BAYESIAN STATISTICS 9,
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)
© Oxford University Press, 2010

Bayesian Variable Selection for Random Intercept Modeling of Gaussian and non-Gaussian Data

Sylvia Frühwirth-Schnatter & Helga Wagner
Department of Applied Statistics and Econometrics, Johannes Kepler Universität Linz, Austria
sylvia.fruehwirth-schnatter@jku.at helga.wagner@jku.at

1. INTRODUCTION

The paper considers Bayesian variable selection for random intercept models both for Gaussian and non-Gaussian data. For Gaussian data the model reads

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\alpha} + \beta_i + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right), \tag{1}$$

where y_{it} are repeated responses observed for N units (e.g. subjects) $i = 1, \ldots, N$ on T_i occasions $t = 1, \ldots, T_i$. $\mathbf{x}_{it}$ is the $(1 \times d)$ design matrix for an unknown regression coefficient $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)'$ of dimension d, including the overall intercept. For each unit, β_i is a subject specific deviation from the overall intercept.

For efficient estimation it is necessary to specify the distribution of heterogeneity $p(\beta_1, \ldots, \beta_N)$. As usual we assume that $\beta_1, \ldots, \beta_N|\boldsymbol{\theta}$ are independent given a random hyper parameter $\boldsymbol{\theta}$ with prior $p(\boldsymbol{\theta})$. Marginally, the random intercepts $\beta_1, \ldots, \beta_N$ are dependent and $p(\beta_1, \ldots, \beta_N)$ acts a smoothing prior which ties the random intercepts together and encourages shrinkage of β_i toward the overall intercept by "borrowing strength" from observations of other subjects. A very popular choice is the following standard random intercept model:

$$\beta_i|Q \sim \mathcal{N}(0, Q), \qquad Q \sim \mathcal{G}^{-1}(c_0, C_0), \tag{2}$$

which is based on assuming conditional normality of the random intercept.

Several papers deal with the issue of specifying alternative smoothing priors $p(\beta_1, \ldots, \beta_N)$, because misspecifying this distribution may lead to inefficient, and for random intercept model for non-Gaussian data, even to inconsistent estimation of the regression coefficient $\boldsymbol{\alpha}$, see e.g. Neuhaus et al. (1992). Recently, Komárek and Lesaffre (2008) suggested to use finite mixture of normal priors for $p(\beta_i|\boldsymbol{\theta})$ to handle this issue. In the present paper we also deviate from the commonly used normal prior (2) and consider more general priors. However, in addition to correct estimation of $\boldsymbol{\alpha}$, our focus will be on Bayesian variable selection.

The Bayesian variable selection approach is commonly applied to a standard regression model where β_i is equal to 0 in (1) for all units and aims at separating

non-zero regression coefficients $\alpha_j \neq 0$ from zero regression coefficients $\alpha_j = 0$. By choosing an appropriate prior $p(\boldsymbol{\alpha})$, it is possible to shrink some coefficients α_r toward 0 and identify in this way relevant coefficients. Common shrinkage priors are spike-and-slab priors (Mitchell and Beauchamp, 1988; George and McCulloch, 1993, 1997; Ishwaran and Rao, 2005), where a spike at 0 (either a Dirac measure or a density with very small variance) is combined in the slab with a density with large variance. Alternatively, unimodal shrinkage priors have been applied like the double exponential or Laplace prior leading to the Bayesian Lasso (Park and Casella, 2008) or the more general normal-gamma prior (Griffin and Brown, 2010); see also Fahrmeir et al. (2010) for a recent review.

Subsequently we consider variable selection for the random intercept model (1). Although this also concerns $\boldsymbol{\alpha}$, we will focus on variable selection for the random effects which, to date, has been discussed only by a few papers. Following Kinney and Dunson (2007), Frühwirth-Schnatter and Tüchler (2008), and Tüchler (2008) we could consider variable selection for the random intercept model as a problem of variance selection. Under prior (2), for instance, a single binary indicator δ could be introduced where $\delta = 0$ corresponds to $Q = 0$, while $\delta = 1$ allows Q to be different from 0. This implicitly implies variable selection for the random intercept, because setting $\delta = 0$ forces all β_i to be zero, while for $\delta = 1$ all random intercepts $\beta_1, \ldots, \beta_N$ are allowed be different from 0.

In the present paper we are interested in a slightly more general variable selection problem for random effects. Rather than discriminating as above between a model where all random effects are zero and a model where all random effects are different from 0, it might be of interest to make unit-specific selection of random effects in order to identify units which are "average" in the sense that they do not deviate from the overall mean, i.e. $\beta_i = 0$, and units which deviate significantly from the "average", i.e. $\beta_i \neq 0$.

In analogy to variable selection in standard regression model, we will show that individual shrinkage for the random effects can be achieved through appropriate selection of the prior $p(\beta_i|\boldsymbol{\theta})$ of the random effects. For instance, if $p(\beta_i|Q)$ is a Laplace rather than a normal prior as in (2) with a random hyperparameter Q, we obtain a Bayesian Lasso random effects models where the smoothing additionally allows individual shrinkage of the random intercept toward 0 for specific units. However, as for a standard regression model too much shrinkage takes place for the non-zero random effects under the Laplace prior. For this reason we investigate alternative shrinkage-smoothing priors for the random intercept model like the spike-and-slab random effects model which is closely related to the finite mixtures of random effects model investigated by Frühwirth-Schnatter et al. (2004) and Komárek and Lesaffre (2008).

2. VARIABLE SELECTION IN RANDOM INTERCEPT MODELS THROUGH SMOOTHING PRIORS

Following standard practice in the econometrics literature, a fixed-effects approach could be applied, meaning that each unit specific parameter β_i is treated just as another regression coefficient and the high dimensional parameter $\boldsymbol{\alpha}^\star = (\boldsymbol{\alpha}, \beta_1, \ldots, \beta_N)$ is estimated from a large regression model without any random intercept:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\alpha}^\star + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right). \tag{3}$$

We could then perform variable selection for $\boldsymbol{\alpha}^\star$ in the large regression model (3), in which case a binary variable selection indicator δ_i is introduced for each random

effect β_i individually. This appears to be the solution to the variable selection problem addressed in the introduction, however, variable selection in (3) is not entirely standard: first, the dimension of $\boldsymbol{\alpha}^\star$ grows with the number N of units; second, an information imbalance between the regression coefficients α_j and the random intercepts β_i is present, because the number of observations is $\sum_{i=1}^{N} T_i$ for α_j, but only T_i for β_i. This make it difficult to choose the prior $p(\boldsymbol{\alpha}^\star)$. Under a (Dirac)-spike-and-slab prior for $p(\boldsymbol{\alpha}^\star)$, for instance, a prior has to be chosen for all non-zero coefficients in $\boldsymbol{\alpha}^\star$. An asymptotically optimal choice in a standard regression model is Zellner's g-prior, however, the information imbalance between α_j and β_i make it impossible to choose a value for g which is suitable for all non-zero elements of $\boldsymbol{\alpha}^\star$.

The information imbalance suggests to choose the prior for the regression coefficients independently from the prior for the random intercepts, i.e. $p(\boldsymbol{\alpha}^\star) = p(\boldsymbol{\alpha})p(\beta_1, \ldots, \beta_N)$. Variable selection for β_i in the large regression model (3) is then controlled through the choice of $p(\beta_1, \ldots, \beta_N)$ which is exactly the same problem as choosing the smoothing in the original random intercept model (1). This motivated us to use common shrinkage priors in Bayesian variable selection as smoothing priors in the random intercept model and to study how this choice effects shrinkage for the random intercept.

Practically all priors have a hierarchical representation where

$$\beta_i|\psi_i \sim \mathcal{N}(0, \psi_i), \qquad \psi_i|\boldsymbol{\theta} \sim p(\psi_i|\boldsymbol{\theta}), \tag{4}$$

$\beta_i|\psi_i$ and $\beta_j|\psi_j$ are independent and $p(\psi_i|\boldsymbol{\theta})$ depends on a hyperparameter $\boldsymbol{\theta}$. The goal is to identify choices of $p(\psi_i|\boldsymbol{\theta})$ which lead to strong shrinkage if many random intercepts are close to zero, but introduce little bias, if all units are heterogeneous.

Note that the marginal distribution

$$p(\beta_i|\boldsymbol{\theta}) = \int p(\beta_i|\psi_i)p(\psi_i|\boldsymbol{\theta})\, d\,\psi_i$$

is non-Gaussian and that the joint density $p(\beta_1, \ldots, \beta_N)$ is smoothing prior in the standard sense only, if at least some components of the hyperparameter $\boldsymbol{\theta}$ are random.

