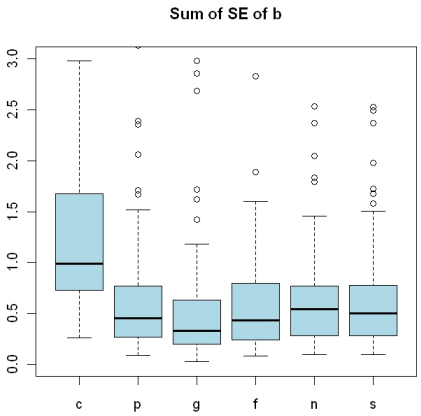


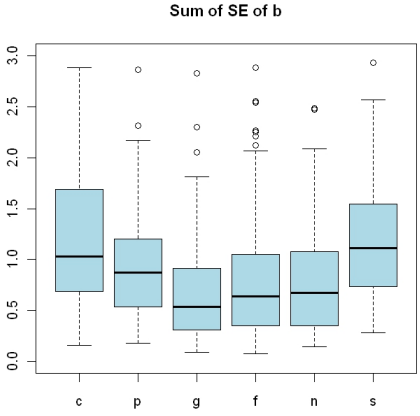
Figure 6.20: Correlated regressors: Box plot of SE of coefficient estimates. Prior variance group c=0.25.



(a) Sum of SE, c=100



(b) Sum of SE, c=1



(c) Sum of SE, c=0.25

Figure 6.21: Correlated regressors: Sum of SE of coefficient estimates for different prior variances.

6.2.2 Variable selection

Correlation among regressors might have an impact on variable selection as e.g. inclusion probability of a strong effect might differ depending on whether it is highly correlated with another strong or a zero effect. Box plots of the inclusion probabilities for different slab variance parameters are shown in figures (6.22), (6.23) and (6.24).
Actually for strong effects the inclusion probability is close to one for all Bayes estimates as well as for the ML estimates based on F-tests in the full model. Using F-tests for the ML estimates weak regressors are only included in 5% of the data sets (compared to 20% in the case of independent regressors). In contrast, under all priors inclusion probabilities of weak effects are slightly higher in the simulation under correlated regressors compared to independent regressors. Zero effects have low inclusion probabilities for both MLE and Bayes estimates.
Whereas for a large prior variance inclusion probabilities are similar under all priors, for smaller prior variance (c=1, c=0.25) we observe a different behavior under g- and fractional prior than under the rest of priors. Whereas for independence prior, SSVS prior and NMIG prior the inclusion probabilities for weak and zero effects increase with smaller prior variance, increase is much smaller under the fractional prior and almost not present under the g-prior.
Figure (6.25) shows NDR rate and FDR rate for correlated regressors. As for independent regressors NDR decreases and FDR increases with increasing prior variance. Again the proportion of misclassified covariates is approximately constant across different prior variances. For large prior variance the proportion of misclassified covariates is slightly smaller than that based on individual F-tests.

We can draw the following conclusions:

- Strong and zero effects are identified correctly by Bayesian variable selection under the priors considered here.
- For a large prior variance the inclusion probabilities of correlated regressors do not differ considerably from those of independent regressors.
- With smaller prior variance the inclusion probabilities of both weak and zero effects increase faster than for independent regressors for all priors except the g-prior.

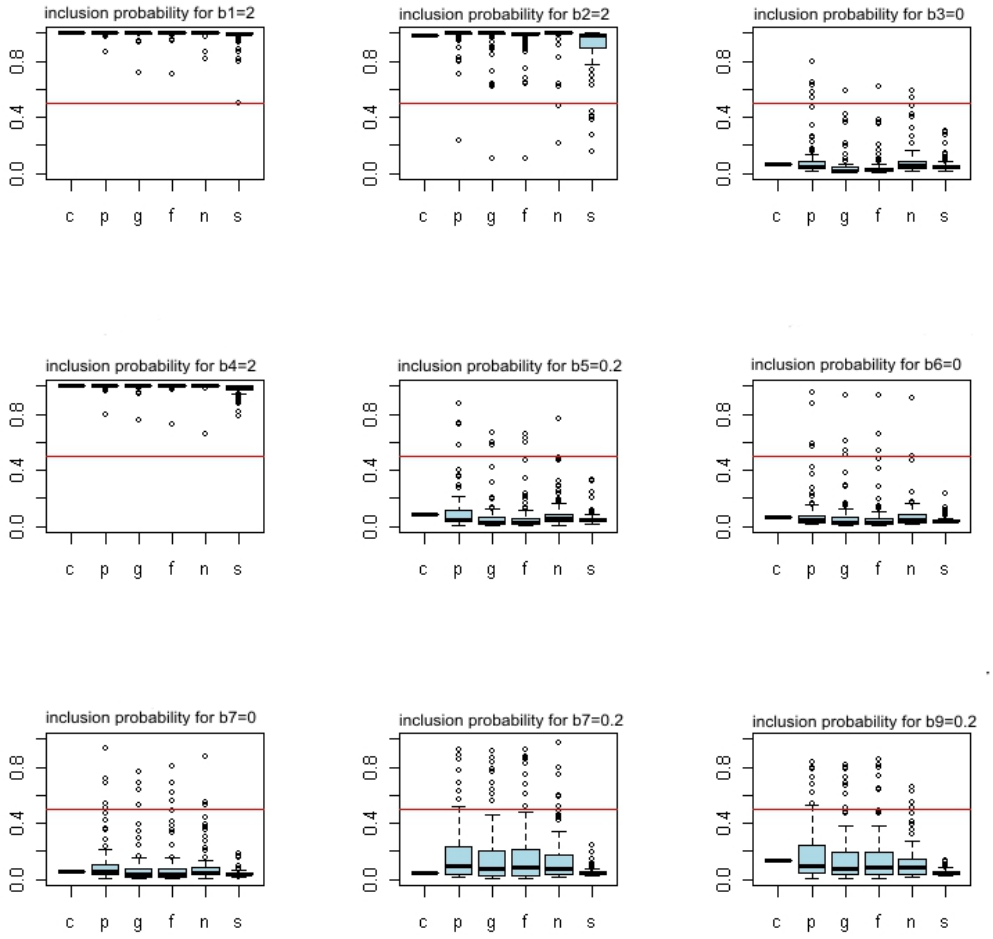


Figure 6.22: Correlated regressors: Box plots of the posterior inclusion probabilities $p(\delta_j = 1|\mathbf{y})$. Prior parameter group c=100.

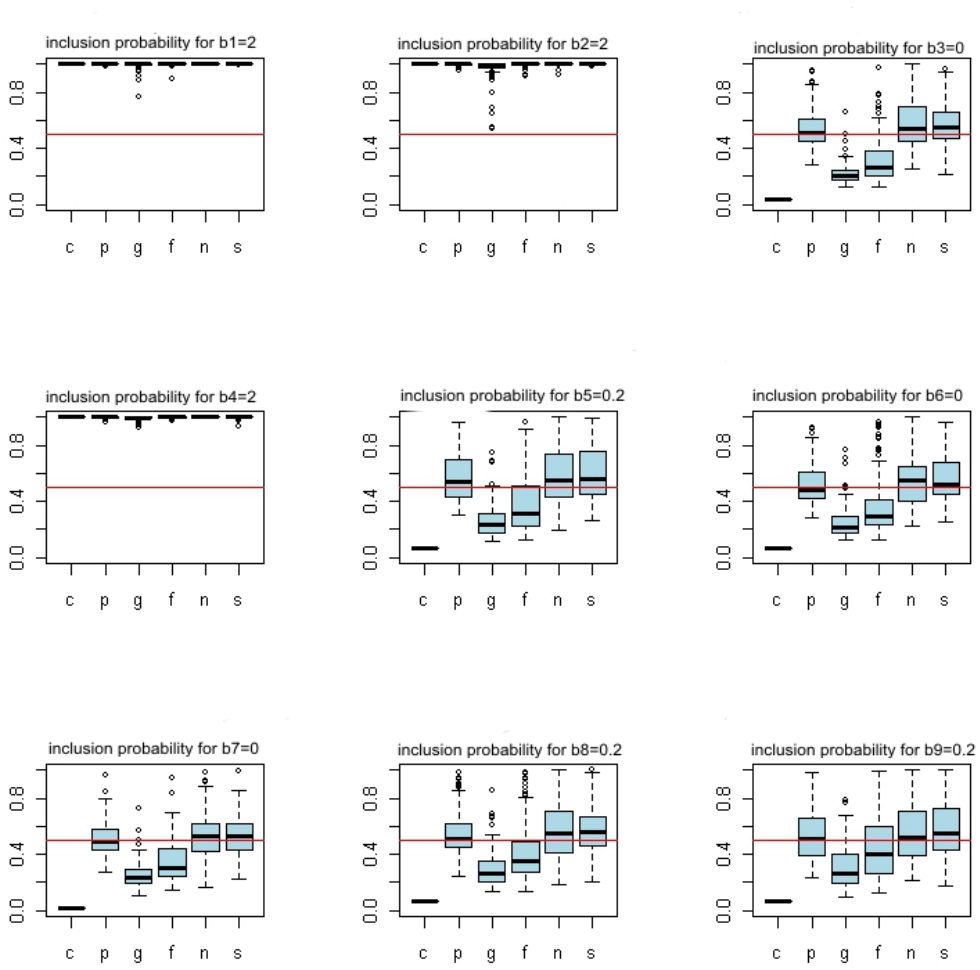


Figure 6.23: Correlated regressors: Box plots of the posterior inclusion probabilities $p(\delta_j = 1|\mathbf{y})$. Prior variance group c=1.

