Modeling high-dimensional categorical data using nonconvex fusion penalties

Jieun Shin

December 9, 2021

Table of Contents

1. Introduction

1. Introduction

- Objective: modeling for estimation in high-dimensional linear models with nominal categorical data.
- Consider the following model relating response vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$ to categorical predictors $X_{ij} \in \{1, \ldots, K_j\}, j = 1, \ldots, p$:

$$Y_i = \mu^0 + \sum_{j=1}^{p} \sum_{k=1}^{K_j} \theta_{jk}^0 \mathbb{1}\{X_{ij} = k\} + \varepsilon_i, \tag{1}$$

  where ε_i are independent zero mean random errors, μ^0 is a global intercept and θ_{jk}^0 is the contribution to the response of the k-th level of the j-th predictor.
- (1) is called ANOVA model.

Motivation: Model for categorical data

- CART (greedily select), random forest (difficulty in interpretation).
- Penalty-based method (ex. LASSO)
- Penalty-based method for categorical data
  - Collapsing and Shrinkage ANOVA (CAS-ANOVA) (Bondell and Reich, 2009)
  - range penalty (Oellker et al., 2015)

Motivation: CAS-ANOVA

- CAS-ANOVA is proposed to estimate the $\boldsymbol{\theta}^0 = (\theta^0_{ij})_{i=1,\dots,p,k=1,\dots,K_j}$ and μ_0:

$$\mathcal{L}(\mu, \boldsymbol{\theta}) := \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - \mu - \sum_{j=1}^{p} \sum_{k=1}^{K_j} \theta^0_{jk} \mathbb{1}\{X_{jk} = k\} \right)^2 \tag{2}$$

  with a penalty of the form

$$\sum_{i=1}^{p} \sum_{k=2}^{K_j} \sum_{l=1}^{k-1} w_{j,kl} |\theta_{jk} - \theta_{jl}|. \tag{3}$$

- The weights $w_{j,kl}$ in (3) can be used to balance the effects of having certain levels of categories more prevalent than others in the data.
- CAS-ANOVA penalty is an 'all-pairs' version of the fused Lasso.
- But, 'all-pairs' penalty have an undesirable preference for unequal group size.

Motivation: range penalty

- range penalty is proposed to estimate the $\boldsymbol{\theta}_j = (\theta_{ij})_{k=1}^{K_j}$:

$$\sum_{j=1}^{K_j-1} |\theta_{j(k+1)} - \theta_{j(k)}| = \max_k \theta_{jk} - \min_k \theta_{jk}, \tag{4}$$

  where $\theta_{j(k)}$ is the k-th smallest entry in $\boldsymbol{\theta}_j$.
- This shrinks the largest and smallest of the estimated coefficients together.
- Because the remaining coefficients lying in between these are unpenalised and so no grouping of the estimates is encouraged.

SCOPE (Sparse Concave Ordering & Penalisation Estimator)

- The proposed penalty, called as SCOPE:

$$\sum_{j=1}^{p} \sum_{k=1}^{K_j - 1} \rho_j(\theta_{j(k+1)} - \theta_{j(k)}) \tag{5}$$

  for concave (and nonconvex) non-decreasing penalty functions ρ_j.

- The nonconvex penalty ρ_j is necessary for shrinkage to sparse solutions to occur (proposition 1).

2. SCOPE methodology

- The goal is to estimate parameters $(\mu^0, \boldsymbol{\theta}^0)$ in model (1):

$$Y_i = \mu^0 + \sum_{j=1}^{p} \sum_{k=1}^{K_j} \theta_{jk}^0 \mathbb{1}\{X_{ij} = k\} + \varepsilon_i,$$

- For any $\boldsymbol{\theta} = (\boldsymbol{\theta}_j)_{j=1}^p \in \mathbb{R}^{K_1} \times \cdots \times \mathbb{R}^{K_p}$, define $\boldsymbol{\theta}_j := (\theta_{jk})_{k=1}^{K_j} \in \mathbb{R}^{K_j}$.
- We impose the following to ensure that $\boldsymbol{\theta}^0$ is identifiable:

$$g_j(\boldsymbol{\theta}_j^0) = 0, \text{ where } g_j(\boldsymbol{\theta}_j) = \sum_{k=1}^{K-j} n_{jk}\theta_{jk} \text{ and } n_{jk} = \sum_{i=1}^{n} \mathbb{1}\{X_{ij} = k\}, \qquad (6)$$

for all $j = 1, \ldots, p$.