3. VARIABLE SELECTION IN RANDOM INTERCEPT MODELS USING SHRINKAGE SMOOTHING PRIORS

This subsection deals with unimodal non-Gaussian shrinkage priors which put a lot of prior mass close to 0, but have heavy tails. Such a prior encourages shrinkage of insignificant random effects toward 0 and, the same time, allows that the remaining random effects may deviate considerably from 0. For such a prior, the posterior mode of $p(\beta_i|\mathbf{y}_i, \boldsymbol{\theta})$ is typically to 0 with positive probability. We call such a prior a non-Gaussian shrinkage prior.

3.1. *Non-Gaussian Shrinkage Priors*

Choosing the inverted Gamma prior $\psi_i|\nu, Q \sim \mathcal{G}^{-1}(\nu, Q)$ leads to the Student-t random intercept model where

$$\beta_i|\nu, Q \sim t_{2\nu}(0, Q/\nu). \tag{5}$$

While this prior has heavy tails, it does not encourage shrinkage toward 0, because the posterior mode of $p(\beta_i|\mathbf{y}_i, \boldsymbol{\theta})$ is different from 0 with probability 1.

Following the usual approach toward regularization and shrinkage in a standard regression model, we choose $\psi_i|Q \sim \mathcal{E}(1/(2Q))$ which leads to the Laplace random intercept model:

$$\beta_i|Q \sim \text{Lap}\left(\sqrt{Q}\right). \tag{6}$$

Since this model may be considered as a Baysian Lasso random intercept model, we expect a higher degree of shrinkage compared to the Student-t random intercept model. In contrast to the Student-t random intercept model, the Laplace prior puts a lot of prior mass close to 0 and allows that also the posterior $p(\beta_i|\mathbf{y}_i, Q)$ has a mode exactly at 0 with positive probability.

Even more shrinkage may be achieved by choosing the Gamma distribution $\psi_i \sim \mathcal{G}(a, 1/(2Q))$ which has been applied by Griffin and Brown (2010) for variable selection in a standard regression model.[1] It appears sensible to extent such a prior to the random effects part. Evidently, the model reduces to the Laplace model for $a = 1$. The marginal density $p(\beta_i|a, Q)$ is available in closed form, see Griffin and Brown (2010):

$$p(\beta_i|a, Q) = \frac{1}{\sqrt{\pi}2^{a-1/2}Q^{2a+1}\Gamma(a)}|\beta_i|^{a-1/2}K_{a-1/2}(|\beta_i|/Q^2), \tag{7}$$

where K is the modified Bessel function of the third kind. The density $p(\beta_i|a, Q)$ becomes more peaked at zero as a decreases.

An interesting special case is obtained for $a = 1/2$ in which case $\psi_i|Q \sim Q\chi^2_{2a}$, or equivalently, $\sqrt{\psi_i} \sim \mathcal{N}(0, Q)$. In this case, the random intercept model may be written in a non-centered version as:

$$z_i \sim \mathcal{N}(0, 1), \tag{8}$$

$$y_{it} = \mathbf{x}^f_{it}\boldsymbol{\alpha} + \sqrt{\psi_i}z_i + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathcal{N}\left(0, \sigma^2_\varepsilon\right). \tag{9}$$

Hence the normal-Gamma prior with $a = 1/2$ is related to Frühwirth-Schnatter and Tüchler (2008) who consider a similar non-centered version of the random effects model, but assume that $\sqrt{\psi_i} \equiv Q$ follows a normal prior.

3.2. *Hyperparameter Settings*

For any of these shrinkage priors hyperparameters are present. All priors depend on a scaling factor Q and some priors depend, additionally, on a shape parameter. We assume for our investigation that any shape parameters is fixed, because these parameters are in general difficult to estimate. For instance, we fix ν in the Student-t prior (5) to a small integer greater than 2. However, we treat Q as a random hyperparameter with prior $p(Q)$.

In standard regression models shrinkage factors like Q are often selected on a rather heuristic basis and held fixed for inference. In the context of random effects, however, this would imply, that the random effects are independent and no smoothing across units takes place. Hence for variable selection in the random

[1] Note that Griffin and Brown (2010) use a different parameterization.

intercept model it is essential to introduce a prior $p(Q)$ for Q, because this turns a shrinkage prior for an individual random intercept into a smoothing prior across the random intercepts.

To make the priors $p(Q)$ for Q comparable among the various types of shrinkage priors introduced in Subsection 3.1, we follow Griffin and Brown (2010) and put an inverted Gamma prior on the variance $v_\beta = \mathrm{Var}(\beta_i|\boldsymbol{\theta})$ of the distribution of heterogeneity:

$$v_\beta \sim \mathcal{G}^{-1}(c_0, C_0)\,, \tag{10}$$

Due to our parameterization $v_\beta = cQ$ for all shrinkage priors, where c is a distribution specific constant, possibly depending on a shape parameter. Conditional on holding any shape parameter fixed, the prior on v_β immediately translates into an inverted Gamma prior for Q:

$$Q \sim \mathcal{G}^{-1}(c_0, C_0/c)\,. \tag{11}$$

For the normal prior (2), $v_\beta = Q$, hence $c = 1$. For the Laplace prior (6) we obtain $v_\beta = 2Q$ and $c = 2$. For the Student-t prior (5) with $v_\beta = Q/(\nu - 1)$ this induces a conditionally inverted Gamma prior for $Q|\nu$ with $c = 1/(\nu - 1)$. For the normal-Gamma prior where $v_\beta = 2aQ$ this leads a conditionally inverted Gamma prior for $Q|a$ with $c = 2a$.

For the standard regression model, Griffin and Brown (2010) choose $c_0 = 2$, in which case $\mathrm{E}(v_\beta|C_0) = C_0$, while the prior variance is infinite. They select C_0 in a data-based manner as the average of the OLS estimators for each regression coefficient. However, this is not easily extended to random-effects models.

For $a = 0.5$, where $\mathrm{E}(\psi_i) = v_\beta = Q$, the non-centered representation (9) suggests the g-type prior $\sqrt{\psi_i} \sim \mathcal{N}\left(0, g_i \sum_{t=1}^{T_i} z_i^2\right)$ where $g_i = 1/T_i$, hence $\mathrm{E}(\psi_i) = \mathrm{E}(z_i^2)$. This suggests to center the prior of v_β at 1 for random effects. This implies choosing $C_0 = 1$, if $c_0 = 2$. By choosing $c_0 = 0.5$ and $C_0 = 0.2275$ as in Frühwirth-Schnatter and Wagner (2008) we obtain a fairly vague prior with prior median equals 1 which does not have any finite prior moments.

3.3. *Classification*

Shrinkage priors have been introduced because they are the Bayesian counterpart of shrinkage estimators which are derived as penalized ML estimators. For known hyperparameters $\boldsymbol{\theta}$ such priors allow for conditional posterior distributions $p(\beta_1, \ldots, \beta_N|\mathbf{y}, \boldsymbol{\theta})$ where the mode lies at 0 for certain random effects β_i. While this enables variable selection in a non-Bayesian or empirical Bayesian framework, it is not obvious, how to classify the random-effects within a fully Bayesian approach, because, as argued earlier, it appear essential to make at least some hyperparameters random.

As mentioned in the introduction, we would like to classify units into those which are "average" ($\delta_i = 0$) and those which are "above average" ($\delta_i = 1$, $\Pr(\beta_i > 0|\mathbf{y})$) and "below average" ($\delta_i = 1$, $\Pr(\beta_i < 0|\mathbf{y})$). This is useful in a context where a random-effects model is used, for instance, for risk assessment in different hospitals or in evaluation different schools.

To achieve classification for shrinkage priors within a fully Bayesian approach some ad hoc procedure has to be applied. Alternatively, shrinkage priors could be selected in such a way that classification is intrinsic in their formulation.

4. VARIABLE SELECTION IN RANDOM INTERCEPT MODELS USING SPIKE-AND-SLAB SMOOTHING PRIORS

Many researchers found spike-and-slab priors very useful in the context of variable selection for regression models (Mitchell and Beauchamp, 1988; George and McCulloch, 1993, 1997; Ishwaran and Rao, 2005). These priors take the form of a finite mixture distribution with two components where one component (the spike) is centered at 0 and shows little variance compared to the second component (the slab) which has considerably larger variance. Spike-and-slab priors can easily be extended to variable selection for random intercept model and lead a two component mixture prior for β_i:

$$p(\beta_i|\omega,\boldsymbol{\theta}) = (1-\omega)p_{\text{spike}}(\beta_i|\boldsymbol{\theta}) + \omega p_{\text{slab}}(\beta_i|\boldsymbol{\theta}). \tag{12}$$

We assume that β_i, $i = 1,\ldots,N$ are independent a priori conditional on the hyperparameters ω and $\boldsymbol{\theta}$.

