Regularization Methods for Categorical Data

Gerhard Tutz
Ludwig-Maximilians-Universität
München

Wien
Oktober 2011

Framework for Univariate Responses

Model for $\mu_i = E(y_i|\mathbf{x}_i)$

$$\mu_i = h(\eta_i) \text{ or } g(\mu_i) = \eta_i$$

with link function g (response function $h = g^{-1}$) and η_i determined by predictors

Structuring of the influential term

- Linear

$$\eta = \beta_0 + x_1\beta_1 + \cdots + x_p\beta_p$$

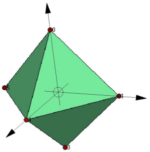
- Additive

$$\eta = \beta_0 + f_{(1)}(x_1) + \cdots + f_{(p)}(x_p),$$

  with unknown functions $f_{(j)}$
- Varying coefficients

$$\eta = \ldots x_j f(u_j) + \ldots$$

Selection Strategies



- Stepwise forward backward
- Lasso for metric predictors

The case of categorical predictors

$$\eta = \beta_0 + x_1\beta_1 + \cdots + x_p\beta_p + f(z_1) + \ldots .$$

For categorical predictor $P \in \{1, \ldots, k\}$ one obtains a linear predictor by using dummy variables.

Various coding schemes available:

0-1-Coding

$$x_{P(j)} = \begin{cases} 1 & \text{if } P = j \\ 0 & \text{otherwise} \end{cases} \qquad j = 1, \ldots, k-1$$

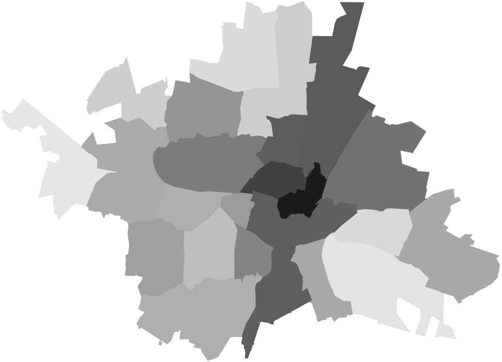
Effect Coding

$$x_{P(j)} = \begin{cases} 1 & \text{if } P = j \\ -1 & \text{if } P = k \\ 0 & \text{otherwise} \end{cases} \qquad j = 1, \ldots, k-1$$

Each categorical predictor increases the number of parameters by $k - 1$

Lasso? Selection depends on coding!

Example: Urban Districts



- Response: monthly rent per m^2.
- Predictors: urban district, decade of construction, number of rooms, floor space, etc.

For categorical predictors

Two cases should be distinguished:

- Unordered factors: Permutation invariance postulated.
- Ordinal predictors: Palindromic invariance postulated.

In both cases the following questions should be answered:

- Which categorical predictors should be included in the model?
  Variable selection
- Which categories within one categorical predictor are to be distinguished?
  Clustering

Reduction to relevant variables/categories necessary since otherwise

- estimates are instable, do not exist or are not unique
- interpretation is harder because too much noise is fitted

(1) Ordinal Predictors

Given predictor x with ordered categories/levels $0, \ldots, K$, let the linear predictor be

$$\eta = \alpha + \beta_0 x_0 + \ldots + \beta_K x_K,$$

with dummy variables $x_0, \ldots, x_K$, i.e.

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

Identifiability is obtained by specifying reference category $k = 0$, so that $\beta_0 = 0$.

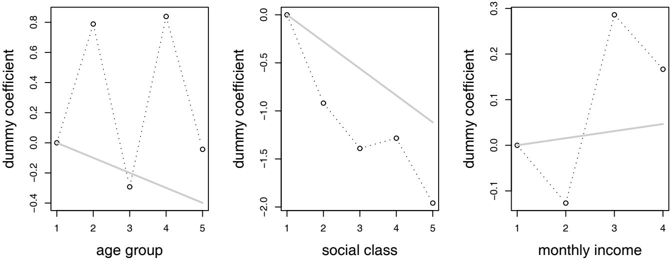
- Since levels are ordered response y is assumed to change slowly between two adjacent levels of x.
- We try to avoid high jumps and prefer a smoother coefficient vector β.

Example: Choice of coffee brand

Logit Model with binary response: cheap discounter or branded product

Explanatory variables: Ordered variables age group, social class, monthly income

Linear model versus full model



Smooth Effects by Penalizing Differences

$\Rightarrow$ Maximization of the penalized log-likelihood

$$l_p(\beta) = -\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) - \frac{\psi}{2}J(\beta),$$

with design matrix X, vector of response values y, and penalty

$$J(\beta) = \textstyle\sum_{k=1}^{K}(\beta_k - \beta_{k-1})^2 = \beta^T U^T U\beta = \beta^T \Omega\beta. \quad U = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ & \ddots & \ddots & \\ 0 & & & 0 \\ 0 & \cdots & -1 & 1 \end{pmatrix}.$$

$\Rightarrow$ For linear model one obtains the generalized ridge estimator with tuning parameter $\lambda = \psi\sigma^2$ and $\Omega = U^T U$

$$\hat{\beta}^* = (X^T X + \lambda\Omega)^{-1} X^T y,$$

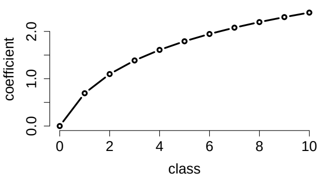
- For GLMs iterative estimation procedure
- Regularization ensures existence of estimates

Bias-Variance

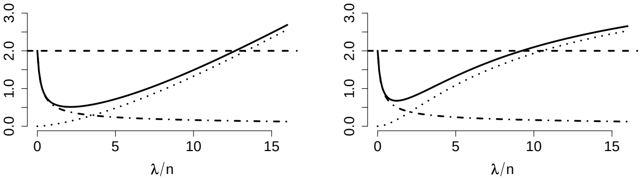
$$\begin{aligned} E(\hat{\beta}^*) &= (X^T X + \lambda\Omega)^{-1} X^T X\beta = \beta - \lambda(X^T X + \lambda\Omega)^{-1}\Omega\beta, \\ V(\hat{\beta}^*) &= \sigma^2 (X^T X + \lambda\Omega)^{-1} X^T X (X^T X + \lambda\Omega)^{-1}. \end{aligned}$$

Illustration

- Balanced designs with n observations in each of $K+1 = 11$ classes, $\sigma^2/n = 0.2$ and coefficient vectors ($\alpha = 0$):



- (squared) bias ($\cdots$), variance ($-\cdot$) and (scalar) MSE ($-$):



Example: Chronic Widespread Pain

- Pain involving several regions of the body, which causes
  - problems in functioning, psychological distress, poor sleep quality, difficulties in activities of daily life,...
- No systematic framework that covers the spectrum of symptoms and limitations of patients with CWP (cf. Cieza et al., 2004).

$\Rightarrow$ ICF - *International Classification of Functioning, Disability and Health* (WHO, 2001) to **define the typical spectrum** of problems of patients with CWP.

