LMU
LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Micha Schneider, Wolfgang Pößnecker, Gerhard Tutz

Variable Selection in Mixture Models with an Uncertainty Component

Technical Report Number 225, 2019
Department of Statistics
University of Munich

http://www.statistik.uni-muenchen.de

Variable Selection in Mixture Models with an Uncertainty Component

Micha Schneider, Wolfgang Pößnecker, Gerhard Tutz
Ludwig-Maximilians-Universität München
Akademiestraße 1, 80799 München

August 6, 2019

Abstract

Mixture Models as CUB and CUP models provide the opportunity to model discrete human choices as a combination of a preference and an uncertainty structure. In CUB models the preference is represented by shifted binomial random variables and the uncertainty by a discrete uniform distribution. CUP models extend this concept by using ordinal response models as the cumulative model for the preference structure. To reduce model complexity we propose variable selection via group lasso regularization. The approach is developed for CUB and CUP models and compared to a stepwise selection. Both simulated data and survey data are used to investigate the performance of the selection procedures. It is demonstrated that variable selection by regularization yields stable parameter estimates and easy-to-interpret results in both model components and provides a data-driven method for model selection in mixture models with an uncertainty component.

Keywords: Mixture Models; Variable Selection; lasso, CUB model; CUP model

1 Introduction

Mixture models are widely used to model heterogeneity in populations. D'Elia and Piccolo (2005) proposed a mixture type model for ordinal responses that accounts for the psychological process of human choices. The model has been investigated and extended in a series of papers for example by Piccolo and D'Elia (2008), Iannario and Piccolo (2012b) and Iannario and Piccolo (2012a). The basic concept of the so-called CUB model is that the choice of a response category is determined by a mixture of feeling and uncertainty. Feeling refers to the deliberate choice of a response category determined by the preferences of a person

while uncertainty refers to the inherent individual's indecision. The first component is modelled by a binomial distribution, the latter by a discrete uniform distribution across response categories. An introduction and overview is given in Piccolo and Simone (2019). The CUP model described in Tutz et al. (2017) and further developed by Tutz and Schneider (2019) extends this concept by using any ordinal model as the cumulative model for the preference structure.
In this type of models the right choice of covariates is essential to get sensible models. Even for a moderate number of covariates simple methods as all-subset selection are too time consuming so that other techniques are in demand. Using penalization techniques as lasso by Tibshirani (1996) can overcome this issue. Previous work on variable selection in mixtures focused on mixtures of normal densities and mixtures where the weights do not depend on covariates. Khalili and Chen (2007) used the lasso approach for mixture models and chose a penalty function which is proportional to the mixture weight. Further work was done by Luo et al. (2008) who propose to penalize the coefficients within and between Gaussian components and Städler et al. (2010) focus on high dimensional settings where $p >> n$. But regularization has not been used to investigate the structure of CUB and CUP models with a specific discrete component and weights that depend on individual-specific covariates. In the following we show how to adopt the lasso framework to CUB and CUP models and compare the approach to a forward selection procedure.

The article is organized as follows. First, in section 2 the models are briefly described. In section 3 we discuss variable selection by a step procedure and the proposed lasso method, followed by section 4 about computational aspects of estimation, initialization and convergence. In section 5 we provide results of a simulation study and in section 6 we use the SHIW and ALLBUS survey to show the applicability of the methods on two real data problems. Finally the results are summarized.

2 Model Class

Let the probability that an individual i chooses the category r from ordered categories $\{1, \ldots, k\}$ given explanatory variables $\boldsymbol{z}_i, \boldsymbol{x}_i$ be composed of the individual's propensity towards uncertainty and preference structure. The mixture distribution has the general form

$$P(R_i = r|\boldsymbol{x}_i) = \pi_i P_M(Y_i = r|\boldsymbol{x}_i) + (1 - \pi_i) P_U(U_i = r), \tag{1}$$

where π_i is the propensity or mixture weight, $P_M(Y_i = r|\boldsymbol{x}_i)$ is a model for the preference, and the uncertainty component $P_U(U_i = r)$ is determined by a uniform distribution with probability $1/k$ for each response category. The uncertainty is assumed to include all kinds of indecision related to the nature of human choices like willingness to respond, lack of time, partial understanding etc. The probability π_i is assumed to be linked to covariates by the logit model

$$\operatorname{logit}(\pi_i) \;=\; \boldsymbol{z}_i^T \boldsymbol{\beta}\,, \qquad i = 1, 2, \ldots, n\,. \tag{2}$$

The CUB and CUP models, used in this article, only vary in the choice of the preference component. The preference structure in CUB models (combination of uncertainty and binomial) is modelled by a shifted binomial distribution $b_r(.)$ with parameter ξ, that is,

$$b_r(\xi_i) = \binom{k-1}{r-1} \xi_i^{k-r} (1 - \xi_i)^{r-1}, \quad r \in \{1, \ldots, k\},$$

where ξ_i is linked to the covariates $\boldsymbol{x}_i^T$ by

$$\operatorname{logit}(\xi_i) \;=\; \gamma_0 + \boldsymbol{x}_i^T \boldsymbol{\gamma}\,, \qquad i = 1, 2, \ldots, n\,. \tag{3}$$

The so called CUP model (combination of uncertainty and preference), described in Tutz et al. (2017), uses any ordinal model. A traditional model is the cumulative logit model

$$\log\left(\frac{P(Y_i \le r | \boldsymbol{x}_i)}{P(Y_i > r | \boldsymbol{x}_i)}\right) = \gamma_{0r} + \boldsymbol{x}_i^T \boldsymbol{\gamma}, \quad r = 1, \ldots, k-1.$$

(see Agresti, 2013; Tutz, 2012). The CUP models are more flexible and can handle complex ordinal data structures. However, the intercept parameters depend on the number of categories k so that more parameters have to be estimated.
Both models use covariates to model the preference structure and the weights. In general the covariates $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$ may be identical, completely different or overlap. It should be mentioned that the omission of the uncertainty component typically yields biased parameter estimates.

3 Variable Selection

Since there are two sets of covariates, variable selection is an major issue in mixture models. Let $\mathcal{X}$ contain all possible variables which can be selected for the two independent sets of $\boldsymbol{z}$ and $\boldsymbol{x}$, which are linked to the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively. It is typically not known which variables are relevant for the weights ($\boldsymbol{z}$) and which for the preference structure ($\boldsymbol{x}$) so that variable selection has to handle two separate effect structures. We propose a variable selection based on penalty terms that are tailored to the problem of selecting variables in two components and compare it with a stepwise procedure.

3.1 Stepwise Variable Selection

Two traditional methods are the forward and backward selection. The latter allows that all available explanatory variables are included in both components and the model complexity is reduced stepwise. Especially in mixture models too many possibly correlated covariates can lead to model degeneracy and convergence problems so that the estimates in the fit are hardly trustworthy or the complete model can not be fitted.
Alternatively, one might use a forward search procedure. Here the selection process starts with a basic model as the intercept model. In the first step all models with one covariate in any part of the model are fitted. Then the model with the strongest improvement in terms of a specific criterion is selected. In the next step the procedure continues with this selected model and all remaining covariates are evaluated. The procedure continues until no improvement is detected. In each step a covariate is assigned to only one of the two variable sets $\boldsymbol{z}$ and $\boldsymbol{x}$. If a covariate is selected for one of the two sets, it is still possible that the same covariate is selected for the other variable set later. Several criteria can be used:

$$\begin{aligned} AIC(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) &= -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) + 2df(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}), \\ BIC(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) &= -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) + log(n)df(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}), \end{aligned}$$

or the likelihood-ratio test with

$$lq = -2[l_0(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) - l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})] \overset{a}{\sim} \chi^2(|df(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) - df_0(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})|),$$

where the likelihood of the previous model is compared to the likelihood of the enlarged model. Since the likelihood-ratio test uses the difference of deviances we refer to it also as “deviance” criterion. That variable is selected that yields the largest improvement in AIC or BIC or the smallest p-value of the likelihood-ratio test. If there are several p-values that are numerically close to zero, the model with the largest deviance difference is selected. When the AIC/BIC does not improve or the p-value of the likelihood-ratio test is larger than 0.05 the forward selection is terminated. The estimation of these models is performed as described in Section 4.1. The initializations and convergence checks are described in detail in section 4.2.
Backward/forward strategies have the disadvantage that they are rather variable. The instability of stepwise regression models was demonstrated, for example, by Breiman (1996). Moreover, the standard errors computed for the final model are not trustworthy because they simply ignore the model search. The larger the available number of variables the more models have to be estimated so that these techniques may not work well for very large data sets.