Note that we are dealing with another variant of the non-Gaussian random effects model considered in Subsection 3.1, however with an important difference. The finite mixture structure of $p(\beta_i|\omega,\boldsymbol{\theta})$ allows to classify each β_i into one of the two components. Classification is based on a hierarchical version of the mixture model (12) which introduces a binary indicator δ_i for each random intercept:

$$\begin{aligned}&\Pr(\delta_i = 1|\omega) = \omega,\\ &p(\beta_i|\delta_i,\boldsymbol{\theta}) = (1-\delta_i)p_{\text{spike}}(\beta_i|\boldsymbol{\theta}) + \delta_i p_{\text{slab}}(\beta_i|\boldsymbol{\theta}).\end{aligned} \tag{13}$$

4.1. *Using Absolutely Continuous Spikes*

As for variable selection in a standard regression model we have to distinguish between two types of spike-and-slab priors. For the first type the distribution modeling the spike is absolutely continuous, hence the marginal prior $p(\beta_i|\omega,\boldsymbol{\theta})$ is absolutely continuous as well. This has certain computational advantages as outlined in Section 5.

The hyperparameters of the component densities are chosen in such a way that the variance ratio r is considerably smaller than 1:

$$r = \frac{V_{\text{spike}}(\beta_i|\boldsymbol{\theta})}{V_{\text{slab}}(\beta_i|\boldsymbol{\theta})} << 1. \tag{14}$$

Strictly speaking, classification is not possible for a prior with an absolutely continuous spike, because $\delta_i = 0$ is not exactly equivalent to $\beta_i = 0$, but indicates only that β_i is "relatively" close to 0 compared to β_is belonging the second component, because $r << 1$. Nevertheless it is common practice to base classification between zero and non-zero coefficients in a regression model on the posterior inclusion probability $\Pr(\delta_i = 1|\mathbf{y})$ and the same decision rule is applied here for the random intercepts.

The standard spike-and-slab prior for variable selection in a regression model is a two component normal mixture, which this leads to a finite Gaussian mixture as random-effects distribution:

$$\beta_i|\omega,Q \sim (1-\omega)\mathcal{N}(0,rQ) + \omega\mathcal{N}(0,Q). \tag{15}$$

Such finite mixtures of random-effects models have been applied in many areas, see Frühwirth-Schnatter (2006, Section 8.5) for some review. They are useful, because

they allow very flexible modeling of the distribution of heterogeneity. We explore in this paper, how they relate to variable selection for random-effects. Note that this prior may be restated in terms of the hierarchical scale mixture prior (4) where ψ_i switches between the two values rQ and Q according to ω.

Ishwaran et al. (2001) and Ishwaran and Rao (2005) introduced the NMIG prior for variable selection in a regression model which puts a spike-and-slab prior on the variance of the prior of the regression coefficients. For random intercept model, this suggests to put a spike-and-slab prior on ψ_i in the hierarchical scale mixture prior (4):

$$\psi_i|\omega, Q \sim (1-\omega)p_{\text{spike}}(\psi_i|r, Q) + \omega p_{\text{slab}}(\psi_i|Q). \tag{16}$$

Based on assuming independence of $\psi_1, \ldots, \psi_N$, this choice leads to a marginal spike-and-slab prior for β_i which is a two component non-Gaussian mixture as in (15).

Ishwaran et al. (2001) and Ishwaran and Rao (2005) choose inverted Gamma distributions both for the spike and the slab in $\psi_i|\omega, Q$, i.e. $\psi_i|\delta_i = 0 \sim \mathcal{G}^{-1}(\nu, rQ)$ and $\psi_i|\delta_i = 1 \sim \mathcal{G}^{-1}(\nu, Q)$. Marginally, this leads to a two component Student-t mixture as spike-and-slab prior for β_i:

$$\beta_i|\omega, Q \sim (1-\omega)t_{2\nu}(0, rQ/\nu) + \omega t_{2\nu}(0, Q/\nu). \tag{17}$$

This mixture prior allows discrimination, however, the spike in (17) does not encourage shrinkage. Hence it makes sense to modify the NMIG prior by choosing other component specific distributions in (16). Choosing the exponential densities $\psi_i|\delta_i = 0 \sim \mathcal{E}(1/(2rQ))$ and $\psi_i|\delta_i = 1 \sim \mathcal{E}(1/(2Q))$ leads to a mixture of Laplace densities as spike-and-slab prior for β_i:

$$\beta_i|\omega, Q \sim (1-\omega)\text{Lap}\left(\sqrt{rQ}\right) + \omega\text{Lap}\left(\sqrt{Q}\right). \tag{18}$$

Note that the corresponding prior $\psi_i|\omega, Q$, being a mixture of exponentials, is unimodal and has a spike at 0, regardless of the choice of ω, Q, and r (Frühwirth-Schnatter, 2006, p.6). Hence it is a shrinkage prior in the spirit of Subsection 3.1 with the additional advantage that it allows classification.

More generally, we may combine in (16) distribution families which lead to shrinkage for the spike and, at the same time, avoid too much smoothing in the slab of the corresponding marginal mixture of β_i. One promising candidate is combining the exponential density $\psi_i|\delta_i = 0 \sim \mathcal{E}(1/(2rQ))$ for the spike with the inverted Gamma density $\psi_i|\delta_i = 1 \sim \mathcal{G}^{-1}(\nu, Q)$ for the slab. This leads to a finite mixture for β_i, where a Laplace density in the spike is combined with a Student-t distribution in the slab:

$$\beta_i|\omega, Q \sim (1-\omega)\text{Lap}\left(\sqrt{rQ}\right) + \omega t_{2\nu}(0, Q/\nu). \tag{19}$$

Because the mixture $\psi_i|\omega, Q$ is truly bimodal and at the same time the Laplace spike in (19) encourages shrinkages of small random effects toward 0, this prior is likely to facilitate discrimination between zero and non-zero random intercepts.

4.2. *Using Dirac Spikes*

A special variant of the spike-and-slab prior is a finite mixture where the spike follows a Dirac measure at 0:

$$p(\beta_i|\omega,\boldsymbol{\theta}) = (1-\omega)\Delta_0(\beta_i) + \omega p_{\text{slab}}(\beta_i|\boldsymbol{\theta}). \tag{20}$$

We call this a Dirac-spike-and-slab prior. The marginal density $p(\beta_i|\omega,\boldsymbol{\theta})$ is no longer absolutely continuous which will have consequences for MCMC estimation in Subsection 5.2. In particular, it will be necessary to compute the marginal likelihood where β_i is integrated out, when sampling the indicators. On the other hand, as opposed to a spike-and-slab prior with an absolutely continuous spike, $\delta_i = 0$ is now equivalent to $\beta_i = 0$, which is more satisfactory from a theoretical point of view.

If the slab has a representation as a hierarchical scale mixture prior as in (4) with $\psi_i \sim p_{\text{slab}}(\psi_i|\boldsymbol{\theta})$, then prior (20) is equivalent to putting a Dirac-spike-and-slab prior directly on ψ_i:

$$p(\psi_i|\omega,\boldsymbol{\theta}) = (1-\omega)\Delta_0(\psi_i) + \omega p_{\text{slab}}(\psi_i|\boldsymbol{\theta}). \tag{21}$$

This makes it possible to combine in (20) a Dirac measure, respectively, with a normal slab ($\psi_i \equiv Q$), with a Student-t slab ($\psi_i \sim \mathcal{G}^{-1}(\nu, Q)$), with a Laplace slab ($\psi_i \sim \mathcal{E}(1/(2Q))$), or with a Normal-Gamma slab (e.g. $\sqrt{\psi_i} \sim \mathcal{N}(0, Q)$).

4.3. *Hyperparameter Settings*

In practical applications of spike-and-slab priors, hyperparameters like ω, Q and r are often chosen in a data based manner and considered to be fixed. However, as mentioned above, for random intercept selection it is sensible to include at least some random hyperparameters, because then the random intercepts $\beta_1, \ldots, \beta_N$ are dependent marginally and $p(\beta_1, \ldots, \beta_N)$ also acts as a smoothing prior across units. Subsequently, we regard the scaling parameter Q and the inclusion probability ω as random hyperparameters, whereas we fix shape parameters in any component density like ν for a Student-t distribution as in Subsection 3.2. Furthermore, under an absolutely continuous spike we fix the ratio r between the variances of the two components in order to guarantee good discrimination.

We use the prior $\omega \sim \mathcal{B}(a_0, b_0)$ for ω, where $a_0/(a_0+b_0)$ is a prior guess of the fraction of non-zero random effects and $N_0 = a_0 + b_0$ is the prior information, usually a small integer. Choosing $a_0 = b_0 = 1$ leads to the uniform prior applied e.g. in Smith and Kohn (2002) and Frühwirth-Schnatter and Tüchler (2008) for covariance selection in random effects models. Making ω random, introduces smoothing also for a Dirac spike, where the random intercepts would be independent, if ω were fixed. Ley and Steel (2009) showed for variable selection in standard regression models that considering ω to be random clearly outperforms variable selection under fixed ω for a Dirac-spike-and-slab prior.

