Bayesian Analysis (2019) **14**, Number 2, pp. 341–369

Bayesian Effect Fusion for Categorical Predictors

Daniela Pauger* and Helga Wagner

Abstract. We propose a Bayesian approach to obtain a sparse representation of the effect of a categorical predictor in regression type models. As this effect is captured by a group of level effects, sparsity cannot only be achieved by excluding single irrelevant level effects or the whole group of effects associated to this predictor but also by fusing levels which have essentially the same effect on the response. To achieve this goal, we propose a prior which allows for almost perfect as well as almost zero dependence between level effects a priori. This prior can alternatively be obtained by specifying spike and slab prior distributions on all effect differences associated to this categorical predictor. We show how restricted fusion can be implemented and develop an efficient MCMC (Markov chain Monte Carlo) method for posterior computation. The performance of the proposed method is investigated on simulated data and we illustrate its application on real data from EU-SILC (European Union Statistics on Income and Living Conditions).

Keywords: spike and slab prior, sparsity, nominal and ordinal predictor, regression model, MCMC, Gibbs sampler.

1 Introduction

In many applications, especially in medical, social or economic studies, potential covariates collected for a regression analysis are categorical, measured either on an ordinal or on a nominal scale. The usual strategy for modelling the effect of a categorical covariate is to define one level as baseline and to use dummy variables for the effects of the other levels with respect to this baseline. Hence, the effect of a categorical covariate is not captured by a single but by a group of regression effects. Including categorical variables as covariates in regression type models can therefore easily lead to a high-dimensional vector of regression effects. Moreover, as only observations with a specific level contribute information on this level effect, estimated effects of rare levels will be associated with high uncertainty.

Many methods have been proposed to achieve sparsity in regression models by identifying regressors with non-zero effects. Whereas frequentist methods, e.g. the lasso (Tibshirani, 1996) or the elastic net (Zou and Hastie, 2005) rely on penalties, Bayesian variable selection methods are based on the specification of appropriate prior distributions, e.g. shrinkage priors (Park and Casella, 2008; Griffin and Brown, 2010) or spike and slab priors (Mitchell and Beauchamp, 1988; George and McCulloch, 1997; Ishwaran and Rao, 2005). However, variable selection methods identify single non-zero regression

*Department of Applied Statistics, Johannes Kepler University, Linz, Austria, daniela.pauger@jku.at

*Department of Applied Statistics, Johannes Kepler University, Linz, Austria, helga.wagner@jku.at

© 2019 International Society for Bayesian Analysis https://doi.org/10.1214/18-BA1096

effects and do not take into account the natural grouping of the set of dummy variables capturing the effect of a categorical covariate.

Moreover, for a categorical covariate a sparser representation of its effect cannot only be achieved by restricting some or all of its level effects to zero but also when some of the levels have the same effect. To address this problem, we propose a Bayesian approach to achieve sparsity by encouraging both shrinkage of non-relevant effects to zero as well as fusion of (almost) identical level effects.

Methods that explicitly address inclusion or exclusion of a whole group of regression coefficients are the group lasso (Yuan and Lin, 2006), the Bayesian group lasso (Raman et al., 2009; Kyung et al., 2010) and the approach of Chipman (1996) who uses spike and slab priors for grouped selection of the set of dummy variables related to a categorical predictor. The recently proposed sparse group lasso (Simon et al., 2013) and Bayesian sparse group selection (Chen et al., 2016) aim at sparsity at the group level as well as within selected groups by selecting non-zero regression effects, but do not consider fusion of effects.

For metric predictors, effect fusion is addressed in Tibshirani et al. (2005) with the fused lasso and in Kyung et al. (2010) with its Bayesian counterpart, the Bayesian fused lasso. Both methods assume some ordering of effects and shrink only effect differences of subsequent effects to zero. Hence, they are not appropriate for nominal predictors where any effect difference should be subject to shrinkage.

Sofar, only a few papers consider effect fusion for nominal predictors. Bondell and Reich (2009) propose a modification of the fused lasso for ANOVA (Analysis of variance) and Gertheiss and Tutz (Gertheiss and Tutz, 2009, 2010; Tutz and Gertheiss, 2016) specify different lasso-type penalties for ordinal and nominal covariates. Recently, Tutz and Berger (2018) propose tree-structured clustering of effects of categorical covariates. In a Bayesian approach, fusion (or merging) of levels of categorical variables is addressed only in Dellaportas and Tarantola (2005). Their goal is to analyse dependence of categorical variables in loglinear models. To search the huge space of models, that can be obtained by collapsing levels of categorical variables, a reversible jump algorithm is employed. This method could be extended to search the space of regression type models where levels of categorical predictors are subject to merging, but we suggest a different approach where Bayesian inference is feasible via a simple Gibbs sampling algorithm.

To allow for effect fusion we specify a joint multivariate Normal prior for all level effects of one covariate with a precision matrix that allows for either almost perfect or low dependence of regression effects. This prior is related to spike and slab prior distributions that have been applied extensively in Bayesian approaches to variable selection. We show that the prior proposed for effect fusion can be derived alternatively by specifying spike and slab prior distributions on all level effects as well as their differences and taking into account their linear dependence. Whereas with the usual variable selection prior, regression effects can be classified as (almost) zero, if assigned to the spike, and as non-zero otherwise, the effect fusion prior allows for intrinsic classification not only of effects but also of effect differences as negligible or relevant. In contrast to a categorical predictor where typically each pair of level effects will be subject to fusion, for an ordinal

predictor the available ordering information can be exploited by restricting fusion to adjacent categories. We show how restriction of effect fusion to specific pairs of effects can be implemented in our framework.

For ease of exposition we will discuss construction of the prior and MCMC (Markov chain Monte Carlo) inference for a Normal linear regression model with categorical covariates. However, the method we propose can be used in any regression type model with a structured additive predictor, where additionally to effects of categorical covariates further effects, e.g. linear or nonlinear effects of continuous covariates, spatial or random effects are included.

The rest of the paper is organised as follows: in Section 2 we introduce the data model and in Section 3 we describe construction of the prior distribution to encourage a sparse representation of the effect of a categorical predictor. Posterior inference is discussed in Section 4 and Section 5 investigates the performance of the method for simulated data. Application of Bayesian effect fusion is illustrated on a real data example in Section 6 and we conclude with Section 7.

2 Model specification

We consider a standard linear regression model with Normal response y and p categorical covariates where covariate h has c_h+1 ordered or unordered levels $0, \dots, c_h$. To represent the effect of covariate h on the response y, we define 0 as the baseline category and introduce dummy variables, $X_{h,k}$, to capture the effect of level k. The regression model is then given as

$$y = \mu + \sum_{h=1}^{p} \sum_{k=1}^{c_h} X_{h,k}\beta_{h,k} + \varepsilon, \tag{1}$$

where μ is the intercept, $\beta_{h,k}$ is the effect of level k of covariate h (with respect to the baseline category 0) and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is the error term.

For an $(n \times 1)$ response vector $\mathbf{y} = (y_1, \dots, y_n)'$ we write the model as

$$\mathbf{y} = \mathbf{1}\mu + \sum_{h=1}^{p} \mathbf{X}_h\boldsymbol{\beta}_h + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right), \tag{2}$$

where $\mathbf{X}_h$ is the $(n \times c_h)$ design matrix for covariate h, $\boldsymbol{\beta}_h$ is the $(c_h \times 1)$ vector of the corresponding regression effects and $\boldsymbol{\varepsilon}$ the $(n \times 1)$ vector of error terms. $\mathbf{1}$ denotes a vector with elements 1 and $\mathbf{I}$ the identity matrix.

3 Prior specification

Bayesian model specification is completed by assigning prior distributions to all model parameters. We assume a prior of the structure

$$p(\mu, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p, \sigma^2) = p(\sigma^2)p(\mu) \prod_{h=1}^{p} p(\boldsymbol{\beta}_h|\tau_h^2, \boldsymbol{\delta}_h)p(\tau_h^2)p(\boldsymbol{\delta}_h),$$

where τ_h^2 and $\boldsymbol{\delta}_h$ denote additional hyperparameters, which are specified below. We assign a flat proper prior $p(\mu) \sim \mathcal{N}(0, M_0)$ to the intercept, and an Inverse Gamma prior $p(\sigma^2) \sim \mathcal{G}^{-1}(s_0, S_0)$ to the error variance. In our analyses we use the standard improper prior $p(\sigma^2) \propto 1/\sigma^2$.

To allow for effect fusion we specify the prior on the regression effects $\boldsymbol{\beta}_h$ hierarchically as

$$\boldsymbol{\beta}_h | \tau_h^2, \boldsymbol{\delta}_h \sim \mathcal{N}\left(\mathbf{0}, \mathbf{B}_{h0}(\boldsymbol{\delta}_h, \tau_h^2)\right), \tag{3}$$

$$\tau_h^2 \sim \mathcal{G}^{-1}(g_{h0}, G_{h0}), \tag{4}$$

with prior covariance matrix of $\boldsymbol{\beta}_h$ given as

$$\mathbf{B}_{h0}(\boldsymbol{\delta}_h, \tau_h^2) = \gamma_h \tau_h^2 \mathbf{Q}_h^{-1}(\boldsymbol{\delta}_h). \tag{5}$$

Here γ_h is a fixed constant, τ_h^2 is a scale parameter and the matrix $\mathbf{Q}_h(\boldsymbol{\delta}_h)$ determines the structure of the prior precision matrix of $\boldsymbol{\beta}_h$. To encourage effect fusion, we let $\mathbf{Q}_h(\boldsymbol{\delta}_h)$ depend on a vector $\boldsymbol{\delta}_h$ of binary indicator variables $\delta_{h,kj}$, which are defined for each pair of level effects $\beta_{h,k}$ and $\beta_{h,j}$ subject to fusion. $\delta_{h,kj} = 1$ indicates that $\beta_{h,k}$ and $\beta_{h,j}$ differ considerable and hence two regression parameters are needed to capture their respective effects whereas for $\delta_{h,kj} = 0$ the effects are almost identical and the two level effects could be fused. To allow also fusion of level effects to 0, i.e. conventional variable selection, we define $\beta_{h,0} = 0$ and include in $\boldsymbol{\delta}_h$ also indicators $\delta_{h,k0}$, $k = 1, \dots c_h$.

The dimension of $\boldsymbol{\delta}_h$ and the concrete specification of $\mathbf{Q}_h(\boldsymbol{\delta}_h)$ depend on which pairs of effects are subject to fusion. For a nominal covariate typically levels are completely unstructured and hence any pair of effects might be fused. We discuss this case where fusion is unrestricted in Section 3.1.

In contrast, for an ordinal covariate the information on the ordering of levels suggests to restrict fusion to adjacent categories (Gertheiss and Tutz, 2009). Restrictions that preclude direct fusion for specified pairs of effects can easily be implemented in the specification of the prior covariance matrix $\mathbf{B}_{h0}(\boldsymbol{\delta}_h, \tau_h^2)$. We present two different ways to incorporate fusion restrictions in Section 3.2 and discuss the case of an ordinal covariate where fusion is restricted to adjacent categories in more detail.

