Gertraud Malsiner-Walli and Daniela Pauger and Helga Wagner

Effect fusion using model-based clustering

Statistical Modelling 2018; **xx(x)**: 1–22

Effect fusion using model-based clustering

Gertraud Malsiner-Walli[1], Daniela Pauger[2] and Helga Wagner[2]
[1]Institute for Statistics and Mathematics, Vienna University for Economics and Business, Austria.
[2]Department of Applied Statistics, Johannes Kepler University Linz, Austria.

Abstract: In social and economic studies many of the collected variables are measured on a nominal scale, often with a large number of categories. The definition of categories can be ambiguous and different classification schemes using either a finer or a coarser grid are possible. Categorization has an impact when such a variable is included as covariate in a regression model: a too fine grid will result in imprecise estimates of the corresponding effects, whereas with a too coarse grid important effects will be missed, resulting in biased effect estimates and poor predictive performance.

To achieve an automatic grouping of the levels of a categorical covariate with essentially the same effect, we adopt a Bayesian approach and specify the prior on the level effects as a location mixture of spiky Normal components. Model-based clustering of the effects during MCMC sampling allows to simultaneously detect categories which have essentially the same effect size and identify variables with no effect at all. Fusion of level effects is induced by a prior on the mixture weights which encourages empty components. The properties of this approach are investigated in simulation studies. Finally, the method is applied to analyse effects of high-dimensional categorical predictors on income in Austria.

Key words: categorical covariate, sparse finite mixture prior, sparsity, MCMC sampling

Received February 2017; **revised** September 2017; **accepted** September 2017

1 Introduction

Researchers in medicine, social and economic sciences routinely collect data measured on a nominal scale as potential predictors in regression models. The usual approach to include such categorical predictors in regression type models is to define one category as the baseline or reference category and use dummy variables for the effects of all other categories with respect to this baseline. Thus, the effect of *one* categorical covariate with $c+1$ categories is captured by a set of c regression coefficients. This leads to several issues. Including such predictors even with a moderate number of categories can easily lead to a high-dimensional vector of regression coefficients. Further, only the subset of observations with a specific covariate level provides information on its effect which may result in high standard errors and unstable estimates for the effects of infrequent levels. These issues become even more

Address for correspondence: Gertraud Malsiner-Walli, Institute for Statistics and Mathematics, WU Vienna University for Economics and Business, Welthandelsplatz 1, AT–1020 Wien, Austria.
E-mail: gmalsine@wu.ac.at

10.1177/1471082X17739058

pronounced if the researcher uses a fine classification grid when categorizing the data. As often the definition of categories is not completely dictated by subject-specific matters, the scientist could categorize observations either finer or coarser when collecting the data. With both strategies she/he could run into problems when categorical variables are used as covariates in a regression model: fine categories can result in only a few subjects per category and imprecise estimates of the corresponding effects, whereas estimated effects using too coarse categories might be biased due to confounding effects of finer categories.

In order to avoid the risk of overlooking substantial differences in level effects, it would be appealing to have a method which allows to start with a large regression model including categories on a very fine classification grid and to obtain a sparser representation of this model during estimation. Sparsity can be achieved whenever the effects of a categorical predictor can be represented by less than c regression effects. Basically there are three different situations, where sparsity is an issue: First, if *all* level effects are zero, the whole covariate can be excluded from the model. Second, if *some* of the level effects are zero, the corresponding levels can be excluded from the model, and finally if some levels have essentially the *same* effect on the response, sparsity is achieved by fusing the effects of these levels.

Usually, sparsity in regression type models is achieved by applying variable selection methods which allow to identify regressors with non-zero effects, that is, lasso (Tibshirani, 1996) or the elastic net (Zou and Hastie, 2005) in the frequentist framework and 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) in the Bayesian framework. However, these methods are not appropriate for categorical covariates as only single level effects are selected or excluded from the model. Approaches that address exclusion of a whole group of regression effects have been proposed by Chipman (1996), Yuan and Lin (2006), Raman et al. (2009), Kyung et al. (2010) and recently by Simon et al. (2013), but none of these approaches allows also for effect fusion.

For metric predictors, effect fusion can be performed by the fused lasso (Tibshirani et al., 2005) and the Bayesian fused lasso (Kyung et al., 2010). Both methods assume some ordering of effects and shrink only effect differences of consecutive levels to zero, and hence are not appropriate for nominal predictors where any pair of level effects should be subject to fusion. Explicit effect fusion for nominal predictors is considered in Bondell and Reich (2009) and by Gertheiss and Tutz (Gertheiss and Tutz, 2009, 2010; Gertheiss et al., 2011; Tutz and Gertheiss, 2016), who specify lasso-type penalties on effects and effect differences. In a Bayesian approach, recently Pauger and Wagner (2017) specified a prior distribution that can be interpreted as a spike and slab prior on effects and effect differences. However, these approaches are limited to covariates with a moderately large number of categories, as for a covariate with $c+1$ categories $\binom{c+1}{2}$ possible differences have to be considered which inflates the large model even more.

An appealing approach for effect fusion which avoids classification of effect differences and allows to fuse effects directly is to use model-based clustering techniques. Basford and McLachlan (1985) considered clustering of treatment effects in analysis of variance (ANOVA) by specifying a finite mixture of Normal components on the observed treatment means and fit the model via an EM algorithm. In regression models, sparse modelling of effects by mixtures is so far primarily used for continuous covariates. Yengo et al. (2014) and Yengo et al. (2016) define a Normal mixture prior for the regression effects and determine the number of components, that is, coefficient groups, using model choice criteria. In a non-parametric framework, MacLehose and Dunson (2010) use an infinite mixture of heavy-tailed double-exponential distributions on the coefficients of continuous predictors to allow groups of coefficients to be shrunk towards the same, possibly non-zero, mean. Only Dunson et al. (2008) consider categorical covariates. They propose a multi-level Dirichlet process prior on the effects of single nucleotide polymorphism (SNP) in genetic association studies. This prior takes the hierarchical structure of the predictors into account and allows clustering of SNPs both within and across genes. However, by considering 22 markers, each with three levels, only a small number of levels is investigated.

Following this line of research we propose to achieve model-based clustering of level effects by specifying a finite Normal mixture prior. Our approach is explicitly designed to address effect fusion for categorical covariates and has several advantageous features.

First, fusing the level effects directly instead of focusing on all effect differences enables us to handle categorical covariates with a large number of categories, for example, 100 or more. Second, the specified mixture prior can be interpreted as a generalization of the standard spike and slab prior (George and McCulloch, 1993) where a spike distribution at zero is combined with a rather flat slab distribution to allow selective shrinkage of effects; see Malsiner-Walli and Wagner (2011) for an overview. We replace the slab distribution by a location mixture distribution with different, non-zero means. This mixture prior allows to shrink non-zero effects to various non-zero values and introduces a natural clustering of the categories: if level effects are assigned to the same mixture component, they are assumed to be (almost) identical and can be fused.

Third, the hyperparameters of the mixture prior are chosen very carefully to achieve the modelling aims. Their specification is based on the data to yield recommendations that are applicable to a wide range of real data situations. The 'fineness' of the estimated level classification can be guided by the size of the specified component variance, with smaller variances inducing a larger number of estimated effect groups. The prior on the mixture weights is specified following the concept of 'sparse finite mixture' (Malsiner-Walli et al., 2016). Specifying a sparsity inducing prior on the weights in an overfitting mixture avoids unnecessary splitting of superfluous components and encourages concentration of the posterior distribution on a sparse cluster solution and thus allows to estimate the number of effect groups from the data.