2. SCOPE methodology

- Let $\boldsymbol{\Theta}_j = \{\boldsymbol{\theta}_j \in \mathbb{R}^{K_j} : g_j(\boldsymbol{\theta}_j) = 0\}$, and let $\boldsymbol{\Theta} = \boldsymbol{\Theta}_1 \times \cdots \times \boldsymbol{\Theta}_p$.
- We construct estimators by minimising over $\mu \in \mathbb{R}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ an objective function of the form

$$\tilde{Q}(\mu, \boldsymbol{\theta}) = \mathcal{L}(\mu, \boldsymbol{\theta}) + \sum_{j=1}^{p} \sum_{k=1}^{K_j - 1} \rho_j(\theta_{j(k+1)} - \theta_{j(k)}), \tag{7}$$

  where $\mathcal{L}$ is the least squares loss function and $\theta_{j(1)} \leq \cdots \leq \theta_{j(K_j)}$ are the order statistics of $\boldsymbol{\theta}_j$.
- Different penalty functions ρ_j are given for each predictor variable in order to help balance the effects of varying numbers of levels K_j.
- The identifiability constraint that $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ ensures that the estimated intercept $\hat{\mu} := \operatorname{argmin}_{\mu} \tilde{Q}(\mu, \boldsymbol{\theta})$ satisfies $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} Y_i$.

Identifiability constraint

- The identifiability constraint has a bearing on the fitted values of regularisation least squares regression.
- Consider the univariate setting with $p = 1$ and the corner point constraint $\theta_1 = 0$.
- Denote $\bar{Y}_k = \frac{1}{n_k} \sum_{i=1}^{n} Y_i \mathbb{1}\{X_i = k\}$.
- Then, the fitted value for an observation with level 1 would simply be the average $\bar{Y}_1$, conciding with that of unpenalised least squares.
- However, that of other level $k \geq 2$ would be subject to regularisation and in general be different to $\bar{Y}_k$.
- This constraint takes into account the prevalence of levels, so the fitted values corresponding to more prevalent levels effectively undergo less regularisation.

- As the estimated intercept μ does not depend on the tuning parameter, we define an objective function:

$$Q(\boldsymbol{\theta}) = \frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - \hat{\mu} - \sum_{j=1}^{p}\sum_{k=1}^{K_j}\theta_{jk}\mathbb{1}\{X_{ij}=k\}\right)^2 \tag{8}$$

$$+\sum_{j=1}^{p}\sum_{k=1}^{K_j-1}\rho_j(\theta_{j(k+1)} - \theta_{j(k)}). \tag{9}$$

- The regularisers $\rho_j : [0,\infty) \to [0,\infty)$ in (9) is taken to be concave (and nonconvex).
- We base the penalties ρ_j on the minimax concave penalty (MCP):

$$\rho(x) = \rho_{\gamma,\lambda}(x) = \int_0^x \lambda\left(1 - \frac{t}{\gamma\lambda}\right)dt,$$

where $(u)_+ = u\mathbb{1}\{u \geq 0\}$ with tuning parameter γ and λ.