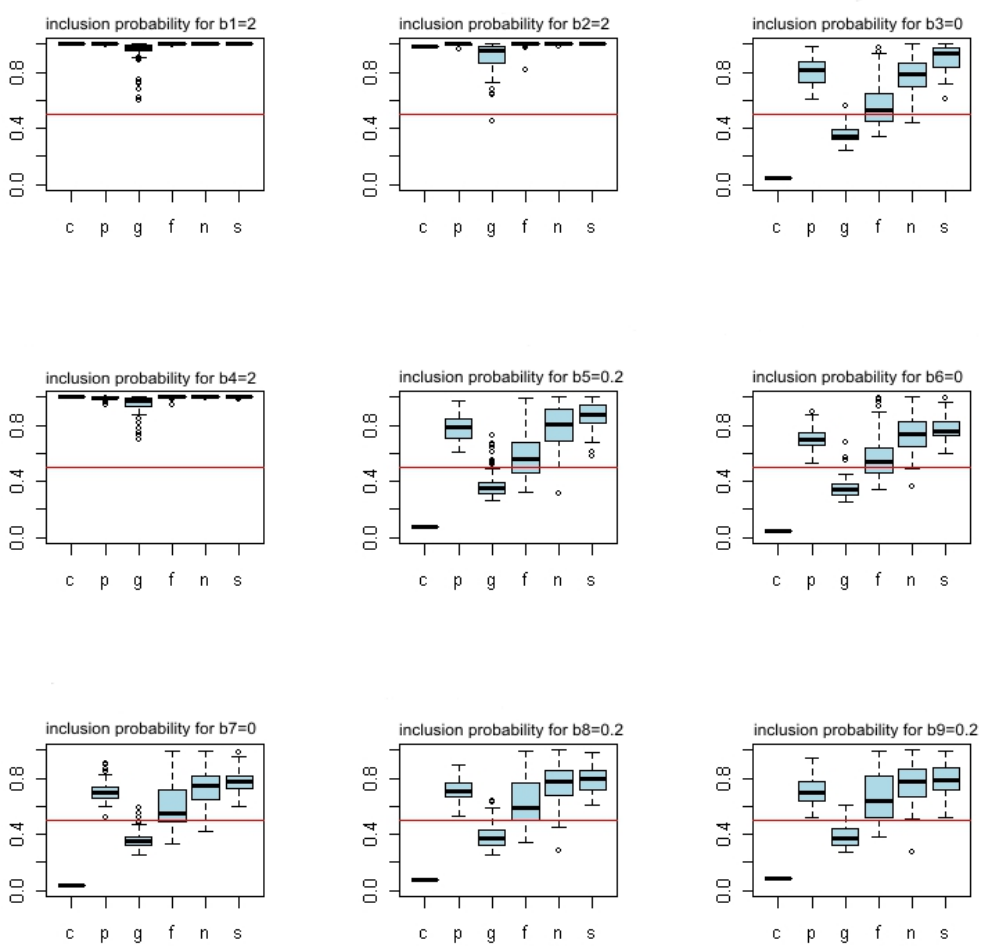
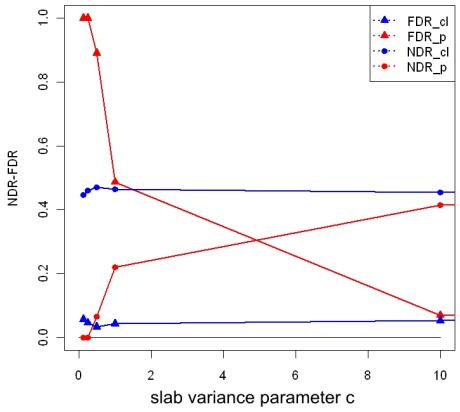
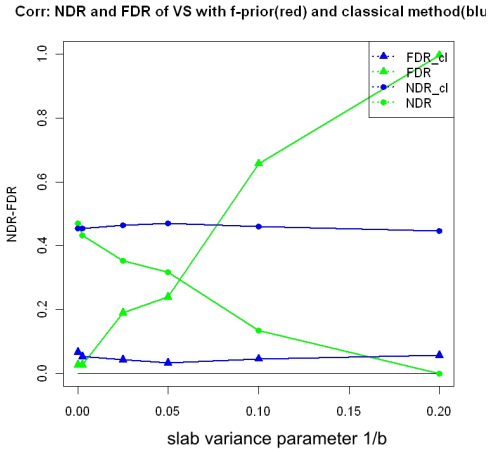


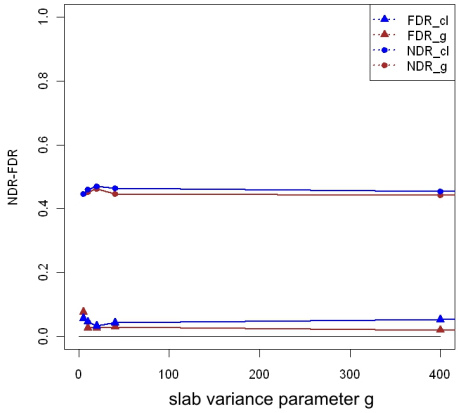
Figure 6.24: Correlated regressors: Box plots of the posterior inclusion probabilities $p(\delta_j = 1|\mathbf{y})$. Prior variance group c=0.25.



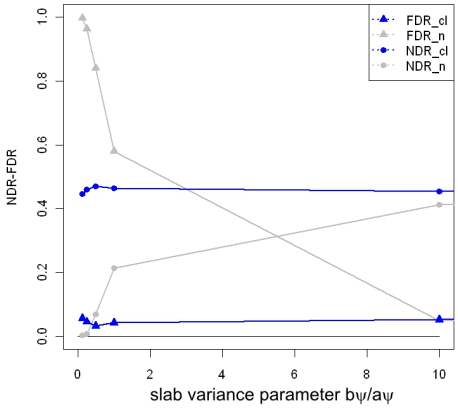
(a) independence prior



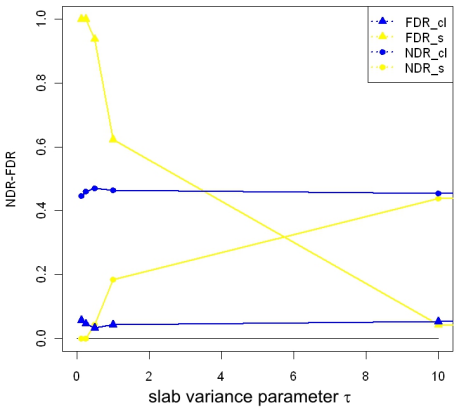
(b) fractional prior



(c) g prior



(d) NMIG prior



(e) SSVS prior

Figure 6.25: Correlated regressors: NDR and FDR for different priors as a function of the prior variance parameters.

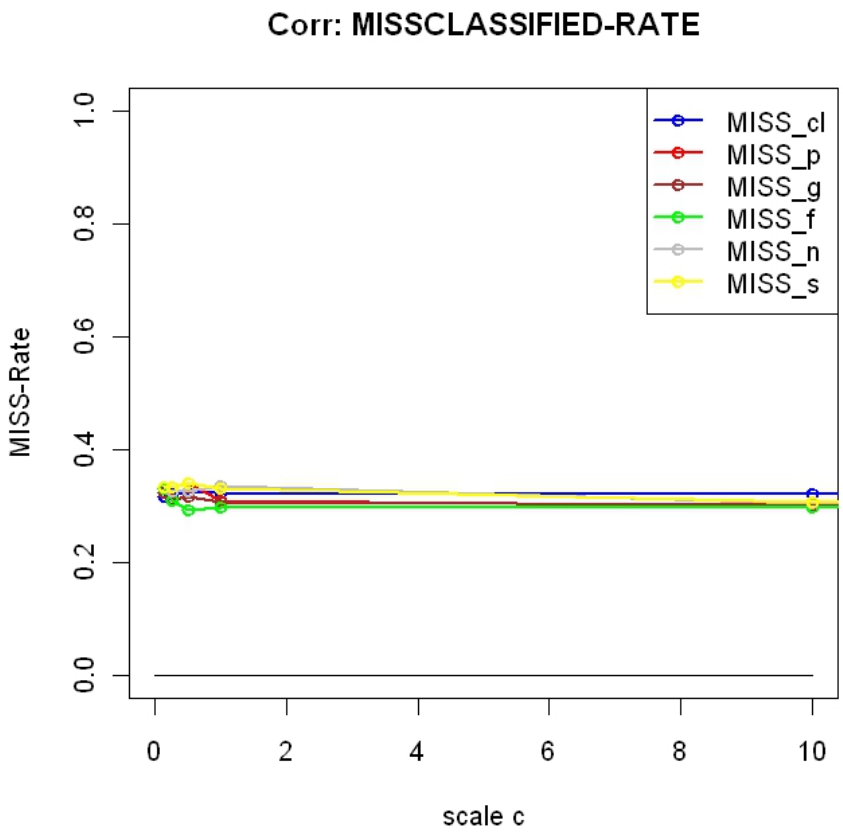


Figure 6.26: Correlated regressors: Proportion of misclassified effects as a function of the prior variance scale.