Fourth, remaining in the framework of finite mixture of Normals and conditionally conjugate priors avoids a computationally intensive estimation as standard Markov chain Monte Carlo (MCMC) methods can be used. The MCMC scheme for posterior inference basically combines a regression and a model-based clustering step, where in both only standard Gibbs sampling steps are needed.

Finally, model selection is based on the posterior draws of the partitions. Two strategies are pursued to select the final partition of the levels, by either selecting the most frequently sampled model or determining the optimal partition of the effects based on their joint posterior fusion probabilities.

The article is organized as follows. In Section 2 the model and the prior distributions for the model parameters are introduced. Details on posterior inference and model selection are given in Section 3. The method is evaluated in a simulation study in Section 4 and applied to a regression model for income data in Austria in Section 5. Finally, Section 6 concludes.

2 Effect clustering prior

We consider a standard linear regression model with observations i, $i = 1, ..., N$, continuous response y and J categorical covariates with categories $0, ..., c_j$, where $j = 1, ..., J$. For each covariate, 0 is defined as the baseline category and X_{jk} denotes the dummy variable corresponding to the kth category of covariate j. Hence, the regression model is given as

$$y_i = \beta_0 + \sum_{j=1}^{J} \sum_{k=1}^{c_j} X_{jk}\beta_{jk} + \epsilon_i, \tag{2.1}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is a Normal error term, β_0 is the intercept and β_{jk}, $k = 1, \ldots, c_j$, is the effect of the kth category of covariate j with respect to the baseline category. We call β_{jk} the ‘level effect’ of category k.

To complete Bayesian model specification, prior distributions have to be assigned to all model parameters. We assume that regression effects are independent between covariates and use a prior of the structure

$$p(\boldsymbol{\beta}, \sigma^2) = p(\beta_0) \prod_{j=1}^{J} p(\boldsymbol{\beta}_j|\boldsymbol{\xi}_j) p(\sigma^2), \tag{2.2}$$

where $\boldsymbol{\beta}_j = (\beta_j 1, \ldots, \beta_{jc_j}$ denotes the regression effects and $\boldsymbol{\xi}_j$ additional hyperparameters for covariate j. A flat Normal prior $\beta_0 \sim \mathcal{N}(0, B_0)$ is assigned to the intercept, and an improper inverse gamma distribution $\sigma^2 \sim \mathcal{G}^{-1}(s_0, S_0)$ with $s_0 = S_0 = 0$ to the error variance.

Our goal is to specify a prior for the level effects of covariate j which allows the identification of effect groups. Therefore, we specify a finite mixture of Normal distributions as a prior on the level effects β_{jk}. In contrast to the popular spike and slab priors employed for selection of regression effects, we use a location mixture of more than two components which have a small variance, that is, all components are spiky.

The prior on a regression effect β_{jk} is specified hierarchically as

$$p(\beta_{jk}) = \sum_{l=0}^{L_j} \eta_{jl} f_{\mathcal{N}}(\beta_{jk}|\mu_{jl}, \psi_j) \tag{2.3}$$

$$\boldsymbol{\eta}_j \sim Dir_{L_j+1}(e_0) \tag{2.4}$$

$$\mu_{j0} = 0 \tag{2.5}$$

$$\mu_{jl} \sim \mathcal{N}(m_{j0}, M_{j0}) \quad \text{for} \quad l = 1, ..., L_j, \tag{2.6}$$

where $L_j + 1$ is the number of Normal mixture components for covariate j with location parameters μ_{jl} and scale parameter ψ_j. For each covariate, the location parameter of the first component μ_{j0} is fixed at 0 to allow identification of categories which have the same effect as the baseline category. If all level effects are assigned to this component, the covariate can be completely excluded from the model. We subsume in $\boldsymbol{\mu}_j = (\mu_{j1}, \ldots, \mu_{jL_j})$ all other component means, which are assumed to be conditionally independent and follow a flat Normal hyperprior with location and scale parameters m_{j0} and M_{j0}. For each covariate, the variance ψ_j is the same for all components in order to ensure that each level effect group has the same dispersion; however ψ_j may vary between covariates. Finally, a symmetric Dirichlet distribution $Dir_{L_j+1}(e_0)$ with parameter e_0 is specified for the mixture weights $\boldsymbol{\eta}_j = (\eta_{j0}, \ldots, \eta_{jL_j})$.

An alternative to our finite mixture approach would be to specify an infinite mixture where a Dirichlet process prior $DP(\alpha)$ is specified on the mixture weights. In this case, the a priori specification of the number of components $L_j + 1$, a well-known limitation of finite mixtures, would not be necessary as it can be estimated from the data. However, we overcome this weakness of finite mixtures by specifying a sparse finite mixture (Malsiner-Walli et al., 2016) as prior on the level effects. This allows to estimate the number of 'true' components through the number of 'non-empty' components in an overfitting mixture. More details on this strategy will be provided in Section 2.1.

Additionally, it has to be pointed out that the clustering behaviour of finite and infinite mixtures is quite different. For infinite mixtures, the a priori expected number of groups when classifying c_j items is proportional to $\alpha \cdot log(c_j)$ (MacLehose and Dunson, 2010; Malsiner-Walli et al., 2016). This means that with increasing number of items c_j also the number of expected clusters increases. In contrast, for a finite mixture prior as proposed here, the a priori number of non-empty groups is asymptotically independent of the number of items c_j (Malsiner-Walli et al., 2016). Hence, using a finite mixture prior for the c_j effects of a categorical predictor seems more suitable, as one would expect that in a hierarchical categorization scheme, there exists an appropriate level of aggregation which is able to capture all relevant effect differences and the number of significant effect sizes would not increase when using a 'finer' classification grid.

2.1 Choice of hyperparameters

The specification of the prior hyperparameters is crucial to achieve our modelling aims. To obtain recommendations that are applicable to a wide range of situations, we take an empirical approach and choose the hyperparameters depending on the data.

The location parameter of the first mixture component μ_{j0} is fixed at 0 in order to allow fusion to the baseline. For the location parameters of all other components μ_{jk}, we specify a Normal hyperprior located at the 'centre' of the effects and with large variance in order to induce only little shrinkage to the prior mean. Thus, we set the mean m_{0j} of the Normal hyperprior to $m_{0j} = mean(\hat{\boldsymbol{\beta}}_j)$ and the variance M_{0j} to the squared range of $\hat{\boldsymbol{\beta}}_j$, that is, $M_{0j} = (\max_k \hat{\beta}_{jk} - \min_k \hat{\beta}_{jk})^2$, where $\hat{\boldsymbol{\beta}}_j$ is the estimated coefficient vector of covariate j under flat prior.

Level effects should be assigned to the same component only if the sizes of their effects are almost identical. Therefore, specification of the component variance ψ_j is crucial as it reflects the notion of *negligible/relevant* effect differences. As the prior on the component variance ψ_j should take into account the scaling of covariates, we allow ψ_j to vary across covariates, but not between levels of one covariate.

We define the component variance ψ_j as some proportion $1/\nu$ from the variation of the estimated level effects $\hat{\boldsymbol{\beta}}_j$ under a flat prior, that is, $\psi_j = \frac{1}{\nu} V_j$, where $V_j = \frac{1}{c_j - 1} \sum_{k=1}^{c_j} (\hat{\beta}_{jk} - \bar{\beta}_j)^2$ and $\bar{\beta}_j = \frac{1}{c_j} \sum_{k=1}^{c_j} \hat{\beta}_{jk}$. With increasing ν, the shapes of the mixture components become more spiky and more distinct groups of level effects will be identified. Thus, ψ_j implicitly controls the 'fineness' of the estimated partition of level effects, and hence the size of the selected model. As mentioned earlier, the component variances are defined covariate-specific in order to account for the dispersion of the level estimates within a covariate. However, the component variances could also be specified globally, that is, with the same spike size for all covariates, if interest lies in defining a 'global' threshold for level effect differences across all covariates.