The ICF consists of $\approx$ 1400 ordinally scaled factors (*variables*), e.g.:

Variable "*walking*" (component "*activities and participation*"):

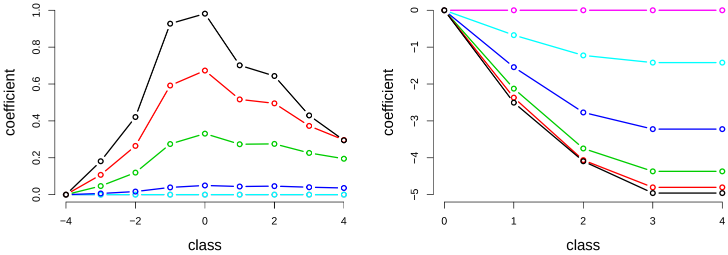
0	1	...	4
no difficulty	mild difficulty	...	complete difficulty

From the ICF categories experts selected the *(Comprehensive)* **ICF Core Set** (67 variables) for CWP (see Cieza et al., 2004).

Some Coefficient Paths

ICF Core Sets → SF36 (Wellness score)

- Environmental factor "*social norms, practices and ideologies*" (left).
- Factor "*walking*" (component "*activities and participation*", right).



Smooth Effects Including Variable Selection: Penalty Approach

For unordered response approaches available.
The Group Lasso (Yuan & Lin, 2006) works with a **Lasso** penalty at the factor level.

For p factors it has the form

$$J_{gl}(\beta) = \sum_{j=1}^{p} \sqrt{df_j} \sqrt{\boldsymbol{\beta}_j^T \boldsymbol{\beta}_j} = \sum_{j=1}^{p} \sqrt{df_j} ||\boldsymbol{\beta}_j||_2$$

where $\boldsymbol{\beta}_j$ refers to the parameter vector of the jth variable.
Thus the group of coefficients collected in $\boldsymbol{\beta}_j$ is shrunk by use of a lasso type penalty

Effects:

- Encourages sparsity at the factor level
- Designed for nominal factors, uses no ordering of categories
- R add-on package `grplasso` (Meier et al., 2008)

Group Lasso for Ordered Categories

Transform the problem with difference penalties

$$(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)+\lambda J(\boldsymbol{\beta}) = (\mathbf{y}-\widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}})^T(\mathbf{y}-\widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}})+\lambda\widetilde{J}(\widetilde{\boldsymbol{\beta}}),$$

with $\widetilde{\mathbf{X}} = (1|\widetilde{\mathbf{X}}_1|\dots|\widetilde{\mathbf{X}}_p)$, $\widetilde{\boldsymbol{\beta}} = (\alpha, \widetilde{\boldsymbol{\beta}}_1^T, \dots, \widetilde{\boldsymbol{\beta}}_p^T)^T$, and $\widetilde{\mathbf{X}}_j = \mathbf{X}_j\mathbf{U}_j^{-1}$, $\widetilde{\boldsymbol{\beta}}_j = \mathbf{U}_j\boldsymbol{\beta}_j$,
New parameters have the form $\widetilde{\beta}_{jr} = \beta_{j,r+1} - \beta_{jr}$

Then the penalty becomes

$$\widetilde{J_{gl}}(\widetilde{\boldsymbol{\beta}}) = \sum_{j=1}^{p} \sqrt{\widetilde{\boldsymbol{\beta}}_j^T \mathbf{I}_j \widetilde{\boldsymbol{\beta}}_j}, \ .$$

Equivalent to predictors given in split-coding

$$\tilde{x}_{A(i)} = \begin{cases} 1 & \text{if } A > i \\ 0 & \text{otherwise} \end{cases}$$

Software for group lasso can be used by appropriate definition of design matrix

$\Rightarrow$ Enforces selection on the factor level including smoothness across categories

Blockwise Boosting

Componentwise L_2-Boosting (Bühlmann, 2006):

- Repeated least squares fitting of residuals.
- In each iteration only one predictors is selected, and the corresponding coefficient updated.

Blockwise Boosting:

- Groups - or *blocks* - of coefficients are updated.
- Blocks are formed by groups of dummy coefficients.
- In each iteration: Regression with difference penalty.
- Coefficients which are never updated remain zero.

$\Rightarrow$ Variable Selection.

Likelihood-based Boosting

Let y_i be from an exponential family distribution with mean $\mu_i = E(y_i|\mathbf{x}_i)$ and the link between the mean and the structuring term specified by

$$\mu_i = h(\eta_i) \quad \text{or} \quad g(\mu_i) = \eta_i$$

1 Initialization

For given data $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, fit the intercept model $\mu^{(0)}(\mathbf{x}) = h(\eta_0)$ by maximizing the likelihood, yielding $\eta^{(0)} = \hat{\eta}_0, \hat{\mu}^{(0)} = h(\hat{\eta}_0)$.

2 Iteration For $l = 0, 1, \ldots$

Fitting step

Fit the model

$$\mu_i = h(\hat{\eta}^{(l)}(\mathbf{x}_i) + \eta(\mathbf{x}_i, \boldsymbol{\gamma}))$$

to data $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, where $\hat{\eta}^{(l)}(\mathbf{x}_i)$ is treated as an offset and the predictor is estimated by fitting the parametrically structured term $\eta(\mathbf{x}_i, \boldsymbol{\gamma})$, obtaining $\hat{\boldsymbol{\gamma}}$

Update step

The improved fit is obtained by

$$\hat{\eta}^{(l+1)}(\mathbf{x}_i) = \hat{\eta}^{(l)}(\mathbf{x}_i) + \hat{\eta}(\mathbf{x}_i, \hat{\gamma}), \quad \hat{\mu}_i^{(l+1)} = h(\hat{\eta}^{(l+1)}(\mathbf{x}_i))$$

For normally distributed response and least squares fitting equivalent to L_2-boosting

Blockwise Boosting of Coefficients

- Parametrically structured term includes only one factor

  For predictor j
  $$\eta(\mathbf{x}_i, \boldsymbol{\gamma}) = \mathbf{x}_j^T \mathbf{b}_j$$

- Penalized fitting

  Fit for all variables $j = 1, \ldots, p$ the one-variable model
  $$\mu_i = h(\hat{\eta}_i^{(l)} + \mathbf{x}_j^T \mathbf{b}_j)$$

  by one step Fisher scoring in the form $\hat{\mathbf{b}}_j^{new} = F_p(\hat{\boldsymbol{\beta}}_j^{(r-1)})^{-1} s_p(\hat{\boldsymbol{\beta}}_j^{(r-1)})$, where F_p is the penalized Fisher matrix, s_p is the penalized score function

  For linear models one uses $\hat{\mathbf{b}}_j = (\mathbf{X}_j^T \mathbf{X}_j + \lambda \boldsymbol{\Omega}_j)^{-1} \mathbf{X}_j^T \mathbf{u}$, where $\mathbf{u}^T = (u_1, \ldots, u_n)$ contains the residuals $u_i = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_j^{(r-1)}, i = 1, \ldots, n$