3.2 Variable Selection by Penalization

We propose to use a version of the lasso (Tibshirani, 1996) that is adapted to the mixture models to obtain a procedure that is not limited by the number of variables and produces stable results. The penalized log-likelihood that is to be maximized is given by

$$l_p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = l(\boldsymbol{\beta}, \boldsymbol{\gamma}) - J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma}),$$

where $l(\boldsymbol{\beta}, \boldsymbol{\gamma})$ denotes the un-penalized log-likelihood and $J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is a specific penalty term that enforces the selection of variables in both model components. Let the vectors $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$ be partitioned into $\boldsymbol{z}_i^T = (\boldsymbol{z}_{i1}^T, \dots, \boldsymbol{z}_{ig}^T)$ and $\boldsymbol{x}_i^T = (\boldsymbol{x}_{i1}^T, \dots, \boldsymbol{x}_{ih}^T)$ such that each components refer to a single variable. For example, the vector $\boldsymbol{z}_{ij}$ can represent all the dummy variables that are linked to the j-th variable, or represent the power functions of the j-th variable if one includes polynomial terms. The corresponding predictors are $\boldsymbol{z}_i^T\boldsymbol{\beta}$ and $\boldsymbol{x}_i^T\boldsymbol{\gamma}$ with corresponding partitioning of the parameter vectors, $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_g^T)$ and $\boldsymbol{\gamma}^T = (\boldsymbol{\gamma}_1^T, \dots, \boldsymbol{\gamma}_h^T)$, respectively. Then the proposed penalty has the form

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \lambda_\beta \sum_{j=1}^{g} \sqrt{df_{\boldsymbol{\beta}_j}} \|\boldsymbol{\beta}_j\|_2 + \lambda_\gamma \sum_{j=1}^{h} \sqrt{df_{\boldsymbol{\gamma}_j}} \|\boldsymbol{\gamma}_j\|_2, \tag{4}$$

where λ_β and λ_γ are the tuning parameters for the selection of $\boldsymbol{x}$ and $\boldsymbol{z}$ variables, respectively. The weights $df_{\boldsymbol{\beta}_j}$ are defined as the number of parameters collected in the corresponding parameter vector $\boldsymbol{\beta}_j$, the weights $df_{\boldsymbol{\gamma}_j}$ are defined in the same way. $\|\|_2$ is the unsquared L_2-Norm so that the penalty enforces the selection of variables in the spirit of the group lasso (Yuan and Lin, 2006) rather than selection of single parameters.

All covariables have to be standardized to ensure that the selection of variables does not depend on their scale. Categorical variables have to be orthonormalized. The parameters $\lambda_\beta, \lambda_\gamma$ can be used to enforce specific selection properties. If $\lambda_\beta \to \infty$ no explanatory variables are included in the mixture component and selection is restricted to the effect of explanatory variables on the structured response. If $\lambda_\gamma \to \infty$ no explanatory variables are included in the structured response part and selection is confined to the mixture component. If no specific structure is pre-specified $\lambda_\beta, \lambda_\gamma$ can take any value and can be chosen in a data driven way. A simplification that is tempting is to set $\lambda_\beta = \lambda_\gamma$. It might be sufficient in some applications but it should be used with care.

To select a certain model the use of a selection criterion is needed. In mixture models cross validation can be very time consuming so that we propose the use of AIC or BIC,

$$\begin{aligned} AIC(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) &= -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) + 2edf(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}), \\ BIC(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) &= -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) + log(n)edf(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}), \end{aligned}$$

where $edf(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ is the effective degrees of freedoms of the mixture model. For each parameter set $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ the effective degrees of freedoms are calculated separately by

$$\begin{aligned} edf(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) &= edf(\hat{\boldsymbol{\beta}}) + edf(\hat{\boldsymbol{\gamma}}) \\ &= 1 + \sum_{j=1}^{g} edf(\hat{\boldsymbol{\beta}}_j) + I + \sum_{j=1}^{h} edf(\hat{\boldsymbol{\gamma}}_j), \end{aligned}$$

where 1 refers to the intercept β_0 and I to the number of intercepts γ_0. The CUB-model consist of $1 + 1$-intercepts and the CUP-model of $1 + (k-1)$-intercepts. g and h denote the number of the penalized variables. Following Yuan and Lin (2006) the effective degrees of freedom of each variable are computed by

$$\begin{aligned} edf(\hat{\boldsymbol{\beta}}_j) &= \mathbb{1}(\|\hat{\boldsymbol{\beta}}_j\|_2 > 0) + (df_{\boldsymbol{\beta}_j} - 1)\frac{\|\hat{\boldsymbol{\beta}}_j\|_2}{\|\hat{\boldsymbol{\beta}}_j^{ML}\|_2}, \\ edf(\hat{\boldsymbol{\gamma}}_j) &= \mathbb{1}(\|\hat{\boldsymbol{\gamma}}_j\|_2 > 0) + (df_{\boldsymbol{\gamma}_j} - 1)\frac{\|\hat{\boldsymbol{\gamma}}_j\|_2}{\|\hat{\boldsymbol{\gamma}}_j^{ML}\|_2}. \end{aligned}$$

If a variable is not penalized the *edf* are identical to $df_{\boldsymbol{\beta}_j}$ and $df_{\boldsymbol{\gamma}_j}$, respectively.

To find the best model the procedure has to be optimized with reference to all sensible combinations of the tuning parameters λ_β and λ_γ. We focus on the BIC criterion to find the best model with the lowest BIC value. A two-dimensional grid of λ-values is investigated and parallelized in the following way. One dimension is kept fixed while the other dimension is varied. By repeating this line search all combinations of tuning parameters are covered. For example, using a 15×15 grid results in a 15 times 1×15 line. The advantage of this approach is that we can use parallized computing architecture but also include the results of the previous model for the initialisation of the current model. This saves computing time and leads to non-degenerated results because the fit of the current model should be close to the fit of the previous model with a slightly different tuning parameter. Nevertheless we still use several random initialisations which are described in Section 4.2 to ensure that the fit is not conditioned on the previous results.
Using a complete random choice of tuning parameter combinations can be parallelized even better, but previous knowledge about model results can not be included easily. Another promising approach is the use of model based optimization as described in Bischl et al. (2017) to replace the more time consuming grid search.

4 Computational Aspects

4.1 Estimation with the EM-Algorithm

The mixture models considered in the previous sections can be estimated by an adapted version of the EM algorithm proposed by Dempster et al. (1977). Given the observed category y_i the likelihood contribution of observation i is

$$Pr(y_i|\boldsymbol{z}_i, \boldsymbol{x}_i) = \pi_i \, P_M(y_i|\boldsymbol{x}_i) + (1-\pi_i) \, P_U(y_i) \quad y_i \in \{1, \dots, k\} \tag{5}$$

yielding the log-likelihood

$$l_i(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \{\log(\pi_i) + \log(P_M(y_i|\boldsymbol{x}_i))\} + \{\log(1-\pi_i) + \log(1/k)\}$$

The corresponding penalized log-likelihood is obtained by including the proposed penalty term yielding

$$\begin{aligned} l_i(\boldsymbol{\beta}, \boldsymbol{\gamma}) &= \{\log(\pi_i) + \log(P_M(y_i|\boldsymbol{x}_i))\} + \{\log(1-\pi_i) + \log(1/k)\} \\ &- \lambda_\beta \sum_{j=1}^{g} \sqrt{df_{\boldsymbol{\beta}_j}} \|\boldsymbol{\beta}_j\|_2 - \lambda_\gamma \sum_{j=1}^{h} \sqrt{df_{\boldsymbol{\gamma}_j}} \|\boldsymbol{\gamma}_j\|_2, \end{aligned}$$

and for all observations

$$\begin{aligned} l_c(\boldsymbol{\beta}, \boldsymbol{\gamma}) &= \sum_{i=1}^{n} [\{\log(\pi_i) + \log(P_M(y_i|\boldsymbol{x}_i))\} + \{\log(1-\pi_i) + \log(1/k)\}] \\ &- \lambda_\beta \sum_{j=1}^{g} \sqrt{df_{\boldsymbol{\beta}_j}} \|\boldsymbol{\beta}_j\|_2 - \lambda_\gamma \sum_{j=1}^{h} \sqrt{df_{\boldsymbol{\gamma}_j}} \|\boldsymbol{\gamma}_j\|_2. \end{aligned}$$

The EM algorithm uses the complete likelihood treating the membership to the uncertainty or structure component as missing data. Let z_i^* take the value 1 if observation i belongs to the structure component and zero if observation i belongs to the uncertainty component. Then the complete penalized log-likelihood is given by

$$\begin{aligned} l_p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = &\sum_{I=1}^{n} z_i^* \{\log(\pi_i) + \log(P_M(y_i|\boldsymbol{x}_i))\} + (1-z_i^*) \{\log(1-\pi_i) + \log(1/k)\} \\ &- \lambda_\beta \sum_{j=1}^{g} \sqrt{df_{\boldsymbol{\beta}_j}} \|\boldsymbol{\beta}_j\|_2 - \lambda_\gamma \sum_{j=1}^{h} \sqrt{df_{\boldsymbol{\gamma}_j}} \|\boldsymbol{\gamma}_j\|_2, \end{aligned}$$

where the probability π_i depends on the individual characteristics by

$$\pi_i = 1/(1 + e^{-\boldsymbol{z}_i^T \boldsymbol{\beta}}).$$

Within the EM algorithm the log-likelihood is iteratively maximized by using an expectation and a maximization step. During the E-step the conditional expectation of the complete log-likelihood given the observed data $\boldsymbol{y}$ and the current estimate $\boldsymbol{\theta}^{(s)} = (\boldsymbol{\beta}^{(s)}, \boldsymbol{\gamma}^{(s)})$,

$$M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \mathrm{E}(l_p(\boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{\theta}^{(s)})$$

has to be computed. Because $l_p(\boldsymbol{\theta})$ is linear in the unobservable data z_i^*, it is only necessary to estimate the current conditional expectation of z_i^*. From Bayes's theorem follows

$$\begin{aligned} E(z_i^*|\boldsymbol{y}, \boldsymbol{\theta}) &= P(z_i^* = 1|y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) \\ &= P(y_i|z_i^* = 1, \boldsymbol{x}_i, \boldsymbol{\theta})P(z_i^* = 1|\boldsymbol{x}_i, \boldsymbol{\theta})/P(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \\ &= \pi_i P_M(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})/(\pi_i P_M(y_i|\boldsymbol{x}_i) + (1-\pi_i)1/k) = \hat{z}_i{}^*. \end{aligned}$$