6.2.3 Efficiency of MCMC

Finally we compare MCMC efficiency and computational effort under the different priors for one data set with correlated regressors (generated as described on page 41). MCMC was run for 10000 iterations with the prior variance parameters $c = 10$ for independence prior, $g = 400$ for g-prior, $b = 1/400$ for fractional prior, $(a_{\psi 0}, b_{\psi 0}) = (5, 50)$ for NMIG prior and $\tau^2 = 10$ for SSVS prior. Inefficiency factor and effective sample size of the posterior inclusion probabilities are computed as described in section 6.1.3.

Figures (6.27) and (6.28) show the autocorrelation function of the inclusion probabilities under the independence prior and NMIG prior. Autocorrelations under the fractional prior, g-prior and SSVS prior are similar to those under the independence prior and are not shown. The autocorrelations are small for all priors except the NMIG prior where the posterior inclusion probabilities are highly autocorrelated. Autocorrelations are similar to those of independent regressors. This means that correlation among regressors has no pronounced effect on the autocorrelation of the inclusion probability chains. This is supported by inefficiency factor and effective sample sizes (ESS), presented in tables (6.9), which are of comparable order as for independent regressors. Again ESS per second is similar for inclusion probabilities under independence prior, g-prior and fractional prior for all regressors (11 to 22). Under the NMIG prior there is a high variation with ESS per second varying from 1 to 295. Under the SSVS prior ESS per second is highest among all priors with values from 29 up to 333, thus outperforming MCMC under all other priors considered.
Finally the number of visits to models during MCMC is compared to the posterior model probability under in tables (6.11), (6.12)and (6.13) respectively. The observed frequency of each model visited during MCMC is a reasonable approximation for the posterior model probability.

We can conclude that

- Correlation between the regressors does not increase autocorrelations in the sampled posterior inclusion probabilities, and inefficiency factor and effective sample sizes are of the same order as for independent regressors.

- Relative frequencies of visits to different models during MCMC are good approximations of posterior model probabilities (for independence prior, g-prior and fractional prior).

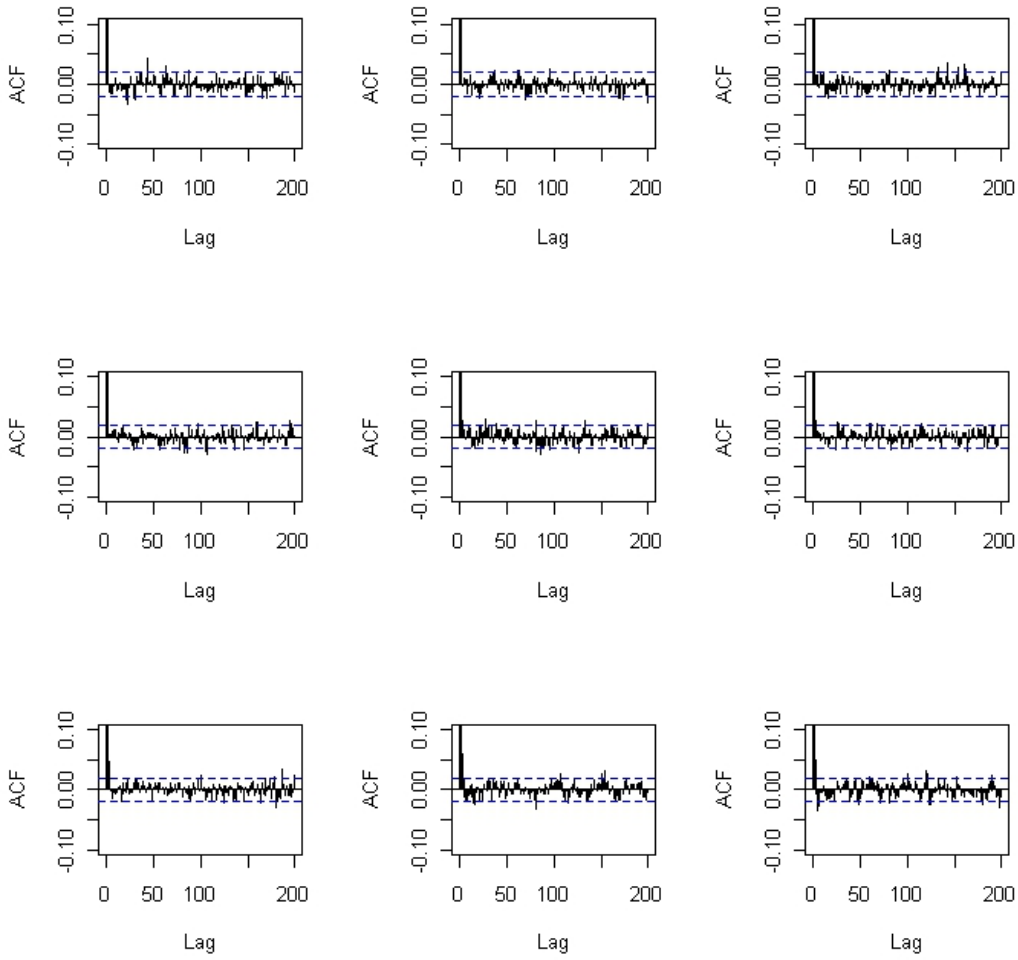


Figure 6.27: Correlated regressors: ACF of the posterior inclusion probabilities $p(\delta = 1|y)$ under the independence prior, prior variance parameter c=10, M=10000.

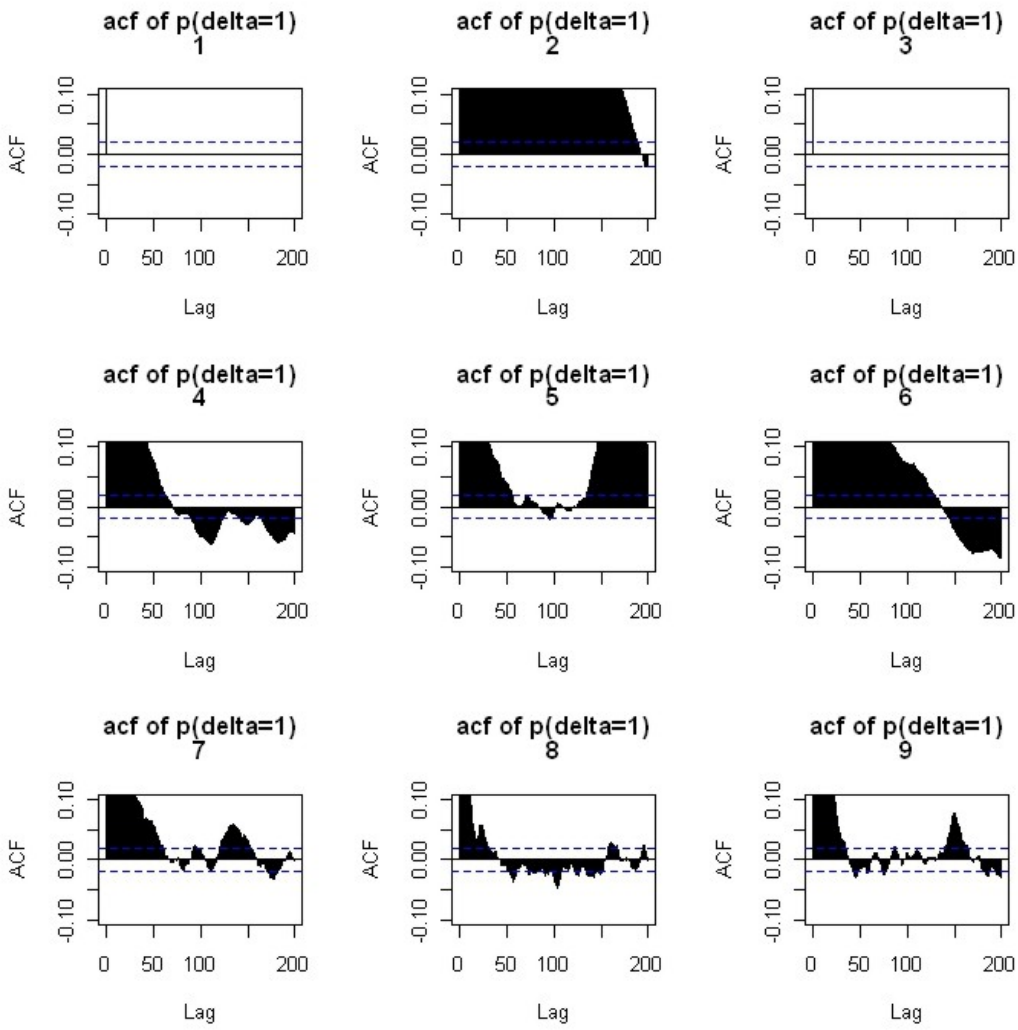


Figure 6.28: Correlated regressors: ACF of the posterior inclusion probabilities $p(\delta = 1|y)$ under the NMIG prior, prior variance parameters $(a_{\psi 0}, b_{\psi 0})$=(5,50), M=10000.