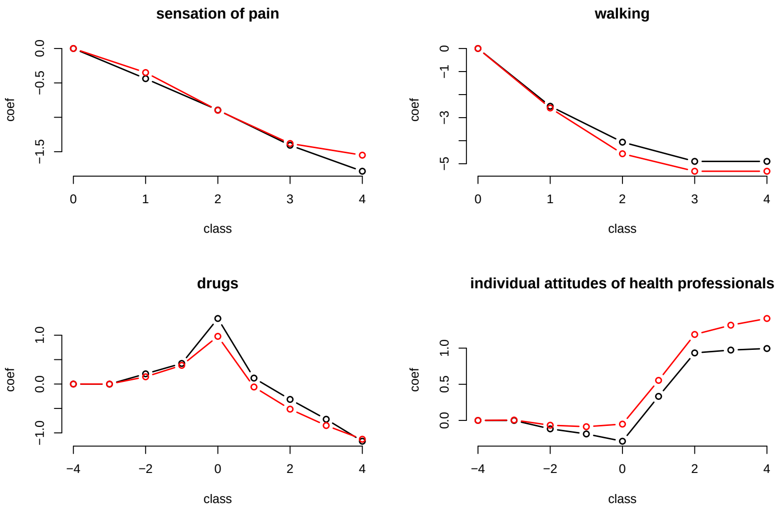
- Selection of block that is updated

  Choose $\hat{j}_r$ such that the deviance or AIC is minimized,

- Update
  $$\boldsymbol{\beta}_{j_r}^{(r)} = \boldsymbol{\beta}_{j_r}^{(r-1)} + \mathbf{b}_j, \quad \boldsymbol{\beta}_j^{(r)} = \boldsymbol{\beta}_j^{(r-1)}, j \neq j_r$$

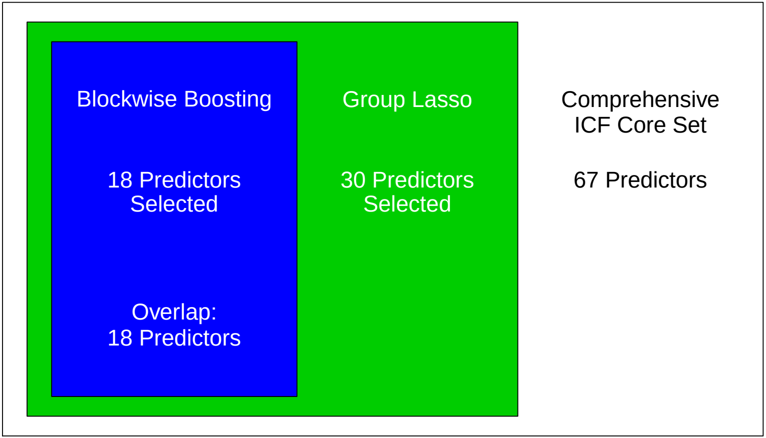
Application to ICF Core Sets

Comparisons Blockwise Boosting / Group Lasso: Some Coefficients



Application to ICF Core Sets

Comparisons Blockwise Boosting / Group Lasso: Selection Results



- The Group Lasso shows a slightly better fit (≈ 7 % lower RSS).

(2) Clustering of Categories for Categorical Predictors

Which categories should be distinguished?

Clustering Ordered Categories

Quadratic penalty is replaced by L_1 difference penalty:

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \sum_{i=1}^{k_j} |\beta_{ji} - \beta_{j,i-1}|$$

- Clustering if some adjacent dummy coefficients are set equal.
- Exclusion if all coefficients belonging to the same predictor are set to zero / equal.

- Equivalent to original Lasso based on split-coding
- Corresponds to blockwise Fused Lasso (Tibshirani et al., 2005).

General Lasso Type Differences

Penalty

- Ordered Predictor

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} w_{il}^{(j)} \sum_{i=1}^{k_j} |\beta_{ji} - \beta_{j,i-1}|$$

- Nominal Predictor

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} w_{il}^{(j)} \sum_{i>l} |\beta_{ji} - \beta_{jl}|$$

Bondell & Reich, 2009 for ANOVA; Gertheiss & Tutz, 2009 for selection

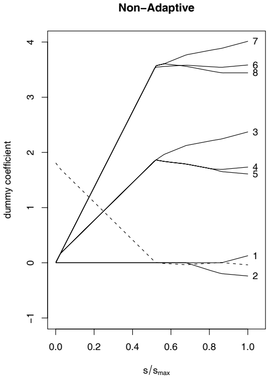
Weights given by

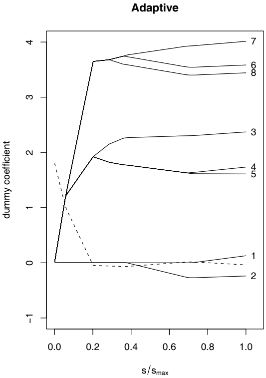
$$w_{il}^{(j)} = w(n_i^{(j)}, n_l^{(j)})|\beta_{ji}^{(LS)} - \beta_{jl}^{(LS)}|^{-1},$$

⇒ Include:

- Dependence on local sample sizes.
- Is adaptive by using consistent estimates (like Zou, 2006).

Illustration ordered case





Large Sample Properties
($p = 1$)

- $\theta = (\theta_{10}, \theta_{20}, \ldots, \theta_{k,k-1})^T$: vector of pairwise differences $\theta_{il} = \beta_i - \beta_l$.
- $\mathcal{C} = \{(i,l) : \beta_i^* \neq \beta_l^*, i > l\}$: set of indices $i > l$ corresponding to differences of (true) dummy coefficients β_i^* which are truly non-zero.
- $\mathcal{C}_n$: estimate of $\mathcal{C}$ with sample size n.
- $\theta_{\mathcal{C}}^*$ / $\hat{\theta}_{\mathcal{C}}$: true / estimated vector of pairwise differences included in $\mathcal{C}$.

Proposition

Suppose $\lambda = \lambda_n$ with $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n \to \infty$, and all class-wise sample sizes n_i satisfy $n_i/n \to c_i$, where $0 < c_i < 1$. Then weights $w_{il} = \phi_{il}(n)|\hat{\beta}_i^{(LS)} - \hat{\beta}_l^{(LS)}|^{-1}$, with $\phi_{il}(n) \to q_{il}$ $(0 < q_{il} < \infty)$ $\forall i, l$, ensure that

(a) $\sqrt{n}(\hat{\theta}_{\mathcal{C}} - \theta_{\mathcal{C}}^*) \to_d N(0, \Sigma)$,

(b) $\lim_{n\to\infty} P(\mathcal{C}_n = \mathcal{C}) = 1$.