This is the posterior probability that the observation y_i belongs to the structure component of the mixture. For the s-th iteration one obtains

$$\begin{aligned} M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = &\underbrace{\sum_{i=1}^{n} \left\{ \hat{z^*}_i^{(s)} \log(\pi_i) + (1 - \hat{z^*}_i^{(s)}) \log(1-\pi_i) \right\} - \lambda_\beta \sum_{j=1}^{g} \sqrt{df_{\boldsymbol{\beta}_j}} \|\boldsymbol{\beta}_j\|_2}_{M_1} \\ &+ \underbrace{\sum_{i=1}^{n} \left\{ \hat{z^*}_i^{(s)} \log(P_M(y_i|\boldsymbol{x}_i) + (1 - \hat{z^*}_i^{(s)}) \log(1/k) \right\} - \lambda_\gamma \sum_{j=1}^{h} \sqrt{df_{\boldsymbol{\gamma}_j}} \|\boldsymbol{\gamma}_j\|_2}_{M_2} \end{aligned}$$

M_1 and M_2 can be estimated independently from each other but most traditional methods, such as Fisher-Scoring, can not be used because the derivatives do not exist. This problem can be solved with the fast iterative shrinkage-thresholding algorithm (FISTA) of Beck and Teboulle (2009) which is implemented in the MRSP package by Pößnecker (2019) and is used for the maximisation problem of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, which can be formulated generally as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmax}}\, l_p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = -\underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmin}}\, l_p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmin}}\, -l(\boldsymbol{\beta}, \boldsymbol{\gamma}) + J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma}). \quad (6)$$

FISTA belongs to the class of proximal gradient methods in which only the unpenalized log-likelihood and its gradient is necessary. The solution for the unknown parameters $\boldsymbol{\theta}$ of the unpenalized log-likelihood in iteration $t+1$ is given by:

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \hat{\boldsymbol{\theta}}^{(t)} + \frac{1}{\nu} \nabla l(\hat{\boldsymbol{\theta}}^{(t)}),$$

where $\nu > 0$ is the inverse stepsize parameter. This estimator converges to the ML estimator so that each update of $\hat{\boldsymbol{\theta}}^{(t)}$ can be considered as an one-step approximation to the ML estimator based on the current iterate. This can be used to define a searchpoint $\boldsymbol{u}$. To motivate the procedure with penalty the equation (6) is reformulated by Lagrange duality to

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathcal{C}}{\operatorname{argmin}}(-l(\boldsymbol{\theta})),$$

where $\mathcal{C} = \{\boldsymbol{\theta} \in \mathbb{R}^d | J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) \leq \lambda\}$ is the constraint region corresponding to $J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$. Given $\boldsymbol{u}$, the proximal operator associated with the penalty $J_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ is then defined by

$$\mathcal{P}(\boldsymbol{u}) = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmin}} \left(\frac{1}{2} \|\boldsymbol{\theta} - \boldsymbol{u}\|^2 + J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \right)$$

and leads to

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \mathcal{P}_{\frac{\lambda}{\nu}}(\hat{\boldsymbol{\theta}}^{(t)} + \frac{1}{\nu} \nabla l(\hat{\boldsymbol{\theta}}^{(t)}).$$

In a first step the penalty is ignored and a step toward the ML estimator via first-order methods creates a search point. Then this search point is projected onto the constraint region C to account for the penalty term. A detailed description is given in Tutz et al. (2015).

For given $\boldsymbol{\theta}^{(s)}$ one computes in the E-step the weights $\hat{z^*}_i^{(s)}$ and in the M-step maximizes $M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$ (or rather M_1 and M_2), which yields the new estimates

$$\boldsymbol{\beta}^{(s+1)} = \operatorname{argmax}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left\{ \hat{z^*}_i^{(s)} \log(\pi_i) + (1 - \hat{z^*}_i^{(s)}) \log(1 - \pi_i) \right\} - \lambda_\beta \sum_{j=1}^{g} \sqrt{df_{\boldsymbol{\beta}_j}} \|\boldsymbol{\beta}_j\|_2$$

$$\boldsymbol{\gamma}^{(s+1)} = \operatorname{argmax}_\gamma \sum_{i=1}^{n} \hat{z^*}_i^{(s)} \log(P_M(y_i|\boldsymbol{x}_i)) - \lambda_\gamma \sum_{j=1}^{h} \sqrt{df_{\boldsymbol{\gamma}_j}} \|\boldsymbol{\gamma}_j\|_2.$$

The E- and M-steps are repeated alternatingly until the difference $l_p(\boldsymbol{\theta}^{(s+1)}) - l_p(\boldsymbol{\theta}^{(s)})$ is small enough to assume convergence. To account for different sizes of the log-likelihood we define

$$\left| \frac{l_p(\boldsymbol{\theta}^{(s+1)}) - l_p(\boldsymbol{\theta}^{(s)})}{rel.tol/10 + |l_p(\boldsymbol{\theta}^{(s+1)})|} \right| < rel.tol$$

as stopping criteria. *rel.tol* is the relative tolerance which has to below a certain value, such as $1e-6$, to assume convergence. λ_β and λ_γ span a two-dimensional grid of tuning parameter space. Dempster et al. (1977) showed that under weak conditions the EM algorithm finds (only) a local maximum of the likelihood function. Hence it is sensible to use meaningful start values to find a good solution of the maximization problem, which is described in the next section 4.2.

4.2 Initialization and Convergence

Using meaningful starting values is a crucial point in mixture models. Misspecified starting values can lead to degenerated results, can be time consuming and can lead to poor estimation results. In the literature several methods were proposed as described in Baudry and Celeux (2015) and Karlis and Xekalaki (2003). In the random setting several random start values are chosen and all models are run until convergence. Then the best fit is selected. In the small EM strategy a large number of short runs are evaluated which do not have to converge completely. Only the model with the best fit is run until full convergence.

We use a special version of the small EM that refers to the model class considered here so that we use several different configurations. The mixture model components are restricted to two components so that for every observation only π_i and its complement $1 - \pi_i$ need to be chosen which has to sum up to 1. From experience we know that the mean weight for the uncertainty component $(1 - \bar{\pi})$ is in most cases between 0.1 and 0.4. By using this information we are able to create meaningful scenarios which are more likely to be close to a realistic solution. The first strategy is to use a fixed weight for all $\pi_i, i = 1, \dots, n$. Here we chose $\pi_i = 0.9$ and $\pi_i = 0.7$ which correspond with a realistic weight for the uncertainty component $(1 - \pi_i)$ of 0.1 and 0.3, respectively.

The second strategy is drawing the weights π_i so that they are not constant for all observations. For example if we choose the value 0.7 and its complement 0.3 we assign randomly one of this two values to π_i. Because of the randomness we repeat the sample strategy at least two times for the chosen value resulting in two weight vectors $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2$. To ensure that we have obtained different realizations we calculate for each observation the quadratic difference between $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ and compute the sum over all observations. If $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ are identical the computed sum is zero so that $\boldsymbol{\pi}_2$ would be replaced by a new random sample. As a rule of thumb the overall sum has to be larger than $0.1 \cdot n$ to accept $\boldsymbol{\pi}_2$ as a valid initialization. Thus, the sample strategy produce several weight vectors for one chosen value. Here we used 0.9 as well as 0.7 leading to four different initializations. Together with the two constant initializations we obtain at least six configurations which are run until small convergence defined as rel.tol < 0.01 or until the maximal numbers of em-iterations equal to 60 depending on which criteria is reached first. The one with the best result is selected and is run until complete convergence (rel.tol $< 1e - 6$ or maximal numbers of em-iterations equal to 200). One E- and one M-step is defined as one em-iteration.

Every time the model is called we use at least these six configurations regardless if we use the stepwise selection or the penalization. In the latter we may also include another weight initialization. As described in section 3.2 we use a line search to find the best tuning parameter combination. Thus from the second position onwards we can use the computed weights of the previous tuning parameter combination as initialization for the current weights. Since at the beginning of each line search less information about a realistic model is available, we use more configurations for initialization. It consists of the constant choice and two samples of the values $0.6, 0.7, 0.8$ and 0.9.

Dempster et al. (1977) showed that the EM-algorithm converges to a local maximum which is measured in this models by a small difference in the (penalized) Likelihood. A priori we have little information about the exact geometrical shape of the likelihood so that in practice several problems may be occur.

It is well known that the speed of convergence is slow near to the maximum. If the density close to the maximum is very flat we experienced that the difference criterion in the (penalized) likelihood may be too strong. So the rule of likelihood difference is supplemented by a maximum number of em-iterations which can be used. Since the number of em-iterations is in most cases a backstop rule, we usually use a higher number of possible em-iterations which we think should usually not be reached. An exception is the initialization part of the algorithm where the algorithm should not run until complete convergence.

In some cases the (penalized) likelihood may jump between several values without approaching a maximum. This can be solved by adjusting the step-size or, if necessary, taking the best values even if the criterion of small differences in likelihood is not completely reached.

If the starting values are too close to the maximum it may happen that the algorithm diverges from the maximum or a good solution. For this case we implemented some checks to ensure that the best composition is used instead of using a solution which is worse but satisfying the criterion of small difference in (penalized) likelihood. During the EM-algorithm we keep the last ten results to be able to jump back to a previous solution. If this problem occurs between different starting values we select the next best solution. On the other hand we also want to allow the algorithm to search for a better solution. So we allow the algorithm to carry on after a dis-improvement of the likelihood in the first six em-iterations. If the algorithm still does not detect a better likelihood we jump back to the best solution found so far.