To make the prior of Q comparable to the prior of Q under the shrinkage priors introduced in Subsection 3.1, we assume that conditional on ω and possibly a fixed shape parameter, the variance $v_\beta = V(\beta_i|Q,\omega)$ follows the same inverted Gamma prior as in (10). Again, v_β is related to Q in a simple way and we derive accordingly a prior for $Q|\omega$. Because we consider only component densities with zero means, we obtain for an absolutely continuous spike,

$$v_\beta = (1-\omega)V_{\text{spike}}(\beta_i|r,Q) + \omega V_{\text{slab}}(\beta_i|Q),$$

where $V_{\text{spike}}(\beta_i|r,Q)$ and $V_{\text{slab}}(\beta_i|Q)$ are linear transformations of the parameter Q. For spikes and slabs specified by different distributions we obtain $V_{\text{spike}}(\beta_i) = c_1 Q r$, $V_{\text{slab}}(\beta_i) = c_2 Q$, and $v_\beta = Q(r(1-\omega)c_1 + \omega c_2)$, where c_1 and c_2 are the distribution specific constants discussed after (11). Therefore,

$$Q|\omega \sim \mathcal{G}^{-1}\left(c_0, C_0/s^*(\omega)\right), \tag{22}$$

with $s^*(\omega) = r(1-\omega)c_1 + \omega c_2$. For density (18), for instance, $s^*(\omega) = 2r(1-\omega) + \omega/(\nu-1)$. If spike and slab have the same distributional form, then $c_1 = c_2 = c$ and we obtain $v_\beta = Q((1-\omega)r + \omega)c$. In this case, $Q|\omega$ has the same form as in (22) with $s^*(\omega) = c((1-\omega)r + \omega)$. Finally, under a Dirac spike $v_\beta = c\omega$. If we define the variance ratio r under a Dirac spike to be equal to 0, we obtain the same prior as in (22) with $s^*(\omega) = c\omega$.

5. COMPUTATIONAL ISSUES

For estimation, we simulated from the joint posterior distribution of all unknown parameters using a Markov chain Monte Carlo (MCMC) sampler. Unknown parameters common to all shrinkage priors are $\boldsymbol{\alpha}$, σ_ε^2, Q, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)$. Additional unknown parameters are $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_N)$ for any prior with a non-Gaussian component densities for $p(\beta_i|\boldsymbol{\theta})$, and the indicators $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)$ for any spike-and-slab priors.

Regardless of the shrinkage prior, the same standard Gibbs step is used to update the regression parameter $\boldsymbol{\alpha}$ and the error variance σ_ε^2 conditional on all remaining parameters. To sample the remaining parameters conditional on $\boldsymbol{\alpha}$ and σ_ε^2 we focus on a model where

$$\tilde{y}_{it} = \beta_i + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right), \tag{23}$$

with $\tilde{y}_{it} = y_{it} - \mathbf{x}_{it}\boldsymbol{\alpha}$. Subsequently $\tilde{\mathbf{y}}_i = (\tilde{y}_{i1}, \ldots, \tilde{y}_{i,T_i})'$.

5.1. *Sampling the random effects distribution*

To sample β_i, ψ_i and Q we use following hierarchical representation of the random effects distribution

$$\beta_i|\psi_i, \delta_i \sim \mathcal{N}\left(0, \tau_i\right), \qquad \tau_i = (\delta_i + (1-\delta_i)r)\psi_i = r(\delta_i)\psi_i, \tag{24}$$

where $\delta_i \equiv 1$, if no mixture structure is present. Note that $\tau_i = \psi_i$ and $\psi_i|\delta_i = 1 \sim p_{\text{slab}}(\psi_i|Q)$ as in in the previous section, whenever $\delta_i = 1$.

For a Dirac spike $r = 0$ for $\delta_i = 0$, hence $\tau_i = 0$. For an absolutely continuous spike, $\tau_i = r\psi_i$ and $\psi_i|\delta_i = 0 \sim p_{\text{spike}}(\psi_i|Q)$, whenever $\delta_i = 0$. Evidently representation (24) slightly differs in the spike from the representation we used earlier, because ψ_i is drawn from the distribution family underlying the spike with scaling factor Q (rather than rQ) and reducing the variance by the factor r takes place when defining τ_i. By defining the latent variances in our MCMC scheme in this slightly modified way we avoid problems with MCMC convergence for extremely small latent variances.

Sampling from $\beta_i|\psi_i, \delta_i, \tilde{\mathbf{y}}_i$ is straightforward, because (23) in combination with (24) constitutes a standard Gaussian random intercept model:

$$\beta_i|\delta_i, \psi_i, \tilde{\mathbf{y}}_i \sim \mathcal{N}\left(B_i \sum_{t=1}^{T_i} \tilde{y}_{it}, \sigma_\varepsilon^2 B_i\right), \qquad B_i^{-1} = T_i + 1/(r(\delta_i)\psi_i). \tag{25}$$

For any Gaussian component density $\psi_i = Q$, hence ψ_i is deterministic given Q. For any non-Gaussian component density ψ_i is sampled from $\psi_i|\beta_i, \delta_i, Q$. The precise form of this posterior depends on the prior $p(\psi_i|\delta_i, Q)$. If $\psi_i|\delta_i, Q \sim \mathcal{G}^{-1}(\nu, Q)$, then

$$\psi_i|\beta_i, \delta_i, Q \sim \mathcal{G}^{-1}\left(\nu + 1/2, Q + \beta_i^2/(2r(\delta_i))\right). \tag{26}$$

If $\psi_i|\delta_i, Q \sim \mathcal{E}(1/(2Q))$, then

$$\psi_i|\beta_i, \delta_i, Q \sim \mathcal{GIG}\left(1/2, 1/Q, \beta_i^2/r(\delta_i)\right), \tag{27}$$

where $\mathcal{GIG}(\cdot)$ is equal to generalized inverse Gaussian distribution. Alternatively, $1/\psi_i$ may be drawn from the inverse Gaussian distribution $\text{InvGau}\left(\sqrt{r(\delta_i)}/(\sqrt{Q}|\beta_i|), Q\right)$.

Note that for a Dirac spike the likelihood $p(\tilde{\mathbf{y}}_i|\delta_i = 0, \beta_i, \sigma_\varepsilon^2)$ is independent from β_i, hence drawing from (25) and (26) or (27) is required only, if $\delta_i = 1$. This saves considerable CPU time, if $\sum_{i=1}^N \delta_i << N$. For $\delta_i = 0$, $\beta_i = 0$, and ψ_i is sampled from the slab, i.e. $\psi_i \sim p_{\text{slab}}(\psi_i|Q)$.

Finally, sampling of $Q|\boldsymbol{\psi}, \boldsymbol{\beta}, \boldsymbol{\delta}$ depends on spike/slab combination. For Laplace mixtures or a Dirac spike with a Laplace slab we obtain with $Q|\boldsymbol{\psi}, \omega \sim \mathcal{G}^{-1}(N + c_0, C_N)$ with:

$$C_N = \frac{C_0}{s^*(\omega)} + \frac{1}{2}\sum_{i=1}^N \psi_i.$$

For Student-t mixtures or a Dirac spike with a Student-t slab

$$Q|\boldsymbol{\psi}, \boldsymbol{\delta}, \omega \sim \mathcal{GIG}\left(\nu N - c_0, 2\sum_{i=1}^N 1/\psi_i, 2C_0/s^*(\omega)\right).$$