Computational Issues

Solution by Quadratic programming
or
Approximate Solution using LARS (much faster)

Vector of pairwise differences is $\theta = (\theta_{10}, \theta_{20}, \ldots, \theta_{k,k-1})^T$ with $\theta_{il} = \beta_i - \beta_l$
Therefore parameters must fulfill restrictions. Since $\theta_{i0} = \beta_i$, one has $\theta_{il} = \theta_{i0} - \theta_{l0}$.

Use adaptive Net Penalty

With Z so that $Z\theta = X\beta$, minimize

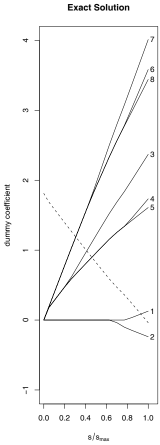
$$\hat{\theta}_{\gamma,\lambda} = (y - Z\theta)^T(y - Z\theta) + \gamma \sum_{i>j>0} (\theta_{i0} - \theta_{j0} - \theta_{ij})^2 + \lambda \sum_{i>j} |\theta_{ij}|.$$

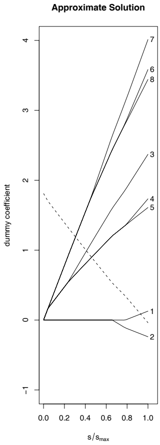
A simple choice of Z is $Z = (X|0)$, since $\theta_{i0} = \beta_i$, $i = 1, \ldots, k$.

The exact solution of the is obtained as the limit

$$\hat{\theta} = \lim_{\gamma \to \infty} \hat{\theta}_{\gamma,\lambda}.$$

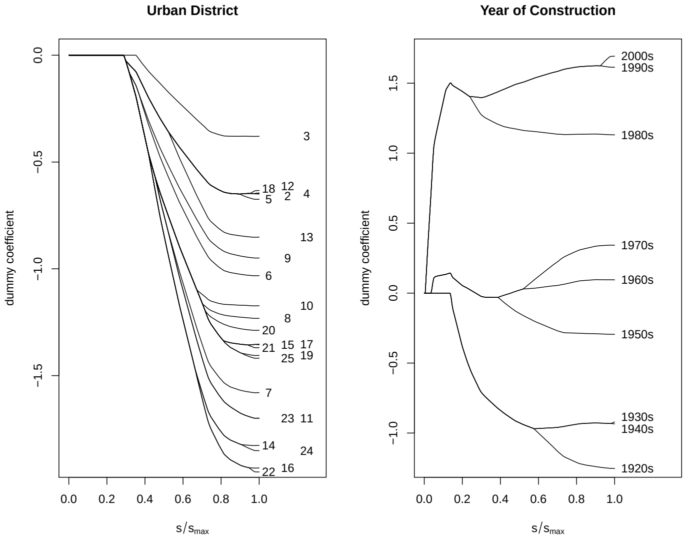
Illustration ordered case





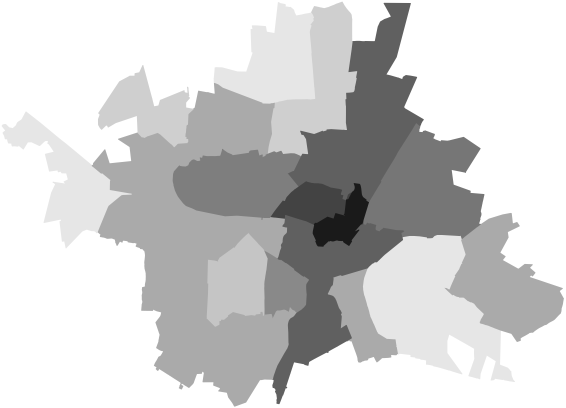
Coefficient Paths for Munich Rent Data

Unordered and Ordered Categories



Rent Data

Some Clustering Results (Adaptive Version with Refitting)



- All in all the estimated model has 32 df (i.e. unique non-zero coefficients).
- The full model has 58 df.

Rent Data

Some results (adaptive version with refitting)

predictor	label	coefficient
urban district	14, 16, 22, 24	-1.931
	11, 23	-1.719
	7	-1.622
	8, 10, 15, 17, 19, 20, 21, 25	-1.361
	6	-1.061
	9	-0.960
	13	-0.886
	2, 4, 5, 12, 18	-0.671
	3	-0.403
number of rooms	4, 5, 6	-0.502
	3	-0.180
	2	0.000
quality of residential area	good	0.373
	excellent	1.444

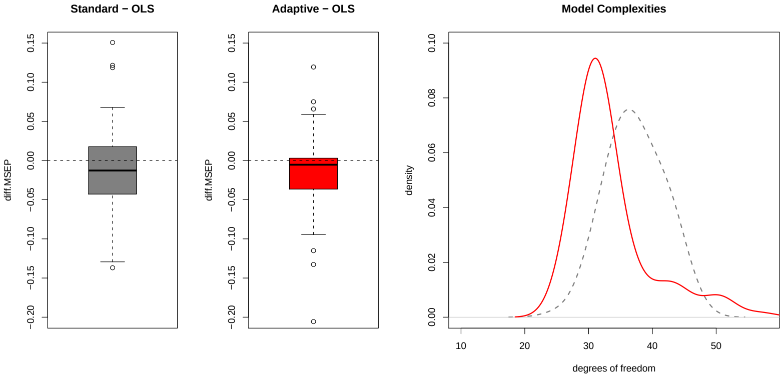
Rent Data

Some results (adaptive version with refitting)

predictor	label	coefficient
year of construction	1920s	-1.244
	1930s, 1940s	-0.953
	1950s	-0.322
	1960s	0.073
	1970s	0.325
	1980s	1.121
	1990s, 2000s	1.624
floor space (m^2)	$[140, \infty)$	-4.710
	$[90, 100)$, $[100, 110)$, $[110, 120)$, $[120, 130)$, $[130, 140)$	-3.688
	$[60, 70)$, $[70, 80)$, $[80, 90)$	-3.443
	$[50, 60)$	-3.177
	$[40, 50)$	-2.838
	$[30, 40)$	-1.733

Rent Data

Prediction accuracies and model complexities (standard/adaptive with refitting)



- Based on random splitting of the data into independent training and test sets (1953/100 observations).
- 100 independent repetitions.