Marginalized over the indicators $\boldsymbol{\delta}_h$, the prior on $\boldsymbol{\beta}_h$ is a mixture of multivariate Normal distributions

$$p(\boldsymbol{\beta}_h | \tau_h^2) = \sum_{\boldsymbol{\delta}_h} p(\boldsymbol{\delta}_h) f_{\mathcal{N}}\left(\boldsymbol{\beta} | \mathbf{0}, \mathbf{B}_{h0}(\boldsymbol{\delta}_h, \tau_h^2)\right),$$

with different component covariance matrices depending on $\boldsymbol{\delta}_h$ and mixture weights $p(\boldsymbol{\delta}_h)$. We discuss specification of the prior on $\boldsymbol{\delta}_h$ in Section 3.3 and the choice of the hyperparameters in Section 3.4.

For notational convenience we drop the covariate index h in the rest of this Section.

3.1 Prior for unrestricted effect fusion

To perform unrestricted effect fusion for a categorical covariate with levels $0, \dots, c$, we introduce a binary indicator δ_{kj} for each pair of effects $0 \leq j < k \leq c$. Thus, the vector $\boldsymbol{\delta}$ subsuming all these indicators is of dimension $d \times 1$ where $d = \binom{c+1}{2}$.

We specify the structure matrix $\mathbf{Q}(\boldsymbol{\delta})$ as

$$\mathbf{Q}(\boldsymbol{\delta}) = \begin{pmatrix} \sum_{j \neq 1} \kappa_{1j} & -\kappa_{12} & \dots & -\kappa_{1c} \\ -\kappa_{21} & \sum_{j \neq 2} \kappa_{2j} & \dots & -\kappa_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ -\kappa_{c1} & -\kappa_{c2} & \dots & \sum_{j \neq c} \kappa_{cj} \end{pmatrix} \tag{6}$$

with diagonal elements q_{kk} given as

$$q_{kk} = \sum_{j \neq k} \kappa_{kj} = \kappa_{k0} + \dots + \kappa_{k,j-1} + \kappa_{k,j+1} + \kappa_{k,c} \qquad k = 1, \dots, c.$$

For $j < k$, κ_{kj} is defined as

$$\kappa_{kj} = \delta_{kj} + r(1 - \delta_{kj}),$$

and $\kappa_{jk} = \kappa_{kj}$ for $j > k$. The value of δ_{kj} determines whether $\kappa_{kj} = 1$ (for $\delta_{kj} = 1$) or $\kappa_{kj} = r$ (for $\delta_{kj} = 0$). r is a fixed large number, which we call precision ratio for reasons explained below. Finally, we set $\gamma = c/2$.

We discuss this specification now in more detail. First, as the structure matrix $\mathbf{Q}(\boldsymbol{\delta})$ determines the prior precision matrix $\mathbf{B}_0^{-1}(\boldsymbol{\delta}, \tau^2)$ up to the scale factor $1/(\gamma\tau^2)$ it has to be symmetric and positive definite. Symmetry of $\mathbf{Q}(\boldsymbol{\delta})$ is guaranteed by definition and positive definiteness as

$$\boldsymbol{\beta}' \mathbf{Q}(\boldsymbol{\delta}) \boldsymbol{\beta} = \sum_{k=1}^{c} \beta_k^2 \kappa_{k0} + \sum_{k=2}^{c} \sum_{j=1}^{k-1} (\beta_k - \beta_j)^2 \kappa_{kj} > 0, \tag{7}$$

if $\boldsymbol{\beta} \neq \mathbf{0}$, see Supplemental Appendix A.1 (Pauger and Wagner (2018)) for a detailed proof.

The diagonal elements q_{kk} determine the prior partial precisions and the off-diagonal elements q_{kj} the prior partial correlations of the level effects:

$$\mathrm{Cor}(\beta_k, \beta_j | \boldsymbol{\beta}_{\setminus kj}) = -\frac{q_{kj}}{\sqrt{q_{jj} q_{kk}}} = \frac{\kappa_{kj}}{\sqrt{q_{jj} q_{kk}}}, \tag{8}$$

$$\mathrm{Prec}(\beta_k | \boldsymbol{\beta}_{\setminus k}) = q_{kk}/(\gamma\tau^2). \tag{9}$$

Thus, depending on the value of the binary indicator δ_{kj}, the prior allows for high (if $\delta_{kj} = 0$) or low (if $\delta_{kj} = 1$) positive prior partial correlation of β_k and β_j.

The prior partial precision $\mathrm{Prec}(\beta_k | \boldsymbol{\beta}_{\setminus k})$ can take one of c different values, depending on the binary indicators involving level k. Subsuming these indicators in $\tilde{\boldsymbol{\delta}}_k =$

$(\delta_{k0}, \ldots, \delta_{k,k-1}, \delta_{k+1,k}, \ldots \delta_{c,k})$, the minimum value for $\text{Prec}(\beta_k|\boldsymbol{\beta}_{\setminus k})$ results for $\tilde{\boldsymbol{\delta}}_k = \mathbf{1}$ as $c/(\gamma\tau^2)$, and its maximum value, attained for $\tilde{\boldsymbol{\delta}}_k = \mathbf{0}$, is $r \cdot c/(\gamma\tau^2)$. Thus the precision ratio r is the ratio of maximum to minimum prior partial precision. With our choice of $\gamma = c/2$ the partial prior precision ranges from $2/\tau^2$ to $r \cdot 2/\tau^2$, and does not depend on the number levels c.

To illustrate the specification of the structure matrix $\mathbf{Q}(\boldsymbol{\delta})$ we consider a covariate with $c = 3$ levels where only one of the indicators δ_{kj}, $0 \leq j < k \leq c$ has the value 0 whereas all others are 1. For a precision ratio of $r = 10000$ and $\delta_{10} = 0$ the structure matrix is given as

$$\mathbf{Q}(\boldsymbol{\delta}) = \begin{pmatrix} 10002 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{pmatrix}.$$

The marginal prior on β_1 is concentrated close to zero, thus encouraging fusion to the baseline category. If $\delta_{31} = 0$, the structure matrix is

$$\mathbf{Q}(\boldsymbol{\delta}) = \begin{pmatrix} 10002 & -1 & -10000 \\ -1 & 3 & -1 \\ -10000 & -1 & 10002 \end{pmatrix}.$$

Hence, the joint prior on (β_1, β_3) is concentrated close to $\beta_1 = \beta_3$ and encourages fusion of these two effects.

The quadratic form given in (7) suggests an interpretation of the effect fusion prior in terms of conditional Normal priors with zero mean and precision proportional to κ_{kj} on all effect differences $\theta_{kj} = \beta_k - \beta_j$, $0 \leq j < k \leq c$. For $\delta_{kj} = 0$ the prior precision is high with $\kappa_{kj} = r$ and hence the effect difference θ_{kj} is concentrated around zero, whereas it is more dispersed for $\delta_{kj} = 1$ where $\kappa_{kj} = 1$. Actually, as we show in Supplemental Appendix A.2 the effect fusion prior specified above can be derived alternatively by first specifying independent spike and slab priors on all effect differences θ_{kj}, and then correcting for the linear restrictions $\theta_{kj} = \theta_{k0} - \theta_{j0}$. Thus, the proposed prior allows for shrinkage of effect contrasts to zero while taking into account their linear dependence.

As all pairwise effect contrasts θ_{kj} are taken into account symmetrically, the effect fusion prior is invariant to the choice of the baseline category, see Supplemental Appendix A.3 for a formal proof. This invariance distinguishes the effect fusion prior from the conventional spike and slab prior used for variable selection, which allows only shrinkage of regression effects $\beta_k = \theta_{k0}$, i.e. effect contrasts with respect to the baseline, but not all other effect contrasts θ_{kj}, where $j > 0$.

Finally, we note that from a frequentist perspective, the effect fusion prior can be interpreted as an adaptive quadratic penalty with either heavy or slight penalization of effect differences, see (7). Similarly, Gertheiss and Tutz (2010) in a frequentist approach use a weighted L_1 penalty on the effect differences, which has the advantage that effect differences are not only shrunken but can actually be set to zero.

3.2 Prior for restricted effect fusion

We consider now the case where due to available information on the structure of levels fusion is restricted to specific pairs of level effects. A prominent example is an ordinal covariate where the ordering of levels suggests to allow only fusion of subsequent level effects β_{k-1} and β_k. A restriction that e.g. β_k and β_j should not be fused can be implemented in our prior in two ways: we can either fix the corresponding indicator at $\delta_{kj} = 1$ or set the corresponding element in the prior precision matrix $\mathbf{Q}(\boldsymbol{\delta})$ to zero. Whereas $q_{kj} = 0$ is a hard restriction which implies conditional independence of β_k and β_j, setting $\delta_{kj} = 1$ is a soft restriction which implies that effects β_k and β_j are still smoothed to each other.

The implementation of soft restrictions is straightforward, but (hard) conditional independence restrictions require slight modifications in the definition of the structure matrix $\mathbf{Q}(\boldsymbol{\delta})$, the vector of indicators $\boldsymbol{\delta}$ and the constant γ. To specify hard fusion restrictions we introduce a vector $\boldsymbol{\zeta}$ of indicators ζ_{kj}, which are defined for each effect difference θ_{kj}. The elements of $\boldsymbol{\zeta}$ are fixed and indicate whether an effect difference is subject to fusion (for $\zeta_{kj} = 1$) or not (for $\zeta_{kj} = 0$). Deviating from unrestricted effect fusion considered in Section 3.1, we define a stochastic indicator δ_{kj} only for those effect differences where $\zeta_{kj} = 1$ and hence the dimension of $\boldsymbol{\delta}$ is $d = \sum_{k=1}^{c} \sum_{0 \le j < k} \zeta_{kj}$.

To allow off-diagonal elements of the prior precision to be zero, we set

$$q_{kj} = \begin{cases} -\kappa_{kj} & \text{if } \zeta_{kj} = 1, \\ 0 & \text{if } \zeta_{kj} = 0, \end{cases}$$

and $q_{jk} = q_{kj}$. Thus, q_{kj} takes the value zero if $\zeta_{kj} = 0$ and $-\kappa_{kj}$ otherwise. Accordingly, the diagonal elements are specified as

$$q_{kk} = \begin{cases} \kappa_{k0} - \sum_{k \neq j} q_{kj} & \text{if } \zeta_{k0} = 1, \\ \quad - \sum_{k \neq j} q_{kj} & \text{if } \zeta_{k0} = 0. \end{cases}$$

As noted above, an important special case is an ordinal covariate, where fusion of effects can be restricted to adjacent categories, see Gertheiss and Tutz (2009). We define

$$\zeta_{kj} = \begin{cases} 1 & \text{for } j = k - 1, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the vector of indicators $\boldsymbol{\delta}$ has only $d = c$ elements, $\boldsymbol{\delta} = (\delta_{10}, \dots \delta_{c,c-1})$ and $\mathbf{Q}(\boldsymbol{\zeta}, \boldsymbol{\delta})$ is a tri-diagonal matrix with elements

$$\mathbf{Q}(\boldsymbol{\zeta}, \boldsymbol{\delta}) = \begin{pmatrix} \kappa_{10} + \kappa_{21} & -\kappa_{12} & \dots & 0 & 0 \\ -\kappa_{21} & \kappa_{21} + \kappa_{32} & \dots & . & 0 \\ 0 & -\kappa_{32} & & \dots & . \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ . & . & \dots & \kappa_{c-1,c-2} + \kappa_{c,c-1} & -\kappa_{c,c-1} \\ 0 & 0 & \dots & -\kappa_{c,c-1} & \kappa_{c,c-1} \end{pmatrix}. \tag{10}$$

In this case, the maximum value of a diagonal element is $q_{kk} = 2r$ and therefore we set $\gamma = 1$.