If a Laplace spike is combined with a Student-t slab, then

$$Q|\boldsymbol{\psi}, \boldsymbol{\delta}, \omega \sim \mathcal{GIG}\left((\nu+1)n_1 - N - c_0, 2\Psi_1, 2C_0/s^*(\omega) + \Psi_0\right),$$

where $\Psi_0 = \sum_{i:\delta_i=0} \psi_i$, $\Psi_1 = \sum_{i:\delta_i=1} 1/\psi_i$, and $n_1 = \sum_{i=1}^N \delta_i$. For normal mixtures $Q|\boldsymbol{\beta}, \boldsymbol{\delta} \sim \mathcal{G}^{-1}(c_0 + N/2, C_N)$ with

$$C_N = \frac{C_0}{s^*(\omega)} + \frac{1}{2}\sum_{i=1}^N {\beta_i}^2/r(\delta_i),$$

while for a Dirac spike with a normal slab $Q|\boldsymbol{\beta}, \boldsymbol{\delta} \sim \mathcal{G}^{-1}(c_0 + n_1/2, C_N)$ with

$$C_N = \frac{C_0}{\omega} + \frac{1}{2}\sum_{i:\delta_i=1} {\beta_i}^2.$$

5.2. *Additional Steps for Spike-and-Slab Priors*

For all spike-and-slab smoothing priors it is possible to sample $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)$ simultaneously, because δ_i, $i = 1, \ldots, N$ are conditionally independent a posteriori given ω. A computational advantage of an absolutely continuous spikes compared to a Dirac spike is that is possible to sample δ_i conditional on β_i, however, we marginalize over ψ_i for non-Gaussian components to improve the efficiency of this step:

$$\Pr(\delta_i = 1|\beta_i, \omega, \boldsymbol{\theta}) \quad = \frac{1}{1+\frac{1-\omega}{\omega} L_i}, \qquad L_i = \frac{p_{\text{spike}}(\beta_i|\boldsymbol{\theta})}{p_{\text{slab}}(\beta_i|\boldsymbol{\theta})}. \tag{28}$$

For a Dirac spike δ_i is drawn without conditioning in the slab on β_i, but conditional on ψ_i (which is equal to Q for a normal slab). Hence

$$\Pr(\delta_i = 1|\psi_i, \tilde{\mathbf{y}}_i, \omega) \quad = \frac{1}{1+\frac{1-\omega}{\omega} R_i}, \qquad R_i = \frac{p(\tilde{\mathbf{y}}_i|\delta_i=0)}{p(\tilde{\mathbf{y}}_i|\psi_i, \delta_i=1)}. \tag{29}$$

Using $\tilde{\mathbf{y}}_i|\delta_i = 0 \sim \mathcal{N}_{T_i}\left(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}\right)$ and $\tilde{\mathbf{y}}_i|\psi_i, \delta_i = 1 \sim \mathcal{N}_{T_i}\left(\mathbf{0}, \mathbf{1}\mathbf{1}'\psi_i + \sigma_\varepsilon^2 \mathbf{I}\right)$ it is possible to work out that

$$2 \log R_i = \log\left(\frac{\sigma_\varepsilon^2 + T_i\psi_i}{\sigma_\varepsilon^2}\right) - \frac{\psi_i}{\sigma_\varepsilon^2 + T_i\psi_i} \sum_{t=1}^{T_i} \tilde{y}_{it}^2/\sigma_\varepsilon^2. \tag{30}$$

Finally, we draw ω from $\omega|\boldsymbol{\delta} \sim \mathcal{B}\left(a_0 + n_1, b_0 + N - n_1\right)$ where $n_1 = \sum_{i=1}^N \delta_i$.

6. EXTENSIONS TO MORE GENERAL MODELS

6.1. *Random Intercept Models for Non-Gaussian Data*

To introduce shrinkage and smoothing priors for non-Gaussian data, any of the distributions for β_i considered in Section 3 and 4 could be combined with a non-Gaussian likelihood depending on a random intercept β_i. A very useful non-Gaussian model is a binary logit model with random effects, where

$$\Pr(y_{it} = 1|\boldsymbol{\alpha}) = \frac{\exp(\mathbf{x}_i\boldsymbol{\alpha} + \beta_i)}{1 + \exp(\mathbf{x}_i\boldsymbol{\alpha} + \beta_i)}. \tag{31}$$

Other examples are count data models where a likelihood based on the Poisson or the negative binomial distribution includes random intercept β_i.

To extend MCMC estimation to such models, data augmentation is applied in such a way that a conditionally Gaussian model results, where the responses z_{it} are not directly observed but are latent variables resulting from data augmentation:

$$z_{it} = \mathbf{x}_{it}\boldsymbol{\alpha} + \beta_i + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathcal{N}\left(0, \sigma_{it}^2\right). \tag{32}$$

For binary data, for instance, data augmentation could be based on Albert and Chib (1993) for probit models, on Frühwirth-Schnatter and Frühwirth (2010) for logit models, while Frühwirth-Schnatter et al. (2009) is useful for repeated count data and binomial data.

Also the structure of the error variance appearing in (32) depends on the distribution of the observed data. Data augmentation leads to $\sigma_{it}^2 = 1$ for the probit

model. Data augmentation for the logit model and the Poisson model involves a finite normal mixture approximation with H components, hence the error variance depends on an additional latent component indicator r_{it} taking values in $\{1,\ldots,H\}$: $\sigma^2_{it} = \sigma^2_{r_{it}}$. Since $\sigma^2_1,\ldots,\sigma^2_H$ are known constants, the error variance is heteroscedastic, but fixed given r_{it}.

We omit the details of the corresponding MCMC sampler, but provide an example of a random-intercept model for binomial data in Subsection 7.2.

6.2. *Bayesian Variable Selection for Mixed-effects Model*

Model (1) is a special case of the more general linear mixed-effects model for modeling longitudinal data (Laird and Ware, 1982), defined by

$$\boldsymbol{\beta}_i|\mathbf{Q} \sim \mathcal{N}_r\left(\mathbf{0},\mathbf{Q}\right), \tag{33}$$

$$y_{it} = \mathbf{x}^f_{it}\boldsymbol{\beta} + \mathbf{x}^r_{it}\boldsymbol{\beta}_i + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathcal{N}\left(0,\sigma^2_{it}\right). \tag{34}$$

$\mathbf{x}^r_{it}$ is the $(1\times r)$ design matrix for the unknown coefficient $\boldsymbol{\beta}_i = (\beta_{i1},\ldots,\beta_{ir})'$ of dimension r. The covariates appearing in $\mathbf{x}^r_{it}$ are called the random effects, because the corresponding regression coefficient $\boldsymbol{\beta}_i$ depends on unit i.

A common approach to variable selection for the random-effects part of a mixed-effects model focuses on the variance of the random-effects (Chen and Dunson, 2003; Frühwirth-Schnatter and Tüchler, 2008; Kinney and Dunson, 2007; Tüchler, 2008). Model specification for the random effects is translated into variable selection for the variances. Consider, for instance, a random coefficient model where $\mathbf{x}^f_{it} = \mathbf{x}^r_{it} = \mathbf{x}_{it}$ and assume, for simplicity, that $\mathbf{Q} = \mathrm{Diag}(Q_1,\ldots,Q_r)$, i.e. $\beta_{ij} \sim \mathcal{N}\left(0,Q_j\right)$, for $j = 1,\ldots,r$. Introduce r binary variable selection indicators $\delta_1,\ldots,\delta_r$. If $\delta_j = 0$, then $Q_j = 0$ and the random effect β_{ij} disappears for all units, leading to a fixed effect of the covariate $x_{j,it}$ equals β_j. On the other hand, if $\delta_j = 1$, then Q_j is unrestricted leading to a random effect of the covariate $x_{j,it}$ equals $\beta_j + \beta_{ij}$.

While this approach is very attractive for potentially high-dimensional random effect models, it might be too simplified for applications with a low-dimensional random effect, like panel data analysis, multilevel analysis or two-way ANOVA applications. For such models, it might be of interest to apply the shrinkage priors introduced in Section 3 and 4 independently to each coefficient β_{ij}.

7. APPLICATIONS

7.1. *Application to Simulated Data*

We generated data with $N = 100$ subjects, $T_i = 10$ replications, and 4 covariates according to the model $y_{it} = \mu + \mathbf{x}_{it}\boldsymbol{\alpha} + \beta_i + \varepsilon_{it}$, $\varepsilon_{it} \sim \mathcal{N}\left(0,\sigma^2_\varepsilon\right)$, where $\mu = 1$, $\boldsymbol{\alpha} = (0.5,-0.5,0.7,-0.7)$, and $\sigma_\varepsilon = 0.5$. The covariates are simulated independently as $x_{it,j} \sim \mathcal{N}\left(0,1\right)$.

Four different data sets were generated with different percentage of non-zero random effects. Data Set 1 has an extremely high fraction of zero random effects: $(\beta_1,\ldots,\beta_5) = (1,1,1,-1.5,-1.5)$, and $\beta_i = 0$ for $i = 6,\ldots,100$. In Data Set 2, half of the random effects are zero, $\beta_i = -4$ for $i = 1,\ldots,5$; $\beta_i = -1$ for $i = 6,\ldots,25$; $\beta_i = 0$ for $i = 25,\ldots,75$ $\beta_i = 1$ for $i = 76,\ldots,95$ and $\beta_i = 4$ for $i = 96,\ldots,100$. For Data Set 3 and 4 all random effects are nonzero, and are drawn independently from the standard normal distribution, $\beta_i \sim \mathcal{N}\left(0,1\right)$ for Data Set 3 and from an Type I extreme value distribution centered at 0 for Data Set 4, i.e. $\beta_i = -\log(-\log U_i))-\gamma$, where U_i is a uniform random numbers and $\gamma = 0.5772$ is Euler's constant.