Figure 1 shows the prior distributions of the level effects of one of the covariates in our application, the covariate `economic sector` with 83 levels, for two values of the component variance ψ_j. One mixture component is centred at zero and the others at the posterior means $\hat{\beta}_{jk}$ under a standard flat Normal prior.

Since the choice of the prior component variance ψ_j influences effect fusion, as an alternative, we consider ψ_j to be random with a hyperprior $\psi_j \sim \mathcal{G}^{-1}(g_0, G_{0j})$. We expect to obtain more robust cluster solutions as the influence of a fixed parameter ψ_j should be mitigated. For a given value of g_0, we choose G_{0j} such that the a priori expected component variance $E(\psi_j) = \frac{G_{0j}}{g_0 - 1}$ matches a desired size, that is, $E(\psi_j) \approx \frac{V_j}{\nu}$, and hence set $G_{0j} = \frac{V_j}{\nu}(g_0 - 1)$. As the variance is given

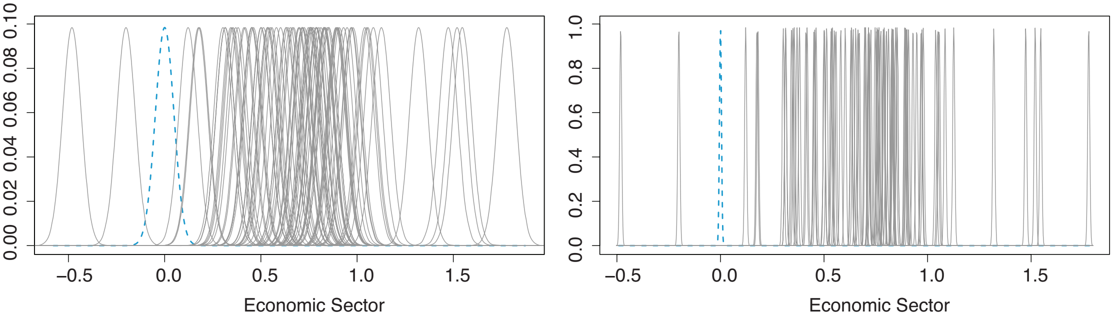

Figure 1 Finite mixture prior on level effects of covariate `economic sector` for two different mixture component variances, $\nu = 10^2$ (left panel) and $\nu = 10^4$ (right panel). One component is centred at zero (blue dashed line), the others at $\hat{\beta}_{jk}$, $k = 1, \dots c_j$, under flat prior

as $V(\psi_j) = E(\psi_j)^2/(g_0 - 2)$, the scale parameter g_0 controls the deviation from the expected value. To allow only for small deviations from the expectation, we set $g_0 = 100$. Thus, a priori, the standard deviation for ψ_j is around $1/10$ of the expected mean. We investigate the influence of the variance parameter ν for fixed variance ψ_j as well as under a hyperprior in the simulation study in Section 4.

We now turn to the specification of the number of mixture components $L_j + 1$. We set $L_j = c_j$ in order to capture the redundant case where all effects are different from each other (and from the baseline). Thus, our prior defines an overfitting mixture model, where the mixture distribution on the level effects has more components than level effects to be estimated. In order to achieve a sparser estimation of the overfitting mixture model by encouraging superfluous components to be empty, we follow Malsiner-Walli et al. (2016), who base their approach on Rousseau and Mengersen (2011).

Rousseau and Mengersen (2011) investigated the asymptotic behaviour of the posterior distribution of an overfitting mixture model and showed that the hyperparameter e_0 of the Dirichlet prior on the mixture weights determines whether superfluous components will be left empty or split in two or more identical components. Asymptotically, if $e_0 < d/2$, where d is the dimension of the component-specific parameter, the posterior expectation of the weights converges to zero for superfluous components. In contrast, for $e_0 > d/2$, the posterior distribution handles overfitting by defining at least two identical components with non-negligible weights. Hence, in order to encourage empty components in the overfitting mixture prior for the level effects, we specify a sparsity inducing prior on the mixture weights $\boldsymbol{\eta}_j$ with $e_0 < d/2$, where $d = 2$ is the dimension of (μ_l, ψ_j). Then, superfluous mixture components should be emptied during MCMC sampling and the sampled partitions concentrate on the model space with sparse solutions.

Following Malsiner-Walli et al. (2016, 2017), Nasserinejad et al. (2017) and Frühwirth-Schnatter (2017), we choose e_0 very small, for example, $e_0 = 0.01$. If $e_0 < 1$, the Dirichlet density is unbounded at the boundaries of its support. As a consequence, much mass is concentrated on weight vectors with only a few large but many small entries. When observations are assigned to the components according to these weight vectors, some of the $L_j + 1$ components will be left empty. Thus, $e_0 \ll 1$ induces a grouping of the coefficients into a few K^* clusters, where K^* is smaller than K with high probability.

3 Posterior inference

The posterior distribution, which results when combining the likelihood derived from equation (2.1) with the prior distribution of $(\boldsymbol{\beta}, \sigma^2)$ specified in equations (2.2)–(2.6), is not of closed form and therefore MCMC methods are used for posterior inference. During MCMC sampling the whole model space will be explored, that is, different clustering solutions for the covariate effects will be visited, which allows to assess model uncertainty and also to determine model averaged estimates.

However, though model averaged estimates of the coefficients may give good results in terms of prediction, researchers are often interested in selection of a *final* model and interpretation of its results. In regression models with categorical predictors, model selection is more involved than in standard variable selection, as the problem is to determine an appropriate clustering of level effects, which means that both the number of clusters as well as the members of each cluster have to be determined. We address this problem in Section 3.3, where we present two different strategies for model selection when clustering the effects of a categorical covariate.

3.1 MCMC sampling

Model estimation is performed through MCMC sampling based on data augmentation (Diebolt and Robert, 1994; Frühwirth-Schnatter, 2006). For each covariate j, latent allocation variables $S_j = S_{j1}, ..., S_{jc_j}$ are introduced to indicate the component a regression effect β_{jk} is assigned to. S_{jk} takes values in $\{0, 1, \ldots, L_j\}$. Conditional on $S_{jk} = l$, the prior distribution for β_{jk} is the Normal mixture component distribution

$$\beta_{jk}|S_{jk} = l \sim \mathcal{N}(\mu_{jl}, \psi_j).$$

MCMC sampling is basically performed by iterating two steps: the regression step, where the level effects and the error variance are sampled conditional on knowing the mixture components the effects are assigned to, and the model-based clustering step, where the parameters of the mixture components and the latent allocation variables are sampled. In the starting configuration, each level effect β_{jk} is assigned to a separate component l, where both the component mean and the effect are estimated under a flat prior. The component located at zero is left empty.

The MCMC sampling scheme iterates the following steps:

Regression steps

1. Sample the regression coefficients $\boldsymbol{\beta}$ conditional on the latent allocation variable $\boldsymbol{S}$ from the Normal posterior $\mathcal{N}(\boldsymbol{b}_N, \boldsymbol{B}_N)$.