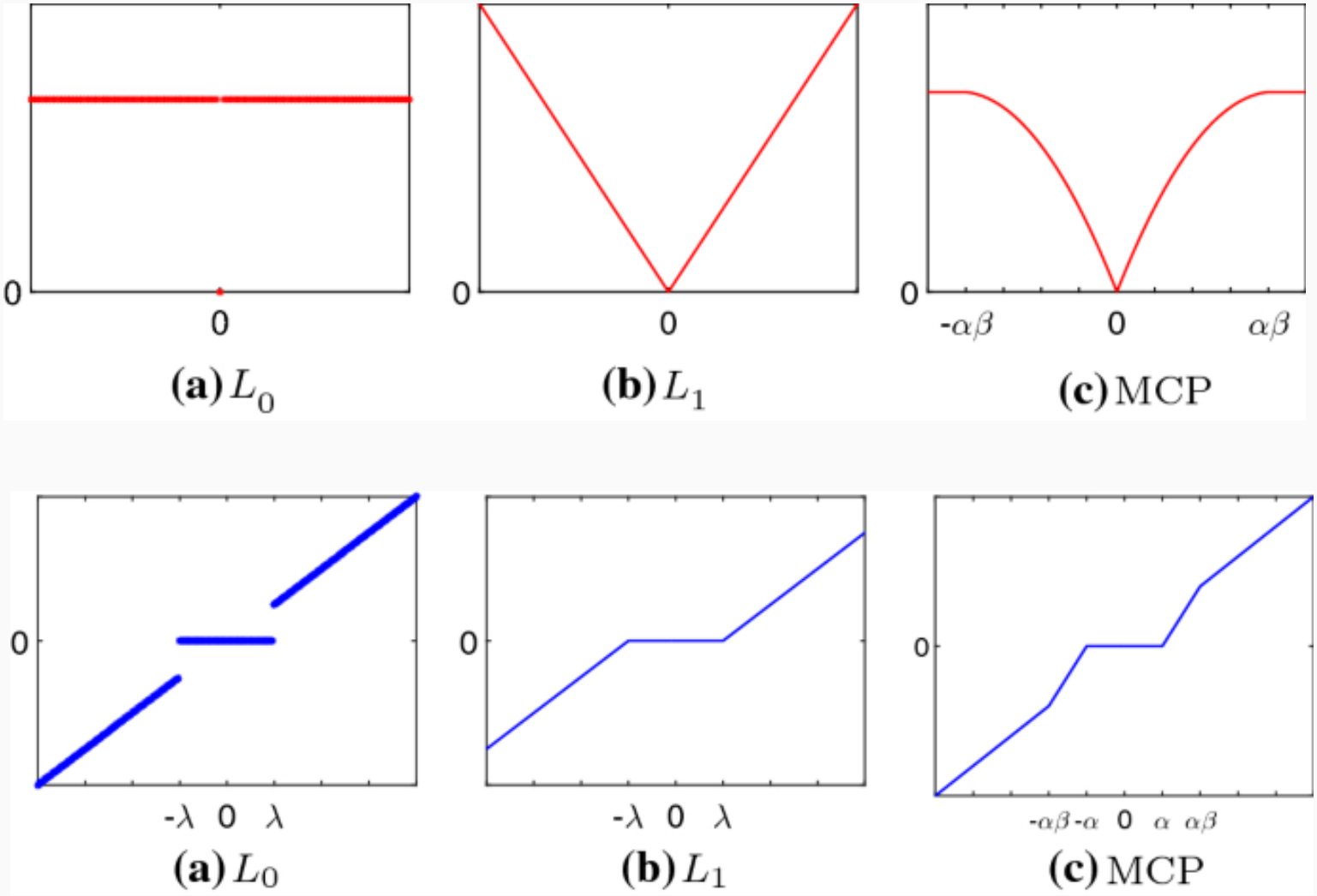

Figure 1: MCP penalty.

Proposition 1

Consider the univariate case with $p = 1$. Suppose the subaverages $(\bar{Y}_k)_{k=1}^K$ are all distinct, and that ρ is convex. Then any minimizer $\hat{\boldsymbol{\theta}}$ of Q has $\hat{\theta}_k \neq \hat{\theta}_l$ for all $k \neq l$ such that

$$\hat{\theta}_{(1)} < \bar{Y}_k - \hat{\mu} < \hat{\theta}_{(K)} \text{ or } \hat{\theta}_{(1)} < \bar{Y}_l - \hat{\mu} < \hat{\theta}_{(K)}.$$

- Proposition 1 discuss that a nonconvex penalty is necessary for fusion to occur.

3. Computation

3.1 Univariate model

- In this section, consider the univariate case, $p = 1$.
- We can be rewrite the least square loss in the following way:

$$\begin{aligned}
&\frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - \hat{\mu} - \sum_{k=1}^{K}\theta_k \mathbb{1}\{X_i = k\}\right) \\
&= \frac{1}{2n}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathbb{1}\{X_i = k\}(Y_i - \hat{\mu} - \theta_k)^2 \\
&= \frac{1}{2n}\sum_{k=1}^{K} w_k(\bar{Y}_k - \hat{\mu} - \theta_k)^2 + \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{1}\{X_i = k\}(Y_i - \bar{Y}_k)^2,
\end{aligned}$$

  where $w_k = n_k/n$.
- Then, the optimisation problem can be written equivalently as

$$\hat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \frac{1}{2}\sum_{k=1}^{K} w_k(\bar{Y}_k - \hat{\mu} - \theta_k)^2 + \sum_{k=1}^{K-1}\rho(\theta_{(k+1)} - \theta_{(k)}). \qquad (10)$$

3.1 Univariate model

- Challenging aspects of the optimisation problem (10):
  1. the presence of the nonconvex ρ.
  2. the presence of the ordered statistics $\theta_{(1)} \leq \cdots \leq \theta_{(K)}$.
- the second is hold by proposition 2, whenever ρ is a concave function.