Table 1: *Comparing the different random effect priors for Data Set 1*

Prior of the random effects	RMSE_μ	$\text{RMSE}_{\boldsymbol{\alpha}}$	RMSE_β	TZDR	TNDR
Normal	0.0185	0.0184	0.133	100	60
Student	0.0058	0.0177	0.117	100	60
Laplace	0.0111	0.0173	0.0992	100	89.5
Normal-spike-normal-slab	0.0132	0.0166	0.0321	100	100
Student-spike-Student-slab	0.0133	0.0165	0.0316	100	100
Laplace-spike-Laplace-slab	0.0133	0.0164	0.0347	100	100
Laplace-spike-Student-slab	0.0131	0.0165	0.0319	100	100
Dirac-spike-normal-slab	0.0133	0.0164	0.0316	100	100
Dirac-spike-Student-slab	0.0132	0.0165	0.0317	100	100
Dirac-spike-Laplace-slab	0.013	0.0164	0.0334	100	100

For Bayesian estimation, we use the improper prior $p(\mu, \sigma_\varepsilon^2, \boldsymbol{\alpha}) \propto 1/\sigma_\varepsilon^2$ for the parameters in the observation equation. The hyperparameters for the inverted Gamma prior for $v_\beta = V(\beta_i|\boldsymbol{\theta})$ are selected as $c_0 = 2$ and $C_0 = 1$ and, for spike-and-slab priors, for the Beta prior for ω as $a_0 = b_0 = 1$. The remaining parameters were chosen as $\nu = 5$ for Student-t component densities and the variance ratio is set to $r = 0.000025$. MCMC was run for 20 000 iterations after a burn-in of 10 000; for spike-and-slab priors in the first 1000 iterations random effects were drawn from the slab only.

Table 2: *Comparing the different random effect priors for Data Set 2*

Prior of the random effects	RMSE_μ	$\text{RMSE}_{\boldsymbol{\alpha}}$	RMSE_β	TZDR	TNDR
Normal	0.0056	0.00761	0.18	100	78
Student	0.0058	0.00743	0.179	100	66
Laplace	0.0117	0.00722	0.176	100	72
Normal-spike-normal-slab	0.0183	0.00963	0.156	94	100
Student-spike-Student-slab	0.0173	0.00954	0.158	94	100
Laplace-spike-Laplace-slab	0.016	0.00904	0.16	92	100
Laplace-spike-Student-slab	0.0149	0.00993	0.151	98	100
Dirac-spike-normal-slab	0.017	0.00971	0.156	94	100
Dirac-spike-Student-slab	0.0166	0.0096	0.157	94	100
Dirac-spike-Laplace-slab	0.0156	0.00901	0.159	92	100

We consider different kinds of criteria to compare the various shrinkage priors. Statistical efficiency with respect to estimating the intercept μ and the regression coefficients $\boldsymbol{\alpha}$ is measured in terms of the root mean squared error $\text{RMSE}_\mu = |\mu - \hat{\mu}|$ and $\text{RMSE}_{\boldsymbol{\alpha}} = \sqrt{||\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}||_2}/\sqrt{d}$, where $d = \dim(\boldsymbol{\alpha}) = 4$. Additionally, we determine the root mean squared error for the random effects as $\text{RMSE}_\beta = (\sum_{i=1}^N (\beta_i - \hat{\beta}_i)^2/N)^{1/2}$. All parameters are estimated in the usual way as average of the corresponding MCMC draws.

Furthermore, in the present context correct classification of truly zero and truly non-zero random effects is important. For spike-and-slab priors variable selection is based on the posterior inclusion probability p_i, i.e. accept $\beta_i \neq 0$ and set $\hat{\delta}_i = 1$, if $p_i \geq 0.5$; otherwise accept $\beta_i = 0$ and set $\hat{\delta}_i = 0$. For an absolutely continuous

spike, we apply the heuristic rule suggested recently by Li and Lin (2010), i.e. accept $\beta_i = 0$ and set $\hat{\delta}_i = 0$ if an $100p\%$ credible interval of β_i covers 0; otherwise accept $\beta_i \neq 0$ and set $\hat{\delta}_i = 1$. A certain difficulty here is the choice of p, because we are dealing with a multiple comparison problem. As in Li and Lin (2010) we choose $p = 0.5$. Aggregate classification measures are the truly-zero-discovery-rate $\text{TZDR} = 100/N_0 \sum_{i \in I_0} I\{\hat{\delta}_i = 0\}$ and the truly-nonzero-discovery-rate $\text{TNDR} = 100/N_1 \sum_{i \in I_1} I\{\hat{\delta}_i = 1\}$, where I_0 and I_1 denote, respectively, the set of observation indices for all truly zero and truly non-zero random effects, and N_0 and N_1 are the corresponding cardinality. Both rates should be as close to 100 percent as possible.

Table 3: *Comparing the different random effect priors for Data Set 3*

Prior of the random effects	RMSE_μ	$\text{RMSE}_{\boldsymbol{\alpha}}$	RMSE_β	TNDR
Normal	0.086	0.0138	0.181	92
Student	0.104	0.0137	0.19	92
Laplace	0.1	0.0138	0.189	91
Normal-spike-normal-slab	0.0835	0.0138	0.179	100
Student-spike-Student-slab	0.106	0.0137	0.191	100
Laplace-spike-Laplace-slab	0.1	0.0137	0.189	100
Laplace-spike-Student-slab	0.0877	0.0138	0.183	100
Dirac-spike-normal-slab	0.0884	0.0138	0.182	100
Dirac-spike-Student-slab	0.107	0.0137	0.191	100
Dirac-spike-Laplace-slab	0.104	0.0137	0.191	100

The results of comparing the different random effect priors are summarized in Table 1 to Table 4. In general, for random effect priors without a mixture structure classification based on confidence regions as in Li and Lin (2010) is less reliable than classification based on spike-and-slab priors. This is even true for Data Set 3, where the normal prior corresponds to the true model, but classification is perfect only for spike-and-slab priors. Even in this case, using a mixture of normals instead of the normal distribution leads to a comparably small loss in efficiency for estimating the regression parameters. These results clearly indicate that spike-and-slab priors are preferable as random effects distribution, if individual variable selection is of interest.

Concerning differences between Dirac and absolutely continuous spikes, we find that there is surprisingly little difference between a spike from the same distribution as the slab and a Dirac spike. Hence, both approaches seem to make sense, although we tend to prefer the Dirac spike for the theoretical reasons outlined above.

The most difficult issue is the choice of the distributions underlying spike-and-slab priors. For Data Set 1, priors based on a Laplace slab perform worse than the other spike-and-slab priors, in particular with respect to RMSE_β which indicates too much shrinkage in the slab. The other spike-and-slab priors yield more or less similar results.

For Data Set 2, a Student-t slab with a Laplace spike yields better results than the other spike-and-slab priors, apart from $\text{RMSE}_{\boldsymbol{\alpha}}$. This prior has, in particular, the best classification rate.

For Data Set 3 priors based on a normal slabs (either with Dirac or normal spike) are better than the other spike-and-slab priors. This is not surprising, because the true random effects distribution is a standard normal distribution. Interestingly, a

Student-t slab with a Laplace spike yields results which are nearly as good as priors with a normal slab, while the remaining priors perform worse.

Also Data Set 4, where the true distribution is equal to the extremely skew Type I extreme value distribution, all priors based on a normal slab outperform the other ones. In addition, we observe quite an influence of the distributions underlying the spike-and-slab prior on the efficiency of estimating the mean μ of the random intercept.

Hence, from this rather limited simulation study we are not able to identify a uniformly best component density and further investigations are certainly necessary.

Table 4: *Comparing the different random effect priors for Data Set 4*

Prior of the random effects	RMSE_μ	$\text{RMSE}_{\boldsymbol{\alpha}}$	RMSE_β	TNDR
Normal	0.0094	0.0137	0.149	95
Student	0.119	0.0139	0.192	93
Laplace	0.251	0.014	0.293	86
Normal-spike-normal-slab	0.091	0.0134	0.176	100
Student-spike-Student-slab	0.183	0.0135	0.237	100
Laplace-spike-Laplace-slab	0.271	0.014	0.311	100
Laplace-spike-Student-slab	0.305	0.0132	0.341	81
Dirac-spike-normal-slab	0.0925	0.0134	0.177	100
Dirac-spike-Student-slab	0.183	0.0136	0.237	100
Dirac-spike-Laplace-slab	0.267	0.0138	0.307	100

7.2. *Application to the Seed Data*

We reconsider the data given by Crowder (1978, Table 3) reporting the number Y_i of seeds that germinated among T_i seeds in $N = 21$ plates covered with a certain root extract. The data are modelled as in Breslow and Clayton (1993) and Gamerman (1997), assuming that Y_i is generated by a binomial distribution, where dependence of the success probability on covariates $\mathbf{x}_i$ is modelled through a logit transform:

$$\begin{aligned} Y_i &\sim \text{BiNom}\left(T_i, \pi_i\right), \\ \log \frac{\pi_i}{1-\pi_i} &= \mathbf{x}_i\boldsymbol{\alpha} + \beta_i, \qquad \beta_i \sim \mathcal{N}\left(0, Q\right). \end{aligned} \tag{35}$$

The covariates are the type of root extract (bean or cucumber), the type of seed (*O. aegyptiaco 73* and *O. aegyptiaco 75*), and an interaction term between these variables. The normally distributed random intercept β_i is added by these authors to capture potential overdispersion in the data.