It is easy to show that this specification of $\mathbf{Q}(\boldsymbol{\zeta}, \boldsymbol{\delta})$ corresponds to a random walk prior on the regression effects:

$$\beta_k = \beta_{k-1} + \theta_k, \qquad \theta_k \sim \mathcal{N}\left(0, \tau^2/\kappa_{k,k-1}\right),$$

with initial value $\beta_0 = 0$. Due to the spike and slab structure, this prior allows for adaptive smoothing, with almost no smoothing for $\delta_{k,k-1} = 1$ and pronounced smoothing for $\delta_{k,k-1} = 0$.

Another special case of a restricted effect fusion prior is the standard spike and slab prior used for variable selection, which encourages only fusion to the baseline, i.e. shrinkage of β_k to $\beta_0 = 0$. In our framework the spike and slab prior is recovered with

$$\zeta_{kj} = \begin{cases} 1 & \text{for } j = 0, \\ 0 & \text{otherwise,} \end{cases}$$

and $\gamma = 1$. Therefore the off-diagonal elements of $\mathbf{Q}(\boldsymbol{\zeta}, \boldsymbol{\delta})$ are zero and $q_{kk} = \kappa_{k0}$.

3.3 Prior on the indicator variables

To perform variable selection the elements of $\boldsymbol{\delta}$ are usually assumed to be conditionally independent a priori with $p(\delta_{kj} = 1) = \omega$, where ω is either fixed or assigned a hyperprior $\omega \sim \mathcal{B}(v_0, w_0)$. This would be possible also for the effect fusion prior, but from a computational point of view a more convenient choice is to set

$$p(\boldsymbol{\delta}) \propto |\mathbf{Q}(\boldsymbol{\delta})|^{-1/2} r^{\sum (1-\delta_{kj})/2}. \tag{11}$$

Thus, the determinant of $\mathbf{Q}(\boldsymbol{\delta})$ cancels out in the joint prior of regression effects and indicators $p(\boldsymbol{\beta}, \boldsymbol{\delta}|\tau^2)$, which results as

$$p(\boldsymbol{\beta}, \boldsymbol{\delta}|\tau^2) = p(\boldsymbol{\beta}|\boldsymbol{\delta}, \tau^2)p(\boldsymbol{\delta}) = (\frac{1}{\gamma\tau^2})^{c/2} \exp\big(-\frac{\boldsymbol{\beta}'\mathbf{Q}(\boldsymbol{\delta})\boldsymbol{\beta}}{2\gamma\tau^2}\big)(\sqrt{r})^{\sum(1-\delta_{kj})}. \tag{12}$$

This prior has attractive features: Firstly, as $p(\boldsymbol{\beta}, \boldsymbol{\delta}|\tau^2)$ can be factorized as

$$p(\boldsymbol{\beta}, \boldsymbol{\delta}|\tau^2) \propto \prod_{k \neq j} (\sqrt{r})^{1-\delta_{kj}} \exp\Big(-\frac{(\beta_k - \beta_j)^2}{2\tau^2\gamma}\big(\delta_{kj} + r(1-\delta_{kj})\big)\Big) \propto \tag{13}$$

$$\propto \prod_{k \neq j} p(\theta_{kj}, \delta_{kj}|\tau^2), \tag{14}$$

the effect differences θ_{kj} and indicators δ_{kj} are jointly independent across all pairs $0 \leq j < k \leq c$ conditional on the scale parameter τ^2. Secondly, conditioning on the corresponding effect difference θ_{kj} the conditional prior of the indicator δ_{kj},

$$p(\delta_{kj} = 1|\boldsymbol{\beta}, \tau^2) = \frac{p\big(\theta_{kj}|\mathcal{N}\left(0, \gamma\tau^2\right)\big)}{p\big(\theta_{kj}|\mathcal{N}\left(0, \gamma\tau^2\right)\big) + p\big(\theta_{kj}|\mathcal{N}\left(0, \gamma\tau^2/r\right)\big)}, \tag{15}$$

is identical for all pairs of indices k, j subject to fusion.

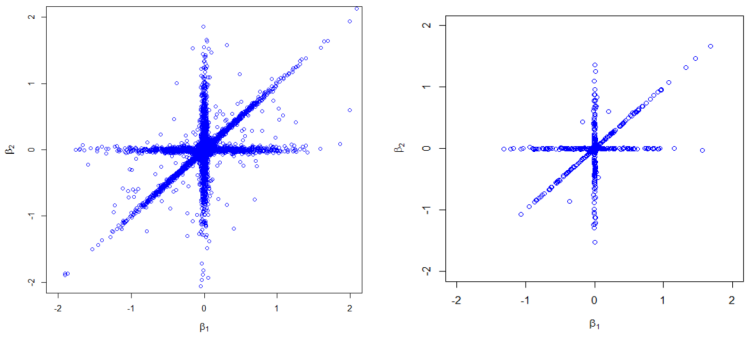

Figure 1: Simulation from the effect fusion prior: Plot of (β_1, β_2) for $c = 3$ and hyper-parameters $g_0 = 5$, $G_0 = 2$, $r = 1000$ (left) and $r = 10000$ (right).

The properties of the prior $p(\boldsymbol{\delta})$ given in (11) depend on the concrete specification of $\boldsymbol{\delta}$. For the special cases of restricted effect fusion discussed above, fusion of subsequent levels of an ordinal covariate with $\boldsymbol{\delta} = (\delta_{10}, \dots \delta_{c,c-1})$ and $\mathbf{Q}(\boldsymbol{\delta})$ specified in (10), and variable selection with $\boldsymbol{\delta} = (\delta_{10}, \dots \delta_{c,0})$ and $\mathbf{Q}(\boldsymbol{\delta}) = \text{diag}(\kappa_{10}, \dots, \kappa_{c0})$, the prior $p(\boldsymbol{\delta})$ is uniform over all 2^c different values of $\boldsymbol{\delta}$,

$$p(\boldsymbol{\delta}) \propto 1,$$

see Supplemental Appendix A.4 for a formal proof.

In contrast, for a nominal covariate with unrestricted effect fusion $p(\boldsymbol{\delta})$ specified in (11) favours sparse models. As $\mathbf{Q}(\boldsymbol{\delta})$ has a more complicated structure in this case (see formula 18 in the Appendix), its determinant is not available in closed form for arbitrary $\boldsymbol{\delta}$, but we can compare the full model, where $\boldsymbol{\delta} = \mathbf{1}$ and the null model where $\boldsymbol{\delta} = \mathbf{0}$ and hence all effects are sampled from the spike component. As $|\mathbf{Q}(\mathbf{0})| = |r\mathbf{Q}(\mathbf{1})|$, we get

$$\frac{p(\boldsymbol{\delta} = \mathbf{0})}{p(\boldsymbol{\delta} = \mathbf{1})} = \frac{|\mathbf{Q}(\mathbf{1})|^{1/2}}{|\mathbf{Q}(\mathbf{0})|^{1/2}} \cdot r^{d/2} = r^{-c/2} r^{d/2} = r^{c(c-1)/2}.$$

Thus, a priori the null model is increasingly favoured over the full model with higher number of levels and higher precision ratio.

Figure 1 shows simulations from the marginal prior $p(\boldsymbol{\beta})$ for a categorical regressor with $c = 3$ levels (except baseline) for $r = 1000$ and $r = 10000$. Due to the symmetry of the prior with respect to all level effects only plots for (β_1, β_2) are presented. The prior is concentrated at configurations where all or a subset of the regression effects is zero, or where two effects are equal, $\beta_j = \beta_k$ and concentration at these configurations increases with the precision ratio r.

Corresponding plots for simulations from the variable selection prior, where fusion is restricted to $\beta_k = 0$, and the ordinal fusion prior with fusion restricted to $\beta_k = \beta_{k-1}$ are shown in Figure 2. Compared to unrestricted effect fusion, sparsity is much less pronounced in these two priors.

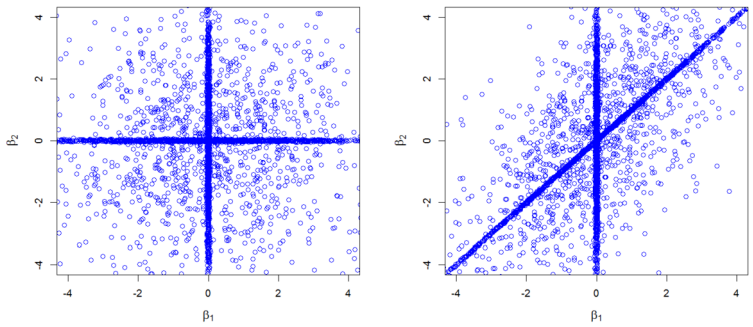

Figure 2: Simulation from the prior: Plot of (β_1, β_2) for $c = 3$ and hyperparameters $g_0 = 5$, $G_0 = 20$ and $r = 10000$. Left: variable selection prior, right: ordinal fusion prior.

3.4 Choice of hyperparameters

To provide a rationale for the choice of the hyperparameters r, g_0 and G_0 we focus on the joint marginal prior distribution of one effect difference and the corresponding indicator $0 \leq j < k \leq c$. As already noted above, the random variables $(\theta_{kj}, \delta_{kj})$ are independent and identically distributed for all pairs of indices k, j conditional on the scale parameter τ^2. Marginalized over δ_{kj} and τ^2 the prior on an effect difference θ_{kj} is a spike and slab distribution, where both the spike and the slab are scaled t-distributions with $2g_0$ degrees of freedom and scale parameter $\sigma = \sqrt{G_0/g_0}$ (for $\delta_{kj} = 1$) and $\sigma/\sqrt{r}$ (for $\delta_{kj} = 0$), respectively. For effect fusion we follow the standard choice in variable selection to set $g_0 = 5$ (Fahrmeir et al., 2010; Scheipl et al., 2012) where tails of spike and slab fat enough to avoid MCMC mixing problems when the indicator δ_{kj} is sampled conditional on θ_{kj}.