	p		g		f		n		s	
regressor	f	ms	f	ms	f	ms	f	ms	f	ms
x1	1.32	1	1.37	3	1.00	1	1.00	1	1.00	1
x2	1.44	1	1.60	3	1.56	3	192.01	193	11.47	15
x3	1.51	5	1.88	1	1.61	3	1.00	1	1.00	1
x4	1.49	3	1.54	3	1.48	1	39.68	69	5.62	15
x5	1.49	5	1.61	3	1.52	3	27.22	61	3.03	5
x6	1.61	7	1.59	5	1.51	1	62.95	101	5.62	11
x7	1.69	3	1.81	3	1.64	3	21.25	43	3.13	9
x8	1.84	5	1.85	3	1.93	5	9.72	17	2.71	7
x9	1.74	3	1.86	3	1.95	7	15.82	37	3.15	11

Table 6.9: Posterior inclusion probability $p(\delta_j = 1|\mathbf{y})$ of correlated regressors: inefficiency factor (f) and number of autocorrelations (ms) summed up for computation the inefficiency factor

	p		g		f		n		s	
regressor	ESS	iter	ESS	iter	ESS	iter	ESS	iter	ESS	iter
x1	7572	17	7298	16	9998	22	10000	295	10000	333
x2	6936	15	6231	14	6390	14	52	1	871	29
x3	6628	14	5305	12	6198	14	10000	295	10000	333
x4	6726	15	6513	14	6770	15	252	7	1780	59
x5	6714	15	6205	14	6597	15	367	10	3303	110
x6	6219	14	6305	14	6623	15	158	4	1777	59
x7	5906	13	5531	12	6099	13	470	13	3192	106
x8	5421	12	5406	12	5187	11	1029	30	3692	123
x9	5740	12	5364	12	5116	11	632	18	3174	105

Table 6.10: Posterior inclusion probability $p(\delta_j = 1|\mathbf{y})$ of correlated regressors: effective sample size (ESS) and ESS per second (iter)

model	x1	x2	x3	x4	x5	x6	x7	x8	x9	h_10000_p	p_mM_p
1	1	1	1	0	0	0	0	0	0	0.298	0.302
2	1	1	1	1	0	0	0	0	0	0.170	0.165
3	1	1	1	0	0	1	0	0	0	0.153	0.151
4	1	1	1	0	1	0	0	0	0	0.073	0.074
5	1	1	1	0	0	0	1	0	0	0.039	0.038
6	1	1	1	1	0	1	0	0	0	0.026	0.026
7	1	1	1	0	0	1	0	0	1	0.018	0.018
8	1	1	1	1	1	0	0	0	0	0.015	0.018
9	1	1	1	0	0	1	1	0	0	0.017	0.016
10	1	1	1	0	1	1	0	0	0	0.015	0.016

Table 6.11: Correlated regressors under independence prior: observed frequencies (h_10000_p) and probability (p_mM_p) of different models.

model	x1	x2	x3	x4	x5	x6	x7	x8	x9	h_10000_g	p_mM_g
1	1	1	1	0	0	0	0	0	0	0.192	0.196
2	1	1	1	0	0	1	0	0	0	0.109	0.106
3	1	1	1	1	0	0	0	0	0	0.091	0.091
4	1	1	1	0	1	0	0	0	0	0.058	0.059
5	1	1	1	0	0	0	1	0	0	0.043	0.043
6	1	1	1	0	0	1	0	0	1	0.026	0.028
7	1	0	1	0	0	0	0	0	0	0.024	0.026
8	1	1	1	1	0	1	0	0	0	0.023	0.024
9	1	1	1	0	0	0	0	1	0	0.023	0.023
10	1	1	1	0	0	0	0	0	1	0.024	0.023

Table 6.12: Correlated regressors under the g-prior: observed frequencies (h_10000_g) and probabilities (p_mM_g) of different models.

model	x1	x2	x3	x4	x5	x6	x7	x8	x9	h_10000_f	p_mM_f
1	1	1	1	0	0	1	0	0	0	0.127	0.129
2	1	1	1	1	0	0	0	0	0	0.107	0.101
3	1	1	1	0	0	0	0	0	0	0.089	0.093
4	1	1	1	0	1	0	0	0	0	0.047	0.051
5	1	1	1	0	0	1	0	0	1	0.050	0.049
6	1	1	1	1	0	1	0	0	0	0.037	0.037
7	1	1	1	0	0	0	1	0	0	0.029	0.031
8	1	1	1	0	0	1	0	1	0	0.025	0.025
9	1	1	1	0	0	1	1	0	0	0.024	0.022
10	1	1	1	1	0	0	1	0	0	0.022	0.022

Table 6.13: Correlated regressors under the fractional prior: observed frequencies (h_10000_f) and probability (p_mM_f) of different models.

model	x1	x2	x3	x4	x5	x6	x7	x8	x9	h_p	h_g	h_f	h_n	h_s	p_mM_p
1	1	1	1	0	0	0	0	0	0	0.298	0.192	0.089	0.299	0.443	0.302
2	1	1	1	1	0	0	0	0	0	0.170	0.091	0.107	0.158	0.104	0.165
3	1	1	1	0	0	1	0	0	0	0.153	0.109	0.127	0.186	0.114	0.151
4	1	1	1	0	1	0	0	0	0	0.073	0.058	0.047	0.069	0.057	0.074
5	1	1	1	0	0	0	1	0	0	0.039	0.043	0.029	0.034	0.043	0.038
6	1	1	1	1	0	1	0	0	0	0.026	0.023	0.037	0.022	0.019	0.026
7	1	1	1	0	0	1	0	0	1	0.018	0.026	0.050	0.020	0.013	0.018
8	1	1	1	1	1	0	0	0	0	0.015	0.016	0.017	0.015	0.010	0.018
9	1	1	1	0	0	1	1	0	0	0.017	0.018	0.024	0.011	0.010	0.016
10	1	1	1	0	1	1	0	0	0	0.015	0.017	0.021	0.016	0.013	0.016

Table 6.14: Correlated regressors: Frequencies of the models for different priors. (p_mM) is the model probability under the independence prior.

Chapter 7

Summary and discussion

Variable selection has become an important challenge in statistical analysis. Statisticians are often faced with a large number of potential regressors; however, usually only a small subset of these really have an influence on the response variable. The goal of variable selection is to distinguish between zero effects and non zero effects. In the Bayesian approach a prior is assigned to regression coefficients, which is a mixture of a 'spike' distribution and a flat 'slab' distribution. The spike allows shrinkage of small effects to zero. The probability that a regressor should be included in the final model is given by the posterior inclusion probability of the regressor estimated by the proportion that the coefficient was assigned to the slab component of the prior during MCMC.
In the literature different prior proposals for spike and slab priors have been discussed, e.g. independence prior, Zellner's g-prior, fractional prior, NMIG-prior and SSVS-prior. The goal of this master thesis was to examine the influence of these different prior proposals on variable selection. In the simulations data sets with 40 observations and 9 regressors with strong effects, weak effects and zero effects were generated. We used settings with independent and highly correlated regressors and compared estimation and variable selection performance. Additionally, the performance of MCMC samples was investigated.

The hypothesis when starting this work was that different priors would lead to different results, since they have different location and variance parameters. It was surprising to find, that this was not the case. The results under the different priors were quite similar if the scale setting of the parameters of the prior variances was at least approximately the same.

During simulations the problem of weak regressors turned up. They were apparently chosen too small to be detected. The question arises whether a threshold exists for the size of influential regressors. It seems that beneath a certain value regressors are nearly never detected, beyond that in almost all the cases. By examining only the posterior inclusion probability of a single regressor in our setting weak effects and zero effects could not be well distinguished. Probably only completely different approaches may be able to achieve this.

