Bayesian effect fusion for categorial predictors

Helga Wagner
joint work with Gertraud Malsiner-Walli and Daniela Pauger

Funded by the Austrian Science Fund (FWF): P25850

Motivation: contributions to private retirement pension

- data on 3077 persons from EU-SILC 2010
- goal: model contributions to private retirement pension
  $\Longrightarrow$ response=log contributions
- categorical covariates
  - age: ordinal, 11 categories (base: 16-20)
  - income class (in quartiles): ordinal, 4 levels (base: 1.quartile)
  - gender: nominal, binary (base: male)
  - child in household: nominal, binary (base: no child)
  - federal states: nominal, 9 levels (base: Upper Austria)
  - employment status: nominal, 4 levels (base: employed)
  - highest education achieved: nominal, 10 levels
    (base: secondary or lower)

Linear regression model

for a categorial predictor with levels $c \in \{0, 1, \ldots, K\}$

- define baseline category (e.g. $c = 0$)
- define dummy variables

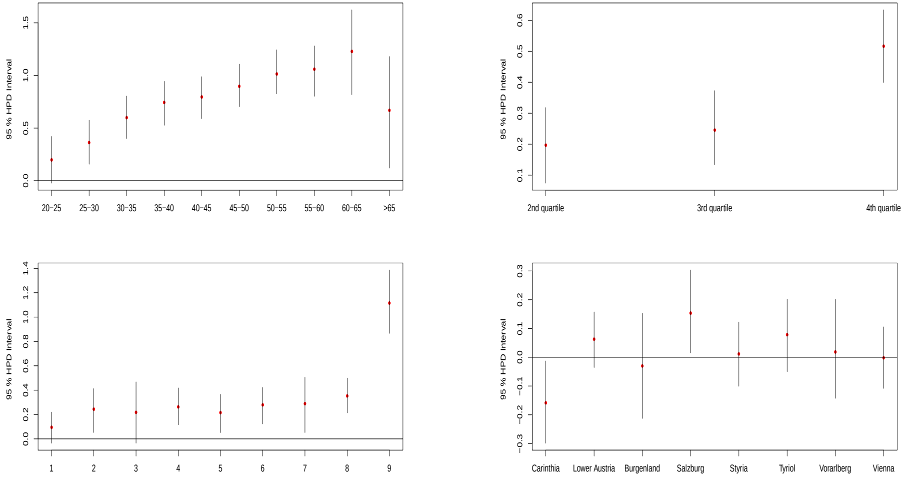
$$x_k = \begin{cases} 1 & \text{if } c = k \\ 0 & \text{otherwise} \end{cases}$$

$$y = \mu + \mathbf{x}'\boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$$

$\mathbf{x} = (x_1, \ldots, x_K)$	design vector/covariate vector,
$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$	regression parameter

effect of one covariate is captured by a set of K regression coefficients

SILC data: 95% HPD-intervals



Effects of age class (upper left), income (upper right), education (lower left) and federal state (lower right)

Sparsity for one categorical predictor

Model

$$y = \mu + \sum_{k=1}^{K} x_k \beta_k + \varepsilon, \qquad \varepsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

Sparsity: effect of the covariate can be modelled by less than K regression coefficients

- All level effects are zero. $\Longrightarrow$ Exclude covariate (group selection).
- Some level effects are zero. $\Longrightarrow$ Select level effects (within-group selection).
- Some levels have the same effect. $\Longrightarrow$ Fuse level effects.

Bayesian modelling

Achieve sparsity via appropriate prior distributions

- covariance mixture of multivariate Normals
  (Pauger and Wagner, 2017)
  - model low or high partial correlation between effects
  - spike and slab prior on effect differences
- model based clustering of level effects
  (Malsiner-Walli et. al., 2017)
  - many spiky Normal components
  - sparsity is achieved by prior on the mixture weights to encourage empty components

Covariance Mixture of Multivariate Normals

$$\beta|\tau^2, \boldsymbol{\delta} \sim \mathcal{N}\left(\mathbf{0}, \frac{K}{2}\tau^2\mathbf{Q}^{-1}(\boldsymbol{\delta})\right)$$
$$\tau^2 \sim \mathcal{G}^{-1}(g_0, G_0)$$