On rare occasions the parameters found may be close to the edge of the parameter space. Especially if almost all estimated mixture weights are close to zero or one. In this case we imposed a threshold of $1e-06$ to prevent the weights of being exact zero or exact one. Nevertheless if all mixture weights are close to one for one of the two components a mixture model may be questionable. In case of doubt we recommend to have a look at the estimated mixture weights.

The difference in the (penalized) likelihood is the main criteria of convergence. Only in the case of non-regular behaviour other criteria may be used. Different starting values not only help to find the best maximum but also help to avoid degenerated results.

5 Simulation

To illustrate whether the two selection methods are able to select the "true" covariates we use simulated data with effects and white noise variables with no effects. For $n = 3000$ observations and $k = 5$ response categories we generate five metric covariates from a standard normal distribution ($N(0, 1)$) and six categorical covariates. We use the same 11 covariates for $\boldsymbol{x}$ and $\boldsymbol{z}$, but the effects differ. The first two columns of Table 1 contain the exact values for β and γ used in the simulation. We want to use almost all possible combinations so that some effects of β and γ are identical and some differ. Also the covariates with no effect are sometimes identical (e.g. `Continous_5`) and in other cases there is an effect for only one of the parameters β and γ (e.g. `Continous_1`+4). In both parameter sets there are two continuous and three categorical covariates with no effects.

We use also relative small parameter values to create a realistic setting and to examine whether the size of the effect may have an impact on the different selection methods. The effects of the continuous covariates are 0.2, 0.3, 1, -1 and 2. Three categorical covariates are binary with the effect strength -0.2, 1 and 0. The other three categorical covariates consist of four, four and five categories. Only for the first of them we use effect sizes different from zero namely 0.2, 0.4 and 0.8. The other multi-categorical variables are white noise. The constants in the CUP model are $-2.391, -1.221, -0.259, 1.023$ and in the CUB model -1.5.

We generate $S = 20$ samples from the CUP- and CUB-model each following the described structure and selected variables with the penalization approach and forward selection. For the CUB and CUP model we present in Table 1 the number of times the covariate was selected depending on the used selection technique and the model. The last row includes the π-deviations, which measure the difference between the estimated individual mixture weights π and the true values, defined by

$$\pi\text{-Deviation} = \frac{1}{S}\sum_{j=1}^{S}\left(\frac{1}{n}\sum_{i=1}^{n}|\pi_{ij} - \hat{\pi}_{ij}|\right),$$

where S is the number of simulated data sets, n the number of observations in each data set, π_{ij} is the true mixture weight of the i-th observation in the j-th simulation, and $\hat{\pi}_{ij}$ is the corresponding estimated mixture weight. We compute the absolute differences on each individual mixture weight and use the average over all observations and all samples as a measurement of discrepancy.

Both the penalization and forward selection technique show good results. Both techniques selected covariates with clear effects (-1,1 and 2) in almost 100% and show worse performance with smaller effect size of the parameters. But the penalization technique selected more often covariates with smaller effect size than the forward selection. For example looking at `Categorical_1` the penalization technique selected these covariates in 30% and 95% of the cases in the CUB model compared with only 10% or 65% of the cases using forward selection. The

TABLE 1: *Result of simulated data*

	Simulated		Selected CUB				Selected CUP			
			Penalize		Forward		Penalize		Forward	
Covariates	β	γ	β	γ	β	γ	β	γ	β	γ
`Continous_1`	0	0.3	0%	100%	0%	100%	0%	100%	0%	100%
`Continous_2`	-1	1	100%	100%	100%	100%	100%	100%	100%	100%
`Continous_3`	2	2	100%	100%	100%	100%	100%	100%	100%	100%
`Continous_4`	0.2	0	60%	0%	20%	0%	40%	25%	15%	0%
`Continous_5`	0	0	5%	0%	0%	0%	5%	5%	0%	0%
`Categorical_1`	-0.2	-0.2	30%	95%	10%	65%	10%	95%	0%	25%
`Categorical_2`	1	1	100%	100%	100%	100%	95%	100%	90%	100%
`Categorical_3`	0	0	0%	0%	0%	0%	0%	15%	0%	0%
`Categorical_4:2`	0.2	0.2	20%	100%	0%	100%	0%	100%	0%	95%
`Categorical_4:3`	0.4	0.4	20%	100%	0%	100%	0%	100%	0%	95%
`Categorical_4:4`	0.8	0.8	20%	100%	0%	100%	0%	100%	0%	95%
`Categorical_5:2-4`	0	0	0%	0%	0%	0%	0%	5%	0%	0%
`Categorical_6:2-5`	0	0	0%	5%	0%	0%	0%	0%	0%	0%
π-Deviation	0		0.044		0.033		0.056		0.037	

same behaviour applies for the CUP model. The selection of the β-parameters, which are linked to the mixture weights, seem to be more difficult for both selection techniques than the selection of the γ-parameters. The `Continous_1` and `Continous_4` are characterized by nearly the same effect size (0.3 and 0.2), but differ very much in their selection frequency. While `Continous_1` was selected for γ in 100% correctly, the covariate `Continous_4` was only selected in 60% at the most correctly for the β-parameter. Similar consequences can be drawn from the covariate `Categorical_4`. The covariate was selected in almost 100% of the cases for γ, but very rarely for β.

Table 2 summarizes the results of Table 1 by investigating how often effects that are zero and effects that are different from zero are detected correctly by the two selection methods. The forward selection technique never selected covariates with a true effect of zero while the penalization approach shows small false positive rates. However, the penalization approach performs distinctly better in detecting variables that have a non-zero effect. Both methods show lower rates in detecting effects for β than for γ.

The computed π-deviations displayed in Table 1 are very small for both selection methods given the average size of the simulated $\pi = 0.8011$ and that both selection methods are not always able to select all covariates correctly. Figure 1 displays the original deviations for all samples resulting in $60,000$ observations in each boxplot. Most of them are very close to zero. The penalty approach shows

Table 2: *Summary of simulated data*

Type	Parameters	Selected CUB		Selected CUP	
		Penalize	Forward	Penalize	Forward
Zero effects	β	1%	0%	1%	0%
	γ	1%	0%	10%	0%
Non-zero effects	β	68%	55%	57%	51%
	γ	99%	94%	99%	87%

higher variability and forward selection seems to yield lower discrepancies than the penalization approach.[1]

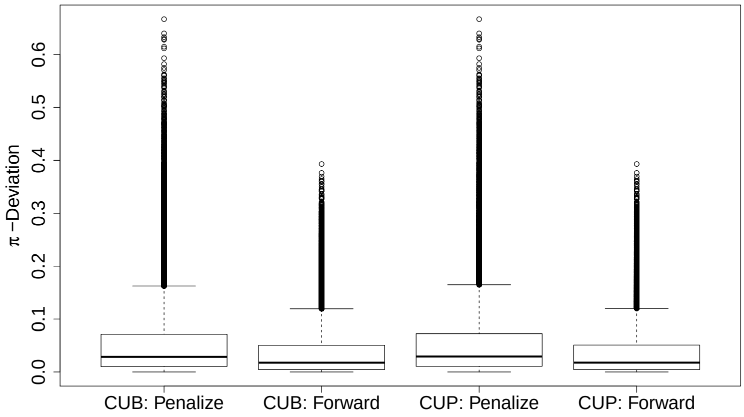

Figure 1: *Simulation: Boxplots of π-Deviations for the different selection methods.*

6 Applications

6.1 Life Well-Being in the Survey on Household Income and Wealth

In the following, the methods are applied to the data from the Survey on Household Income and Wealth (SHIW) by the Bank of Italy, which are earlier used by Gambacorta and Iannario (2013). The data set consists of 3816 respondents from the wave of 2010. The response is the happiness index indicating the overall life well-being measured on a Likert Scale from 1 (very unhappy) to 10 (very happy). 25 covariates as, for example, age, marital status, area of living and educational degree are included in the model selection.

[1] Note that the π_{ij}-differences of the penalization approach based on the penalized estimates.

First we describe the used penalization approach and then the forward selection. Then both techniques are compared and some parameter interpretations are given.

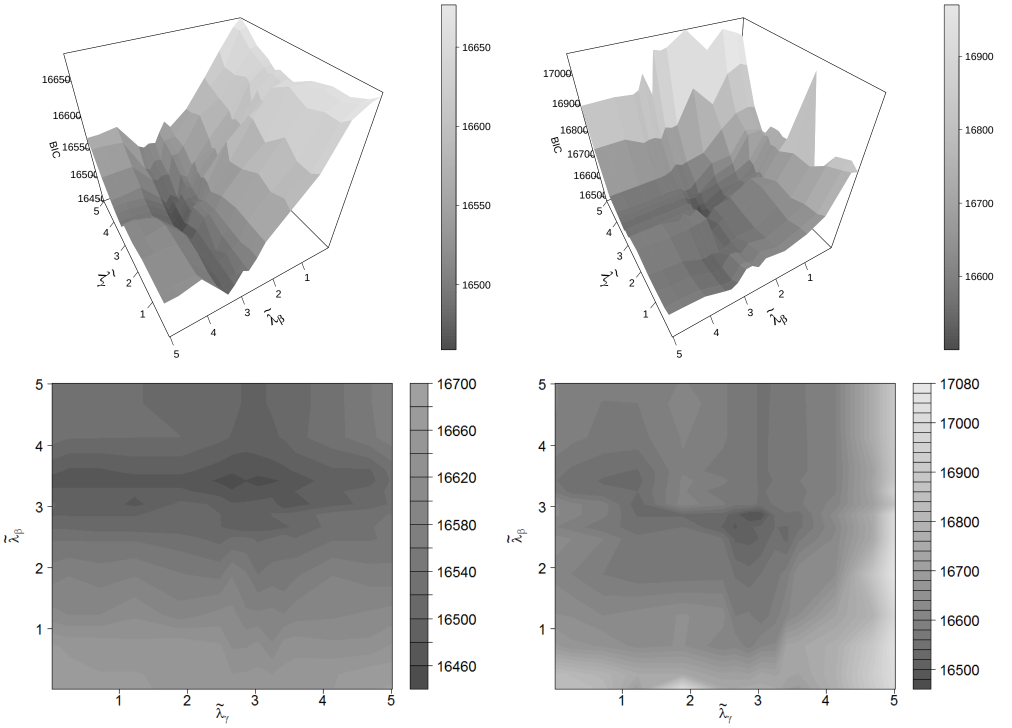

FIGURE 2: *SHIW: Grid of lambda values to find the best model for CUB (left) and CUP model (right).*