Our modelling goal is to allow for fusion of level effects with negligible difference while level effects with relevant difference should be modelled with seperate effects. To avoid misclassification of relevant differences as negligible, which we call *false negatives* or of negligible effects as relevant, i.e. *false positives*, the scale parameter G_0 and the precision ration r could be chosen by specifying the conditional fusion probability

$$P(\delta_{kj} = 0|\theta_{kj}) = \frac{1}{1 + \frac{t_{2g_0}(\frac{\theta_{kj}}{\sigma})}{\sqrt{r}t_{2g_0}(\sqrt{r}\frac{\theta_{kj}}{\sigma})}}$$

for two values of θ_{kj}.

Figure 3 shows plots of the fusion probabilities as a function of the effect difference for various values of r and G_0 with $g_0 = 5$. For fixed θ_{kj} the fusion probability $P(\delta_{kj} = 0|\theta_{kj})$ decreases with r and increases with G_0 which suggests to choose a large value for r large and a small value for G_0. However for smaller values of G_0 shrinkage effect differences to zero is more pronounced even under the slab, which might hamper detection of small effect differences and hence G_0 has to be chosen carefully to represent the scale of relevant effects.

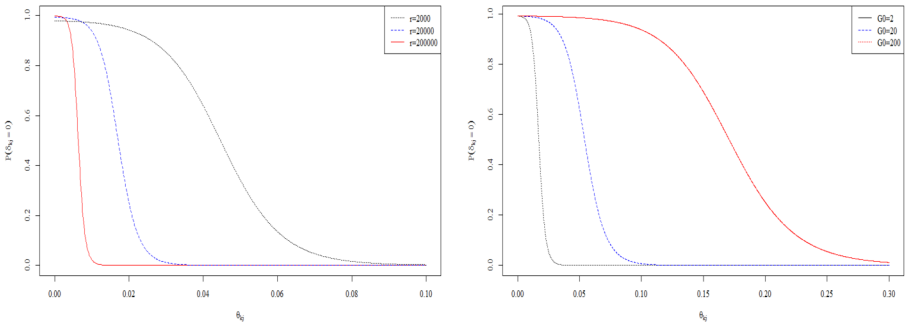

Figure 3: Fusion probabilities for $g_0 = 5$. Left: $G_0 = 2$ and different values of r, right: $r = 20000$ and different values of G_0.

We will investigate different values for both hyperparameters G_0 and r in the simulation study in Section 5.3. As suggested by a referee, we investigate also a hyperprior on G_0, where a convenient choice is an exponential prior $G_0 \sim \mathcal{E}(\lambda)$ with mean λ. We tried also to put a hyperprior on r, which did however not prove useful as the additional flexibility resulted in all effect differences being assigned to the spike component.

4 Posterior inference

Our goal is posterior inference for the parameters of linear regression model

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

with prior

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

Here the regressor matrix is $\mathbf{X} = [\mathbf{1}, \mathbf{X}_1, \ldots, \mathbf{X}_p]$ and the vector of regression effects is $\boldsymbol{\beta} = (\mu, \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_p')'$. The vector $\boldsymbol{\delta} = (\boldsymbol{\delta}_1', \ldots, \boldsymbol{\delta}_p')'$ stacks the vectors of binary indicators $\boldsymbol{\delta}_h$, $h = 1, \ldots, p$ and $\boldsymbol{\tau}^2 = (\tau_1^2, \ldots, \tau_p^2)'$ subsumes the scale parameters τ_h^2. The prior for $\boldsymbol{\beta}$ is multivariate Normal, $\boldsymbol{\beta}|\boldsymbol{\delta}, \boldsymbol{\tau}^2 \sim \mathcal{N}\left(\mathbf{0}, \mathbf{B}_0(\boldsymbol{\delta}, \boldsymbol{\tau}^2)\right)$ with covariance matrix $\mathbf{B}_0(\boldsymbol{\delta}, \boldsymbol{\tau}^2)$, which is block-diagonal with elements M_0 (for the intercept) and $\mathbf{B}_{h0}(\boldsymbol{\delta}_h, \tau_h^2)$, $h = 1, \ldots, p$.

The resulting posterior distribution is proper even for the partial improper prior with $p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$ under conditions given by a theorem of Sun et al. (2001), see Supplemental Appendix B.1 for details.

Posterior inference can be accomplished by sampling from the posterior distribution using MCMC methods. For the prior distributions specified above posterior inference for the model parameters is feasible by a Gibbs sampler, where the full conditionals are standard distributions. The sampling scheme is outlined in Section 4.1. Model averaged estimates of the parameters can be obtained as the means of the MCMC draws but often the goal is to select a final model with eventually fused levels. In Section 4.2,

we discuss how this model selection problem can be addressed in Bayesian decision theoretic approach by choosing an appropriate loss function.

4.1 MCMC scheme

We initialise MCMC by choosing starting values for the error variance σ^2, the indicators $\boldsymbol{\delta}$ and the scale parameters $\boldsymbol{\tau}^2$ and compute the prior covariance matrix $\mathbf{B}_0(\boldsymbol{\delta}, \boldsymbol{\tau}^2)$.

MCMC then iterates the following steps:

(1) Sample the regression coefficients $\boldsymbol{\beta}$ from the full conditional Normal posterior $p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{\delta}, \boldsymbol{\tau}^2, \mathbf{y})$.

(2) Sample the error variance σ^2 from the full conditional Inverse Gamma distribution $p(\sigma^2|\boldsymbol{\beta}, \mathbf{y})$.

(3) For $h = 1, \ldots, p$: Sample the scale parameter τ_h^2 from the full conditional Inverse Gamma distribution $p(\tau_h^2|\boldsymbol{\beta}_h, \boldsymbol{\delta}_h)$.

(4) For all $h \in \{1, \ldots, p\}$ with a hyperprior specified on G_h: Sample G_h from $p(G_h|\lambda_h, \tau_h^2)$

(5) For $h = 1, \ldots, p$: Sample $\boldsymbol{\delta}_h$ independently from its full conditional posterior $p(\boldsymbol{\delta}_h|\boldsymbol{\beta}_h, \tau_h^2)$.

(6) Update the prior covariance matrix $\mathbf{B}_0(\boldsymbol{\delta}, \boldsymbol{\tau}^2)$.

Due to the hierarchical structure of the prior, the posterior of $\boldsymbol{\delta}_h$ in sampling step (5) depends only on $\boldsymbol{\beta}_h$ and τ_h^2 and as discussed in Section 3.3 elements of $\boldsymbol{\delta}_h$ are conditionally independent given τ_h^2 and $\boldsymbol{\beta}_h$. Therefore all binary indicators $\delta_{h,kj}$ can be sampled independently from $p(\delta_{h,kj} = 1|\theta_{kj} = \beta_{h,k} - \beta_{h,j}, \tau_h^2)$, which is given in (15).

Full details of the sampling steps are provided in Supplemental Appendix B.2. The sampling scheme is implemented in the `R` package `effectFusion` (Pauger et al., 2016).

Compared to posterior inference with a standard Normal prior on the regression effects $\boldsymbol{\beta}$, only steps (3)–(6) have to be added under the effect fusion prior. These steps are fast for nominal covariates that are typical in applications with 2–50 levels. However, as all pairwise differences are assessed in each sweep of the sampler computation times become prohibitive for covariates with 100 or more levels. For a data set of $n = 10000$ observations and one nominal covariate 1000 MCMC iterations take 1.5 seconds for a covariate with $c = 20$ levels, but roughly 25 min. for $c = 100$ levels on a standard laptop with Intel i7-5600U processor with 2.60 GHz and 16 GB RAM. More details on computation times are given in Supplemental Appendix B.3.

As already noted above, due to the hierarchical structure of the effect fusion prior the full conditionals of the parameters $\boldsymbol{\delta}_h, \tau_h^2$ and G_h depend only on the regression effects $\boldsymbol{\beta}_h$ and not on the data likelihood. Thus, implementation of effect fusion for categorical covariates is straightforward in any type of regression model with an additively structured linear predictor.

4.2 Model selection

If the goal is to select a final model, e.g. to be used for prediction, a Bayesian decision theoretic approach requires to choose an appropriate loss function. A particularly appealing loss function for effect fusion is a special case Binder's loss (Binder, 1978) which is used in Lau and Green (2007) for Bayesian model based clustering of observations. Binder's loss considers pairs of items and penalizes incorrect clustering, which occurs when two items which should not be clustered are assigned to the same cluster or when two items which should be clustered are assigned to different clusters. For effect fusion incorrect clustering corresponds to classifying an effect difference falsely as negative or falsely as positive, respectively.

Binder's loss is given as

$$\mathcal{L}(\mathbf{z}, \mathbf{z}^*) = \sum_{j \neq k} \left(\ell_1 \mathbf{I}_{\{z_k = z_j\}} \mathbf{I}_{\{z_k^* \neq z_j^*\}} + \ell_2 \mathbf{I}_{\{z_k \neq z_j\}} \mathbf{I}_{\{z_k^* = z_j^*\}} \right),$$

where $\mathbf{z}$ denotes the true and $\mathbf{z}^*$ the proposed clustering and the constants ℓ_1 and ℓ_2 are misclassification costs. For $\ell_1 = \ell_2$ the expected posterior loss results as

$$\mathrm{E}(\mathcal{L}(\mathbf{z}, \mathbf{z}^*)|\mathbf{y}) = \sum_{j \neq k} |\mathbf{I}_{\{z_k^* = z_j^*\}} - \pi_{kj}|, \tag{16}$$

where $\pi_{kj} = P(z_k = z_j|\mathbf{y})$ denotes the (k, j) element of the posterior similarity matrix, see Fritsch and Ickstadt (2009). The Bayes optimal action, i.e. the clustering which minimizes the expectation of Binder's loss, can be determined by minimizing

$$\sum_{j \neq k} \mathbf{I}_{\{z_k^* = z_j^*\}} \left(\frac{1}{2} - \pi_{kj} \right). \tag{17}$$

For this minimization problem Lau and Green (2007) use an algorithm based on integer programming, which is implemented in the function `minbinder` (`R` package `mcclust`).

We determine the optimal fusion model with respect to the expected posterior Binder loss (16) for each covariate C_h separately and approximate the elements $\pi_{h,kj}$ of the corresponding posterior similarity matrix from M MCMC draws (after burnin) by

$$\hat{\pi}_{h,kj} = \frac{1}{M} \sum_{m=1}^{M} \delta_{h,kj}^{(m)}.$$

Finally, we refit the selected model with dummy-coded regression coefficients for the fused levels under a flat Normal prior $\mathcal{N}(0, \mathbf{I}B_0)$.