In the literature often the median probability model is recommended, e.g. in Barbieri and Berger (2004). This means that the selected model includes variables with posterior inclusion probability exceeding 0.5. In the light of our results, the cut-off point of 0.5 seems at least questionable. As we have observed, the size of the variance of the slab component controls the result of variable selection, in the sense that a small prior variance encourages regressor inclusion. Posterior inclusion probabilities increase with smaller prior variance. So by choosing the prior variance small enough, every small coefficient can achieve a posterior indicator of more than 0.5 for a given data set.

In our simulation setting relatively small data sets were used with 40 observations only, therefore it was not reasonable to take much more than 9 regressors. This small setting was chosen because the influence of the respective priors is more pronounced in small data sets with little information than in large informative data sets. However, to compare computational speed further simulation studies with considerably larger data sets would be of interest.

Regarding the efficiency of MCMC under different priors, it was astonishing to observe how fast NMIG prior and SSVS prior work. The CPU time needed was about 1/10 of the time needed by independence, fractional and g-prior. However, the high autocorrelations displayed under the NMIG prior was also an unexpected and unfavourable result. In contrast, the performance of variable selection with the SSVS prior was surprisingly good, it acts both fast and sparsely autocorrelated.

Appendix A

Derivations

A.1 Derivation of the posterior distribution of $\boldsymbol{\beta}$

The posterior distribution $p(\boldsymbol{\beta}, \sigma^2|\mathbf{y})$ of the parameters $(\boldsymbol{\beta}, \sigma^2)$ is proportional to the product of likelihood function and prior distribution, see formula (1.7). It can be simplified by the following derivation, by ignoring constants not depending on $\boldsymbol{\beta}$ and σ^2.

$$
\begin{aligned}
p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) &\propto p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2) \\
&\propto \frac{1}{(\sigma^2)^{N/2}} \exp(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})) \cdot \\
&\quad \frac{1}{(\sigma^2)^{(k+1)/2}|\mathbf{B}_0|^{1/2}} \exp(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\mathbf{b}_0)'\mathbf{B}_0^{-1}(\boldsymbol{\beta}-\mathbf{b}_0)) \cdot \\
&\quad \frac{S_0^{s_0}}{\Gamma(s_0)} \frac{1}{(\sigma^2)^{s_0+1}} \exp(-\frac{S_0}{\sigma^2}) = \\
&= \frac{1}{(\sigma^2)^{(k+1)/2}} \frac{S_0^{s_0}}{\Gamma(s_0)} \frac{1}{|\mathbf{B}_0|^{1/2}} \frac{1}{(\sigma^2)^{N/2+s_0+1}} \cdot \\
&\quad \exp(-\frac{1}{2\sigma^2}((\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta}-\mathbf{b}_0)'\mathbf{B}_0^{-1}(\boldsymbol{\beta}-\mathbf{b}_0))) \exp(-\frac{S_0}{\sigma^2}) = \\
&= const \frac{1}{(\sigma^2)^{(k+1)/2}} \frac{1}{(\sigma^2)^{N/2+s_0+1}} \cdot \\
&\quad \exp(-\frac{1}{2\sigma^2}(\mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{B}_0^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{B}_0^{-1}\mathbf{b}_0 + \mathbf{b}_0'\mathbf{B}_0^{-1}\mathbf{b}_0)) \exp(-\frac{S_0}{\sigma^2}) = \\
&= const \frac{1}{(\sigma^2)^{(k+1)/2}} \frac{1}{(\sigma^2)^{N/2+s_0+1}} \cdot \\
&\quad \exp(-\frac{1}{2\sigma^2}(\mathbf{y}'\mathbf{y} + \boldsymbol{\beta}' \underbrace{(\mathbf{X}'\mathbf{X} + \mathbf{B}_0^{-1})}_{\mathbf{B}_N^{-1}} \boldsymbol{\beta} - 2\boldsymbol{\beta}'(\mathbf{X}'\mathbf{y} + \mathbf{B}_0^{-1}\mathbf{b}_0) + \mathbf{b}_0'\mathbf{B}_0^{-1}\mathbf{b}_0)) \exp(-\frac{S_0}{\sigma^2}) =
\end{aligned}
$$

$$
\begin{aligned}
&= const\frac{1}{(\sigma^2)^{(k+1)/2}}\frac{1}{(\sigma^2)^{N/2+s_0+1}}\cdot \\
&\quad \exp(-\frac{1}{2\sigma^2}(\mathbf{y}'\mathbf{y}+\boldsymbol{\beta}'\mathbf{B}_N^{-1}\boldsymbol{\beta}-2\boldsymbol{\beta}'\mathbf{B}_N^{-1}\underbrace{\mathbf{B}_N(\mathbf{X}'\mathbf{y}+\mathbf{B}_0^{-1}\mathbf{b}_0)}_{\mathbf{b}_N}+\mathbf{b}_0'\mathbf{B}_0^{-1}\mathbf{b}_0))\exp(-\frac{S_0}{\sigma^2}) = \\
&= const\frac{1}{(\sigma^2)^{(k+1)/2}}\frac{1}{(\sigma^2)^{N/2+s_0+1}}\cdot \\
&\quad \exp(-\frac{1}{2\sigma^2}(\mathbf{y}'\mathbf{y}+\boldsymbol{\beta}'\mathbf{B}_N^{-1}\boldsymbol{\beta}-2\boldsymbol{\beta}'\mathbf{B}_N^{-1}\mathbf{b}_N+\mathbf{b}_0'\mathbf{B}_0^{-1}\mathbf{b}_0))\exp(-\frac{S_0}{\sigma^2}) = \\
&= const\frac{1}{(\sigma^2)^{(k+1)/2}}\frac{1}{(\sigma^2)^{N/2+s_0+1}}\cdot \\
&\quad \exp(-\frac{1}{2\sigma^2}(\mathbf{y}'\mathbf{y}+(\boldsymbol{\beta}-\mathbf{b}_N)'\mathbf{B}_N^{-1}(\boldsymbol{\beta}-\mathbf{b}_N)-\mathbf{b}_N'\mathbf{B}_N^{-1}\mathbf{b}_N+\mathbf{b}_0'\mathbf{B}_0^{-1}\mathbf{b}_0))\exp(-\frac{S_0}{\sigma^2}) = \\
&= const\frac{1}{(\sigma^2)^{(k+1)/2}}\frac{1}{(\sigma^2)^{N/2+s_0+1}}\cdot \\
&\quad \exp(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\mathbf{b}_N)'\mathbf{B}_N^{-1}(\boldsymbol{\beta}-\mathbf{b}_N)-\frac{1}{\sigma^2}(\underbrace{\frac{1}{2}(\mathbf{y}'\mathbf{y}-\mathbf{b}_N'\mathbf{B}_N^{-1}\mathbf{b}_N+\mathbf{b}_0'\mathbf{B}_0^{-1}\mathbf{b}_0)+S_0}_{S_N})) = \\
&= const\frac{1}{(\sigma^2)^{(k+1)/2}}\exp(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\mathbf{b}_N)'\mathbf{B}_N^{-1}(\boldsymbol{\beta}-\mathbf{b}_N))\underbrace{\frac{1}{(\sigma^2)^{N/2+s_0+1}}}_{\frac{1}{(\sigma^2)^{s_N+1}}}\exp(-\frac{S_N}{\sigma^2}) = \\
&\propto f_N(\boldsymbol{\beta};\mathbf{b}_N,\mathbf{B}_N\sigma^2)f_{G^{-1}}(\sigma^2;s_N,S_N)
\end{aligned}
$$

with parameters

$$
\begin{aligned}
\mathbf{B}_N &= (\mathbf{X}_1'\mathbf{X}_1+\mathbf{B}_0^{-1})^{-1} \\
\mathbf{b}_N &= \mathbf{B}_N(\mathbf{X}_1'\mathbf{y}+\mathbf{B}_0^{-1}\mathbf{b}_0) \\
s_N &= s_0+N/2 \\
S_N &= S_0+\frac{1}{2}(\mathbf{y}\mathbf{y}+\mathbf{b}_0\mathbf{B}_0^{-1}\mathbf{b}_0-\mathbf{b}_N\mathbf{B}_N^{-1}\mathbf{b}_N)
\end{aligned}
$$

A.2 Derivation of the full conditionals of $\boldsymbol{\beta}$ and σ^2

The full conditional distributions of $\boldsymbol{\beta}$ and σ^2 are proportional to the joint posterior $p(\boldsymbol{\beta},\sigma^2|\mathbf{y})$. By ignoring factors not depending on the parameter of interest, it can be seen, that the full conditionals of $\boldsymbol{\beta}$ and σ^2 are proportional to a multivariate Gaussian distribution and a inverse gamma distribution, respectively.