Proposition 2

Consider the univariate optimisation (10) with ρ any concave function such that a minimiser $\boldsymbol{\theta}$ exists. For k, l, If we have $\bar{Y}_k > \bar{Y}_l$, then $\hat{\theta}_k \geq \hat{\theta}_l$.

- This observation substantially simplifies the optimisation: after re-indexing such that $\bar{Y}_1 \leq \cdots \leq \bar{Y}_K$, we re-express (10) as,

$$\hat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}:\theta_1 \leq \cdots \leq \theta_K}{\operatorname{argmin}} \frac{1}{2} \sum_{k=1}^{K} w_k (\bar{Y}_k - \hat{\mu} - \theta_k)^2 + \sum_{k=1}^{K-1} \rho(\theta_{(k+1)} - \theta_{(k)}). \tag{11}$$

- Then, we denote the intermediate functions to structure the algorithm:

$$f_1(\theta_1) = \frac{1}{2} w_1 (\bar{Y}_1 - \hat{\mu} - \theta_1)^2, \tag{12}$$

$$f_k(\theta_k) = \min_{\theta_{k-1}:\theta_{k-1} \leq \theta_k} \{f_{k-1}(\theta_{k-1}) + \rho(\theta_k - \theta_{k-1})\} + \frac{1}{2} w_k (\bar{Y}_k - \hat{\mu} - \theta_k)^2, \tag{13}$$

$$b_k(\theta_k) = \underset{\theta_{k-1}:\theta_{k-1} \leq \theta_k}{\operatorname{sargmin}} \{f_{k-1}(\theta_{k-1}) + \rho(\theta_k - \theta_{k-1})\}, \tag{14}$$

for $k = 2, \ldots, K$, sargmin means to the smallest minimiser in the case that it is not unique.

- Since Proposition 3 holds that this will be unique, we will assume uniqueness in optimization problem.

Proposition 3

The set of $(\bar{Y}_k)_{k=1}^K$ that yields distinct solutions to (10) has Lebesgue measure zero as a subset of $\mathbb{R}^K$.

- $\hat{\theta}_K$ is the minimiser of the univariate objective function f_K: for $k \geq 2$,

$$f_k(\theta_k) = \min_{(\theta_1,\ldots,\theta_{k-1})^T:\theta_1\leq\cdots\theta_{k-1}\leq\theta_k} \left\{ \frac{1}{2}\sum_{l=1}^{k} w_l(\bar{Y}_l - \hat{\mu} - \theta_l)^2 + \sum_{l=1}^{k-1}\rho(\theta_{l+1}-\theta_l) \right\}.$$

- And we have $\hat{\theta}_k = b_{k+1}(\hat{\theta}_{k+1})$ for $k = K-1,\ldots,1$, generally.
- Given f_k and b_k for $k \geq 2$, we can iteratively compute $\hat{\theta}_K, \hat{\theta}_{K-1}, \ldots, \hat{\theta}_1$.

Computation of f_k and $b_2, \ldots, b_K$

Lemma 4

For each k,

1. f_k is continuous, coercive and piecewise quadratic with finitely many pieces;
2. b_k is piecewise linear with finitely many pieces;
3. for each $\theta_{k+1} \in \mathbb{R}$, if a minimiser $\tilde{\theta}_k = \tilde{\theta}_k(\theta_{k+1})$ of $\theta_k \mapsto f_k(\theta_k) + \rho(\theta_{k+1} - \theta_k)$ over $(-\infty, \theta_{k+1}]$ satisfies $\tilde{\theta}_k < \theta_{k+1}$, then f_k must be differentiable at $\tilde{\theta}_k$.