5 Simulation study

To investigate the performance of the proposed method we conducted a simulation study with a similar set-up as in Gertheiss and Tutz (2010) where we compare model

averaged results (*FusBMA*) and results of the finally selected model (*FusSel*) under the effect fusion prior with respect to parameter estimation, correct effect fusion and predictive performance to various other approaches: penalized regression (*Penalty*), the Bayesian lasso (*BLasso*), the Bayesian elastic net (*BEN*), the group lasso (*GLasso*), the sparse group lasso (*SGL*) and the Bayesian Sparse Group Lasso (*BSGS*). Additionally, we include Bayesian regularization via graph Laplacian (*GLap*), proposed in Liu et al. (2014), where the prior is also specified directly on the elements of the prior precision matrix, however with the goal to identify conditional independence by shrinking off-diagonal elements to zero. A list of the papers introducing these methods and the related `R` packages is given in Supplemental Appendix C.1. For comparison we also fit the full model (*Full*) with separate dummy variables for each level and the true model (*True*), i.e. the model where the correct fusion is assumed to be known.

5.1 Simulation set-up

We generated 100 data sets with $n = 500$ observations from the Gaussian linear regression model (2) with intercept $\mu = 1$, a standard Normal error $\varepsilon \sim \mathcal{N}(0, 1)$ and fixed design matrix $\mathbf{X}$. We use four ordinal and four nominal predictors, where two regressors have eight and two have four categories for each type of covariate (ordinal and nominal). Regression effects are set to $\boldsymbol{\beta}_1 = (0, 1, 1, 2, 2, 4, 4)$ and $\boldsymbol{\beta}_3 = (0, -2, -2)$ for the ordinal, to $\boldsymbol{\beta}_5 = (0, 1, 1, 1, 1, -2, -2)$ and $\boldsymbol{\beta}_7 = (0, 2, 2)$ for the nominal covariates, and $\boldsymbol{\beta}_h = \mathbf{0}$ for $h = 2, 4, 6, 8$. Levels of the predictors are generated with probabilities $(0.1, 0.1, 0.2, 0.05, 0.2, 0.1, 0.2, 0.05)$ and $(0.1, 0.4, 0.2, 0.3)$ for regressors with eight and four levels, respectively.

To perform effect fusion, we specify a Normal prior with variance $B_0 = 10000$ on the intercept and the improper prior $p(\sigma^2) \propto 1/\sigma^2$ (which corresponds to an Inverse Gamma distribution with parameters $s_0 = S_0 = 0$) on the error variance σ^2. For each covariate h, the hyperparameters are set to $G_{h0} = 20$ and $r = 20000$, but we investigate also various other values for both parameters and an exponential hyperprior on G_{h0} with $\lambda = \mathrm{E}(G_{h0}) = 2$ in Section 5.3.

MCMC is run for 10000 iterations after burnin of 5000 for each data set to obtain model averaged estimates (*FusBMA*) and to perform model selection. Models *Full* and *True* and the refit of the selected model *FusSel* are estimated from 3000 MCMC iterations (after a burnin of 1000) under a flat Normal prior $\mathcal{N}(0, \mathbf{I}B_0)$ with $B_0 = 10000$ on the regression coefficients. The tuning parameters of the frequentist methods *Penalty* and *GLasso* are selected automatically via cross-validation in the corresponding `R` packages. For *SGL* we choose the penalty parameter via cross-validation in the range from 0.00005 to 0.05. For the Bayesian methods, we use the default prior parameter settings in the code (for *GLap*) and the `R` packages `monomvn` and `EBglmNet` and estimate the regression coefficients by the posterior means. For *BSGS* which is tailored to sparse group selection of numeric regressors in a model with no intercept the recommendation in Lee and Chen (2015) to demean the response is not useful in our setting and hence we set the prior inclusion probability for the intercept to 0.99 (which due to implementation specifics resulted in a higher posterior inclusion probability than setting this value to 1) and used the default value of 0.5 for all other covariates.

5.2 Simulation results

We first compare the different methods with respect to estimation of the regression effects. Figures 4 and 5 show boxplots of the mean squared estimation error (MSE), which is defined for data set i and covariate h as

$$MSE_h^{(i)} = \frac{1}{c_h} \sum_{k=1}^{c_h} (\hat{\beta}_{h,k}^{(i)} - \beta_{h,k})^2,$$

for ordinal covariates ($h = 1, \ldots, 4$) and nominal covariates ($h = 5, \ldots, 8$), respectively.

No method outperforms the others consistently in all data sets for all covariates. The mean MSE (averaged over all 100 data sets) is lower for both versions of Bayesian effect fusion, *FusBMA* and *FusSel*, than in the model *Full* and only slightly higher than in the model *True* for all covariates. *FuseBMA* and *FuseSel* perform similar and outperform all other methods with respect to the mean MSE for the ordinal covariates 1 and 3 and for all nominal covariates (5–8). However they are outperformed in some data sets by *BLasso* and *BSGS* for the ordinal covariates 2 and 4 (both with no effect on the response) and by *BEN* for covariate 4. Though overall performance of *FusBMA* and *FusSel* is very good, for single data sets the MSE can be even higher than for the model *Full*. This occurs when levels with actually different effects are fused, e.g. for some of the 8 levels of the nominal covariate 5.

As noted above the overall performance of *FusBMA* and *FusSel* is very similar in this simulation. Some differences can be seen for covariates 2 and 4 where *FusSel* performs better than *FusBMA* in most data sets, however occasional outliers are more extreme. This is not surprising, as *FusSel* depends only on the selected model and hence the MSE will be lower than for *FusBMA* if the correct model is selected and higher otherwise.

The frequentist method for effect fusion, *Penalty*, which uses a global penalty parameter across all covariates yields a lower MSE than the model *Full* for covariates with no effect and for ordinal covariates, but only small improvements for nominal covariates.

BLasso which aims at shrinkage of effects to zero performs very well for covariates with no effect but also good for the other covariates 1, 3, 5 and 7. The performance of *BEN* is similar, however slightly worse than that of *BLasso* for all covariates. As expected, *GLasso* which aims at sparsity at the group level performs well for covariates with no effect but the MSE values are similar to those of the model *Full* for covariates with non-zero effects. *GLap* which is designed for a different goal yields small improvements for covariates with no effect compared to the model *Full* and performs similar for covariates with non-zero effects. Finally, both *SGL* and *BSGS*, which aim at sparsity at the group level as well as within groups of effects outperform *Full* for covariates with no effect and to a smaller extent also for the nominal covariates 5 and 7, where some level effects are zero. However, *SGL* performs worse then *Full* for covariates 1 and 3, where the ordinal structure is not taken into account. *BSGS* performs almost as good as *True* and similar to Bayesian effect fusion (slightly better for covariates 2 and 4, slightly worse for covariates 6 and 8) for covariates with no effects. For nominal and ordinal covariates with non-zero effects it is outperformed by *FusBMA* and *FusSel* but performs similar as the two other Bayesian methods *BLasso* and *BEN*.

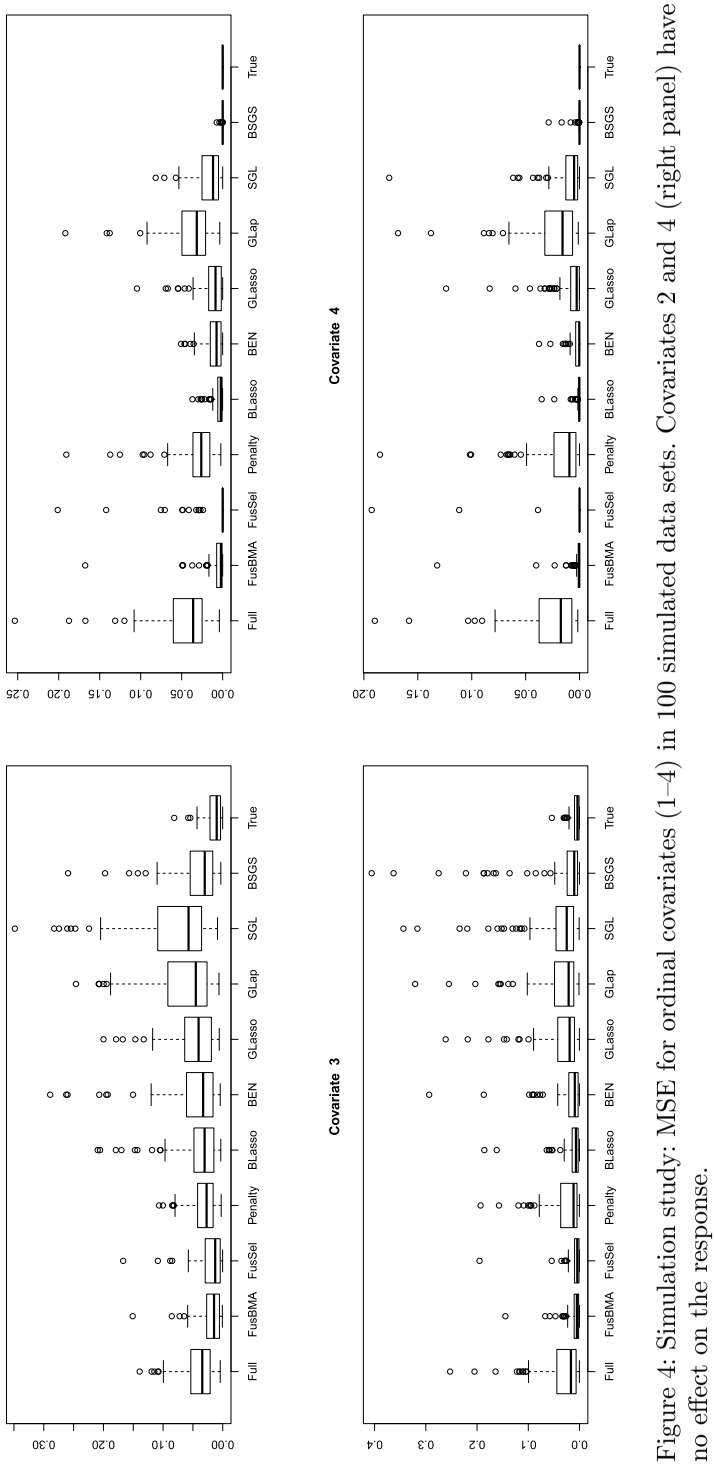

Figure 4: Simulation study: MSE for ordinal covariates (1–4) in 100 simulated data sets. Covariates 2 and 4 (right panel) have no effect on the response.