6.1.1 Penalization

To illustrate the proposed penalization we use both the CUB and the CUP-model. A 15 times 15 grid of λ_β and λ_γ-values is used to find the best combination of the tuning parameters regarding to the lowest BIC-value. The tuning parameters are transformed by $\log(\lambda + 1)$ because they were created on a logarithm scale and to avoid very large negative values when λ-values are close to zero. Figure 2 shows the results of the 225 models each for the CUB-model on the left hand side and for the CUP-model on the right. If both tuning parameters are zero (right corner) an unpenalized model is estimated. In this case all available covariates are included. On the opposite corner (left) the model is close to an intercept model.

In this application the BIC-surfaces for the two models are quite different. In the CUB-model the choice of the tuning parameter for the $\boldsymbol{\beta}$-covariates seems to be more important than the choice of the preference covariates. So it is advisable to use a smaller grid to find the λ_β-value than the λ_γ-value. In the CUP-model both dimensions of the tuning parameters seem to be more equally important.

The lowest BIC value was found at 16450 with $\log(\lambda_\beta+1) \approx 3.42$ and $\log(\lambda_\gamma+1) \approx 2.66$ in the CUB-model and at 16478 with $\log(\lambda_\beta + 1) = \log(\lambda_\gamma + 1) \approx 2.86$ for the CUP-model. The tuning parameters are not the same but are found in a similar region. Choosing only identical λ-values leads to a slightly worse BIC of 16462 in the CUB-model but with the same selected variables. This is consistent with the nature of lasso regularization which not only selects covariables but also shrinks variables towards zero. It is not unusual that new variables do not enter the model at every grid point in both model components. In general a grid of several tuning parameters should be used, but in this application the restriction on $\lambda_\beta = \lambda_\gamma$ would be sufficient.

To get a better understanding of the mechanism of the variable selection we cut Figure 2 into slices and look at the development of both coefficient sets β and γ. Because of the two-dimensional grid one dimension is fixed to the selected λ-value and the other varies from high penalty (5.02) to low penalty (1.89). The lower the penalty the more parameters enter the model. Each line type in the coefficient paths stands for one parameter group. Because of the penalty term there are some parameters which are selected in both parameter set as for example marital status or area of living and others which are only selected in one of the two sets.

Figures 3 and 4 display the results for the CUB- and CUP-model, respectively. In the first and second row the development of the γ- and β-parameter are displayed. In the third row the resulting boxplots of the weights π are shown. The weights are calculated by using the individual characteristics and estimated β-coefficients. In the first column λ_γ is fixed to the best λ_γ-value and λ_β varies. So the effect of penalization of the β-parameters specifying the weights are shown for β, γ and the weights. In the second column λ_β is fixed and λ_γ varies so that the penalty for the parameters determining the weights do not change.

In the CUB-model two different λ-values are found at $\log(\lambda_\beta + 1) \approx 3.42$ and $\log(\lambda_\gamma+1) \approx 2.66$ to receive the lowest BIC value. On the left column in Figure 3 the λ_γ-parameter is fixed at 2.66 and the penalty for the β varies.

Looking at the β-coefficients in the left column shows that at 5.02 no covariates are selected and the model for the weights only consists of the intercept. The π_i-values are 0.534 for all observations because no individual covariable is present. By adding covariables to the model the weights π_i are adjusted by the individual characteristics of persons and change individually. However the median of the distribution stays almost the same. The more covariables enter the model the variance increase so that the discriminatory power increase, too. But as we can see from Figure 2 using much variables in the β-part increase the BIC-values so that in this case the better discriminatory power does not compensate the higher number of variables. The best trade off between number of variables and model fit according to BIC is found at 3.42. While the β-coefficients are changing the

γ-parameters, displayed in the upper left corner, stay nearly constant.

When λ_β is fixed at 3.42 and only the penalty for the structure component λ_γ changes, as displayed in the right column, the graphs are swapped. Now the coefficients for β are nearly constant while more and more γ-coefficients enter the model. The weights are almost constant. Note that at the maximum of λ_γ already parameters are non-zero. In contrary to a flexible λ_β there is not a pure intercept model for $log(\lambda_\gamma + 1) = 5.02$.

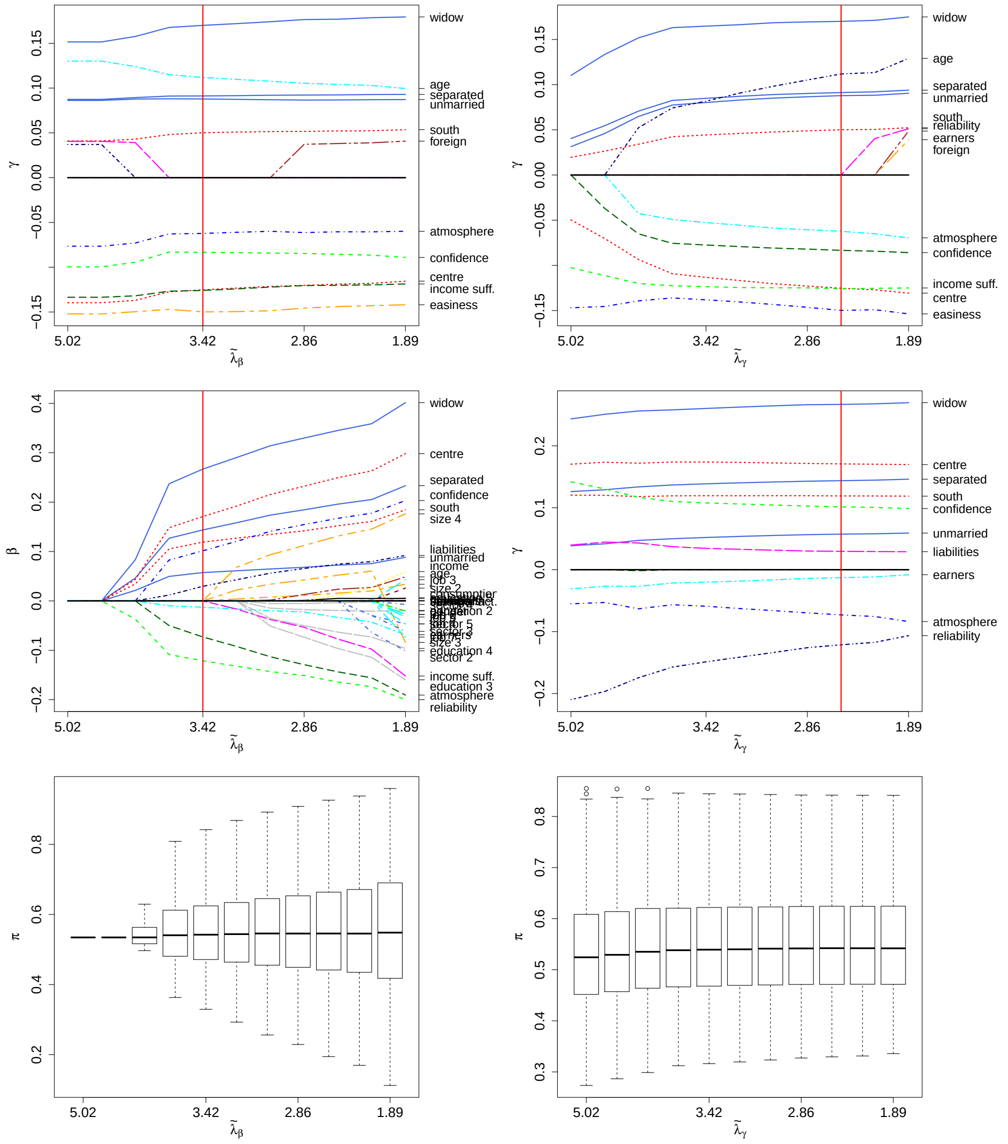

FIGURE 3: *SHIW: Standardized coefficient paths of β and γ and π for fixed lambda (left) and fixed c.lambda (right) in the CUB model.*