Subsequently, the binomial model (35) is estimated by recovering the full binary experiment as in Frühwirth-Schnatter and Frühwirth (2007). Any observation Y_i from model (35) is equivalent with observing T_i repeated measurements y_{it} from a binary model with random effects,

$$\Pr(y_{it} = 1|\boldsymbol{\alpha}) = \frac{\exp(\mathbf{x}_i\boldsymbol{\alpha} + \beta_i)}{1 + \exp(\mathbf{x}_i\boldsymbol{\alpha} + \beta_i)},$$

where $y_{it} = 1$, $1 \le t \le Y_i$, and $y_{it} = 0$, for $Y_i < t \le T_i$. Hence we are dealing with repeated measurement in a logit model with a random intercept.

Table 5: *Seed Data; Variable and covariance selection in the full random coefficient model using Tüchler (2008)*

		`const` ($j=1$)	`root` ($j=2$)	`seed` ($j=3$)	`root*seed` ($j=4$)
$\Pr(\alpha_j \neq 0 \mid \mathbf{y})$		0.969	0.975	0.431	0.895
$\Pr(Q_{1j} \neq 0 \mid \mathbf{y})$	`const`	0.243	0.005	0.006	0
$\Pr(Q_{2j} \neq 0 \mid \mathbf{y})$	`root`	0.005	0.044	0.021	0.002
$\Pr(Q_{3j} \neq 0 \mid \mathbf{y})$	`seed`	0.006	0.021	0.05	0.002
$\Pr(Q_{4j} \neq 0 \mid \mathbf{y})$	`root*seed`	0	0.002	0.002	0.055

Table 6: *Seed Data; Variable and covariance selection in the random intercept model using log marginal likelihoods (based on Frühwirth-Schnatter and Wagner (2008))*

k	$\mathcal{M}_k$	logit ($Q=0$)	$\beta_i \sim \mathcal{N}(0, Q)$
1	`const`	-578.50	-555.78
2	`const`, `root`	-553.11	-551.35
3	`const`, `seed`	-579.18	-556.11
4	`const`,`root*seed`	-580.05	-556.77
5	`const`, `root`, `seed`	-553.46	-551.58
6	`const`, `root`, `root*seed`	**-550.58**	**-550.32**
7	`const`, `seed`, `root*seed`	-578.47	-556.59
8	`const`, `root`, `seed`, `root*seed`	-552.06	-551.49

4.2.1 *Variable and Covariance Selection*

First, we consider the full random effects model where all covariates are included and $\boldsymbol{\beta}_i \sim \mathcal{N}_d(\mathbf{0}, \mathbf{Q})$. We consider variable and covariance selection as in Tüchler (2008) based on a spike-and-slab prior for the regression coefficients and the Cholesky factors of $\mathbf{Q}$ where a fractional normal prior is used for the non-zero coefficients. In terms of the elements of $\mathbf{Q}$ this prior means, for instance, that, marginally, the diagonal elements Q_{jj} follow a χ^2_1 distribution. Table 5 reports marginal inclusion probabilities for all regression coefficients and we find that the covariable `seed` may be eliminated from the full model. The same table reports also marginal inclusion probabilities for the elements of the covariance matrix $\mathbf{Q}$. All elements of this matrix but Q_{11} have a practically zero probability of being non-zero, meaning that all effects but the intercept are fixed with very high probability. This leaves either a logit random intercept model or a standard logit model as possible model specifications. Evidence for the random intercept model is not overwhelming, but not practically zero either.

Frühwirth-Schnatter and Wagner (2008) computed marginal likelihoods for these data in order to perform variable selection and testing for the presence of a random intercept model. The results are reproduced in Table 6 and confirm Table 5, although a different prior was used. To make model comparison through marginal likelihoods feasible, the improper prior $p(\boldsymbol{\alpha}, Q) \propto 1/\sqrt{Q}$ used by Gamerman (1997) was substituted by the proper priors $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the usual inverted Gamma prior $Q \sim \mathcal{G}^{-1}(c_0, C_0)$ where $c_0 = 0.5$ and $C_0 = 0.2275$. Among all models considered, a random intercept model where the covariable `seed` is eliminated has the largest marginal likelihood, however, evidence in comparison to a model with the

same predictors, but no random intercept is pretty weak, with the posterior probabilities of both models being roughly the same.

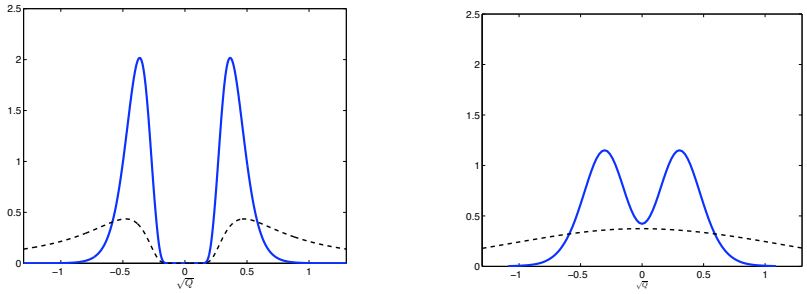

Figure 1: Estimated marginal posterior density $p(\pm\sqrt{Q}|\mathbf{y})$ (bold line) under the inverted Gamma prior $Q \sim \mathcal{G}^{-1}(0.5, 0.2275)$ (left hand side) and under the normal prior $\pm\sqrt{Q} \sim \mathcal{N}(0, 1)$ (right hand side) for a model excluding the covariable `seed`; the dashed lien corresponds to the prior

To get more insight how the prior on Q effects posterior inference, Figure 1 compares the posterior distribution of $\pm\sqrt{Q}$ under the usual inverted Gamma prior $Q \sim \mathcal{G}^{-1}(0.5, 0.2275)$ with the normal prior $\pm\sqrt{Q} \sim \mathcal{N}(0, 1)$ which corresponds to a χ^2_1 distribution for Q or, $Q \sim \mathcal{G}(0.5, 0.5)$. This figure clearly indicates that the inverted Gamma prior assigns zero probability to values close to 0, bounding the posterior distribution away from 0, while the χ^2_1 prior allows the posterior distribution to take values close to zero. For the χ^2_1 prior, the ratio of the prior over the posterior ordinate at 0, also known as Savages density ratio, is an estimator of the Bayes factor of a model without and with heterogeneity, see e.g. McCulloch and Rossi (1991). This ratio is roughly 1 which is in line with the evidence of Table 6 although a different prior was used in this table.

7.2.2. *Individual Random Effects Selection*

Since these results from pure covariance selection are rather inconclusive concerning the presence (or absence) of a random intercept in the logit model we consider individual random effects selection using the shrinkage priors introduced in this paper. We consider a random intercept model where the covariable `seed` is eliminated and use the prior $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, 100\mathbf{I})$ for the regression coefficients. The hyperparameters for the inverted Gamma prior for $v_\beta = V(\beta_i|\boldsymbol{\theta})$ are selected as $c_0 = 2$ and $C_0 = 1$ and, for spike-and-slab priors, for the Beta prior for ω as $a_0 = b_0 = 4$. The remaining parameters were chosen as $\nu = 5$ for Student-t component densities and the variance ratio is set to $r = 0.000025$. MCMC was run for 20 000 iterations after a burn-in of 10 000; for spike-and-slab priors in the first 1000 iterations random effects were drawn from the slab only.