2. Sample the error variance σ^2 from its full conditional posterior distribution $\mathcal{G}^{-1}(s_N, S_N)$.

Model-based clustering steps

3. For $j = 1, ..., J$ sample the component weights $\boldsymbol{\eta}_j$ from the Dirichlet distribution $Dir(e_{j0}, e_{j1}, \dots, e_{jL_j})$.
4. For $j = 1, ..., J; l = 1, \dots, L_j$ sample the mixture component means μ_{jl} from their Normal posterior $\mathcal{N}(m_{jl}, M_{jl})$.
5. If a hyperprior is specified on ψ_j, sample the mixture component variances ψ_j from their inverse gamma posterior $\mathcal{G}^{-1}(g_{jN}, G_{jN})$ for $j = 1, ..., J$; otherwise this step is omitted.
6. Sample the latent allocation indicators $\boldsymbol{S}$ from their full conditional posterior

$$P(S_{jk} = l|\beta_{jk}, \boldsymbol{\mu}_j, \psi_j) \propto \eta_{jl} f_{\mathcal{N}}(\beta_{jk}|\mu_{jl}, \psi_j).$$

More details on the sampling steps are given in Appendix A. The sampling scheme is implemented in the `R` package `effectFusion` (Pauger et al., 2016) which is available on CRAN.

3.2 Model-averaged estimates

MCMC draws approximate the whole posterior distribution taking into account model uncertainty: for example, for a regression effect β_{jk}, the posterior is the mixture distribution

$$p(\beta_{jk}|\mathbf{y}) = \sum_m p(\beta_{jk}|\mathbf{y}, \mathcal{M}^{(m)}) p(\mathcal{M}^{(m)}|\mathbf{y}),$$

where the mixture components are model-specific posterior distributions and the mixture weights are the posterior model probabilities $p(\mathcal{M}^{(m)}|\mathbf{y})$. Hence, the mean over all MCMC draws for β_{jk} should be a robust, model-averaged estimator. Its predictive performance is investigated in Section 4.

3.3 Model selection

Before performing model selection, generally the samples from the mixture model have to be identified. In the Bayesian framework, identification of a finite mixture model requires handling the 'label switching' problem (Redner and Walker, 1984) which is caused by the invariance of representation (2.3) with respect to reordering the components:

$$p(\beta_{jk}) = \sum_{l=0}^{L_j} \eta_{jl}\, f_{\mathcal{N}}(\beta_{jk}|\mu_{jl}, \psi_j)$$
$$= \sum_{l=0}^{L_j} \eta_{j\rho(l)}\, f_{\mathcal{N}}(\beta_{j\rho(l)}|\mu_{j\rho(l)}, \psi_j),$$

where ρ is an arbitrary permutation of $\{0, \ldots, L_j\}$. Practically, it may happen, that during MCMC sampling, the labels associated with the components change, which impedes component-specific inference from the MCMC output. The label switching problem is usually solved by post-processing the MCMC output in order to obtain a unique labelling of the draws. We avoid solving the label switching problem by basing model selection on the information whether a pair of level effects is assigned to the same or to different clusters. For each iteration m and each covariate j, we construct the $(L_j + 1) \times (L_j + 1)$ matrix $\boldsymbol{M}_j^{(m)}$ with entry 1, if the two corresponding levels g and h belong to the same cluster, and 0 otherwise, this is,

$$M_{j,gh}^{(m)} = I_{\{S_{jg}^{(m)} = S_{jh}^{(m)}\}}.$$

This matrix is independent of the component labelling and therefore invariant to label switching. It contains the clustering information for covariate j, this is, all information regarding number of effect groups and group memberships.

After MCMC sampling, there are several options to summarize the posterior clustering distribution and to select a final partition of the level effects of covariate j. One possibility is to choose the partition $\boldsymbol{M}_j$ that was selected most often during MCMC sampling. Since the parameter e_0 of the Dirichlet distribution is specified very small, according to Rousseau and Mengersen (2011) 'true' clusters should not be split. The posterior distribution will concentrate on parsimonious partitions of the effects and the number of clusters will depend only on the specified spike variance size. Thus, the posterior mode estimate, that is, the model sampled most frequently during MCMC sampling should be a good choice for the final model.

Another option to select the final partition is to average the matrix $\boldsymbol{M}_j^{(m)}$ over all N_m MCMC iterations yielding the matrix $\boldsymbol{C}_j = \frac{1}{N_m} \sum_{m=1}^{N_m} \boldsymbol{M}_j^{(m)}$. Its entries $C_{j,gh}$ correspond to the relative frequency with which effects of two levels g and h are assigned to the same cluster and approximate the posterior probability that β_{jg} and β_{jh} are members of the same cluster. Hence, each matrix $\boldsymbol{C}_j$ can be interpreted as a 'similarity' matrix: a value of $C_{j,gh}$ close to 1 indicates that the two level effects are almost identical. To find a clustering of the level effects which corresponds most closely to the similarity matrix, we follow Molitor et al. (2010) and use k-medoids clustering.

Similar to k-means clustering, k-medoids clustering aims at clustering points by minimizing the distances between points assigned to a cluster and the point defined as the centre of the cluster. k-medoids always chooses a data point as centre of a cluster ('medoid') and works with arbitrary distance metrics between the data points. This feature makes it attractive for our approach since the similarity matrix can easily be transformed to a distance matrix $\mathbf{D}_j = \mathbf{1} - \mathbf{C}_j$, where $\mathbf{1}$ is a matrix with elements 1. We use the clustering algorithm Partitioning Around Medoids (PAM) proposed by Kaufman and Rousseeuw (2005) which yields an optimal partition for a specified number of clusters. The final partition is chosen by comparing partitions with different numbers of clusters by their silhouette coefficients (Rousseeuw, 1987). The definition of the silhouette coefficient is given in Appendix B.

An advantage of this approach is that clusters of effects are correctly identified even if distances are high, that is, joint inclusion probabilities are rather small. This can happen if the number of categories is large and the strong overlapping of the mixture components induces a frequent switching of the levels between the components, so that the inclusion probability of any two level effects become small, and the most frequent model is not a good representative of the sampled models. However, a drawback of this approach is that the silhouette coefficient cannot be computed for a one-cluster solution. Therefore, with this strategy it is not possible to identify the case where all level effects are assigned to the zero component and the corresponding predictor can be excluded from the model.

4 Simulation study

A sparser representation of the effects of a categorical covariate is possible when (a) *some* or (b) *all* of the levels have no effect or (c) some levels have the same effect and hence can be fused. To investigate the performance of the proposed prior distribution in these situations, we perform a simulation study where categorical covariates with moderate as well as large number of levels represent the various types of sparsity. We evaluate both model selection strategies proposed in Section 3.3, that is, using either the most frequent sampled partition or the partition selected by performing PAM and the silhouette coefficient, with respect to correct model selection. Further, we determine estimation accuracy and predictive performance of the estimates based on the selected models as well as the model averaged estimates.

4.1 Set-up

We define a regression model according to (2.1) with four independent categorical predictors, the first three predictors having 10 and the forth 100 categories. All categories have uniform prior class probabilities. The level effects of the first covariate have three different values ($\boldsymbol{\beta}_1 = (0, 0, 0, 0.5, 0.5, 0.5, 1, 1, 1)$), for the second covariate only one level has a non-zero effect on the outcome ($\boldsymbol{\beta}_2 = (0, 0, 0, 0, 0, 0, 0, 0, 1)$), the levels of the third variable have no effect at all, and levels of the last covariate have six different effects $(0, 0.5, 1, 1.5, 2, 2.5)$ roughly equally distributed among the levels. The intercept β_0 is set to zero. A total of 100 datasets, each consisting of $n = 4\,000$ observations, a random design matrix and a Normal error $\varepsilon \sim \mathcal{N}(0, 0.5)$, are generated. The regression model with prior specifications as described in Section 2.1 and a flat prior on the intercept is fitted to the datasets. In order to investigate the influence of the component variance, the simulations are performed with varying sizes of the variance parameter ν, that is, $\nu = 10, 10^2, \ldots, 10^6$, and fixed as well as random component variance specifications.