- $\mathbf{Q}(\boldsymbol{\delta})$ determines the structure of the prior precision matrix, depending on $\boldsymbol{\delta}$
- $\boldsymbol{\delta}$ is a vector of binary indicators
- τ^2 is a scale factor

Prior for unrestricted effect fusion

- δ_{kj} defined for each pair of effects $0 \leq j < k \leq K$
  $\Longrightarrow \boldsymbol{\delta}$ has $\binom{K+1}{2}$ elements
- prior precision matrix

$$\mathbf{Q}(\boldsymbol{\delta}) = \begin{pmatrix} \sum_{j \neq 1} \kappa_{1j} & -\kappa_{12} & \cdots & -\kappa_{1K} \\ -\kappa_{21} & \sum_{j \neq 2} \kappa_{2j} & \cdots & -\kappa_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ -\kappa_{K1} & -\kappa_{K2} & \cdots & \sum_{j \neq K} \kappa_{Kj} \end{pmatrix}$$

- elements κ_{kj} depend on the corresponding indicator

$$\kappa_{kj} = \begin{cases} 1 & \text{if } \delta_{kj} = 1 \\ r >> 1 & \text{if } \delta_{kj} = 0 \end{cases}$$

and $\kappa_{jk} = \kappa_{kj}$ for $j > k$.

Structure matrix; Examples

$K = 4$: $\boldsymbol{\delta} = (\delta_{10}, \delta_{20}, \ldots, \delta_{40}, \ldots \delta_{43})$
$r = 1000$

- $\delta_{10} = 0$

$$\mathbf{Q}(\delta) = \begin{pmatrix} \mathbf{1003} & -1 & -1 & -1 \\ -1 & 4 & -1 & -1 \\ -1 & -1 & 4 & -1 \\ -1 & -1 & -1 & 4 \end{pmatrix}$$

- $\delta_{21} = \delta_{34} = 0$

$$\mathbf{Q}(\delta) = \begin{pmatrix} \mathbf{1003} & \mathbf{-1000} & -1 & -1 \\ \mathbf{-1000} & \mathbf{1003} & -1 & -1 \\ -1 & -1 & \mathbf{1003} & \mathbf{-1000} \\ -1 & -1 & \mathbf{-1000} & \mathbf{1003} \end{pmatrix}$$

Properties of the effect fusion prior

- results from spike and slab prior on effect contrasts
  - set $\beta_0 = 0$ and define effect contrasts

    $$\theta_{kj} = \beta_k - \beta_j \quad \text{for } 0 \le j < k \le K$$

  - spike and slab prior

    $$\theta_{kj} \sim \delta_{kj}\mathcal{N}\left(0, \tau^2\right) + (1 - \delta_{kj})\mathcal{N}\left(0, \frac{\tau^2}{r}\right)$$

  - determine marginal prior for $\boldsymbol{\beta} = (\theta_{10}, \dots, \theta_{k0})$ taking into account for linear dependence

    $$\theta_{kj} = \theta_{k0} - \theta_{j0}$$

- all pairs are taken into account in the same way $\Longrightarrow$ prior is invariant to choice of the baseline

Marginal prior on regression effects

- joint prior on the indicators

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

  computationally convenient as $|\mathbf{Q}(\boldsymbol{\delta})|$ cancels out in the joint prior

- prior concentrates at
  - $\beta_k = 0$
  - $\beta_k = \beta_j$

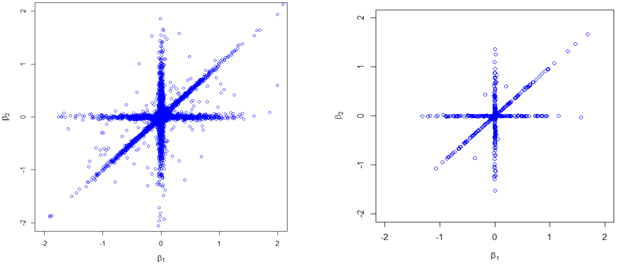


Figure: Simulation from the effect fusion prior: Plot of (β_1, β_2) for $c = 3$ and hyperparameters $g_0 = 5$, $G_0 = 2$, $r = 1000$ (left) and $r = 10000$ (right)