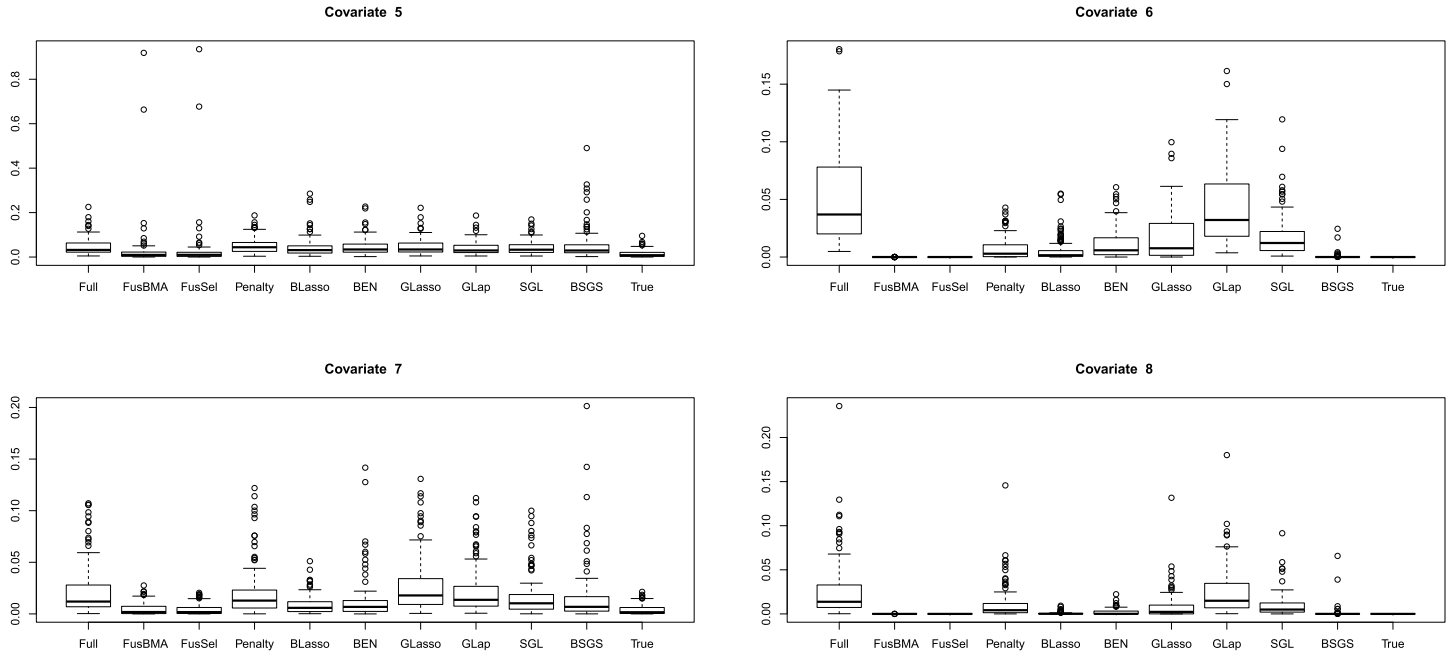

Figure 5: Simulation study: MSE for nominal covariates (5–8) in 100 simulated data sets. Covariates 6 and 8 (right panel) have no effect on the response.

To evaluate the predictive performance of Bayesian effect fusion, we generate a new sample of $n^* = 500$ observations $z_j,\ j = 1, \ldots, n^*$ from the linear regression model (2) with fixed regressors $\tilde{\mathbf{x}}_j$ and the same parameters as in the simulated data sets. Predictions for these new observations are computed using the estimates from each of the original data sets as $\hat{z}_j^{(i)} = \tilde{\mathbf{x}}_j \hat{\boldsymbol{\beta}}^{(i)}, i = 1, \ldots, 100$.

The mean squared prediction errors (MSPE) defined for each data set as

$$\text{MSPE}^{(i)} = \frac{1}{n^*} \sum_{j=1}^{n^*} (z_j - \hat{z}_j^{(i)})^2, \quad i = 1, \ldots, 100$$

are shown in Figure 6. The predictive performance of *FusBMA* and *FusSel* is almost as good as for the model *True* with correctly fused effects, and is considerably better than for all competing methods in most data sets. The Bayesian methods *BLasso* and *BSGS* perform similar with respect to the mean (averaged over 100 data sets) and slightly better than *Penalty*, *BEN* and *GLasso*. Finally, MSPEs for *SGL* and *GLap* are similar to those of model *Full*.

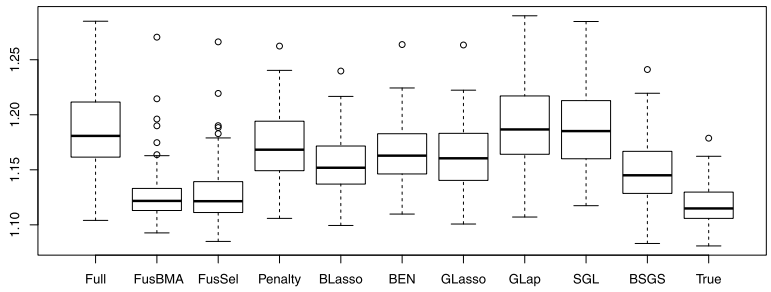

Figure 6: Simulation study: MSPE of 500 new observations. Predictions are based on the estimates from 100 simulated data sets.

To evaluate and compare the performance of the methods with respect to model selection, we report for each covariate the true positive rate (TPR), the true negative rate (TNR), the positive predictive value (PPV) and the negative predictive value (NPV), see Supplemental Appendix C.2 for detailed definitions. If fusion is completely correct, all four values are equal to 100% but TPR and PPV are not defined for covariates where all effects are zero.

For Bayesian effect fusion (*Fusion*) we perform model selection as described in Section 4.2, for the other methods we consider two level effects as identical if the posterior mean of their difference is smaller or equal to 0.01. Tables 1 and 2 report the averages of these statistics in the 100 simulated data sets for each covariate separately. Bayesian effect fusion clearly outperforms all other methods with respect to identifying categories with the same effect with averaged TNR higher than 95% for all covariates. This comes at the cost of occasionally missing a non-zero effect difference and hence an average TPR slightly lower than 100%.

Covariate	Method	TPR	TNR	PPV	NPV
1	Fusion	99.7	95.2	95.3	99.8
	Penalty	100.0	18.8	48.9	100.0
	BLasso	100.0	6.5	44.7	100.0
	BEN	100.0	19.0	48.4	100.0
	GLasso	100.0	2.8	43.7	100.0
	GLap	100.0	2.8	43.6	100.0
	SGL	100.0	8.2	45.2	100.0
	BSGS	100.0	12.0	46.3	100.0
2	Fusion	-	98.3	-	100.0
	Penalty	-	21.3	-	100.0
	BLasso	-	22.3	-	100.0
	BEN	-	54.4	-	100.0
	GLasso	-	20.6	-	100.0
	GLap	-	2.7	-	100.0
	SGL	-	18.6	-	100.0
	BSGS	-	83.9	-	100.0
3	Fusion	100.0	99.0	99.3	100.0
	Penalty	100.0	24.5	44.8	100.0
	BLasso	100.0	24.5	43.5	100.0
	BEN	100.0	47.0	52.7	100.0
	GLasso	100.0	10.0	36.7	100.0
	GLap	100.0	6.5	35.5	100.0
	SGL	100.0	19.0	40.0	100.0
	BSGS	100.0	31.0	45.7	100.0
4	Fusion	-	99.0	-	100.0
	Penalty	-	21.3	-	100.0
	BLasso	-	48.0	-	100.0
	BEN	-	62.0	-	100.0
	GLasso	-	33.0	-	100.0
	GLap	-	7.0	-	100.0
	SGL	-	19.3	-	100.0
	BSGS	-	74.7	-	100.0

Table 1: Simulation study: Model selection results for ordinal covariates. TPR, TNR, PPV and NPV are averaged over 100 data sets.

5.3 Influence of hyperparameters

In this section, we investigate the sensitivity of model selection under the effect fusion prior with respect to the hyperparameters. While false positives result in a loss of estimation efficiency, false negatives will yield biased effect estimates and poor predictive performance, and hence the goal is to avoid *false negatives* while keeping *false positives* at a moderate level. We report false negative rates, $\text{FNR} = 1 - \text{TPR}$ and false positive rates $\text{FPR} = 1 - \text{TNR}$ for various values of G_0 and fixed $r = 20000$ in Table 3 and for various values of r and fixed $G_0 = 20$ in Table 4. Results in Table 3 indicate that

Covariate	Method	TPR	TNR	PPV	NPV
5	Fusion	99.1	98.8	99.5	98.5
	Penalty	100.0	17.2	75.3	100.0
	BLasso	100.0	6.6	72.9	100.0
	BEN	100.0	12.4	74.1	100.0
	GLasso	100.0	4.5	72.4	100.0
	GLap	100.0	4.4	72.4	100.0
	SGL	100.0	7.0	72.9	100.0
	BSGS	100.0	7.1	72.9	100.0
6	Fusion	-	100.0	-	100.0
	Penalty	-	51.4	-	100.0
	BLasso	-	26.8	-	100.0
	BEN	-	53.8	-	100.0
	GLasso	-	19.5	-	100.0
	GLap	-	3.9	-	100.0
	SGL	-	19.2	-	100.0
	BSGS	-	83.2	-	100.0
7	Fusion	100.0	99.5	99.8	100.0
	Penalty	100.0	17.5	71.6	100.0
	BLasso	100.0	21.5	72.5	100.0
	BEN	100.0	43.0	78.3	100.0
	GLasso	100.0	5.0	68.0	100.0
	GLap	100.0	3.5	67.6	100.0
	SGL	100.0	16.0	71.0	100.0
	BSGS	100.0	31.0	75.1	100.0
8	Fusion	-	100.0	-	100.0
	Penalty	-	27.5	-	100.0
	BLasso	-	48.0	-	100.0
	BEN	-	70.2	-	100.0
	GLasso	-	34.3	-	100.0
	GLap	-	4.8	-	100.0
	SGL	-	21.2	-	100.0
	BSGS	-	80.0	-	100.0

Table 2: Simulation study: Model selection results for nominal covariates. TPR, TNR, PPV and NPV are averaged over 100 data sets.

increasing G_0 from 0.2 to 200 has little effect on FNR but yields lower FPR for ordinal predictors (covariates 1–4), and has little effect on FPR but leads to higher FNR for nominal predictors (covariates 5–8). We conclude that for nominal predictors $G_0 = 2$ is a good choice which allows to detect also small effect differences whereas for ordinal predictors we suggest to choose a larger value for G_0, e.g. $G_0 = 20$. An exponential hyperprior with $\mathrm{E}(G_0) = 2$ performs similar to a fixed value of $G_0 = 2$ with slightly lower FNR and FPR for nominal covariates but higher FPR for ordinal covariates.

Table 4 reports FNR and FPR for values of the precision ratio r from $2 \cdot 10^2$ to $2 \cdot 10^5$. Whereas for ordinal predictors FNR and FPR change only little with r, for

	$G_0 = 0.2$		$G_0 = 2$		$G_0 = 20$		$G_0 = 200$		$G_0 \sim \mathcal{E}(2)$	
h	FNR_h	FPR_h	FNR_h	FPR_h	FNR_h	FPR_h	FNR_h	FPR_h	FNR_h	FPR_h
1	0.0	13.0	0.3	10.5	0.3	4.7	0.7	1.7	0.3	9.0
2	-	18.9	-	7.4	-	1.7	-	0.3	-	11.7
3	0.0	11.0	0.0	7.5	0.0	1.0	0.0	0.5	0.0	5.0
4	-	14.3	-	5.0	-	1.0	-	0.3	-	8.7
5	0.4	1.2	0.7	1.0	0.9	1.2	81.0	0.0	0.5	0.9
6	-	0.0	-	0.0	-	0.0	-	0.0	-	0.0
7	0.0	3.0	0.0	2.0	0.0	0.5	0.0	0.0	0.0	1.5
8	-	0.0	-	0.0	-	0.0	-	0.0	-	0.0

Table 3: Simulation study: Model selection results for $r = 20,000$ and various values of G_0.