MCMC sampling is run for 15 000 iterations after a burn-in of 15 000. The final model is chosen by employing both model selection strategies suggested in Section 3.3. The selected models are then refitted under a flat Normal prior $\mathcal{N}(0, \boldsymbol{I}B_0)$ with $B_0 = 10\,000$ on all level effects. For the refit, MCMC is run for 3000 iterations after a burn-in of 1000.

In order to compare the different final models, two model choice criteria, the Deviance Information Criterion (DIC), proposed by Spiegelhalter et al. (2002), and the BICmcmc, suggested by Frühwirth-Schnatter (2011), are performed. Both measures rely on the MCMC output and can be easily computed. BICmcmc is determined from the largest log-likelihood value observed across the MCMC draws. Whereas the classical BIC is independent from the prior, BICmcmc depends also on the prior of the regression parameters.

4.2 Model selection results

The model selection results are evaluated by reporting the estimated number of level effect groups. Additionally, the clustering quality is assessed by calculating the adjusted Rand index (Hubert and Arabie, 1985), the error rate, the false negative and the false positive rate.

The adjusted Rand index (AR) allows to quantify the similarity between the true and estimated partition of the level effects (Hubert and Arabie, 1985). It is a corrected form of the Rand index (Rand, 1971), adjusted for chance agreement. A value of 1 corresponds to perfect agreement between two partitions, whereas an adjusted Rand index of 0 corresponds to results no better than expected by randomly drawing two partitions, each with a fixed number of clusters and a fixed number of elements in each cluster. A formal definition of the index can be found in Appendix B.

The error rate (err) of the clustering result is the number of misclassified categories divided by all categories. It should be as small as possible. Since interest mainly lies in avoiding incorrect fusion of categories rather than unnecessary splitting of 'true' groups, additionally false negative rate (FNR) and false positive rate (FPR) are reported. They are defined as

$$FNR = \frac{FN}{TP + FN} \qquad FPR = \frac{FP}{TN + FP},$$

where FN is the number of levels incorrectly fused, FP is the number of levels incorrectly split, and TN and TP are the number of levels fused and split correctly, respectively.

Table 1 shows the clustering results for all four covariates using both model selection strategies, that is, the most frequent model ('most') and the model selected using PAM ('pam'), for fixed component variance ψ_j and $\nu = 10^3$. 'freq' reports the number of iterations (out of 15 000) where the most frequent model is sampled, and 'groups' reports the estimated number of clusters. All results are averaged over 100 datasets. Obviously, sparsity is achieved for all covariates. The true number of clusters is correctly identified for both strategies, except for covariate 3, where 'pam' is not able to select the one-cluster solution with all level effects being 0. Also 'most' has some difficulty to fuse all levels to the baseline. However, using a broader variance by setting ν to 10 or 10^2, fusion to the baseline is perfect for this variable (Table C.3). The selected partitions under both model selection strategies show high values of AR and low error rate indicating that the identified clusters capture the true group structure of level effects well. Notably, fusion is almost perfect also for the 100 categories of covariate 4, with an average error rate of err = 0.04.

In order to compare our clustering results to those obtained following the approach proposed by Gertheiss and Tutz (2010) and Oelker et al. (2014), we use the R package **gvcm.cat** to fit a regression model with a regularizing penalty term on the level effect differences. The penalty parameter is chosen via cross-validation. Table 2 reports the classification results. The approach yields large models where level effects are fused very cautiously, resulting in small AR and FNR values and high values for err and FPR.

To investigate the impact of the component variances ψ_j on model selection, we ran MCMC for various values of ν for fixed as well as random component variance ψ_j. In Table 3, we report the results for covariate 4, which is of special interest due to its large number of levels (results for all other covariates are reported in Tables C.1–C.3).

Table 1 Model selection results for fixed ψ with $\nu = 10^3$. Comparison of the two model selection strategies 'most' and 'pam'. The first three variables have 10 categories, the 4th variable 100 categories. FNR is not defined for variable 3

		groups			AR		err		FPR		FNR	
Var	freq	true	most	pam	most	pam	most	pam	most	pam	most	pam
1	14 844	3	3.0	3.0	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
2	14 970	2	2.0	2.0	1.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00
3	11 897	1	1.8	2.0	0.26	0.00	0.23	0.28	0.33	0.41	–	–
4	11 044	6	6.0	6.2	0.91	0.90	0.04	0.04	0.08	0.08	0.02	0.02

Table 2 Penalty approach: Model selection results

Var	true	groups	AR	err	FPR	FNR
1	3	9.0	0.12	0.60	0.91	0.00
2	2	7.7	0.03	0.66	0.92	0.00
3	1	7.2	0.00	0.74	0.92	–
4	6	59.8	0.05	0.83	0.96	0.01

Table 3 Model selection results for variable 4, 100 categories, true number of groups is 6

			groups		AR		Err		FPR		FNR	
	ν	freq	most	pam	most	pam	most	pam	most	pam	most	pam
Fixed	10	9	3.8	3.4	0.52	0.53	0.49	0.49	0.06	0.04	0.21	0.20
	10^2	54	6.1	6.1	0.91	0.93	0.04	0.03	0.07	0.06	0.02	0.01
	10^3	11 044	6.0	6.2	0.91	0.90	0.04	0.04	0.08	0.08	0.02	0.02
	10^4	8 077	6.7	7.0	0.87	0.86	0.09	0.09	0.14	0.15	0.02	0.01
	10^5	7 159	11.0	11.6	0.69	0.68	0.28	0.30	0.40	0.42	0.01	0.01
	10^6	6 800	19.6	20.6	0.46	0.44	0.50	0.53	0.65	0.68	0.00	0.00
Random	10	9	3.7	3.4	0.52	0.53	0.49	0.49	0.06	0.04	0.21	0.20
	10^2	46	4.2	4.3	0.62	0.64	0.35	0.30	0.07	0.10	0.14	0.12
	10^3	38	4.7	4.8	0.70	0.73	0.26	0.21	0.06	0.09	0.11	0.08
	10^4	42	4.9	4.9	0.73	0.73	0.24	0.21	0.06	0.09	0.09	0.08
	10^5	44	4.8	4.9	0.72	0.73	0.25	0.21	0.05	0.09	0.10	0.08
	10^6	45	4.8	4.9	0.72	0.74	0.24	0.20	0.06	0.09	0.10	0.08

For fixed ψ_j, as expected, the number of identified groups increases with ν as the spike variance ψ_j decreases. To detect the 'true' effect clusters, a good choice for ν is a value in the range of $\nu = 10^2$ to $\nu = 10^3$, also AR and err are good for this choice. Larger values of ν lead to a finer classification of the level effects. The number of estimated effect groups increases up to 20 for the very small spike variance ($\nu = 10^6$), with $AR = 0.46$ and err 0.50. However, the relatively high values of FPR and low values of FNR indicate that groups are split into subgroups, while almost no levels of truly different groups are combined to new groups.

When a hyperprior on the component variances is specified as described in Section 2.1, the true number of effects is captured well for variables 1 to 3, where the true number of clusters is at most three (see Tables C.1–C.3). However, for covariate 4 with six different effects, the true number of effect groups is underestimated, regardless of the employed model selection strategy, see Table 3. Although for larger values of ν more (splitted) groups would be expected, the number of estimated groups does not increase. This result suggests that a hyperprior on the component variance cannot be recommended, if a larger number of level effect groups is expected.