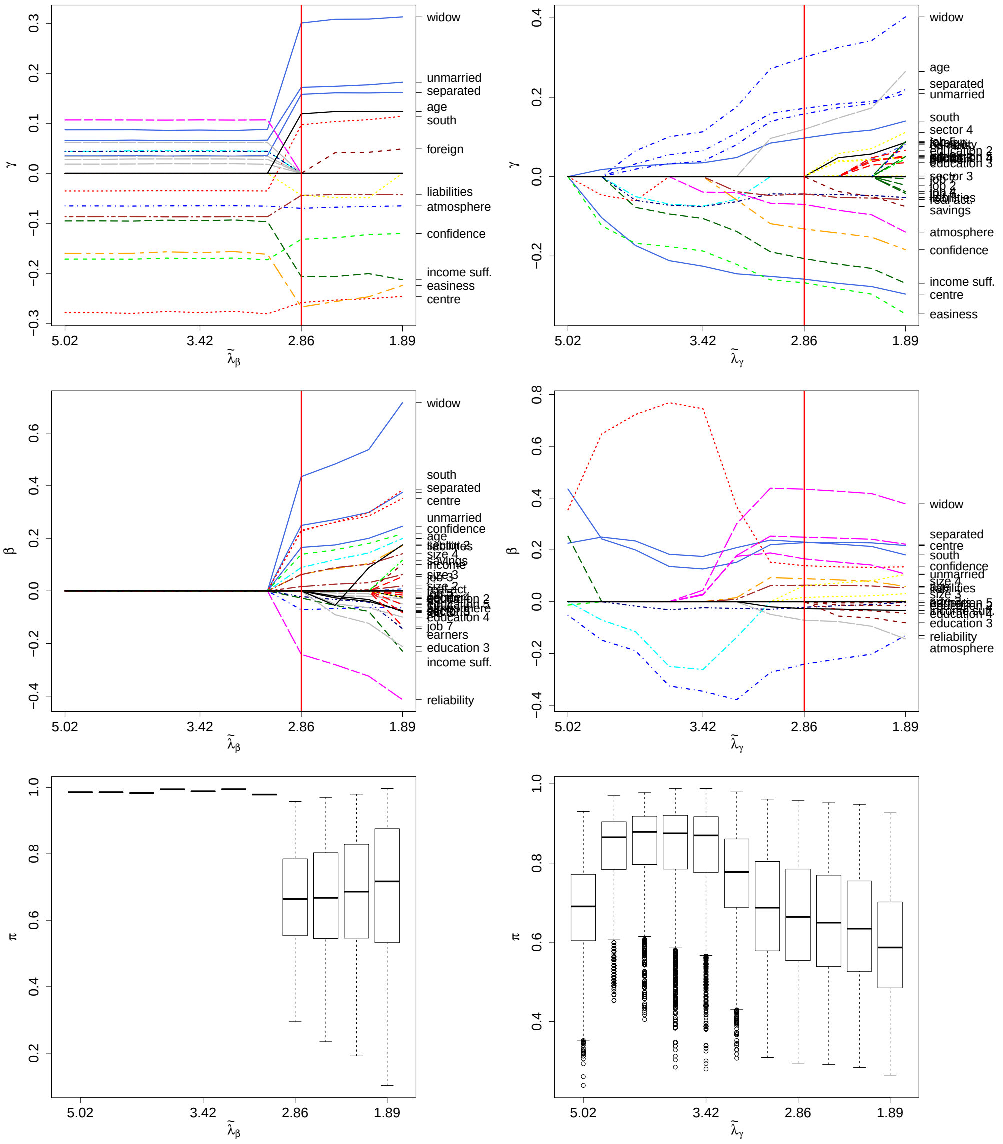

FIGURE 4: *SHIW: Standardized coefficient paths of β and γ and π for fixed lambda (left) and fixed c.lambda (right) in the CUP model.*

The behavior in the CUP model is different. The left column of Figure 4 shows the results for λ_γ fixed at 2.86 and a flexible λ_β-parameter. The first time β-Parameter entering the model is much later than in the CUB models. Until 2.86 a pure intercept model is fitted where nearly no uncertainty component is present, because the π-values are close to 1. At 2.86 some parameters are non-zero and the marginal median weight declines to 0.664. Then again the variance enlarge with more covariates but the marginal median does not change much. At 2.86 the coefficients for γ also change even if the penalty is not changed for γ. That's may be the result of the very different weights which are used for the structured component. Before and after this cutpoint the coefficients of γ are nearly constant.

The results for a fix λ_β at 2.86 is displayed in the right column of Figure 4. With less penalty more and more γ-coefficients enter the model. Even though λ_β does not change, the β-coefficients are not constant and consequently also the weights π_i change substantially.

Both the CUB- and CUP-model detect a reasonable combination of parameter and the CUB-model seem to be more stable than the CUP-model in this application.

6.1.2 Forward Selection

Using forward selection no choice of tuning parameters is necessary. Figure 5 displays the forward selection process for the CUB (left) and CUP-model (right). The y-axis shows the value of the used criteria and the x-axis the selected variables. In the case of the likelihood-ratio test we display the estimated deviance as well as the corresponding p-values. The selected variable is the result of estimating several models and choosing the variable with the greatest impact at that stage of the selection process. The last covariate on the x-axis on the right is the first one which is not selected and where the algorithm stopped. At the beginning the reduction is mostly the highest. The criteria seem to have a great impact on how and which variables are chosen. In the CUB model on the left hand side referring to BIC results in a sparer model than using the likelihood-ratio test or deviance. Not only the number of variables but also the order of selected variables are different. The model constructed by the deviance includes also all variables from the smaller model selected by the BIC. In the CUP case the deviance criterion surprisingly results in a sparer model than using the BIC criterion. However, there are some variables which are only included in one of the models. For example gender and income is only selected in the model with the deviance criterion. The selection process between the CUP and CUB model seems to be also different. Some covariates are selected in both models by both criteria and some are only available in a certain model.

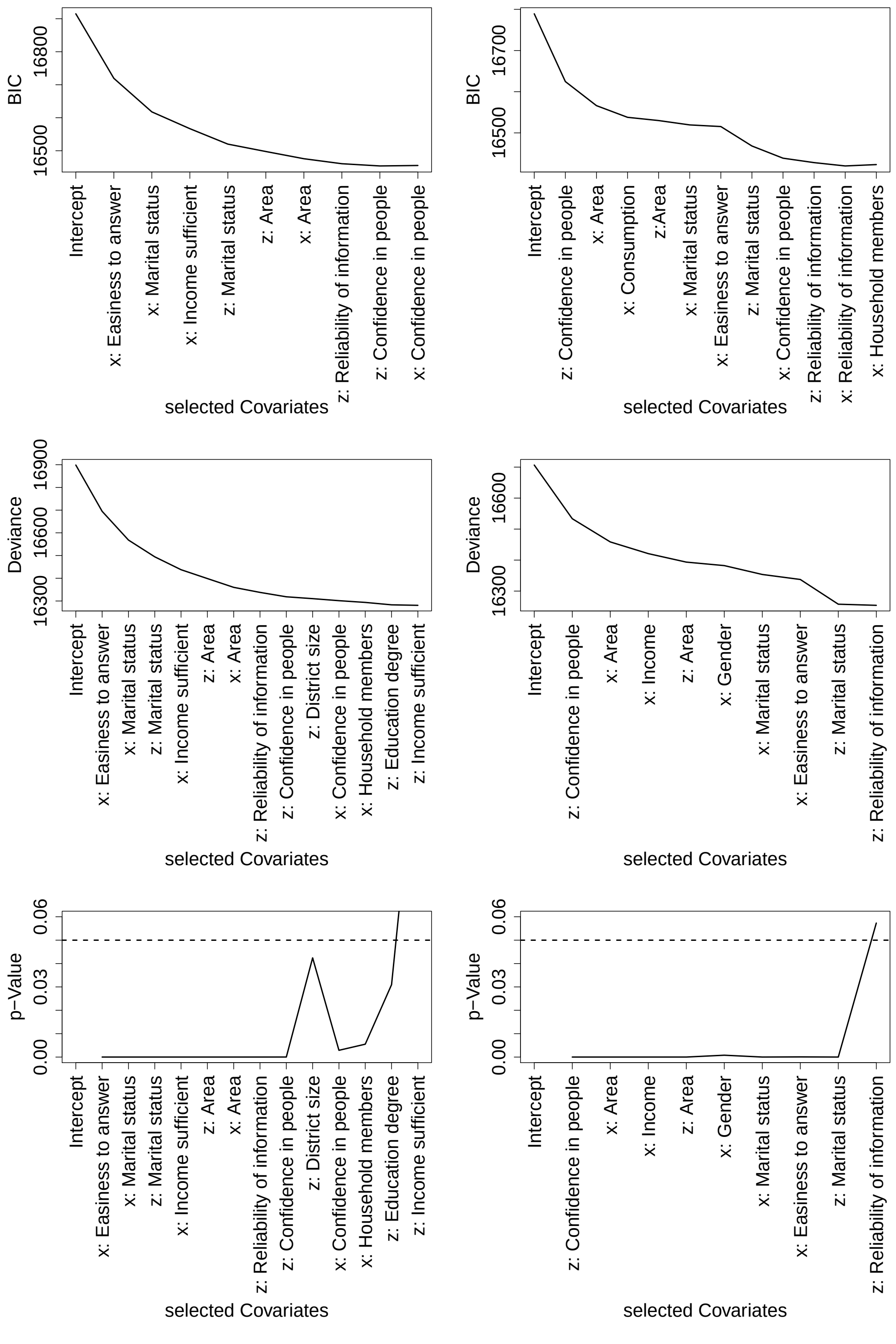

Figure 5: *SHIW: Forward selection for the CUB (left) and CUP model (right).*

6.1.3 Comparison of the Selection Approaches

Table 3 compares both selection methods concerning different selection criteria. For each criterion the value and (effective) degrees of freedom are given. The first entry 16450 is the BIC value which results of a variable selection via the penalization approach for the CUB model with the BIC as optimization criterion followed by the effective degrees of freedom. The next entry 16288 is the AIC value of the same selection technique but optimized according to AIC. Thus each column represents a different model search. In five of the six settings the penalization approach reach a lower value of the selection criteria than the forward selection. In all cases the penalization methods selects larger models than the forward selection.