Generalizations to Non-Normal Outcomes

Example: Wisconsin breast cancer database (Wolberg & Mangasarian, 1990)

- Instances are to be classified as **benign** ($y = 0$) or **malignant** ($y = 1$)
- Available covariates are cytological characteristics as
  - marginal adhesion,
  - bare nuclei,
  - mitoses,
  - ...
- Predictors are graded on a **1 to 10 scale** at the time of sample collection, with 1 being the closest to normal tissue and 10 the most anaplastic.
- We fit a **logistic regression** model using **penalized likelihood** estimation.
- Minimize the penalized negative log-likelihood

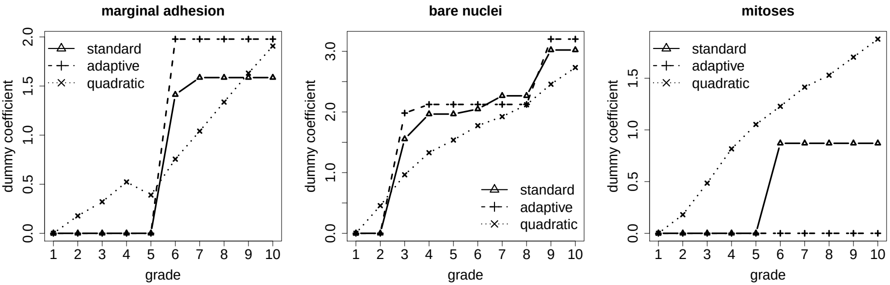
$$-l_p(\beta) = -l(\beta) + \lambda J(\beta).$$

Generalizations to Non-Normal Outcomes

Example: Wisconsin breast cancer database (Wolberg & Mangasarian, 1990)

Some estimated coefficient functions (cf. Stelz, 2010):

- standard/adaptive L_1-regularization (using R package `glmpath` (Park & Hastie, 2007)),
- quadratic difference penalty for smooth modeling (using R package `ordPens` (Gertheiss, 2010)).



Numerical Experiments

Simulation Design

- Setting with 8 predictors (intercept $\alpha = 1$):

type	no. of levels	true dummy coefficients
nominal	4	$(0, 2, 2)'$
nominal	8	$(0, 1, 1, 1, 1, -2, -2)'$
nominal	4	$(0, 0, 0)'$
nominal	8	$(0, 0, 0, 0, 0, 0, 0)'$
ordinal	4	$(0, -2, -2)'$
ordinal	8	$(0, 1, 1, 2, 2, 4, 4)'$
ordinal	4	$(0, 0, 0)'$
ordinal	8	$(0, 0, 0, 0, 0, 0, 0)'$

- Standard normal error.
- Training set size $n = 500$.
- 100 simulation runs.
- Independent test set ($n = 1000$).
- Compare ordinary least squares (ols), standard, adaptive version, with/without refitting.
  Refitting means the selected coefficients are fitted in the last step - selection of tuning parameters refers to the whole procedure.

Errors of Parameter Estimates and Prediction:

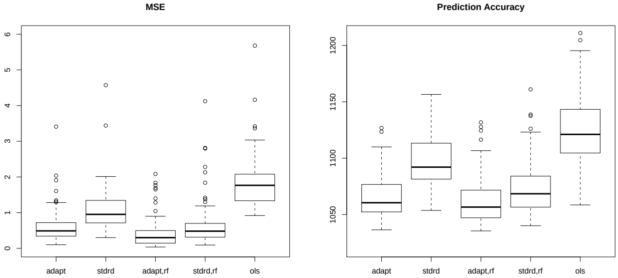
- MSE of parameter estimates.
- Prediction Accuracy: Empirical sum of squared test set errors.

Variable Selection and Clustering Performance:

- False Positive Rates / FPR:
  - Variable Selection: Any dummy coefficient of a pure noise factor is set to non-zero.
  - Clustering / Identifying Differences: A difference within a non-noise factor which is truly zero is set to non-zero.
- False Negative Rates / FNR:
  - Variable Selection: All dummy coefficients of a truly relevant factor are set to zero.
  - Clustering / Identifying Differences: A truly non-zero difference is set to zero.

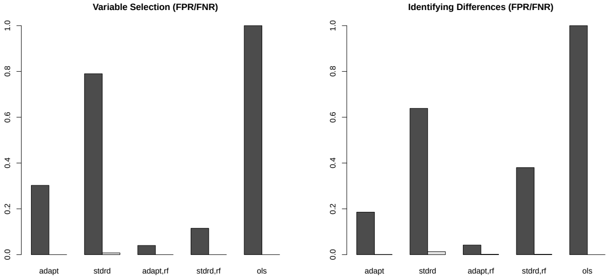
Numerical Experiments

Errors of Parameter Estimates and Prediction



Numerical Experiments

Variable Selection and Clustering Performance



(3) Varying-Coefficient Models

Varying-coefficient models (Hastie & Tibshirani, 1993) offer a quite flexible framework for regression modeling.

In a linear model, with one **effect modifier u**:

$$y = \beta_0(u) + x_1\beta_1(u) + \ldots + x_p\beta_p(u) + \epsilon,$$

with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$.

$\hookrightarrow$ Functions $\beta_j(u)$ are allowed to vary with the effect modifier u.

Usually **metric/continuous** effect modifiers u are investigated, and $\beta_j(u)$ are modeled as smooth functions.

Two questions should be answered:

(1) **Variable selection**, i.e. selecting relevant predictors x_j.
$\hookrightarrow$ Determine if $\beta_j(u) = 0$.

(2) **Identify varying coefficients** $\beta_j(\cdot)$.
$\hookrightarrow$ Determine if $\beta_j(u)$ is a constant or not.

Given continuous u, penalty approaches have been used to answer (one of) these questions:

(1) Wang et al. (2008), Wang & Xia (2009);

(2) Leng (2009).

In this talk:

- **Categorical** effect modifier u.
- **Penalty approach** that accounts for **both (1) and (2)**.

Categorical Effect Modifiers

- For categorical $u \in \{1, \ldots, k\}$ the varying functions have the form

$$\beta_j(u) = \sum_{r=1}^{k} \beta_{jr} I(u = r).$$

- The model with p predictors contains $(p+1)k$ parameters:

$$y = \sum_{r=1}^{k} \beta_{0r} I(u = r) + \sum_{r=1}^{k} x_1 \beta_{1r} I(u = r) + \ldots + \sum_{r=1}^{k} x_p \beta_{pr} I(u = r) + \epsilon$$

- On level r of u:

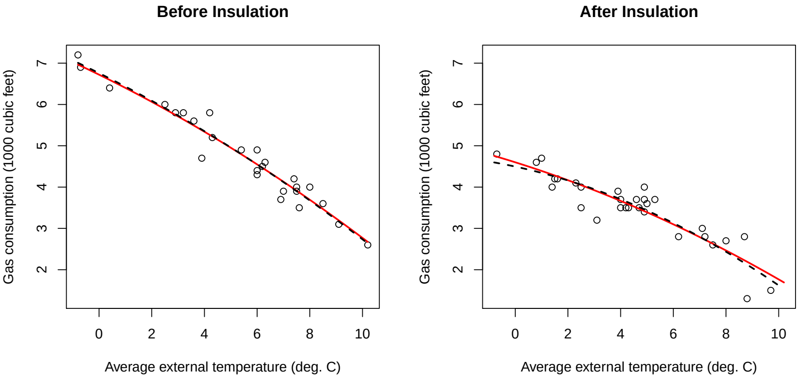
$$y = \beta_{0r} + x_1 \beta_{1r} + \ldots + x_p \beta_{pr} + \epsilon$$

An Illustrative Example

Whiteside's insulation data (Hand et al., 1994; Venables & Ripley, 2002)

$u \in \{1, 2\} = \{\text{Before}, \text{After}\}$:

$$E(y|x, u = r) = \beta_{0r} + x\beta_{1r} + x^2\beta_{2r}$$



Penalized Estimation

Minimization of the penalized least-squares criterion

$$\hat{\beta} = \text{argmin}_{\beta}\, Q_p(\beta),$$

with

$$\begin{aligned} Q_p(\beta) &= \sum_{i=1}^{n}\left(y_i - \beta_0(u_i) - \sum_{j=1}^{p} x_{ij}\beta_j(u_i)\right)^2 + \lambda J(\beta) \\ &= (y - Z\beta)^T(y - Z\beta) + \lambda J(\beta), \end{aligned}$$

$y = (y_1, \ldots, y_n)^T$ and $\beta = (\beta_1^T, \ldots, \beta_k^T)^T$,
with $\beta_r = (\beta_{0r}, \beta_{1r}, \ldots, \beta_{pr})^T$.