- Properties 1 and 2 above permit exact representation of f_k and b_k with finitely many quantities.
- Then, the key task is to form the collection of intervals and corresponding coefficients of quadratic functions for

$$g_k(\theta_{k+1}) := \min_{\theta_k : \theta_k \leq \theta_{k+1}} \{f_k(\theta_k) + \rho(\theta_{k+1} - \theta_k)\}$$

  given a similar piecewise quadratic representation of f_k.
- And also the same for the linear functions composing b_k.

Computation of f_k and $b_2, \ldots, b_K$ (Cont.)

- To use property 3 above, in computing $g_k(\theta_{k+1})$ we can separately search for minimisers at stationary points in $(-\infty, \theta_{k+1})$ and compare the corresponding function values with $f_k(\theta_{k+1})$.
- Need only consider potential minimisers at points of differentiability will simplify things.

Computation of f_k and $b_2, \ldots, b_K$ (Cont.)

- Let $I_{k,1}, \ldots, I_{k,m(k)}$ are intervals that partition $\mathbb{R}$ (closed on the left).
- Let $q_{k,1}, \ldots, q_{k,m(k)}$ are corresponding quadratic functions such that $f_k(\theta_k) = q_{k,r}(\theta_k)$ for $\theta_k \in I_{k,r}$.
- Let

$$\tilde{q}_{k,r}(\theta_k) = \begin{cases} q_{k,r}(\theta_k), & \text{if } \theta_k \in I_{k,r}, \\ \infty, & \text{otherwise.} \end{cases}$$

- we express f_k as $f_k(\theta_k) = \min_r \tilde{q}_{k,r}(\theta_k)$.
- Let

$$\tilde{\rho}_1(x) := -\gamma\lambda^2\{1 - x/(\lambda)\}^2 + \gamma\lambda^2/2 \text{ if } 0 \leq x < \gamma\lambda \text{ and } \infty \text{ otherwise,}$$
$$\tilde{\rho}_2(x) := \gamma\lambda^2/2 \text{ if } x \geq \gamma\lambda \text{ and } \infty \text{ otherwise,}$$

Then $\rho(x) = \min_t \tilde{\rho}_t(x)$ for $x \geq 0$.

Computation of f_k and $b_2, \ldots, b_K$ (Cont.)

- Let D_k be the set of points at which f_k is differentiable.
- We then have using Lemma 4 (3) that

$$g_k(\theta_{k+1}) = \min_{\theta_k : \theta_k \leq \theta_{k+1}} \{\min_r \tilde{q}_{k,r}(\theta_k) + \min_t \tilde{\rho}_t(\theta_{k+1} - \theta_k)\} \tag{15}$$

$$= \min[\tilde{\min}_{\theta_k \in D_k : \theta_k < \theta_{k+1}} \min_{r,t} \{\tilde{q}_{k,r}(\theta_k) + \tilde{\rho}_t(\theta_{k+1} - \theta_k)\}, f_k(\theta_{k+1})] \tag{16}$$

$$= \min[\min_{r,t} \tilde{\min}_{\theta_k \in D_k : \theta_k < \theta_{k+1}} \{\tilde{q}_{k,r}(\theta_k) + \tilde{\rho}_t(\theta_{k+1} - \theta_k)\}, f_k(\theta_{k+1})], \tag{17}$$

where $\tilde{\min}$ denotes the minimum if it exists and ∞ otherwise.

Algorithm

- Let $J_{k,1}, \ldots, J_{k,n(k)}$ as intervals and $p_{k,1}, \ldots, p_{k,n(k)}$ as corresponding quadratic function.
- Define the active set at x as $A(x) = \{r : x \in I_{k,r}\}$, for each $x \in \mathbb{R}$.
- the endpoints of the intervals $J_{k,r}$ are the points where the active set.
- Let $r(x)$ be the index such that $g_k(x) = p_{k,r(x)}(x)$.
- Consider for each $r \in A(x) \setminus \{r(x)\}$, the horizontal coordinate x' of the first intersection beyond x (if it exists) between $p_{k,r}$ and $p_{k,r(x)}$.
- Denote $N(x)$ as the collection of all such tuples (x', r), the intersection set at x.

Algorithm

Initialize: $x_{old}, r_{old} = r(x_{old})$ and $x_{cur} = x_{old}$.