TABLE 3: *SHIW: Comparison of selection methods*

model	method	criteria					
		BIC		AIC		Deviance	
		value	(e)df	value	(e)df	value	(e)df
CUB	penalize	**16450**	21.33	**16288**	42.46	**16192**	58.99
	forward	16453	16	16335	21	16283	25
CUP	penalize	16479	35.41	**16178**	64.27	**16038**	78.44
	forward	**16420**	26	16389	24	16257	24

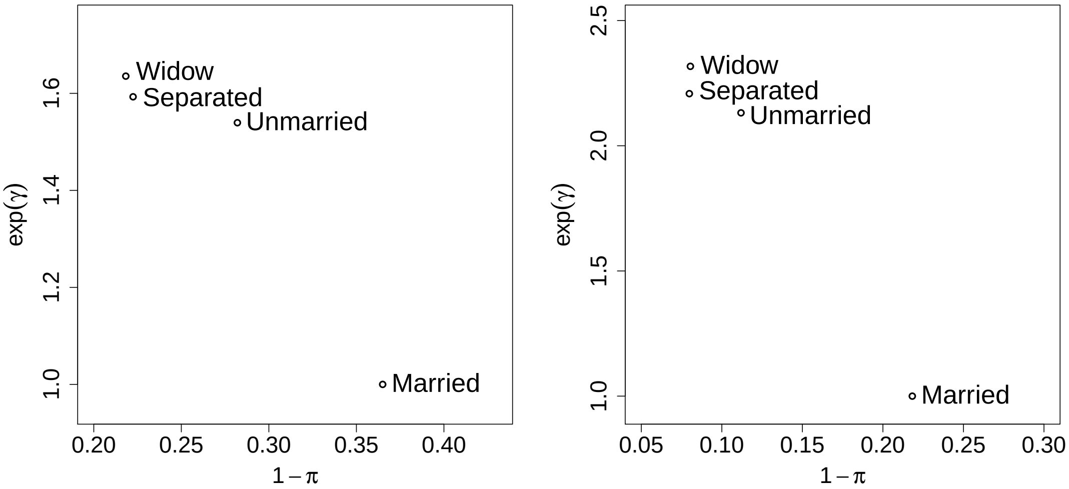

FIGURE 6: *SHIW: Effects of the categorical covariates marital status in CUB- (left) and CUP-Modell (right) in the structure and uncertainty component.*

6.1.4 Parameter Interpretation

For illustration we use both models and selection techniques optimized according to BIC. Using the penalization approach we refitted the models to avoid shrinked coefficients. Note that in this case the goodness-of-fit measurements may be slightly changed, too. Table 4 shows the result for the CUB model and Table 5 for the CUP model. As already mentioned the number of variables are smaller using forward selection than the penalization approach in both models. The effect sizes are similar and show always the same direction. In both components the CUP-model select more variables than the CUB-model.

Figure 6 illustrates the effects of marital status in the CUP- and CUB-model. It is not possible to compare the values of the γ-parameter directly because the models are too different. But for both models the marital status "widow" corresponds to high values of unhappiness and high certainty (small $1 - \pi$). In contrast, the status "married" indicates happiness but a large amount of uncertainty in the response. The order of the marital categories is almost the same in CUB- and CUP-model, but the connected uncertainty is for them higher in the CUB-model than in the CUP-model. This is consistent with the overall behavior of the CUB-model predicting a higher uncertainty than the CUP-model.

Table 4: *SHIW: Coefficients of the chosen (refitted) CUB model*

	Refitted Penalized model		Forward Selection	
Covariates	Concomitant(β)	Structure(γ)	Concomitant(β)	Structure(γ)
Constant	0.554	0.734	0.538	0.586
Marital status: Unmarried	0.381	0.431	0.489	0.368
Marital status: Separated	0.698	0.466	0.834	0.400
Marital status: Widow	0.722	0.492	1.174	0.560
Area: Centre of Italy	0.528	-0.259	0.936	-0.255
Area: South of Italy	0.273	0.100	0.412	0.071
Confidence in people	0.042	-0.042	0.093	
Interview atmosphere	-0.050	-0.038		
Income sufficient		-0.113		-0.126
Age (centered)		0.005		
Easiness to answer		-0.088		-0.133
Income earners	-0.018			
Reliability of information	-0.073		-0.193	
Financial liabilities	0.004			

Table 5: *SHIW: Coefficients of the chosen (refitted) CUP model*

	Refitted Penalized model		Forward Selection	
Covariates	Concomitant(β)	Structure(γ)	Concomitant(β)	Structure(γ)
Constant	1.276		0.682	
Marital status: Unmarried	0.796	0.757	0.801	0.402
Marital status: Separated	1.169	0.792	1.088	0.361
Marital status: Widow	1.160	0.840	1.429	0.649
Area: Centre of Italy	0.783	-0.514	0.930	-0.680
Area: South of Italy	0.514	0.230	0.633	0.180
Confidence in people	0.055	-0.056	0.128	-0.208
Interview atmosphere	-0.051	-0.048		
Income sufficient	-0.023	-0.170		
Age (centered)	0.006	0.008		
Easiness to answer		-0.167		-0.282
Income earners	-0.029			
Reliability of information	-0.139		-0.211	0.046
Financial liabilities	0.040	-0.012		
Foreign		0.203		
Real activity		-0.002		
District size Cat2	0.045			
District size Cat3	0.027			
District size Cat4	0.271			
Family Consumption				-0.009

6.2 Enrichment of Cultural Life by Foreigners in the German General Social Survey

The German General Social Survey (ALLBUS) provided by the GESIS-Leibniz-Institut für Sozialwissenschaften (2017) collects data on behavior, attitudes and social structure in Germany. In 2016 a big focus was on attitudes towards migrants, foreigners and religious groups. The 3490 participants were asked to rate on a 7-point scale whether foreigners enrich the German cultural life from “Completely disagree” (1) to “Completely agree” (7). The data set consist of over 700 possible variables. We restricted ourself to the 43 most meaningful variables which would still result in over 200 parameters for the complete model because of a large number of categorical variables and the two parameter sets $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

We applied both proposed methods. For the penalization we used a 19 times 19 grid of λ_β and λ_γ-values to deduce the best combination of the tuning parameters regarding to the lowest BIC-value. The result of this procedure is displayed in Figure 7. White areas in the contour plots correspond with higher BIC-values than being able to be displayed in this figure. In this application the surface of

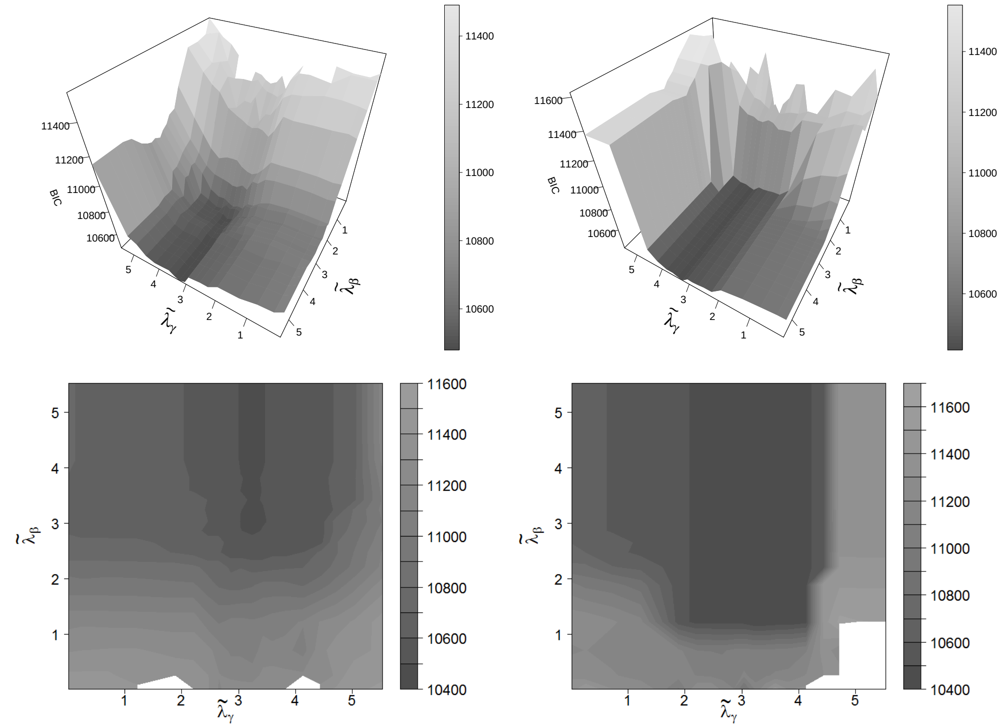

Figure 7: *ALLBUS: Grid of lambda values to find the best model for CUB (left) and CUP model (right).*

both models are quite similar. But in the CUP-model the transition from low to higher BIC-values is sharper than in the CUB-model even though in both models the same grid is used. In the CUB-model the lowest BIC was found at 10471 for $\log(\lambda_\beta + 1) \approx 5.02$ and $\log(\lambda_\gamma + 1) \approx 3.245$. The CUP-model detected the lowest BIC-value at 10408 with $\log(\lambda_\beta + 1) \approx 3.25$ and $\log(\lambda_\gamma + 1) \approx 3.42$. In both the CUB and the CUP models no covariables are selected in the $\boldsymbol{\beta}$-component. This results in a pure intercept model for the weights which are constant for all individuals. The mean mixture weight $(1 - \bar{\pi})$ is 0.0004 for the CUP-model and 0.33 for the CUB-model. If there are no covariables in $\boldsymbol{\beta}$ selected, the intercepts of the cumulative model γ_{0r} in the CUP-model seem to be able to capture the constant probability of the uniform distribution for all individuals resulting in a mixture weight for the uncertainty component close to zero. Moreover the BIC is lower than in the CUB-model with a much higher weight for the uncertainty component.