Prior for restricted fusion

- no direct fusion of level effects k and j: set $\kappa_{kj} = 0$
- examples
  - ordinal covariate: fusion restricted to adjacent categories

$$\kappa_{kj} = 0 \qquad j \neq k-1$$

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

  - variable selection: restrict fusion to the baseline

$$\kappa_{kj} = 0 \qquad j \neq 0$$

$$\mathbf{Q}(\delta) = \begin{pmatrix} \kappa_{10} & 0 & \dots & 0 & 0 \\ 0 & \kappa_{20} & \dots & . & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \kappa_{k,0} \end{pmatrix}$$

Posterior inference

1. MCMC: start with $\boldsymbol{\delta} = \mathbf{1}$
   - sample $\boldsymbol{\beta}$
     - compute the prior precision matrix $\mathbf{Q}(\boldsymbol{\delta})$
     - sample $\boldsymbol{\beta}$ from the conditional Normal posterior $\mathcal{N}(\mathbf{b}_n, \mathbf{B}_n)$
   - compute the effect differences $\boldsymbol{\theta}$
   - sample δ_{kj} from $p(\delta_{kj}|\theta_{kj}, \tau^2)$
   - sample τ^2
2. model selection: minimization of Binder's loss

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

   where $\mathbf{z}$ is the true and $\mathbf{z}^*$ the proposed clustering
3. refit of the selected model (with fused levels)

Simulation Study

Set-up:

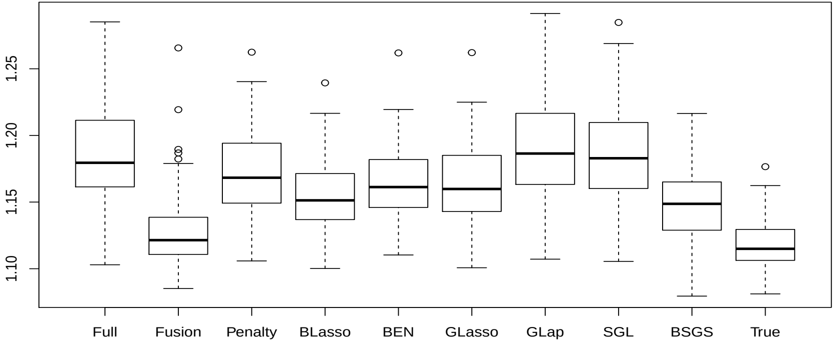
- 100 data sets of size $N = 500$
- four ordinal $C_1, \ldots, C_4$ and four nominal predictors $C_5, \ldots, C_8$
- covariate effects
  - relevant ordinal: $\beta_1 = (0, 1, 1, 2, 2, 4, 4)$, $\beta_3 = (0, -2, -2)$
  - relevant nominal: $\beta_5 = (0, 1, 1, 1, 1, -2, -2)$, $\beta_7 = (0, 2, 2)$
  - irrelevant: $\beta_2 = \beta_6 = (0, 0, 0, 0, 0, 0, 0)$, $\beta_4 = \beta_8 = (0, 0, 0)$

Results

- both zero and non-zero effect differences are identified well
- lower averaged MSE (than in the full model and other methods)
- predictive performance only slightly worse than in the true model

Simulation: Predictive Performance

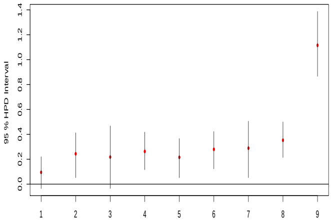
- new data set of 500 observations from the regression model
- prediction of new observations using the selected model and parameters estimated in each of the 100 data sets

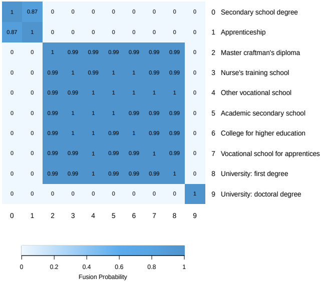


Mean squared prediction error

EU-SILC: Results

- effect of education





- final model: same fit but sparser
  - 11 regression effects (full model: 35)
  - $\hat{\sigma}^2 = 0.829$ almost identical to the full model (0.826)
  - BIC: 8240.69 (full model: 8402.00)