The ith row of design matrix Z is $((1, x_i^T)I(u_i = 1), \ldots, (1, x_i^T)I(u_i = k))$.

But: classical penalties are not designed for categorical effect modifiers.

Categorical Effect Modifiers

Penalized estimation

- Nominal u:

$$J(\beta) = \sum_{j=0}^{p} \sum_{r>s} |\beta_{jr} - \beta_{js}| + \sum_{j=1}^{p} \sum_{r=1}^{k} |\beta_{jr}|, \text{ or}$$

$$J(\beta; \psi) = \psi \sum_{j=0}^{p} \sum_{r>s} |\beta_{jr} - \beta_{js}| + (1-\psi) \sum_{j=1}^{p} \sum_{r=1}^{k} |\beta_{jr}|$$

- Ordinal u:

$$J(\beta) = \sum_{j=0}^{p} \sum_{r=2}^{k} |\beta_{jr} - \beta_{j,r-1}| + \sum_{j=1}^{p} \sum_{r=1}^{k} |\beta_{jr}|, \text{ or}$$

$$J(\beta; \psi) = \psi \sum_{j=0}^{p} \sum_{r=2}^{k} |\beta_{jr} - \beta_{j,r-1}| + (1-\psi) \sum_{j=1}^{p} \sum_{r=1}^{k} |\beta_{jr}|$$

Large Sample Properties

as before

Suppose $0 \leq \lambda < \infty$ has been fixed, and all class-wise sample sizes n_r satisfy $n_r/n \to c_r$, where $0 < c_r < 1$.

- The non-adaptive estimator $\hat{\beta}$ is consistent in terms of $\lim_{n\to\infty} P(||\hat{\beta} - \beta^*||^2 > \epsilon) = 0$ for all $\epsilon > 0$, if β^* denotes the vector of true coefficient functions $\beta_j(u)$, resp. true β_{jr}.
- No consistency in terms of variable selection and the identification of relevant differences $\hat{\beta}_{jr} - \hat{\beta}_{js}$.

Choose $\lambda = \lambda_n$ with $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n \to \infty$.

- Adaptive version for selection and fusion consistency.

Given nominal u, we employ the adaptive penalty

$$J(\beta) = \sum_{j=0}^{p} \sum_{r>s} w_{rs(j)} |\beta_{jr} - \beta_{js}| + \sum_{j=1}^{p} \sum_{r=1}^{k} w_{r(j)} |\beta_{jr}|,$$

with adaptive weights (similarly to Zou, 2006)

$$w_{rs(j)} = \phi_{rs(j)}(n) |\hat{\beta}_{jr}^{(LS)} - \hat{\beta}_{js}^{(LS)}|^{-1} \text{ and } w_{r(j)} = \phi_{r(j)}(n) |\hat{\beta}_{jr}^{(LS)}|^{-1},$$

with $\hat{\beta}_{jr}^{(LS)}$ denoting the ordinary least squares estimator of β_{jr}.

- $\phi_{rs(j)}(n) \rightarrow q_{rs(j)}$ and $\phi_{r(j)}(n) \rightarrow q_{r(j)}$ respectively, with $0 < q_{rs(j)}, q_{r(j)} < \infty$.
- $\phi_{rs(j)}(n)$ and $\phi_{r(j)}(n)$ will usually be fixed, for example as ψ and $(1 - \psi)$.

- $\beta_{-0,r} = (\beta_{1r}, \ldots, \beta_{pr})^T$,
- $\delta_j = (\beta_{j2} - \beta_{j1}, \beta_{j3} - \beta_{j1}, \ldots, \beta_{jk} - \beta_{j,k-1})^T$, $j = 0, \ldots, p$.
- $\beta_{-0}^T = (\beta_{-0,1}^T, \ldots, \beta_{-0,k}^T)$, $\delta^T = (\delta_0^T, \ldots, \delta_p^T)$, and $\theta^T = (\beta_{-0}^T, \delta^T)$.
- $\mathcal{C}$ the set of indices corresponding to entries of θ which are truly non-zero, $\mathcal{C}_n$ the estimate with sample size n.
- $\theta_{\mathcal{C}}^*$ the true vector of θ-entries included in $\mathcal{C}$, and $\hat{\theta}_{\mathcal{C}}$ the corresponding estimate.

Suppose $\lambda = \lambda_n$ with $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n \to \infty$, and all class-wise sample sizes n_r satisfy $n_r/n \to c_r$, where $0 < c_r < 1$. Then the adaptive penalty ensures

(a) Asymptotic normality: $\sqrt{n}(\hat{\theta}_{\mathcal{C}} - \theta_{\mathcal{C}}^*) \to_d N(0, \Sigma)$.

(b) Selection/fusion consistency: $\lim_{n\to\infty} P(\mathcal{C}_n = \mathcal{C}) = 1$.

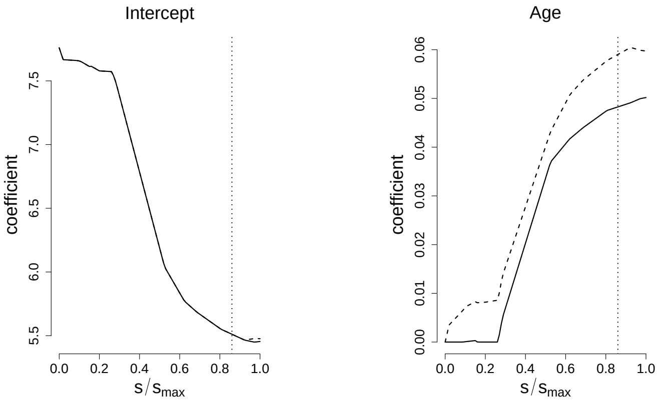
Response:	Monthly income	in Euro
Predictors:	Age	in years between 21 and 60
	Job tenure	in months
	Body height	in cm
	Gender	male/female
	Married	no/yes
	Abitur (≈ A-levels)	no/yes
	Blue-collar worker	no/yes