Table 4 shows that all models with fixed or random component variance outperform the full model with respect to the BICmcmc; models with a fixed component variance outperform the full model even in terms of DIC unless the component variance is large (i.e., for $\nu = 10$). Thus, with a 'reasonable' variance, that is, ν between 10^2 and 10^5, a good fit of the models can be obtained.

Finally, accuracy and predictive performance of the approach is evaluated by computing the mean squared error (MSE) of the coefficient estimates and the mean squared predictive error (MSPE). The results are compared to those of the full model, the true model and using penalized ML-estimates, and are reported in Appendix C.

Table 4 Model choice criteria for the selected models using the model selection strategies 'most' and 'pam', the penalty approach ('penalty'), and fitting the true and the full model under flat prior

		BICmcmc		DIC	
	ν	most	pam	most	pam
True		8 501		8 445	
Fixed	10^1	9 154	9 089	9 043	9 113
	10^2	8 644	8 643	8 579	8 587
	10^3	8 643	8 645	8 581	8 582
	10^4	8 646	8 648	8 570	8 571
	10^5	8 655	8 657	8 535	8 536
	10^6	8 735	8 731	8 527	8 527
Random	10	9 154	9 091	9046	9 114
	10^2	8 915	8 850	8 799	8 864
	10^3	8 836	8 771	8 716	8 782
	10^4	8 814	8 768	8 713	8 759
	10^5	8 826	8 768	8 713	8 772
	10^6	8 826	8 768	8 713	8 772
Penalty		9 365		8 692	
Full		9 579		8 703	

5 Application

We illustrate the proposed approach for effect fusion in an application to data from EU-Statistics on Income and Living Conditions (SILC) survey 2010 in Austria. Relying on a questionnaire, the EU-SILC data are the main source for statistics on income distribution and social inclusion at the European level, see Statistics Austria (http://www.statistik.at/web_de/frageboegen/private_haushalte/eu_silc/index.html). We use a linear regression model to analyse the effects of socio-demographic variables on the (log-transformed) annual income and aim at identifying levels of categorical covariates which account for income differences.

As potential regressors, we consider the continuous covariate `age` (as linear and squared term) and categorical predictors such as `gender`, `citizenship`, `federal state` of residence in Austria, highest `education level` a person achieved, the economic `sector` a person is working and the `job function`.

The economic `sector` is classified using the classification scheme NACE (statistical classification of economic activities in the European Community), whereas `job function` is determined by using a two-level scheme. Both classifications have a hierarchical structure with 21 and 5 categories on the first level and 84 and 25 categories on the second level of aggregation, respectively. The definition of the categories for both aggregation levels is given in Appendix D. We use the finer second levels of aggregation and specify the effect fusion prior on the categories to achieve a sparser representation of the effects.

We standardize the response y and restrict the analysis to observations of full-time employees with a minimum annual income of EUR 2 000. After removing observations with missing values in the response or the predictors, the dataset consists of observations from 3 865 people. As baseline categories we choose the categories with the lowest labels in the classification schemes, except `federal state` where the baseline is `Upper Austria`. Figure 2 shows the 95% HPD intervals for the level effects under a flat Normal prior.

We fit regression models with prior specifications as described in Section 2, with fixed and random component variances, $\nu = 10, \ldots, 10^6$, and perform model selection as described in Section 3.3. MCMC sampling is run for 15 000 iterations after a burn-in of 25 000 iterations. Table 5 reports the estimated number of effect groups for each of the categorical covariates under both model selection strategies and for the different variance specifications. Additionally, we report the results when fitting a regularized regression using `gvcm.cat` ('pen'). Finally, in order to evaluate the selected models, the BICmcmc of the refitted models is shown.

For fixed ψ_j, as expected, the number of effect clusters increases if the component variances decreases. Again, both model selection strategies yield similar clustering results. BICmcmc is smallest for $\nu = 10^4$ with 2 effect groups for `citizen` and `federal state`, 5 for `education`, and 7 and 6 effect groups for `sector` and

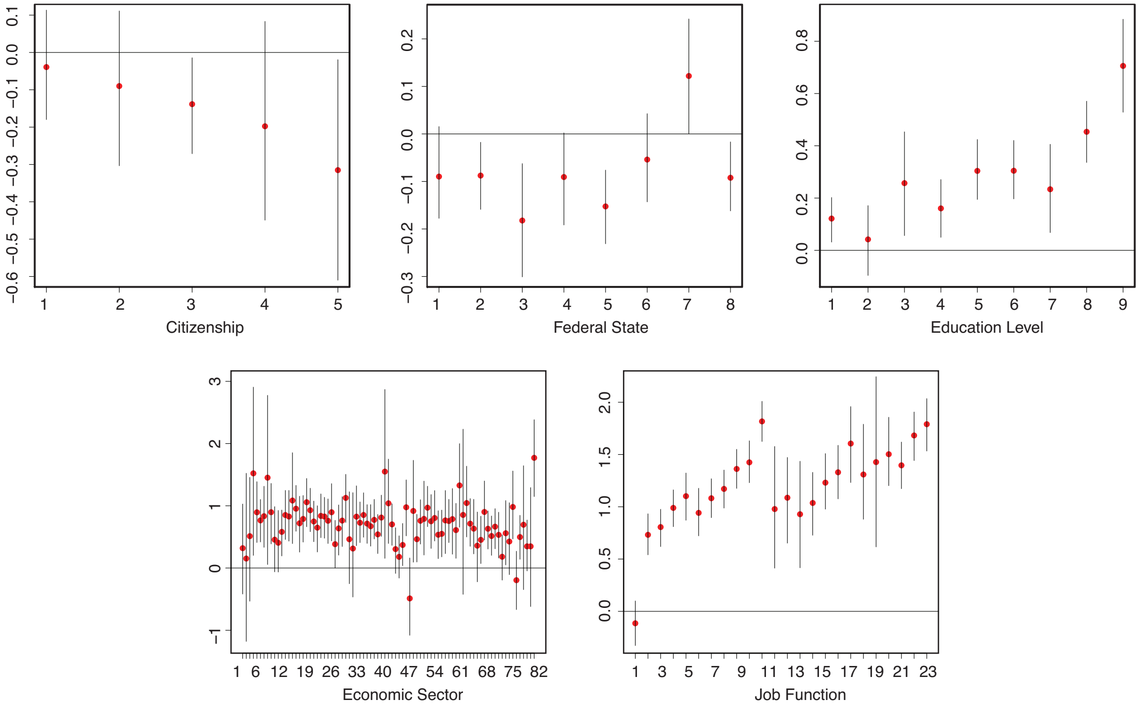

Figure 2 SILC data, posterior means and 95% HPD intervals under flat prior

Table 5 SILC data: estimated number of level groups for the categorical covariates and BICmcmc for various scaling factors ν, with fixed and random component variances ψ_j

		Citizen		Federal State		Education		Sector		Job Function		*BICmcmc*	
	ν	most	pam	most	pam	most	pam	most	pam	most	pam	most	pam
Fixed	10^1	2	2	2	2	2	2	3	2	4	3	8 774	8 520
	10^2	2	2	2	2	3	3	3	2	5	5	8 294	8 275
	10^3	1	3	3	2	3	5	4	4	5	5	8 193	8 165
	10^4	2	2	2	2	5	5	7	7	6	6	8 114	8 117
	10^5	2	2	4	4	5	5	13	13	8	8	8 174	8 171
	10^6	3	3	4	4	6	6	17	20	11	11	8 231	8 255
Random	10^1	2	2	2	2	2	2	3	2	3	2	9 204	9 189
	10^2	2	2	2	2	3	3	3	2	4	3	8 621	8 451
	10^3	2	2	2	2	3	3	3	2	4	3	8 563	8 451
	10^4	2	2	3	3	3	3	3	3	4	3	8 559	8 448
	10^5	2	2	3	3	3	3	3	2	4	3	8 559	8 440
	10^6	2	2	3	3	3	3	3	2	4	3	8 558	8 440
Pen	–	5		8		9		7		21		8 445	
Full	–	6		9		10		84		25		9 047	

job function, respectively. The posterior means and the 95%HPD intervals of the refitted model are plotted in Figure 3.