$$
\begin{aligned}
p(\boldsymbol{\beta}|\sigma^2, \mathbf{y}) &\propto p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) = \\
&\quad f_N(\boldsymbol{\beta}; \mathbf{b}_N, \mathbf{B}_N\sigma^2) f_{G^{-1}}(\sigma^2; s_N, S_N) \\
&\propto f_N(\boldsymbol{\beta}; \mathbf{b}_N, \mathbf{B}_N\sigma^2) \text{ since } f_{G^{-1}}(\sigma^2; s_N, S_N) \text{ does not depend on } \boldsymbol{\beta}
\end{aligned}
$$

$$
\begin{aligned}
p(\sigma^2|\boldsymbol{\beta}, \mathbf{y}) &\propto p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) \\
&\propto \frac{1}{(\sigma^2)^{(\underbrace{N/2 + (k+1)/2 + s_0}_{s_N} + 1)}} \cdot \\
&\quad exp(-\frac{1}{\sigma^2} \underbrace{(\frac{1}{2}((\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \mathbf{b}_0)'\mathbf{B}_0^{-1}(\boldsymbol{\beta} - \mathbf{b}_0)) + S_0)}_{S_N}) \\
&\propto f_{G^{-1}}(\sigma^2|s_N, S_N)
\end{aligned}
$$

A.3 Derivation of the ridge estimator

To obtain the ridge estimator $\hat{\boldsymbol{\beta}}_{ridge}$ the derivative of the penalized residual sum of squares (PRSS)

$$\text{PRSS}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{\beta}$$

with respect to $\boldsymbol{\beta}$ is set to zero:

$$
\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\beta}}(PRSS) &= \frac{\partial}{\partial\boldsymbol{\beta}}((\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{\beta}) \\
&= \frac{\partial}{\partial\boldsymbol{\beta}}(\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}'\boldsymbol{\beta}) \\
&= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + 2\lambda\boldsymbol{\beta} \\
\frac{\partial}{\partial\boldsymbol{\beta}}(PRSS) &= 0 \Leftrightarrow \mathbf{X}'\mathbf{y} - (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta} = 0 \Leftrightarrow \hat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}.
\end{aligned}
$$

Appendix B

R codes

B.1 Estimation and variable selection under the independence prior

```
#Gibbs sampling to perform variable selection and coefficient
estimation under independence prior on the regression coefficients


#parameter of the function:

#X ... design matrix of centered coefficients, without the intercept

#y ... centered dependent variable

#m ... number of iterations

#y_orig ... dependent variable (without centering)

#c ... scaling parameter for the covariance matrix of the
coefficient prior

#values returned by the function:

#b ... matrix with estimated coefficients by each iteration of the
algorithm, the first column is the estimated intercept

#s2 ... vector of estimated error variance

#de ... matrix with the indicator variables: d[i,j]=1 means that
regressor xj is included in the model in iteration i

#post_delta_EQ_1 ... matrix with the posterior inclusion
```

```
probabilities of the regressors p(delta_j=1|y)



regIndPrior_J<-function(X,y,m,y_orig,c){
   N=length(X[,1])                #number of observations
   k=length(X[1,])                #number of columns=k (without intercept!)

   #prior constants:
   A0=diag(k)*c
   a0=matrix(0,k)
   s0=0                           #prior parameter for s2~G-1(s0,S0)
   S0=0

   #for storing the results:      #attention: actually the dimension of the models
                                  # is k(without intercept)
   s2=matrix(0,m)
   b=matrix(0,m,k)
   de=matrix(0,m,k)
   post_delta_EQ_1=matrix(0,m,k)
   mu=c()
   delta=c(rep(FALSE,k))          #starting value for delta

   for(i in 1:m){

     #step1: sampling each component of delta conditional delta without_j

     sam=sample(1:k,k)             #random permutation of the column numbers 1,2...k

     for(l in 1:k){
       log_ml_yd_j=c()
       log_p_delta=c()                     #log(p(delta))
       log_post_delta=c()

       for(h in 0:1){
         delta[sam[l]]=h==1                # delta[sampled order] is set 0 or 1 alternately

         log_ml_yd_j[h+1]=marlik_P(a0,A0,s0,S0,y,X,delta)
         log_p_delta[h+1]=lbeta(sum(delta)+1,k-sum(delta)+1)
         log_post_delta[h+1]=log_ml_yd_j[h+1]+log_p_delta[h+1]
       }
       max_ml=max(log_post_delta)
       e=exp(log_post_delta-max_ml)
       post_delta=e/sum(e)

       delta[sam[l]]=sample(c(0,1),1,replace=TRUE,post_delta)

       delta=delta==TRUE;
```

```
      post_delta_EQ_1[i,sam[l]]=post_delta[2]
    }

    de[i,]=delta       #storing the chosen model

    #step2:sample s2 conditional on delta  from the G-1(sN_d,SN_d) distribution:

    if(sum(delta)){                      #calculating the new constants only
                                         #if at least one column/row is in the design-matrix
      X_d=X[,delta]
      a0_d=a0[delta,]
      if(is.matrix(A0[delta,delta])){
        A0_d=A0[delta,delta]
      }else {
        A0_d=matrix(A0[delta,delta])
      }
      invA0_d=solve(A0_d)
      AN_d=solve(t(X_d)%*%X_d+invA0_d)
      aN_d=AN_d%*%(t(X_d)%*%y+invA0_d%*%a0_d)
      sN=s0+(N-1)/2
      SN_d=S0+0.5*(t(y)%*%y+t(a0_d)%*%invA0_d%*%a0_d-t(aN_d)%*%solve(AN_d)%*%aN_d)
    } else{
      AN_d=1
      A0_d=1
      sN=s0+(N-1)/2
      SN_d=S0+0.5*(t(y)%*%y)
    }

    s2[i]=1/rgamma(1,sN,SN_d)

    #step3:sample the intercept mu from N(yquer,S2/N):
    mu[i]=rnorm(1,mean(y_orig),sqrt(s2[i]/N))

    #step4:sample the reg.coefficients b[,i] in one block from N(aN_d,AN_d*s2)
    if(sum(delta)){
      b[i,delta]=mvrnorm(1,aN_d,s2[i]*AN_d)
    }else{
      b[i,]= rep(0,k)
    }
  }

  b=cbind(mu,b)
  Liste=list(b=b,s2=s2,de=de,post_delta_EQ_1=post_delta_EQ_1)
  return(Liste)
}

#computation of the marginal likelihood p(y|delta)for a given delta:
marlik_P<- function(a0,A0,s0,S0,y,X,delta){
  N=length(X[,1])
```

```
  if(sum(delta)){                    #calculating the new constants only if
    X_d=X[,delta]                    #at least one column/row is in the design-matrix
    a0_d=a0[delta,]
    if(is.matrix(A0[delta,delta])){
      A0_d=A0[delta,delta]
    }else {
      A0_d=matrix(A0[delta,delta])
    }
    invA0_d=solve(A0_d)
    AN_d=solve(t(X_d)%*%X_d+invA0_d)
    aN_d=AN_d%*%(t(X_d)%*%y+invA0_d%*%a0_d)
    sN=s0+(N-1)/2
    SN_d=S0+0.5*(t(y)%*%y+t(a0_d)%*%invA0_d%*%a0_d-t(aN_d)%*%solve(AN_d)%*%aN_d)

    log_ml_yd=-0.5*log(N)-(N-1)*0.5*(log(2*pi))+0.5*(log(det(AN_d)))-0.5*(log(det(A0_d)))+
              +lgamma(sN)-lgamma(s0+1)+s0*(log(S0+1))-sN*(log(SN_d));
  } else{
    AN_d=matrix(1,1,1)
    A0_d=matrix(1,1,1)
    sN=s0+(N-1)/2
    SN_d=S0+0.5*(t(y)%*%y)

    log_ml_yd=-0.5*log(N)-(N-1)*0.5*(log(2*pi))+0.5*(log(det(AN_d)))-0.5*(log(det(A0_d)))+
              +lgamma(sN)-sN*(log(SN_d));
  }
  return(log_ml_yd)
}
```