The estimated posterior means of the random effects are plotted in Figure 2, while Table 7 summarizes individual random effects selection. All priors find that a considerable fraction of the random effects are 0, meaning that only for a few units unobserved heterogeneity is present. This clearly explains why pure variance

Table 7: *Seed data; units where 0 is not included in the 50% credible interval are marked with* **x** *for shrinkage priors; for the remaining priors the estimated posterior inclusion probabilities* $\Pr(\delta_i = 1|\mathbf{y})$ *are reported (bold numbers correspond to accepting* $\beta_i \neq 0$*).*

	Shrinkage Priors			Continuous Slab			Dirac Slab		
Unit	$\mathcal{N}$	t_{10}	Lap	$\mathcal{N}$	t_{10}	Lap	$\mathcal{N}$	t_{10}	Lap
1	**x**	**x**	**x**	0.47	0.43	0.44	0.44	0.45	0.46
2				0.29	0.27	0.26	0.24	0.24	0.29
3	**x**	**x**	**x**	**0.50**	0.45	0.45	0.44	0.46	0.48
4	**x**	**x**	**x**	**0.65**	**0.62**	**0.57**	**0.58**	**0.59**	**0.60**
5				0.34	0.32	0.32	0.31	0.32	0.35
6				0.43	0.41	0.39	0.39	0.39	0.42
7				0.32	0.29	0.31	0.28	0.28	0.32
8	**x**	**x**		0.46	0.43	0.39	0.39	0.42	0.44
9				0.44	0.37	0.34	0.34	0.35	0.37
10	**x**	**x**	**x**	**0.68**	**0.60**	**0.61**	**0.57**	**0.58**	**0.58**
11				0.44	0.36	0.35	0.35	0.35	0.38
12				0.43	0.41	0.37	0.38	0.38	0.40
13				0.31	0.25	0.31	0.28	0.28	0.33
14				0.39	0.36	0.36	0.34	0.34	0.38
15	**x**	**x**	**x**	**0.61**	**0.56**	**0.60**	**0.55**	**0.57**	**0.57**
16				**0.56**	**0.50**	0.44	0.49	**0.50**	**0.51**
17	**x**	**x**		**0.62**	**0.59**	**0.54**	**0.58**	**0.59**	**0.59**
18				0.32	0.27	0.32	0.28	0.28	0.32
19				0.34	0.30	0.32	0.29	0.29	0.33
20	**x**	**x**		**0.52**	0.42	0.44	0.45	0.45	0.47
21				0.43	0.41	0.40	0.36	0.36	0.39
$\#\{\beta_i \neq 0\|\mathbf{y}\}$	8	8	5	7	5	4	4	5	5

selection based on deciding whether $Q = 0$ or not is too coarse for this data set. Among the shrinkage priors, the Laplace prior leads to the strongest degree of shrinkage and $\beta_i = 0$ is rejected only for 5 units. There is quite an agreement across all shrinkage priors for several units that $\beta_i \neq 0$, while for others units the decision depends on the prior, in particular, if the inclusion probability is around 0.5. What

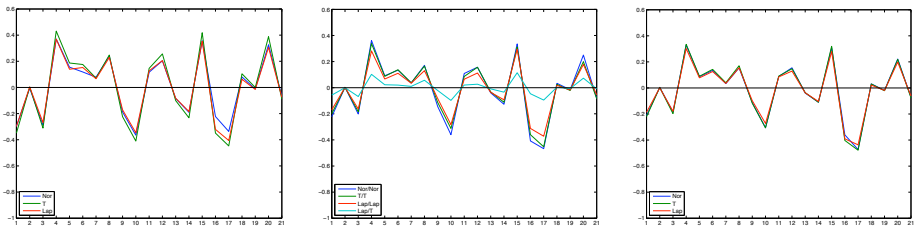

Figure 2: Seed data; Estimated posterior mean $\mathrm{E}(\beta_i|\mathbf{y})$ for the various random effects. Left: Shrinkage priors, middle: continuous spikes, right: Dirac spikes.

8. CONCLUDING REMARKS

Variable selection problems arise for more general latent variable models than the random intercept model considered in this paper and some examples were already mentioned in Section 6. Other examples are variable selection in non-parametric regression (Shively et al., 1999; Smith and Kohn, 1996; Kohn et al., 2001), structured additive regression models (Belitz and Lang, 2008) and in state space models (Shively and Kohn, 1997; Frühwirth-Schnatter and Wagner, 2010). Typically, these problems often concern the issue of how flexible the model should be.

Variable selection in time-varying parameter models and in more general state space models, for instance, has been considered by Shively and Kohn (1997) and Frühwirth-Schnatter and Wagner (2010). In these papers, variable selection for the time-varying latent variables is reduced to a variable selection for the variance of the innovations in the state equation. The resulting procedure discriminates between a model where a certain component of the state variable either remains totally dynamic and possibly changes at each time point and a model where this component is constant over the whole observation period. To achieve more flexibility for these type of latent variable models, it might be of interest to apply the shrinkage priors discussed in this paper to the innovations independently for each time point. This allows to discriminate time points where the state variable remains constant from time points where the state variable changes. However, we leave this very promising approach for future research.

REFERENCES

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88**, 669–679.

Belitz, C. and S. Lang (2008). Simultaneous selection of variables and smoothing parameters in structured additive regression models. *Comput. Statist. Data Anal.* **53**, 61–81.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88**, 9–25.

Chen, Z. and D. Dunson (2003). Random effects selection in linear mixed models. *Biometrics* **59**, 762–769.

Crowder, M. J. (1978). Beta-binomial ANOVA for proportions. *Appl. Statist.* **27**, 34–37.

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

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer.

Frühwirth-Schnatter, S. and R. Frühwirth (2007). Auxiliary mixture sampling with applications to logistic models. *Comput. Statist. Data Anal.* **51**, 3509–3528.

Frühwirth-Schnatter, S. and R. Frühwirth (2010). Data augmentation and MCMC for binary and multinomial logit models. In T. Kneib and G. Tutz (Eds.), *Statistical Modelling and Regression Structures – Festschrift in Honour of Ludwig Fahrmeir*, pp. 111–132. Heidelberg: Physica-Verlag.

Frühwirth-Schnatter, S., R. Frühwirth, L. Held, and H. Rue (2009). Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data. *Statist. Computing* **19**, 479–492.

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

Frühwirth-Schnatter, S., R. Tüchler, and T. Otter (2004). Bayesian analysis of the heterogeneity model. *Journal of Business & Economic Statistics 22*, 2–15.

Frühwirth-Schnatter, S. and H. Wagner (2008). Marginal likelihoods for non-Gaussian models using auxiliary mixture sampling. *Comput. Statist. Data Anal.* **52**, 4608–4624.

Frühwirth-Schnatter, S. and H. Wagner (2010). Stochastic model specification search for Gaussian and partially non-Gaussian state space models. *J. Econometrics* **154**, 85–100.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statist. Computing* **7**, 57–68.

George, E. I. and R. McCulloch (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88**, 881–889.

George, E. I. and R. McCulloch (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–373.

Griffin, J. E. and P. J. Brown (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* **5**, 171–188.

Ishwaran, H., L. F. James, and J. Sun (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *J. Amer. Statist. Assoc.* **96**, 1316–1332.

Ishwaran, H. and J. S. Rao (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *Ann. Statist.* **33**, 730–773.

Kinney, S. K. and D. B. Dunson (2007). Fixed and random effects selection in linear and logistic models. *Biometrics* **63**, 690–698.

Kohn, R., M. Smith, and D. Chan (2001). Nonparametric regression using linear combinations of basis functions. *Statist. Computing* **11**, 313–322.

Komárek, A. and E. Lesaffre (2008). Generalized linear mixed model with a penalized Gaussian mixture as a random effects distribution. *Comput. Statist. Data Anal.* **52**, 3441–3458.

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

Ley, E. and M. F. J. Steel (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, 651–674.

Li, Q. and N. Lin (2010). The Bayesian elastic net. *Bayesian Analysis* **5**, 151–170.

McCulloch, R. and P. E. Rossi (1991). A Bayesian approach to testing the arbitrage pricing theory. *J. Econometrics* **49**, 141–168.

Mitchell, T. J. and J. J. Beauchamp (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83**, 1023–1036.

Neuhaus, J. M., W. W. Hauck, and J. D. Kalbfleisch (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* **79**, 755–762.

Park, T. and G. Casella (2008). The Bayesian Lasso. *J. Amer. Statist. Assoc.* **103**, 681–686.

Shively, T. S. and R. Kohn (1997). A Bayesian approach to model selection in stochastic coefficient regression models and structural time series models. *J. Econometrics* **76**, 39–52.

Shively, T. S., R. Kohn, and S. Wood (1999). Variable selection and function estimation in additive nonparametric regression using a data-based prior. *J. Amer. Statist. Assoc.* **94**, 777–794.

Smith, M. and R. Kohn (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75**, 317–343.

Smith, M. and R. Kohn (2002). Parsimonious covariance matrix estimation for longitudinal data. *J. Amer. Statist. Assoc.* **97**, 1141–1153.

Tüchler, R. (2008). Bayesian variable selection for logistic models using auxiliary mixture sampling. *J. Comp. Graphical Statist.* **17**, 76–94.