To visualize the cluster solutions for different values of ν, the estimated effects of the (refitted) selected models for variable job function are plotted in Figure 4. Obviously with decreasing spike variance, the clustering of the level effects gets 'finer'. With a higher resolution of the effects (e.g., $\nu \geq 10^5$), an interesting structure is revealed: as levels are ordered by hierarchy function within each contract type (see Table D.1), obviously effects are fused across contract types. This structure would have been missed by using the coarser classification level, whereas on the other hand even for the very fine resolution with $\nu = 10^6$, the number of estimated effects is less than half compared to the full model.

With a hyperprior on the component variance ψ_j, the selected models are very sparse and the number of effect clusters is almost constant, see Table 5. This is in agreement with the results from the simulation study, and the considerably higher values of BICmcmc indicate that a prior with fixed component variances should be preferred.

6 Discussion

In this article, we proposed to specify a finite Normal mixture prior on the level effects of a categorical predictor to obtain a sparse representation of these effects. The mixture specification allows to shrink non-zero effects to different non-zero locations and introduces a natural clustering of the level effects. Level effects assigned to the same mixture component are fused, that is, their effects are replaced by the

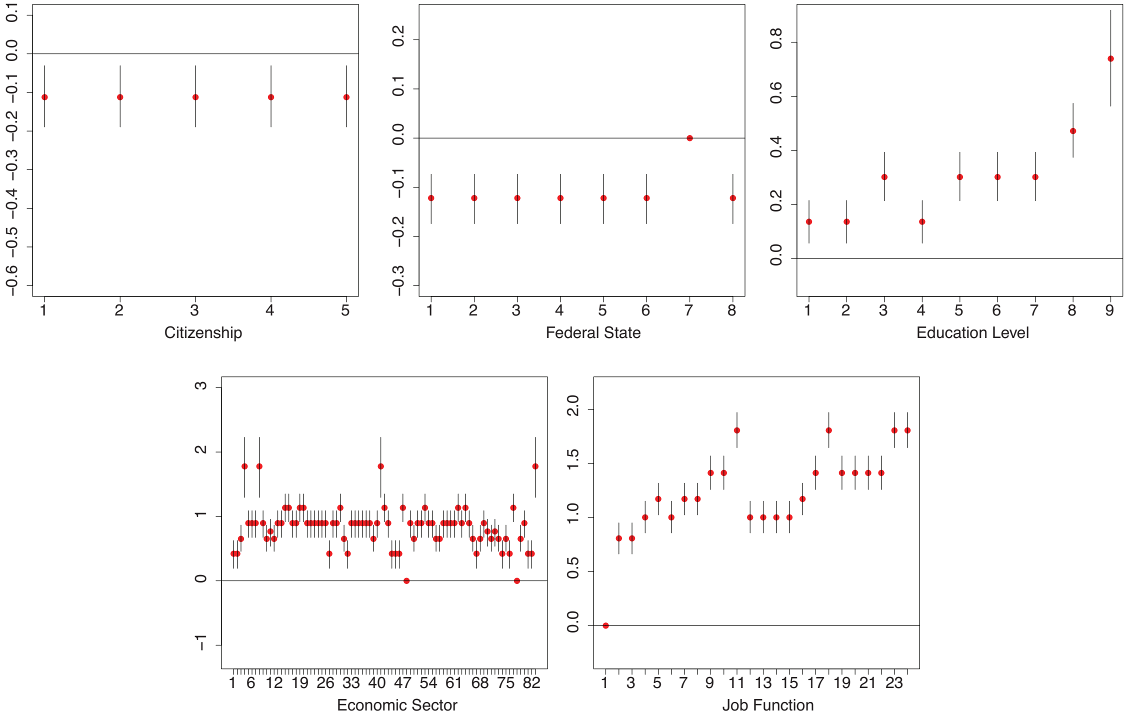

Figure 3 SILC data, posterior means and 95% HPD intervals under mixture prior with $\nu = 10^4$ and fixed variance specification

same joint effect. The number of components as well as their locations are treated as unknown and estimated from the data. A sparse prior on the mixture weights helps to avoid unnecessary splitting of non-empty components and to concentrate the posterior distribution on a sparse cluster solution. The number of estimated level groups can be guided by the size of the component variances, with a smaller variance inducing a larger number of estimated effect groups.

We noted that surprisingly the specification of a hyperprior on the component variances did not work well. In contrast to the common clustering of known data points, we aim at clustering of regression effects, which are not fixed but have to be estimated themselves. Assigning an effect to a mixture component corresponds to selecting a particular prior distribution for its estimation, and hence has an impact on its value in the next parameter estimation step. Thus, additional uncertainty is introduced when clustering regression effects. This leads to the estimation of large component variances and only few effect groups, if the component variances are allowed to be random. Therefore, we recommend to fix the variances of the mixture components and investigate the resolution of level effects with different values. To select the final model, model choice criteria can be used. A strength of our approach is that the spike variance specification can vary across

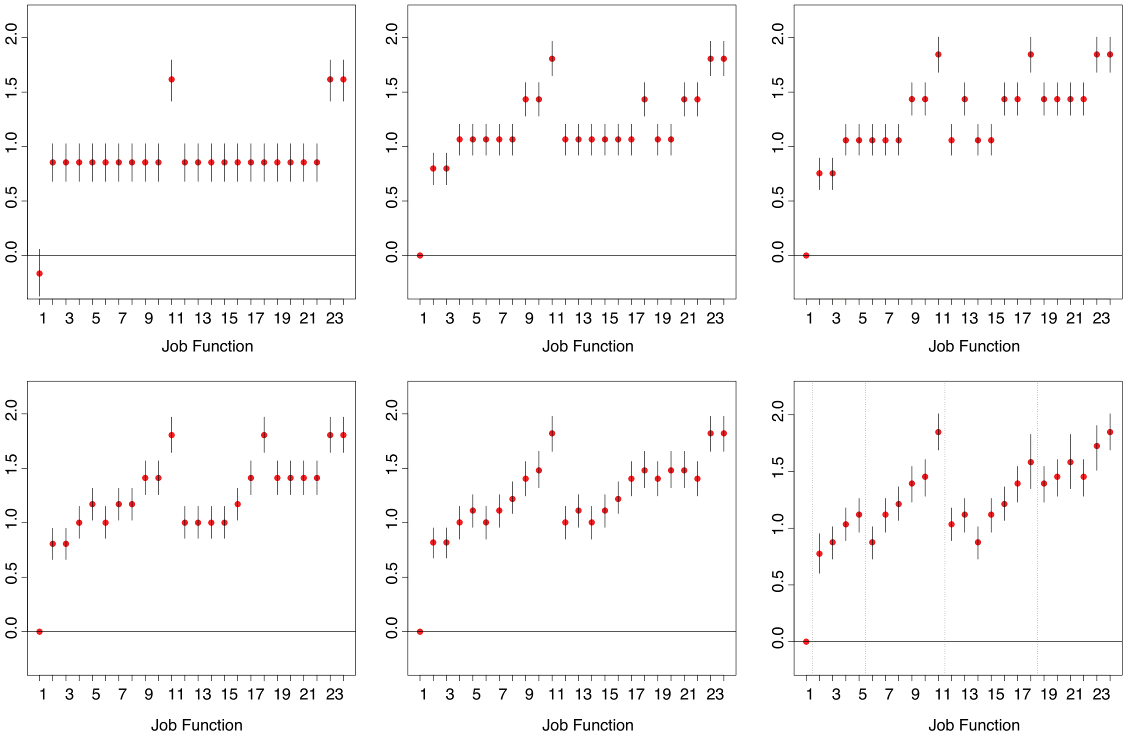

Figure 4 SILC data, variable 'job function': Level estimates with 95% HPD intervals for various $\nu = 10, \ldots, 10^6$ (from top left), by selecting the most frequent model and with fixed component variance ψ_j. In the last plot on the bottom right, the dotted lines indicate the 5 categories of the first level of aggregation, see Table D. 1

the variables, which allows the researcher to obtain a 'finer' clustering for effects of particular interest.