B.2 Estimation and variable selection under Zellner's g-prior

```
#Gibbs sampling to perform variable selection and coefficient
estimation under Zellner's g-prior on the regression coefficients


#parameter of the function:

#X ... design matrix of centered coefficients, without the intercept

#m ... number of iterations

#y ... dependent variable (without centering)

#g ... scaling parameter for the covariance matrix of the
coefficient prior
```

```
#values returned by the function:

#b ... matrix with estimated coefficients by each iteration of the
algorithm, the first column is the estimated intercept

#s2 ... vector of estimated error variance

#de ... matrix with the indicator variables: d[i,j]=1 means that
regressor xj is included in the model in iteration i

#post_delta_EQ_1 ... matrix with the posterior inclusion
probabilities of the regressors p(delta_j=1|y)


regGPrior_J<-function(X,y,g,m){
  N=length(X[,1])
  k=length(X[1,])

  #prior constants:
  s0=0
  S0=0

  #for storing the results:         #Attention:actually the dimension of the models is k(without intercept)
   s2=matrix(0,m)
   b=matrix(0,m,k)
   de=matrix(0,m,k)
   post_delta_EQ_1=matrix(0,m,k)
   mu=c()

  #starting value for delta
   delta=c(rep(FALSE,k))

 #starting  Gibbs-Sampling:
  for(i in 1:m){
    #cat("g: ",i,"\n")
    #step1: sampling each component of delta conditional delta without_j

    sam=sample(1:k,k)            #random permutation of the column numbers 1,2...k(without intercept)
    for(l in 1:k){
      log_ml_yd_j=c()
      log_p_delta=c()                              #log(p(delta))
      log_post_delta=c()

      for(h in 0:1){
        delta[sam[l]]=h==1
        log_ml_yd_j[h+1]=marlik_Z(y,X,delta,g);
        log_p_delta[h+1]=lbeta(sum(delta)+1,k-sum(delta)+1)          #for hierarchical prior
```

```
        log_post_delta[h+1]=log_ml_yd_j[h+1]+log_p_delta[h+1]
      }
      max_ml=max(log_post_delta)
      e=exp(log_post_delta-max_ml)
      post_delta=e/sum(e)

      delta[sam[l]]=sample(c(0,1),1,replace=TRUE,post_delta)
      delta=delta==TRUE;
      post_delta_EQ_1[i,sam[l]]=post_delta[2]
    }
    de[i,]=delta       #storing the chosen model

    #step2:sample s2 from the G-1((N-1)/2,S/2)distribution
    if(sum(delta)){
      X_d=X[,delta]
      S=sum((y-mean(y))^2)-g*t(y-mean(y))%*%X_d%*%solve(t(X_d)%*%X_d)%*%t(X_d)%*%(y-mean(y))/(g+1)
    }else{
      S=sum((y-mean(y))^2)
    }
    s2[i]=1/rgamma(1,(N-1)/2,S/2)

    #step3: sample mu from N(mean(y),s2/N)
    mu[i]=rnorm(1,mean(y),sqrt(s2[i]/N))

    #step4: sample b from N(g*solve(t(Xc_d)%*%Xc_d)%*%t(Xc_d)%*%(y-mean(y))/(1+g);solve(t(Xc_d)%*%Xc_d)*g*s2/(1+g))
    if(sum(delta)){
      b[i,delta]=mvrnorm(1,g*solve(t(X_d)%*%X_d)%*%t(X_d)%*%(y-mean(y))/(1+g),g*s2[i]*solve(t(X_d)%*%X_d)/(1+g))
    }else{
      b[i,]=rep(0,k)
    }
  }
  b=cbind(mu,b)
  Liste=list(b=b,s2=s2,de=de,post_delta_EQ_1=post_delta_EQ_1)
  return(Liste)
}


marlik_Z <- function(y,X,delta,g){
  N=length(X[,1])
  g=g
  if(sum(delta)){
    X_d=X[,delta]
    S=sum((y-mean(y))^2)-g*t(y-mean(y))%*%X_d%*%solve(t(X_d)%*%X_d)%*%t(X_d)%*%(y-mean(y))/(g+1)
  }else{
    S=sum((y-mean(y))^2)
  }

  log_ml_yd=-(sum(delta))/2*log(1+g)+lgamma((N-1)/2)-0.5*log(N)-(N-1)/2*log(pi)-(N-1)*0.5*log(S)
  return(log_ml_yd)
```

```
}
```

B.3 Estimation and variable selection under the fractional prior

```
#Gibbs sampling to perform variable selection and coefficient
estimation under fractional prior on the regression coefficients


#parameter of the function:

#X ... design matrix of centered coefficients, without the intercept

#m ... number of iterations

#y ... dependent variable (without centering)

#f ... scaling parameter for the covariance matrix of the
coefficient prior (f=1/b)

#s0,S0 ... prior parameters for the error variance s2~G-1(s0,S0)


#value returned by the function:

#b ... matrix with estimated coefficients by each iteration of the
algorithm, the first column is the estimated intercept

#s2 ...vector of estimated error variance

#de ... matrix with the indicator variables: d[i,j]=1 means that
regressor xj is included in the model in iteration i

#post_delta_EQ_1 ... matrix with the posterior inclusion
probabilities of the regressors p(delta_j=1|y)


regFracPrior_J<-function(y,X,f,m){
  N=length(X[,1])                #number of observations
  k=length(X[1,])

  #prior constants:
  s0=0
  S0=0

  #for storing the results:
  s2=matrix(0,m)
```

```
b=matrix(0,m,k)
de=matrix(0,m,k)
post_delta_EQ_1=matrix(0,m,k)
mu=c()

#starting value for delta
delta=c(rep(FALSE,k))

#starting  Gibbs-Sampling:
for(i in 1:m){
  #cat("f: ",i,"\n")
  #step1: sampling each component of delta conditional delta without_j

  sam=sample(1:k,k)                      #random permutation of the column numbers 1,2...k

  for(l in 1:k){
    log_ml_yd_j=c()
    log_p_delta=c()                              #log(p(delta))
    log_post_delta=c()

    for(h in 0:1){
      delta[sam[l]]=h==1
      log_ml_yd_j[h+1]=marlik_F(s0,S0,y,X,delta,f)

      #log_p_delta[h+1]=sum(delta)*log(p)+(k-sum(delta))*log(1-p)
      log_p_delta[h+1]=lbeta(sum(delta)+1,k-sum(delta)+1)

      log_post_delta[h+1]=log_ml_yd_j[h+1]+log_p_delta[h+1]
    }
    max_ml=max(log_post_delta)
    e=exp(log_post_delta-max_ml)
    post_delta=e/sum(e)

    delta[sam[l]]=sample(c(0,1),1,replace=TRUE,post_delta)
    delta=delta==TRUE;
    post_delta_EQ_1[i,sam[l]]=post_delta[2]

  }
  de[i,]=delta       #storing the chosen model

  #step2:sample s2 from the G-1(sN,SN_d) distribution:
  sN=s0+(N-1)*(1-f)/2
  if(sum(delta)){
    X_d=X[,delta]
    AN_d=solve(t(X_d)%*%X_d)
    aN_d=AN_d%*%(t(X_d)%*%y)
    SN_d=S0+0.5*(1-f)*(t(y-mean(y))%*%(y-mean(y))-t(aN_d)%*%solve(AN_d)%*%aN_d)
  }else{
    SN_d=S0+0.5*(1-f)*(t(y-mean(y))%*%(y-mean(y)))
```

```
    }
    s2[i]=1/rgamma(1,sN,SN_d)

    #step3:sample the intercept mu from N(yquer,S2/N):
     mu[i]=rnorm(1,mean(y),sqrt(s2[i]/N))

    #step4:sample the reg.coefficients b[,i] in one block from N(aN_d,AN_d*s2)
     if(sum(delta)){
       b[i,delta]=mvrnorm(1,aN_d,s2[i]*AN_d)
     }else{
       b[i,]= rep(0,k)
     }
  }
  b=cbind(mu,b)
  Liste=list(b=b,s2=s2,de=de,post_delta_EQ_1=post_delta_EQ_1)
  return(Liste)
}

marlik_F <- function(s0,S0,y_orig,X,delta,f){
  N=length(X[,1])
  f=f
  sN=s0+(N-1)*(1-f)/2
  if(sum(delta)){
    X_d=X[,delta]
    AN_d=solve(t(X_d)%*%X_d)
    aN_d=AN_d%*%t(X_d)%*%(y-mean(y))
    SN_d=S0+0.5*(1-f)*(t((y-mean(y)))%*%(y-mean(y))-t(aN_d)%*%solve(AN_d)%*%aN_d)
  }else{
    SN_d=S0+0.5*(1-f)*(t((y-mean(y)))%*%(y-mean(y)))
  }
  log_ml_yd=-(N-1)*(1-f)*0.5*(log(2*pi))+(sum(delta)/2)*log(f)+
              +lgamma(sN)-sN*(log(SN_d));
  return(log_ml_yd)
}
```