Sparse finite mixture prior

Model

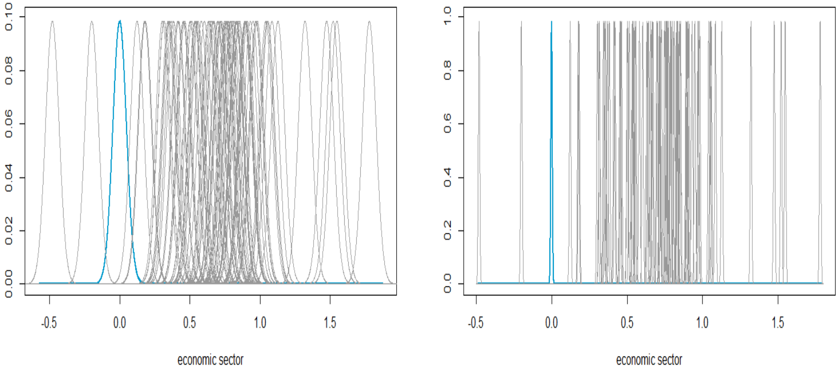
$$y = \mu + \sum_{k=1}^{K} x_k \beta_k + \varepsilon$$

- prior distribution on the regression effects

$$p(\beta_k) = \sum_{l=0}^{L} \eta_l \, p(\beta_k | \mathcal{N}(\mu_l, \psi_l))$$
$$\boldsymbol{\eta} \sim \mathcal{DIR}(e_0, \ldots, e_0)$$
$$\mu_0 = 0; \quad \mu_l \sim \mathcal{N}(m_{l0}, M_{l0}) \quad l = 1, \ldots L \text{ (e.g. } L = K)$$

- sparsity is achieved by small e_0 (e.g. 0.001)

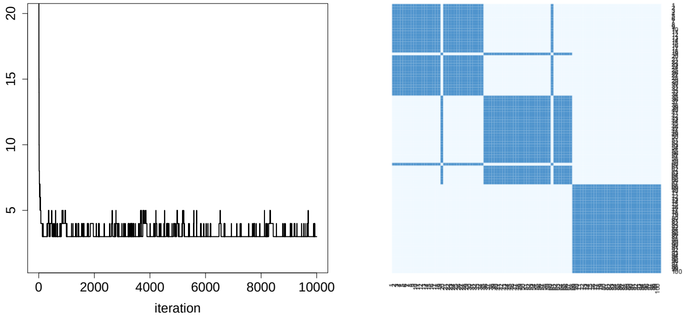
Sparse finite mixture priors



Finite mixture prior for level effects of covariate `economic sector` $\psi = 100$ (left plot) and $\psi = 10000$ (right plot). One component is centred at zero (dashed), the others at $\hat{\beta}_{jk}$, $k = 1, \ldots c_j$.

Simulation study

- Setup similar to the application in order to tune the prior parameters.
- $N = 4000$, covariates: x_1 (10 categories), x_2 (100 categories).



Simulation study, one data set: Trace plot of the number of nonempty groups during MCMC sampling for variable x_2 (left), and visualization of the most frequent model (right).

Conclusions

- prior distributions for effect fusion
  - covariance mixture of multivariate Normals
    spike and slab prior distribution on effect contrasts
  - finite mixture prior
    location mixture of normals with small variance
- Bayesian estimation
  - feasible by MCMC methods
  - "add on" for regression type models with normal priors
- pros
  - covariance mixture: simple implementation of restricted fusion
  - sparse finite mixture prior allows finer resolution
- R-package `effectFusion`

Future research

- extension to generalized linear models (straightforward)
- sparse modelling in more general models
  - multinomial logit models

$$P(Y = r) = \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta}_r)}{1 + \sum_{s=1}^{R} \exp(\boldsymbol{x}'\boldsymbol{\beta}_s)} \qquad r = 1, \ldots, R$$

    β_{rk} is the effect of predictor category k on response category r

    - ★ sparsity with respect to the predictor $\beta_{rk} = \beta_{rk'}$
    - ★ sparsity with respect to the response $\beta_{rk} = \beta_{r'k}$
  - item response models with differential item functioning
  - generalized regression models for location, scale and shape

References

- Malsiner-Walli, G., Pauger, D. and Wagner, H. (2017). Effect Fusion Using Model- Based Clustering. *Statistical Modelling*, accepted.
- Pauger, D. and Wagner, H. (2017). Bayesian Effect Fusion for Categorical Predictors. *https://arxiv.org/abs/1703.10245*.
- Pauger, D., Wagner, H., and Malsiner-Walli, G. (2016). `effectFusion`: Bayesian Effect Fusion for Categorical Predictors. *http://www.R-project.org/*.