We investigated two different model selection strategies. We selected either the model sampled most frequently or applied the PAM clustering algorithm (Kaufman and Rousseeuw, 2005) to the matrix of posterior inclusion probabilities and selected the final model maximizing the silhouette coefficient of the obtained clusterings. Both strategies have shown to perform similar. An advantage of the first strategy is that a one-group solution can also be selected, which is not possible for the 'PAM' strategy, but the latter is robust against the switching of single effects between groups.

The approach works well even if the number of categories is high, for example, around 100. For Gaussian response regression models, the computational effort is low as a standard Gibbs sampling scheme can be used for MCMC estimation. The sampling scheme is implemented in the `R` package `effectFusion` (Pauger et al., 2016) which is available on CRAN. However, the method is not at all restricted to Gaussian regression models. It can be easily implemented as an 'add-on' to an MCMC sampling scheme for any regression type model with a multivariate Normal prior on the regression effects, as in each MCMC iteration only the steps for model-based clustering as well as the update of the prior parameters of the regression effects are required.

Acknowledgements

This research was financially supported by the Austrian Science Fund FWF (P25850, V170 and P28740) and the Austrian National Bank (Jubiläumsfond 14663). We want to thank Bettina Grün from the Department of Applied Statistics, Johannes Kepler University Linz, for many fruitful discussions.

References

Basford KE and McLachlan GJ (1985) Cluster analysis in a randomized complete block design. *Communication in Statistics: Theory and Methods*, **14**, 451–63.

Bondell HD and Reich BJ (2009) Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics*, **65**, 169–77.

Chipman H (1996) Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, **1**, 17–36.

Diebolt J and Robert CP (1994) Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society B*, **56**, 363–75.

Dunson DB, Herring AH and Engel SM (2008) Bayesian selection and clustering of polymorphisms in functionally related genes. *Journal of the American Statistical Association*, **103**, 534–46.

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

——— (2011) Label switching under model uncertainty. In K. Mengerson, C. Robert and D. Titterington, eds. *Mixtures: Estimation and Application*, pages 213–39. Chichester: Wiley.

Frühwirth-Schnatter S (2017) *From here to infinity-sparse finite versus Dirichlet process mixtures in model-based clustering*. arXiv preprint arXiv:1706.07194.

George EI and McCulloch RE (1993) Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–89.

——— (1997) Approaches for Bayesian variable selection. *Statistica Sinica*, **7**, 339–73.

Gertheiss J, Hogger S, Oberhauser C and Tutz G (2011) Selection of ordinally scaled independent variables with application to international classication of functioning score sets. *Journal of Royal Statistical Society, Series C*, **60**, 377–95.

Gertheiss J and Tutz G (2009) Penalized regression with ordinal predictors. *International Statistical Review*, **77**, 345–65.

——— (2010) Sparse modelling of categorical explanatory variables. *The Annals of Applied Statistics*, **4**, 2150–80.

Griffin JE and Brown PJ (2010) Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, **5**, 171–88.

Hubert L and Arabie P (1985) Comparing partitions. *Journal of Classication*, **2**, 193–218.

Ishwaran H, and Rao JS (2005). Spike and Slab Variable Selection: Frequentist and Bayesian Strategies. *The Annals of Statistics*, **33**, 730–73.

Kaufman L and Rousseeuw PJ (2005) *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ: Wiley Series in Probability and Mathematical Statistics.

Kyung M, Gill J, Ghosh M and Casella G (2010) Penalized regression, standard errors, and

Bayesian lasso. *Bayesian Analysis*, **5**, 369–12.

MacLehose RF and Dunson DB (2010) Bayesian semiparametric multiple shrinkage. *Biometrics*, **66**, 455–62.

Malsiner-Walli G, Frühwirth-Schnatter S and Grün B (2016) Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing*, **26**, 303–24.

——— (2017) Identifying mixtures of mixtures using Bayesian estimation. *Journal of Computational and Graphical Statistics*, **26**, 285–95.

Malsiner-Walli G and Wagner H (2011) Comparing spike and slab priors for Bayesian variable selection. *Austrian Journal of Statistics*, **40**, 241–64.

Mitchell TJ and Beauchamp JJ (1988) Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, **83**, 1023–32.

Molitor J, Papathomas M, Jerrett M and Richardson S (2010) Bayesian profile regression with an application to the national survey of children's health. *Biostatistics*, **11**, 484–98.

Nasserinejad K, van Rosmalen J, de Kort W and Lesaffre E (2017) Comparison of criteria for choosing the number of classes in Bayesian finite mixture models. *PloS one*, **12**, e0168838. DOI: https://doi.org/10.1371/journal.pone.0168838

Oelker MR, Gertheiss J and Tutz G (2014) Regularization and model selection with categorical predictors and effect modifiers in generalized linear models. *Statistical Modelling*, **14**, 157–77.

Park T and Casella G (2008) The Bayesian lasso. *Journal of the American Statistical Association*, **103**, 681–86.

Pauger D and Wagner H (2017) Bayesian effect fusion for categorical predictors. URL arXiv:1703.10245

Pauger D, Wagner H and Malsiner-Walli G (2016) *effectFusion: Bayesian effect fusion for categorical predictors*, R package version 1.0. URL https://cran.r-project.org (last accessed on 23 November 2017).

Raman S, Fuchs TJ, Wild PJ, Dahl E and Roth V (2009) The Bayesian group-lasso for analyzing contingency tables. In L. Bottou and M. Littman, eds. *Proceedings of the 26th Annual International Conference on Machine Learning. ICML* 2009, 881–88 pages. Montreal: Omnipress.

Rand WM (1971) Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**, 846–50.

Redner RA and Walker HF (1984) Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, **26**, 195–239.

Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53–65.

Rousseau J and Mengersen K (2011) Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of Royal Statistical Society, Series B*, **73**, 689–710.

Simon N, Friedman J, Hastie T and Tibshirani R (2013) A sparse-group lasso. *Journal of Computational and Graphical Statistics*, **22**, 231–45.

Spiegelhalter DJ, Best NG, Carlin BP and Van Der Linde A (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 583–639.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society, Series B*, **58**, 267–88.

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**, 91–108.

Tutz G and Gertheiss J (2016) Regularized regression for categorical data. *Statistical Modelling*, **16**, 161–200.

Yengo L, Jacques J and Biernacki C (2014) Variable clustering in high dimensional linear regression models. *Journal de la Societe Francaise de Statistique*, **155**, 38–56.

Yengo L, Jacques J, Biernacki C and Canouil M (2016) Variable clustering in high-dimensional linear regression: The R package clere. *The R Journal*, **8**, 92–106.

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.

Zou H and Hastie T (2005) Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society, Series B*, **67**, 301–20.