B.4 Estimation and variable selection under the SSVS prior

```
#Gibbs sampling to perform variable selection and coefficient
estimation under SSVS prior on the regression coefficients


#parameter of the function:

#X ... design matrix of centered coefficients, without the intercept

#y ... dependent variable
```

```
#m ... number of iterations

#tau2 ... prior parameter for alpha~nu*N(0,tau2)+(1-ny)*N(0,cTau2)

#values returned by the function:

#b ... matrix with estimated coefficients by each iteration of the
algorithm, the first column is the estimated intercept

#s2 ... vector of estimated error variance

#ny ... matrix with the indicator variables: ny[i,j]=1 means that on
iteration i coefficient alpha_j is sampled from the the flat
distribution N(0,ny_j*psi2)

#post_delta_EQ_1 ... matrix with the posterior inclusion
probabilities of the regressors p(ny_j=1|y)


regSSVS<-function(y,X,m,tau2){
  N=length(X[,1])
  k=length(X[1,])

  #prior constants:
  c=1/1000

  #matrices for storing the results:
  mu=c()                              #vector for the intercept
  b=matrix(0,m,k)                     #matrix for regression coefficients (without intercept)
  ny=matrix(0,m,k)                    #matrix for the mixture components
  w=matrix(0,m)                       #matrix for the mixture weights
  post_delta_EQ_1=matrix(0,m,k)
  s2=c()

  #starting values:
  mu[1]=1
  b[1,]=rep(1,k)
  ny[1,]=rep(1,k)
  w[1]=0.5
  s2[1]=5

  #MCMC-steps:
  for(i in 2:m) {
    #print(i)
    for(j in 1:k){

      #(i)sampling the mixture component ny[i,j]:
      p0=(1-w[i-1])*dnorm(b[i-1,j],0,sqrt(c*tau2))
```

```
      p1=w[i-1]*dnorm(b[i-1,j],0,sqrt(tau2))
      ny[i,j]=sample(c(0,1),1,prob=c(p0,p1))
      post_delta_EQ_1[i,j]=p1/(p0+p1)
    }

    #(ii)sampling the mixture weigth w[i]:
    w[i]=rbeta(1,1+sum(ny[i,]==1),1+sum(ny[i,]==0))

    #(iii)sampling the regressor coefficients
    D=diag(tau2*(ny[i,]+(1-ny[i,])*c))
    AN=solve(t(X)%*%X*(1/(s2[i-1]))+solve(D))
    aN=AN%*%t(X)%*%(y-mean(y))*(1/s2[i-1])
    b[i,]=mvrnorm(1,aN,AN)

    #(iv)sampling the errot variance s2:
    sN=0+(N-1)/2
    SN=0+0.5*(t((y-mean(y))-X%*%t(t(b[i,])))%*%((y-mean(y))-X%*%t(t(b[i,]))))
    s2[i]=1/rgamma(1,sN,SN)

    #(v)sampling the common mean mu:
    mu[i]=rnorm(1,mean(y),s2[i]/N)

  }
  b=cbind(mu,b)
  Liste=list(b=b,ny=ny,s2=s2,post_delta_EQ_1=post_delta_EQ_1)
  return(Liste)
}
```

B.5 Estimation and variable selection under the NMIG prior

```
#Gibbs sampling to perform variable selection and coefficient
estimation under NMIG prior on the regression coefficients


#parameter of the function:

#X ... design matrix of centered coefficients, without the intercept

#y ... dependent variable

#m ... number of iterations

#apsi0,bpsi0 ... prior parameter for psi~G-1{apsi0,bpsi0)
```

```
#value returned by the function:

#b ... matrix with estimated coefficients by each iteration of the
algorithm, the first column is the estimated intercept

#s2 ... vector of estimated error variance

#ny_w ... matrix with the indicator variables: ny_w[i,j]=1 means
that on iteration i coefficient alpha_j is sampled from the the flat
distribution N(0,ny_j*psi2)


#post_delta_EQ_1 ... matrix with the posterior indicator
probabilities of the regressors p(delta_j=1|y)


regNMIG<-function(y,X,m,apsi0,bpsi0){
  N=length(X[,1])
  k=length(X[1,])


  #prior constants:
  s0=0.000025
  s1=1


  #matrices for storing the results:
  mu=c()                              #vector for the intercept
  b=matrix(0,m,k)                     #matrix for regression coefficients (without intercept)
  ny=matrix(0,m,k)                    #matrix for the mixture components
  psi2=matrix(0,m,k)                  #matrix for the scales
  w=matrix(0,m)                       #matrix for the mixture weights
  s2=c()
  post_delta_EQ_1=matrix(0,m,k)

  #starting values:
  mu[1]=1
  b[1,]=rep(1,k)
  ny[1,]=rep(1,k)
  psi2[1,]=rep(1,k)
  w[1]=0.5
  s2[1]=5

  #MCMC-steps:
  for(i in 2:m) {
    #cat("nmig: ",i,"\n")
    for(j in 1:k){

      #(i)sampling the mixture component ny[i,j]:
```

```
      f0=(1-w[i-1])*dnorm(b[i-1,j],0,sqrt(s0*psi2[i-1,j]))
      f1=w[i-1]*dnorm(b[i-1,j],0,sqrt(s1*psi2[i-1,j]))

      ny[i,j]=sample(c(s0,s1),1,prob=c(f0,f1))

      post_delta_EQ_1[i,j]=f1/(f0+f1)


      #(ii)sampling the scale psi2[i,j]:
      psi2[i,j]=1/rgamma(1,apsi0+0.5,bpsi0+(b[i-1,j]*b[i-1,j])/(2*ny[i,j]))
    }

    #(iii)sampling the mixture weight w[i]:
    w[i]=rbeta(1,1+sum(ny[i,]==s1),1+sum(ny[i,]==s0))

    #(iv)sampling the regressor coefficients
    D=diag(psi2[i,])%*%diag(ny[i,])
    AN=solve(t(X)%*%X*(1/(s2[i-1]))+solve(D))
    aN=AN%*%t(X)%*%(y-mean(y))*(1/s2[i-1])
    b[i,]=mvrnorm(1,aN,AN)

    #(v)sampling the error variance s2:
    sN=0+(N-1)/2
    SN=0+0.5*(t((y-mean(y))-X%*%t(t(b[i,])))%*%((y-mean(y))-X%*%t(t(b[i,]))))
    s2[i]=1/rgamma(1,sN,SN)

    #(vi)sampling the common mean mu:
    mu[i]=rnorm(1,mean(y),sqrt(s2[i]/N))

  }
  b=cbind(mu,b)
  ny_w=(ny==1)
  Liste=list(b=b,ny_w=ny_w,psi2=psi2,s2=s2, post_delta_EQ_1= post_delta_EQ_1)
  return(Liste)
}
```

Bibliography

Barbieri, M. and O.J. Berger (2004). Optimal predictive model selection. *Annals of Statistics 32*(3), 870–897.

Brown, P.J., M. Vannucci and T.Fearn (2002). Bayes model averaging with variable selection of regression. *Journal of the Royal Statistical Society, Series B 64*(3), 519–536.

Fahrmeir, L., A. Hammerle and G. Tutz (1996). *Multivariate statistische Verfahren*. De Gruyter. Berlin.

Fahrmeir, L., T. Kneib and Susanne Konrath (2010). Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing 20*(2), 203–219.

Fahrmeir, L., T. Kneib and S. Lang (2007). *Regression. Modelle, Methoden und Anwendungen*. Springer. Heidelberg.

Frühwirth-Schnatter, S. and R. Tüchler (2008). Bayesian parsimonious covariance estimation for hierarchical linear mixed models. *Statistics and Computing 18*(1), 1–13.

George, E. and Y. Maruyama (2010). Fully Bayes model selection with a generalized g-prior. Conference 'Frontiers of Statistical Decisions Making and Bayesian Analysis',University of Texas, San Antonio. March 17-20, 2010.

George, E.I. and R.E. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the Amercan Statistical Association 88*(423), 881–889.

Geyer, C.J. (1992). Practical Markov Chain Monte Carlo. *Statistical Science 7*(4), 473–511.

Ishwaran, H. and J. Rao (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *Journal of the American Statistical Association 98*(1), 438–455.

Kass, R.E. and A.E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*(1), 773–795.

Konrath, S., T. Kneib and L. Fahrmeir (2008). *Bayesian regularisation in structured additive regression models for survival data*. Technical report.

Ley, E. and M.F.J. Steel (2007). Jointness in Bayesian variable selection with application to growth regression. *Journal of Macroeconomics 29*(1), 476–493.

Liang, F., R. Paulo, G. Molina, C.A. Clyde and J.O. Berger (2008). Mixtures of g Priors for Bayesian Variable Selection. *Journal of the American Statistical Association 103*(1), 410–423.

Maruyama, Y. and E. George (2008). A g-prior extension for p>n. http://arxiv.org/abs/0801.4410v1. Submitted on 29 Jan 2008.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society B 57*(1), 99–138.

O'Hara, R. B. and M. J. Sillanpää (2009). Review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis 4*(1), 85–118.

Swartz, M.D., R.K. Yu and S. Shete (2008). Finding factors influencing risk: Comparing Bayesian stochastic search and standard variable selection methods applied to logistic regression models of cases and controls. *Statistics in Medicine 27*(29), 6158–6174.

Tibshirani, R. (1996). Regression shrinkage and selection via lasso. *Journal of the Royal Society B 58*(1), 267–288.

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 90*(1), 233–243.