Using the forward selection leads to models with covariates in both mixture components. Figure 8 displays the selection process for the CUB and CUP model, respectively. Furthermore the selected covariates are quite different between the CUB and the CUP model. In the first case "foreign literature", "age", "houshold income" and "party membership" are selected for $\boldsymbol{z}$ whereas in the CUP model only “age” and “eastwest” were chosen which also results in quite different mixture weights π. The questions of the selected covariates can be found in the appendix.

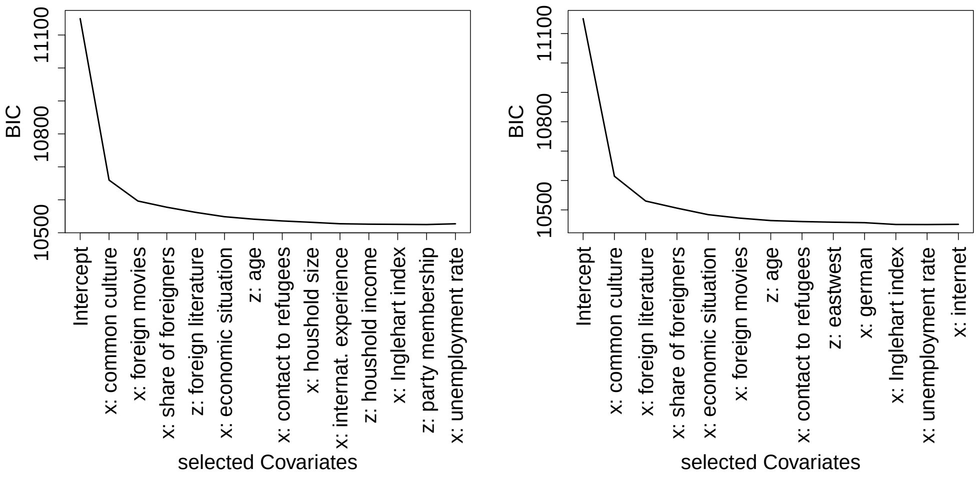

FIGURE 8: *Allbus: Forward selection for the CUB (left) and CUP model (right).*

Table 6 summarizes the results for this application. For both models the BIC value is smaller when using the penalization approach than the forward selection. Also both selection techniques differ not very much in the estimated average mixture weight $\bar{\pi}$ especially in the CUB model, the models are quite different. Using penalization results in larger models but without β effects except of the intercept. On the other hand the selected β-coefficients using forward selection seem to have not enough impact to reduce the BIC in an reasonable way. The lowest BIC value was detected for the penalized CUP model with mixture weight of 0.9996 which is almost a pure cumulative model without uncertainty component.

TABLE 6: *Allbus: Comparison of selection methods*

model	method	BIC	No β	No γ	$\bar{\pi}$
CUB	penalize	**10470**	0	26	0.6747
	forward	10524	4	15	0.6273
CUP	penalize	**10408**	0	27	0.9996
	forward	10450	2	17	0.8742

This application shows that the penalization approach leads here to lower BIC values as the forward selection and stable results even if no β-effects are selected. Furthermore the best combination of tuning parameters is quite different from the previous application so that the best tuning parameter combination has to be estimated for each application separately.

7 Concluding Remarks

We have shown how to adapt the group lasso framework for mixture models with an uncertainty component and compared it to the forward selection. As demonstrated in the simulation section both methods show good performance in selecting the true covariates. The methods allow to decide which variables should be included in the uncertainty part of the model and/or in the preference part of the model. Since often covariates are only included in one of the model components, the model complexity can be reduced substantially. Although forward selection often yields sparser models variable selection via stepwise procedures has some drawbacks. The procedure is rather variable and time-consuming when the number of covariates increases, and often yields higher goodness-of-fit measurements than the penalization approach. Penalization is more flexible and can be used in very high dimensional settings.

It is seen from the applications to real data problems that the choice of the selection method and the optimization criterion determine which final model is chosen. In the Survey on Household Income and Wealth some variables as "marital status" and "area of living" were always selected. Regularization methods yield information on the importance of covariates by visualization of coefficient paths. Also nonparametric bootstrap samples might be a possibility to evaluate how often a covariate is selected. However, including the search for the best tuning-parameter combination without restrictions will lead to huge computing time. One possibility to save computing time would be the restriction on the tuning parameters to be equal. In the first application this restriction would have been sufficient. However, further research is necessary to derive a general rule.

Acknowledgements: Thanks to Maria Iannario and Domenico Piccolo who provided the SHIW data (STAR Programme at University of Naples Federico II).

References

Agresti, A. (2013). *Categorical Data Analysis, 3d Edition*. New York: Wiley.

Baudry, J.-P. and G. Celeux (2015). Em for mixtures - initialization requires special care. https://hal.inria.fr/hal-01113242.

Beck, A. and M. Teboulle (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences 2*(1), 183–202.

Bischl, B., J. Richter, J. Bossek, D. Horn, J. Thomas, and M. Lang (2017). mlrmbo: A modular framework for model-based optimization of expensive black-box functions. *ArXiv e-prints 1703.03373*.

Breiman, L. (1996). Heuristics of instability and stabilisation in model selection. *Annals of Statistics 24*, 2350–2383.

D'Elia, A. and D. Piccolo (2005). A mixture model for preference data analysis. *Computational Statistics & Data Analysis 49*, 917–934.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B 39*, 1–38.

Gambacorta, R. and M. Iannario (2013). Measuring job satisfaction with CUB models. *Labour 27*(2), 198–224.

GESIS-Leibniz-Institut für Sozialwissenschaften (2017). *Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2016*, Volume 2.1.0. GESIS Datenarchiv Köln.

Iannario, M. and D. Piccolo (2012a). CUB models: Statistical methods and empirical evidence. In R. Kennett and S. Salini (Eds.), *Modern Analysis of Customer Surveys: with applications using R*, pp. 231–258. New York: Wiley.

Iannario, M. and D. Piccolo (2012b). Investigating and modelling the perception of economic security in the survey of household income and wealth. In M. S. C. Perna (Ed.), *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, pp. 237–244. Berlin: Springer.

Karlis, D. and E. Xekalaki (2003). Choosing initial values for em algorithm for finite mixtures. *Computational Statistics and Data Analysis 41*, 577–590.

Khalili, A. and J. Chen (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association 102*(479), 1025–1026.

Luo, R., H. Wang, and C. Tsai (2008). On mixture regression shrinkage and selection via the mr-lasso. *International Journal of Pure and Applied Mathematics 46*, 403–414.

Piccolo, D. and A. D'Elia (2008). A new approach for modelling consumers' preferences. *Food Quality and Preference 19*, 247–259.

Piccolo, D. and R. Simone (2019). The class of cub models: statistical foundations, inferential issues and empirical evidence. *Statistical Methods & Applications. online published*.

Pößnecker, W. (2019). MRSP: Multinomial response models with structured penalties. R package version 0.6.11,
https://github.com/WolfgangPoessnecker/MRSP.

Städler, N., P. Bühlmann, and S. van de Geer (2010). L1-penalization for mixture regression models. *Test 19*, 209–256.

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

Tutz, G. (2012). *Regression for Categorical Data*. Cambridge: Cambridge University Press.

Tutz, G., W. Pößnecker, and L. Uhlmann (2015). Variable selection in general multinomial logit models. *Computational Statistics & Data Analysis 82*, 207 – 222.

Tutz, G. and M. Schneider (2019). Flexible uncertainty in mixture models for ordinal responses. *Journal of Applied Statistics 46*(9), 1582–1601.

Tutz, G., M. Schneider, M. Iannario, and D. Piccolo (2017). Mixture models for ordinal responses to account for uncertainty of choice. *Advances in Data Analysis and Classification 11*, 281–305.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B 68*, 49–67.

A Appendix

Variable description of some selected covariates of the ALLBUS data:

- `Common culture`: It is better for a country, if all persons belong to a common culture?
  "completely agree", "rather agree", "rather disagree", "completely disagree"

- `Economic situation`: How do you evaluate the current economic situation in Germany?
  "very good", "good","partly good/ partly bad", "bad", "very bad"
- `Foreign literature`: Do you read - at least occasionally - newspapers, magazines or books in a foreign language?
  "yes", "no"
- `Foreign movies`: Do you watch - at least occasionally - television broadcast or movies in a foreign language without subtitles?
  "yes", "no"
- `Contact to refugees`: Have you had direct personal contact with refugees?
  "yes", "no"
- `Internat. experience`: Have you stayed during your life for more than three months in a foreign country?
  "yes", "no"
- `Internet`: ...Do you use at least occasionally the internet for private purposes?
  "yes", "no"
- `Household size`: ...Do other persons than you live in this household?
  "yes", "no, I live alone"
- `Party membership`: ...Are you member of a political party?
  "yes", "no"
- `German`: German citizen
  "yes, only", "yes, too","no"
- `Eastwest`: Living region
  "Old Federal states", "Newly-formed German states"
- `Inglehart Index`: Computed from several questions:
  "postmaterialist", "postmaterialist mix", "materialist mix", "materialist"
- `Age`: Age of the respondent
- `Household income`: Equivalised disposable income
- `Share of foreigners`: Share of foreigners in living region
- `Unemployment rate`: Unemployment rate in living region