	$r = 200$		$r = 2{,}000$		$r = 20{,}000$		$r = 200{,}000$	
h	FNR_h	FPR_h	FNR_h	FPR_h	FNR_h	FPR_h	FNR_h	FPR_h
1	0.3	2.7	0.3	4.2	0.3	4.7	0.3	3.0
2	-	0.9	-	1.4	-	1.7	-	1.7
3	0.0	0.5	0.0	1.5	0.0	1.0	0.0	2.0
4	-	0.3	-	0.3	-	1.0	-	1.7
5	100.0	0.0	92.0	0.0	0.9	1.2	0.8	1.9
6	-	0.0	-	0.0	-	0.0	-	0.6
7	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.5
8	-	0.0	-	0.0	-	0.0	-	0.0

Table 4: Simulation study: Model selection results for $G_0 = 20$ and various values of r.

nominal covariates low values of r encourage too much fusion of effects and hence result in a high FNR. Both FNR and FPR are small for $r = 2 \cdot 10^4$ and $r = 2 \cdot 10^5$. As MCMC mixing is adequate for both values, we suggest to choose r in this range.

6 Real data example

As an illustration of Bayesian effect fusion on real data, we model contributions to private retirement pension in Austria. The data were collected in the European household survey EU-SILC (European Union Statistics on Income and Living Conditions) 2010 in Austria. We use a linear regression model to analyse the effects of socio-demographic variables on the (log-transformed) annual contributions to private retirement pensions. As potential regressors we consider `gender` (binary, 1=female/0=male), `age group` (ordinal with eleven levels), `child` in household (binary, 1=yes/0=no), `income class` (in quartiles of the total data set, i.e. ordinal with four levels), `federal state` of residence in Austria (nominal with nine levels), highest attained level of `education` (nominal with ten levels) and `employment status` (nominal with four levels). We restrict the analysis to observations without missing values in either regressors or response and a minimum annual contribution of EUR 100. Hence, the final data set used for our analysis comprises data of 3077 persons.

We standardised the response and fit a regression model including all potential covariates. Results reported in Table 5 (left panel) indicate that several levels of covariate `education` have a similar effect and most level effects of `federal state` are close to zero, which suggests that a sparser model might be adequate for these data.

To specify the effect fusion prior, we chose the hyperparameters with $r = 50000$, $g_{h0} = 5$, $G_{h0} = 2$ for nominal and $G_{h0} = 20$ for ordinal predictors and used the improper prior $p(\sigma^2) = \frac{1}{\sigma^2}$. MCMC was run for 50000 iterations after a burn-in of 30000 with the first 500 draws of the burnin drawn from the unrestricted model where all elements of $\boldsymbol{\delta}$ are 1. The median of the integrated autocorrelation times over all regression effects was 11.9 (range 1–104) and integrated autocorrelation times were lower than 30 for effects of all covariates except age. We ran several chains from different starting values which yielded essentially the same results.

Figure 7 shows the estimated posterior means of the pairwise fusion probabilities $1 - \hat{\pi}_{h,kk-1}$ for the ordinal covariate `age group`. The estimated fusion probability is higher than 0.5 (dotted line) for five levels, which indicates that age categories could be fused as follows: categories 20–25 and 25–30 to 20–30, categories 35–40 and 40–45 to 35–45 and finally the four categories 50–55, 55–60, 60–65 and >65 to a new category >50.

For a nominal covariate we suggest to visualize the pairwise fusion probabilities in a heatmap. Figure 8 shows the corresponding heatmap for the covariate `education`, where all pairwise fusion probabilities are displayed. Darker colours indicate higher fusion probabilities and values in the diagonal (which represent fusion probability of a category with itself) are always one. Thus, only three levels are required to capture the effect of education (secondary school and apprenticeship; doctoral degree; all remaining levels) and thus the number of effects to be estimated reduces from nine to two.

Model averaged parameter estimates under the effect fusion prior (i.e. *FusBMA*) are reported in the middle panel of Table 5. The narrow model averaged 95% – HPD intervals around zero for the effects of covariates `child, federal state` and `employment status` indicate that there is low uncertainty on the irrelevance of these effects. Encouraging fusion of level effects by the effect fusion prior also reduces uncertainty on the level effects of `education` compared to the full model. However, uncertainty on effects can even increase under the effect fusion prior, see e.g. the effect of age category 25–30, for which fusion probabilities to the category 20–25 are close to 0.5 and hence there is high uncertainty whether these categories should be fused or not.

Next, we select the final model for all covariates based on the estimated pairwise fusion probabilities as described in Section 4.2. Covariates `child, federal state` and `employment status` are completely excluded from the model, whereas some of the levels are fused for covariates `age group, income class` and `education`. Thus, the final model has only 11 regression effects compared to 35 in the full model. The right panel of Table 5 shows results from the selected model using flat priors (*FusSel*). Regression effects are similar to the model averaged estimates and the posterior mean of the error variance $\hat{\sigma}^2 = 0.829$ is almost identical to that of the full model, where $\hat{\sigma}^2 = 0.826$.

	Full model Posterior mean	95% HPD interval	Model averaged Posterior mean	95% HPD interval	Selected model Posterior mean	95% HPD interval
Intercept	-1.15	(-1.38 – -0.91)	-1.05	(-1.29 – -0.80)	-1.10	(-1.31 – -0.90)
Age						
20–25	0.20	(-0.03 – 0.42)	0.20	(-0.01 – 0.49)	0.33	(0.14 – 0.52)
25–30	0.36	(0.15 – 0.57)	0.30	(0.00 – 0.54)		
30–35	0.60	(0.40 – 0.80)	0.62	(0.36 – 0.88)	0.65	(0.45 – 0.86)
35–40	0.74	(0.53 – 0.95)	0.74	(0.49 – 0.98)		
40–45	0.80	(0.60 – 1.00)	0.77	(0.54 – 1.00)	0.82	(0.62 – 1.01)
45–50	0.90	(0.70 – 1.10)	0.87	(0.61 – 1.12)	0.93	(0.74 – 1.13)
50–55	1.01	(0.80 – 1.23)	0.96	(0.73 – 1.20)		
55–60	1.06	(0.81 – 1.30)	0.97	(0.73 – 1.21)		
60–65	1.23	(0.83 – 1.62)	0.97	(0.72 – 1.25)	1.05	(0.85 – 1.24)
> 65	0.67	(0.14 – 1.22)	0.75	(0.10 – 1.18)		
Female	-0.24	(-0.32 – -0.17)	-0.25	(-0.31 – -0.18)	-0.25	(-0.31 – -0.18)
Child	0.00	(-0.07 – 0.07)	0.00	(-0.02 – 0.02)	-	-
Income						
2nd quartile	0.20	(0.07 – 0.32)	0.23	(0.01 – 0.35)		
3rd quartile	0.25	(0.13 – 0.37)	0.23	(0.08 – 0.35)	0.23	(0.12 – 0.35)
4th quartile	0.52	(0.40 – 0.64)	0.53	(0.38 – 0.66)	0.54	(0.42 – 0.66)
Federal State						
Carinthia	-0.16	(-0.30 – -0.01)	0.00	(-0.01 – 0.01)	-	-
Lower Austria	0.06	(-0.03 – 0.16)	0.00	(-0.01 – 0.01)	-	-
Burgenland	-0.03	(-0.21 – 0.15)	0.00	(-0.01 – 0.01)	-	-
Salzburg	0.15	(0.01 – 0.30)	0.00	(-0.01 – 0.01)	-	-
Styria	0.01	(-0.10 – 0.13)	0.00	(-0.01 – 0.01)	-	-
Tyrol	0.08	(-0.06 – 0.20)	0.00	(-0.01 – 0.01)	-	-
Vorarlberg	0.02	(-0.15 – 0.19)	0.00	(-0.01 – 0.01)	-	-
Vienna	0.00	(-0.11 – 0.10)	0.00	(-0.01 – 0.01)	-	-

Table 5: EU-SILC data: Posterior means and 95%-HPD intervals of regression effects and error variance in the full model, model averaged and in the selected model.

	Full model Posterior mean	95% HPD interval	Model averaged Posterior mean	95% HPD interval	Selected model Posterior mean	95% HPD interval
Education						
Apprenticeship, trainee	0.09	(-0.04 – 0.23)	0.02	(-0.01 – 0.17)	0.00	-
Master craftman's diploma	0.24	(0.06 – 0.43)	0.22	(0.14 – 0.36)		
Nurse's training school	0.22	(-0.04 – 0.47)	0.22	(0.14 – 0.35)		
Other vocational school (medium level)	0.26	(0.11 – 0.42)	0.22	(0.14 – 0.36)		
Academic secondary school (upper level)	0.22	(0.05 – 0.37)	0.22	(0.14 – 0.35)	0.21	(0.14 – 0.28)
College for higher vocational education	0.28	(0.12 – 0.43)	0.22	(0.14 – 0.35)		
Vocational school for apprentices	0.29	(0.06 – 0.51)	0.22	(0.14 – 0.36)		
University, academy: first degree	0.35	(0.20 – 0.49)	0.22	(0.14 – 0.35)		
University: doctoral studies	1.12	(0.85 – 1.37)	1.00	(0.76 – 1.25)	1.03	(0.80 – 1.26)
Employment status						
Unemployed	-0.10	(-0.34 – 0.12)	0.00	(-0.01 – 0.01)	-	-
Retired	-0.15	(-0.37 – 0.07)	0.00	(-0.01 – 0.01)	-	-
Not-working (other reason)	0.01	(-0.11 – 0.14)	0.00	(-0.01 – 0.01)	-	-
Variance σ^2	0.826	(0.787 – 0.870)	0.830	(0.789 – 0.872)	0.829	(0.790 – 0.870)

Table 5: (Continued.)

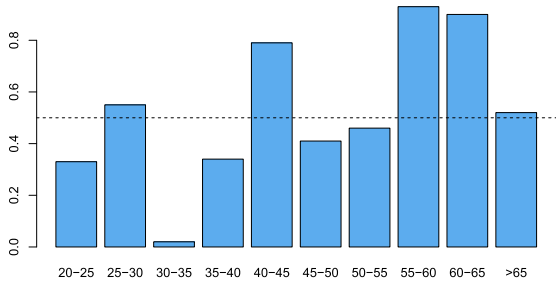

Figure 7: EU-SILC data, covariate `age group`: estimated probabilities for fusion with preceding level.