1. Given $r(x_{cur})$, compute $N(x_{cur})$ and set $(x_{int}, r_{int}) = \text{argmin}_{(x,r)\in N(x_{cur})} x$.
2. If there are no changes in the active set between x_{cur} and x_{int}, we have found the next knot point at x_{int} and $r_{int} = r(x_{int})$.
3. If instead the active set changes, move x_{cur} to the leftmost change point. We have that $r(x) = r_{old}$ for $x \in [x_{old}, x_{cur})$. To determine if $r(x)$ changes at x_{cur}, we check if
   - 3.1 r_{old} leaves the active set at x_{cur}, so $r_{old} \notin A(x_{cur})$, or
   - 3.2 some r_{new} enters the active set at x_{cur} and beats r_{old}, so $r_{new} \in A(x_{cur}) \setminus A(x_{old})$ and $p_{k,r_{new}}(x_{cur} + \epsilon) < p_{k,r_{old}}(x_{cur} + \epsilon)$ for $\epsilon > 0$ sufficiently small.

If either hold x_{cur} is a knot and $r(x_{cur})$ may be computed via $r(x_{cur}) = \text{argmin}_{r\in A(x_{cur})} p_{k,r}(x_{cur})$.

If neither hold, we conclude that $r(x_{cur}) = r_{old}$ and go to step 1 once more.

Computation of f_k and $b_2, \ldots, b_K$ (Cont.)