Model:

$$
\begin{aligned}
\log(\text{Income}) &= \beta_0(\text{Gender}) + \beta_1(\text{Gender})\text{Age} + \beta_2(\text{Gender})\text{Age}^2 \\
&+ \beta_3(\text{Gender})\text{Tenure} + \beta_4(\text{Gender})\text{Height} \\
&+ \beta_5(\text{Gender})\text{Married} + \beta_6(\text{Gender})\text{Abitur} \\
&+ \beta_7(\text{Gender})\text{Blue-collar} + \epsilon.
\end{aligned}
$$

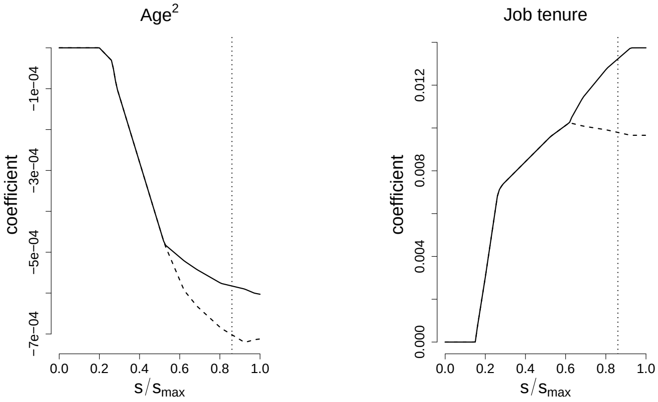
Income Data

Coefficient paths I (adaptive estimator with fixed $\psi = 0.5$)



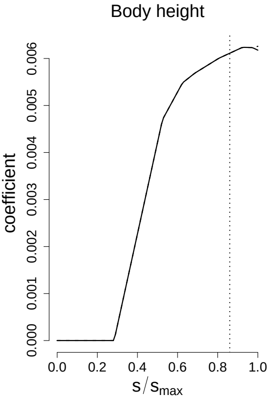
Income Data

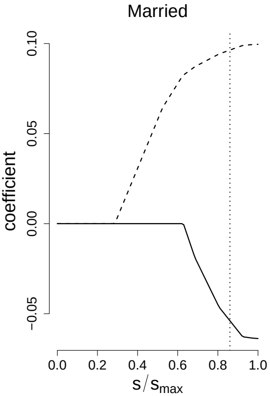
Coefficient paths II (adaptive estimator with fixed $\psi = 0.5$)



Income Data

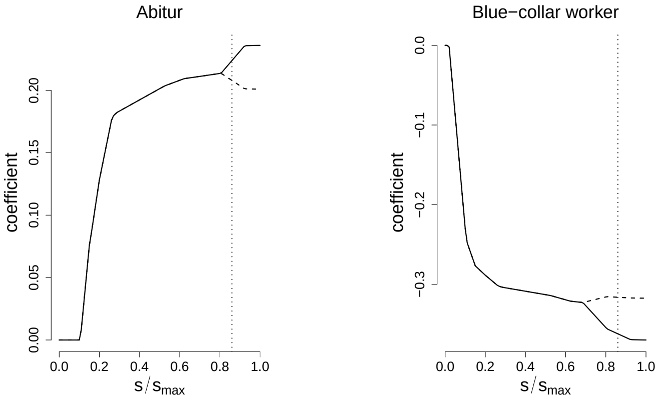
Coefficient paths III (adaptive estimator with fixed $\psi = 0.5$)





Income Data

Coefficient paths IV (adaptive estimator with fixed $\psi = 0.5$)



(4) Multinomial Response Models

For data $(Y_i, \mathbf{x}_i), i = 1, \ldots, n$, with $Y_i \in \{1, \ldots, p\}$ denoting the response variable and $\mathbf{x}_i$ the predictor, the multinomial logit model specifies

$$P(Y_i = r|\mathbf{x}_i) = \frac{\exp(\beta_{r0} + \mathbf{x}_i^T \boldsymbol{\beta}_r)}{\sum_{s=1}^{k} \exp(\beta_{s0} + \mathbf{x}_i^T \boldsymbol{\beta}_s)} = \frac{\exp(\eta_{ir})}{\sum_{s=1}^{k} \exp(\eta_{is})},$$

with predictor

$$\eta_{ir} = \beta_{r0} + \mathbf{x}_i^T \boldsymbol{\beta}_r,$$

where $\boldsymbol{\beta}_r^T = (\beta_{r1}, \ldots, \beta_{rp})$.

More generally in the linear predictor category-specific variables $\mathbf{w}_{i1}, \ldots, \mathbf{w}_{ik}$ can be included yielding the predictor

$$\eta_{ir} = \beta_{r0} + \mathbf{x}_i^T \boldsymbol{\beta}_r + (\mathbf{w}_{ir} - \mathbf{w}_{ik})^T \boldsymbol{\alpha}, \qquad r = 1, \ldots, k-1.$$

Penalized log-likelihood approach maximizes

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \lambda J(\boldsymbol{\beta}).$$

Straightforward use of the lasso uses

$$J(\boldsymbol{\beta}) = \sum_{r=1}^{k-1} ||\boldsymbol{\beta}_r||_1 = \sum_{r=1}^{k-1} \sum_{j=1}^{p} |\beta_{rj}|,$$

(Friedman et al, 2010).

Drawback:

- Single effects are selected, no variable selection because one variable has k-1 effects

With the focus on variable selection one collects all the parameters linked to variable j in $\boldsymbol{\beta}_{.j}^T = (\beta_{1j}, \dots, \beta_{k-1,j})$. We propose the penalty

$$
\begin{aligned}
J(\boldsymbol{\beta}) =& \gamma \sum_{j=1}^{p} s(k-1)||\boldsymbol{\beta}_{\cdot j}||_2 + (1-\gamma)s(1)||\boldsymbol{\alpha}|| \\
&= \gamma \sum_{j=1}^{p} s(k-1)(\beta_{1j}^2 + \cdots + \beta_{k-1,j}^2)^{1/2} + (1-\gamma)\sum_{j=1}^{L} s(1)|\alpha_j|,
\end{aligned}
$$

where γ is an additional tuning parameter that balances the penalty on the global and the category-specific variables, and $s(m) = m^{1/2}$ accounts for the number of penalized parameters within one term.

Minimization by appropriate block coordinate ascent algorithm.