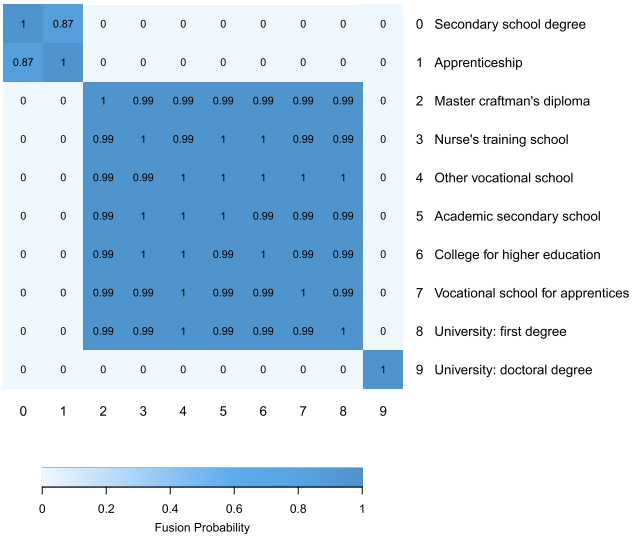

Figure 8: EU-SILC data, covariate `education`: Estimated pairwise fusion probabilities of level effects (baseline: category 0). Darker colours indicate higher fusion probabilities.

7 Conclusion

In this paper, we present a method that allows for sparse modelling of the effects of categorical covariates in Bayesian regression models. Sparsity is achieved by excluding irrelevant predictors and by fusing levels which have essentially the same effect on the response. To encourage effect fusion, we propose a finite mixture of Normal prior distributions with component specific precision matrices, that allow for almost perfect or almost zero partial dependence of level effects. Alternatively, this prior can be derived by specifying spike and slab prior distributions on all level effect contrasts associated with one covariate and taking the linear restrictions among them into account. The structure of this prior easily allows to incorporate prior information that restricts direct

fusion to specific pairs of level effects. This property is of particular interest for ordinal covariates where fusion usually will be restricted to effects of subsequent levels.

Posterior inference for all model parameters is straightforward using MCMC methods. To select one final model we suggest to determine the clustering of level effects which minimizes Binder's loss based on the estimated posterior means of the pairwise fusion probabilities. Simulation results show that the proposed method automatically excludes irrelevant predictors and outperforms competing methods in terms of correct model selection, coefficient estimation as well as predictive performance.

The proposed method for Bayesian effect fusion is not restricted to linear regression models with categorical predictors but can be applied in more general regression type models e.g. generalised linear models, with a structured additive predictor that also contains other types of effects, e.g. nonlinear or spatial effects. Implementation of effect fusion requires little adaption of an MCMC scheme for posterior sampling of any Bayesian regression type model with a Gaussian prior on the regression effects as only two Gibbs sampling steps and the update of the structure matrix have to be added in each MCMC sweep.

A certain drawback of the method is that all pairwise effect differences have to be determined and classified in each MCMC sampling step to construct the prior covariance matrix and hence the computational effort can be prohibitive for nominal covariates with a large number of levels. If we consider each combination of spike and slab distribution on effect differences as a different model, then for a nominal covariate with $c+1$ levels the procedure searches a space of $2^{c(c+1)/2}$ different models. This number is much larger than the number of possible clusterings, which is given by the Bell number $B(c+1)$ of order c. A search in the model space of possible clusterings could be performed by employing a reversible jump MCMC algorithm as in Dellaportas and Tarantola (2005). Another option to determine a model with eventually fused level effects was recently proposed in Malsiner-Walli et al. (2018). The authors use model-based clustering methods and specify a finite mixture prior with many spiky components on the level effects. The prior on the mixture weights encourages empty components and thus some level effects will be assigned to the same mixture component. Due to the spikyness of the components, almost identical effects are assigned to the same component which suggests to fuse effects according to the clustering solution. This approach works well for nominal covariates but does not easily allow incorporation of fusion restrictions.

Supplementary Material

Supplementary Material of "Bayesian Effect Fusion for Categorical Predictors" (DOI: 10.1214/18-BA1096SUPP; .pdf).

References

Binder, D. A. (1978). "Bayesian Cluster Analysis." *Biometrika*, 65: 31–38. MR0501592. doi: https://doi.org/10.1093/biomet/65.1.31. 353

Bondell, H. D. and Reich, B. J. (2009). "Simultaneous Factor Selection and Collapsing Levels in ANOVA." *Biometrics*, 65: 169–177. MR2665858. doi: https://doi.org/10.1111/j.1541-0420.2008.01061.x. 342

Chen, R.-B., Chu, C.-H., Yuan, S., and Wu, Y. N. (2016). "Bayesian Sparse Group Selection." *Journal of Computational and Graphical Statistics*, 25: 665–683. MR3533632. doi: https://doi.org/10.1080/10618600.2015.1041636. 342

Chipman, H. (1996). "Bayesian Variable Selection with Related Predictors." *The Canadian Journal of Statistics*, 1: 17–36. MR1394738. doi: https://doi.org/10.2307/3315687. 342

Dellaportas, P. and Tarantola, C. (2005). "Model Determination for Categorical Data with Factor Level Merging." *Journal of Royal Statistical Society, Series B*, 67: 269–283. MR2137325. doi: https://doi.org/10.1111/j.1467-9868.2005.00501.x. 342, 366

Fahrmeir, L., Kneib, T., and Konrath, S. (2010). "Bayesian Regularisation in Structured Additive Regression: A Unifying Perspective on Shrinkage, Smoothing and Predictor Selection." *Statistics and Computing*, 20: 203–219. MR2610773. doi: https://doi.org/10.1007/s11222-009-9158-3. 350

Fritsch, A. and Ickstadt, K. (2009). "Improved Criteria for Clustering Based on the Posterior Similarity Matrix." *Bayesian Analysis*, 4: 367–392. MR2507368. doi: https://doi.org/10.1214/09-BA414. 353

George, E. and McCulloch, R. (1997). "Approaches for Bayesian Variable Selection." *Statistica Sinica*, 7: 339–373. 341

Gertheiss, J. and Tutz, G. (2009). "Penalized Regression with Ordinal Predictors." *International Statistical Review*, 345–365. 342, 344, 347

Gertheiss, J. and Tutz, G. (2010). "Sparse Modelling of Categorical Explanatory Variables." *The Annals of Applied Statistics*, 4: 2150–2180. MR2829951. doi: https://doi.org/10.1214/10-AOAS355. 342, 346, 353

Griffin, J. and Brown, P. J. (2010). "Inference with Normal-Gamma Prior Distributions in Regression Problems." *Bayesian Analysis*, 5: 171–188. MR2596440. doi: https://doi.org/10.1214/10-BA507. 341

Ishwaran, H. and Rao, S. J. (2005). "Spike and Slab Variable Selection; Frequentist and Bayesian Strategies." *Annals of Statistics*, 33: 730–773. MR2163158. doi: https://doi.org/10.1214/009053604000001147. 341

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). "Penalized Regression, Standard Errors, and Bayesian Lassos." *Bayesian Analysis*, 5(2): 369–412. MR2719657. doi: https://doi.org/10.1214/10-BA607. 342

Lau, J. W. and Green, P. J. (2007). "Bayesian Model Based Clustering Procedures." *Journal of Computational and Graphical Statistics*, 16: 526–558. MR2351079. doi: https://doi.org/10.1198/106186007X238855. 353

Lee, K.-J. and Chen, R.-B. (2015). "BSGS: Bayesian Sparse Group Selection." *The R Journal*, 7(2): 122–132. 354

Liu, F., Chakraborty, S., Li, F., Liu, Y., and Lozano, A. C. (2014). "Bayesian Regularization via Graph Laplacian." *Bayesian Analysis*, 9(2): 449–474. MR3217003. doi: https://doi.org/10.1214/14-BA860. 354

Malsiner-Walli, G., Pauger, D., and Wagner, H. (2018). "Effect Fusion Using Model-Based Clustering." *Statistical Modelling*, 18(2): 175–196 MR3770129. doi: https://doi.org/10.1177/1471082X17739058. 366

Mitchell, T. and Beauchamp, J. J. (1988). "Bayesian Variable Selection in Linear Regression." *Journal of the American Statistical Association*, 83: 1023–1032. MR0997578. 341

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103(482): 681–686. MR2524001. doi: https://doi.org/10.1198/016214508000000337. 341

Pauger, D. and Wagner, H. (2018). "Supplementary Material of "Bayesian Effect Fusion for Categorical Predictors"." *Bayesian Analysis*. doi: https://doi.org/10.1214/18-BA1096SUPP. 345

Pauger, D., Wagner, H., and Malsiner-Walli, G. (2016). "effectFusion: Bayesian Effect Fusion for Categorical Predictors." http://www.R-project.org/. 352

Raman, S., Fuchs, T. J., Wild, P. J., Dahl, E., and Roth, V. (2009). "The Bayesian Group-Lasso for Analyzing Contingency Tables." In *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML 2009, Montreal. 342

Scheipl, F., Fahrmeir, L., and Kneib, T. (2012). "Spike-and-Slab Priors for Function Selection in Structured Additive Regression Models." *Journal of the American Statistical Association*, 107(500): 1518–1532. MR3036413. doi: https://doi.org/10.1080/01621459.2012.737742. 350

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). "A Sparse-Group Lasso." *Journal of Computational and Graphical Statistics*, 22(2): 231–245. MR3173712. doi: https://doi.org/10.1080/10618600.2012.681250. 342

Sun, D., Tsutakawa, R. K., and He, Z. (2001). "Propriety of Posteriors with Improper Priors in Hierarchical Linear Mixed Models." *Statistica Sinica*, 11: 77–95. MR1820002. 351

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso." *Journal of Royal Statistical Society, Series B*, 58(1): 267–288. MR1379242. 341

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Kneight, K. (2005). "Sparsity and Smoothness via the Fused Lasso." *Journal of Royal Statistical Society, Series B*, 67(1): 91–108. MR2136641. doi: https://doi.org/10.1111/j.1467-9868.2005.00490.x. 342

Tutz, G. and Berger, M. (2018). "Tree-Structured Clustering in Fixed Effects Models." *Journal of Computational and Graphical Statistics*, published online. 342

Tutz, G. and Gertheiss, J. (2016). "Regularized Regression for Categorical Data." *Statistical Modelling*, 16(3): 161–200. MR3515384. doi: https://doi.org/10.1177/1471082X16642560. 342

Yuan, M. and Lin, Y. (2006). "Model Selection and Estimation in Regression with Grouped Variables." *Journal of Royal Statistical Society, Series B*, 68: 49–67. MR2212574. doi: https://doi.org/10.1111/j.1467-9868.2005.00532.x. 342

Zou, H. and Hastie, T. (2005). "Regularization and Variable Selection via the Elastic Net." *Journal of Royal Statistical Society, Series B*, 67: 301–320. MR2137327. doi: https://doi.org/10.1111/j.1467-9868.2005.00503.x. 341

Acknowledgments

This work was financially supported by the Austrian Science Fund (FWF) via the research project number P25850 'Sparse Bayesian modelling for categorical predictors'.