Example Spatial Election Theory

Response is party

- Christian Democratic Union (CDU: 1)
- Social Democratic Party (SPD: 2)
- Green Party (3)
- Liberal Party (FDP: 4)
- Left Party (Die Linke: 5)

Global Predictors

- age, political interest (1: less interested 0: very interested),
- religion (1: evangelical, 2: catholic, 3: otherwise),
- regional provenance (west; 1: former West Germany, 0: otherwise),
- gender (1: male, 0: female),
- union (1: member of a union 0: otherwise),
- satisfaction with the functioning of democracy (democracy; 1: not satisfied 0: satisfied),
- unemployment (1: currently unemployed, 0: otherwise),
- high school degree (1: yes, 0: no)

Category-specific predictors are distances between position of the voter and the perceived position of the party on

- attitude toward immigration of foreigners
- attitude toward the use of nuclear energy
- positioning on a left-right scale

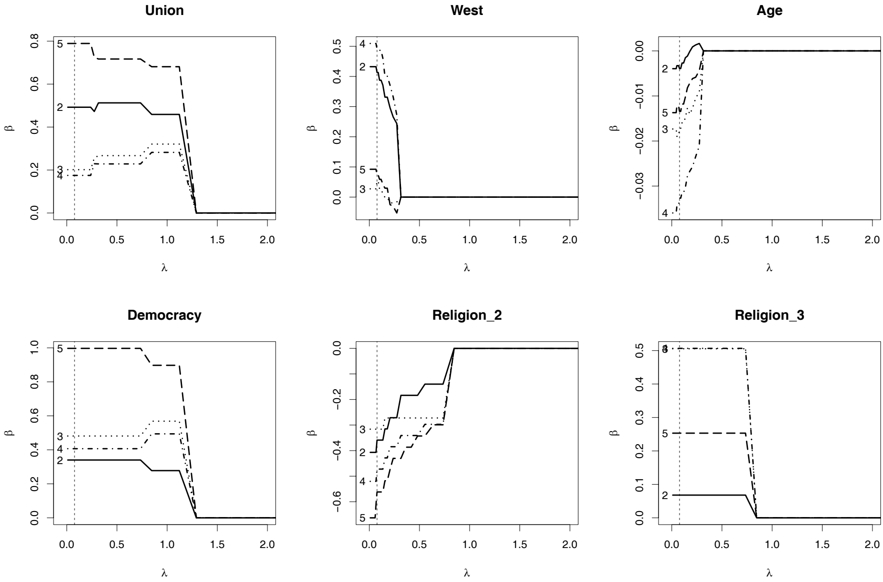


Figure: Coefficient buildups for selected global variables of party choice data.

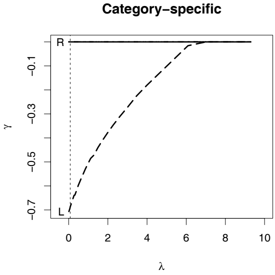


Figure: Coefficient buildups for category-specific variables of party choice data (L denotes left right scale, R denotes the rest).

Summary

- Common shrinking methods are typically designed for metric predictors.
- In case of categorical covariates penalties must be modified.
- Quadratic regularization for smooth modeling of ordinal predictors.
- L_1-penalization of pairwise differences of dummy coefficients allows for:
  - **Variable Selection**.
  - **Clustering** of categories $\leftrightarrow$ Identification of relevant differences/jumps.
- Sparser representations of varying-coefficient models with categorical effect modifiers via penalizing absolute differences and L_1-norms of coefficients.
- Simulation studies and real-world data evaluation showed:
  - **Model complexity** can be **reduced**, which facilitates interpretation.
  - **Estimation accuracy** can be **increased**.
- Appropriate Penalization allows Variable Selection in Multinomial Response Models.

References

Tutz, G (2011) Regression for Categorical Data. Cambridge University Press, to appear

Tutz, G., Binder, H. (2006): Generalized additive modelling with implicit variable selection by likelihood based boosting. *Biometrics*, 62, 961-971.

Gertheiss, J., Hogger, S., Oberhauser, C. & Tutz, G. (2011): Selection of ordinally scaled independent variables with applications to international classification of functioning core sets, *Applied Statistics* (to appear).

Gertheiss, J. & Tutz, G. (2009b): Penalized regression with ordinal predictors, *International Statistical Review*, 77, 345–365.

Gertheiss, J. & Tutz, G. (2010): Sparse modeling of categorial explanatory variables, *The Annals of Applied Statistics*, 4, 2150–2180.

Gertheiss, J. & Tutz, G. (2011): Regularization and model selection with categorial effect modifiers, *Statistica Sinica* (to appear).

Gertheiss, J. (2010): *ordPens: Selection and/or Smoothing of Ordinal Predictors*, R package version 0.1-4 (test version).

Tutz, G. & Uhlmann, L. (2011): Variable Selection for the Multinomial Logit Model with L_1Type Penalization (Preprint)

- True model on level $u = 1$:

$$y = -\mathbf{1} - \mathbf{2}x_1 + \mathbf{2}x_2 + 0x_3 + 0x_4 + 0x_5 + 0x_6 + 0x_7 + 0x_8 + \epsilon,$$

- on level $u = 2$:

$$y = +\mathbf{1} - \mathbf{4}x_1 + \mathbf{2}x_2 + \mathbf{2}x_3 + 0x_4 + 0x_5 + 0x_6 + 0x_7 + 0x_8 + \epsilon,$$

- on level $u = 3$:

$$y = +\mathbf{1} + \mathbf{2}x_1 + \mathbf{2}x_2 + \mathbf{2}x_3 - \mathbf{4}x_4 + 0x_5 + 0x_6 + 0x_7 + 0x_8 + \epsilon,$$

- on level $u = 4$:

$$y = -\mathbf{1} + \mathbf{1}x_1 + \mathbf{2}x_2 + \mathbf{3}x_3 - \mathbf{4}x_4 - \mathbf{2}x_5 + 0x_6 + 0x_7 + 0x_8 + \epsilon.$$

- Data: balanced design with respect to u, training set size $n = 400$, independent test set ($n = 1200$), $x_j \sim U[0, 1]$ (iid), $\epsilon \sim N(0, 2)$ (iid), 100 simulation runs.

Errors of Parameter Estimates and Prediction:

- Empirical **MSE** of parameter estimates.
- **Prediction Accuracy**: Empirical sum of squared test set errors.

Variable Selection and Fusion Performance:

- **Sensitivity**:
  - Variable Selection: Proportion of relevant variables which are selected.
  - Fusion / Identifying Differences: Proportion of relevant differences between coefficients which are set to non-zero.
- **Specificity**:
  - Variable Selection: Proportion of noise variables which are not selected.
  - Fusion / Identifying Differences: Proportion of zero differences which are set to zero.

We compare:

- ordinary least squares (ols) estimation,
- L_1-regularization standard/adaptive version with fixed $\psi = 0.5$ or flexible ψ,
- forward selection based on AIC/BIC.

method	MSE	MSEP
ols	11.380 (.380)	2.219 (.011)
stdrd, fixed ψ	7.500 (.240)	2.163 (.010)
stdrd, flex. ψ	8.183 (.455)	2.173 (.010)
adapt, fixed ψ	6.920 (.334)	2.149 (.010)
adapt, flex. ψ	7.091 (.334)	2.151 (.010)
forward select, AIC	9.755 (.414)	2.191 (.011)
forward select, BIC	10.856 (.698)	2.215 (.016)

Variable selection and fusion performance