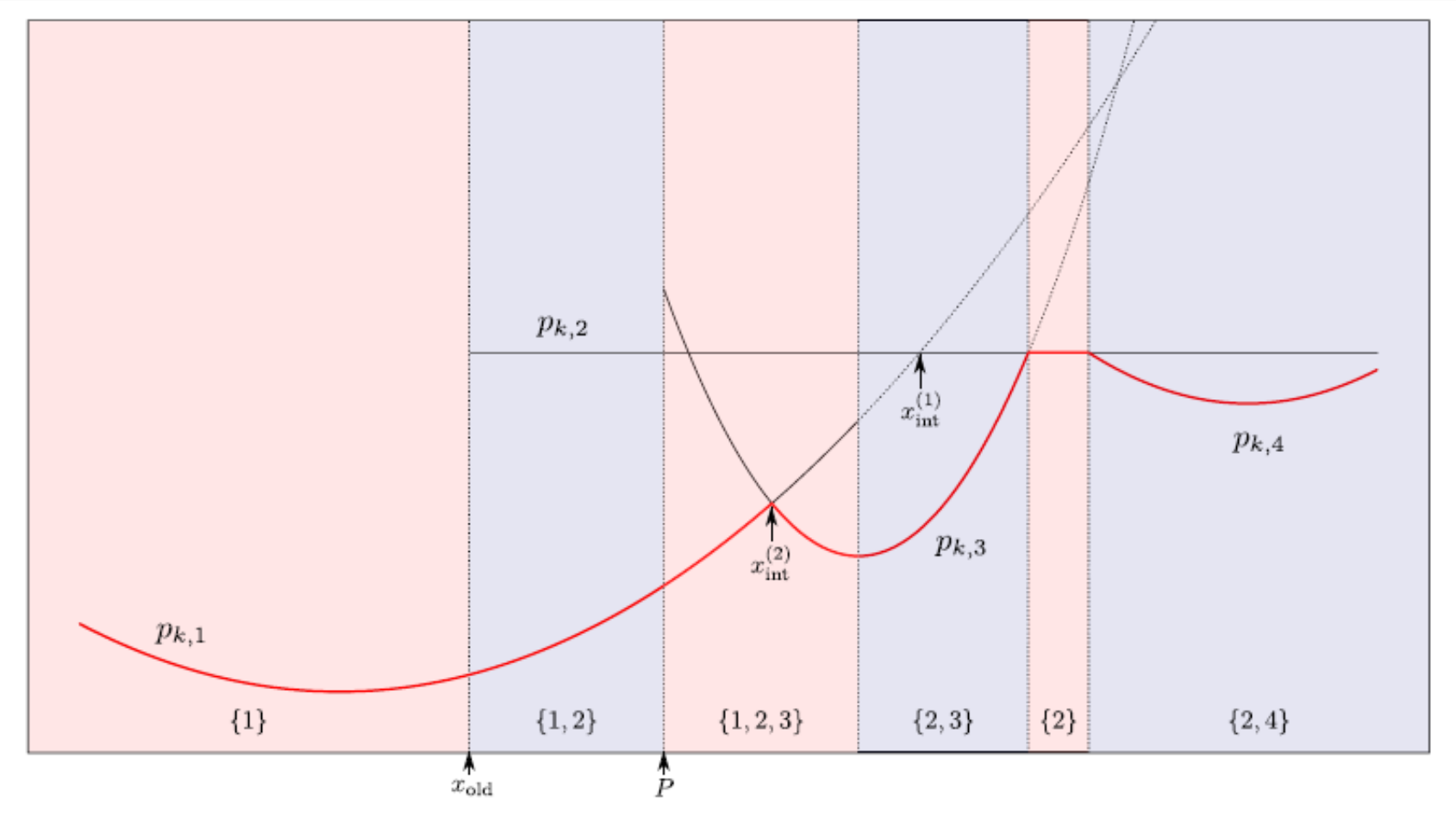

Figure 2: Illustration of the optimisation problem and our algorithm.

3.2 Multivariate model

- To solve multivariate problem, we minimise the objective $Q(\boldsymbol{\theta})$ written by

$$Q(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - \hat{\mu} - \sum_{j=1}^{p} \sum_{k=1}^{K_j} \theta_{jk} \mathbb{1}\{X_{ij} = k\} \right)^2 + \sum_{j=1}^{p} \sum_{k=1}^{K_j - 1} \rho_j(\theta_{j(k+1)} - \theta_{j(k)})$$

  for using block coordinate descent.
- This has been shown empirically to be a successful strategy for minimising objectives for high-dimensional regression with nonconvex penalties such as the MCP.
- We iteratively minimise the objective $Q(\boldsymbol{\theta})$ over $\boldsymbol{\theta}_j := (\theta_{jk})_{k=1}^{K_j} \in \boldsymbol{\Theta}_j$ keeping all other parameters fixed.

Algorithm

For give (γ, λ), and initial estimate $\hat{\theta}^{(0)} \in \boldsymbol{\Theta}$, repeat the following:

1. Initialise $m = 1$, and set for $i = 1, \ldots, n$

$$R_i = Y_i - \hat{\mu} - \sum_{l=1}^{p} \sum_{k=1}^{K_j} \hat{\theta}_{lk}^{(m-1)} \mathbb{1}\{X_{ij} = k\}.$$

2. For $j = 1, \ldots, p$, compute

$$R_i^{(j)} = R_i + \sum_{k=1}^{K_j} \hat{\theta}_{jk}^{(m-1)} \mathbb{1}\{X_{ij} = k\} \quad \text{for each } i,$$

$$\hat{\boldsymbol{\theta}}_j^{(m)} = \underset{\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j}{\operatorname{argmin}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(R_i^{(j)} - \sum_{k=1}^{K_j} \theta_{jk} \mathbb{1}\{X_{ij} = k\} \right)^2 + \left(\sum_{k=1}^{K_j - 1} \rho_j(\theta_{j(k+1)} - \theta_{j(k)}) \right) \right\}$$

$$R_i = R_i^{(j)} - \sum_{k=1}^{K_j} \hat{\theta}_{jk}^{(m)} \mathbb{1}\{X_{ij} = k\} \quad \text{for each } i.$$

3. Increment $m \to m + 1$.

3.2 Multivariate model

- Finally, we define a blockwise optimum of Q to be any $\hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$, such that for each $j = 1, \ldots, p$,

$$\hat{\boldsymbol{\theta}}_j \in \underset{\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j}{\operatorname{argmin}} Q(\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_{j-1}, \boldsymbol{\theta}_j, \hat{\boldsymbol{\theta}}_{j+1}, \ldots, \hat{\boldsymbol{\theta}}_p). \tag{18}$$

- The tuning parameter γ and λ are selected by cross-validation over a grid of (γ, λ).

4. Numerical experiments

Simulation setting

- The way to randomly generate simulation data according to the model (1) with the covariates X_{ij}.

1. draw $(W_{ij})_{j=1}^{p}$ from $N_p(\mathbf{0}, \Sigma)$ where the covariance matrix Σ had ones on the diagonal.
2. The off-diagonal elements of Σ were chosen such that $U_{ij} := \Phi^{-1}(W_{ij})$ had $\text{corr}(U_{ij}, U_{ik}) = \rho$ for $j \neq k$.
3. The marginally uniform U_{ij} were then quantised this to give $X_{ij} = \lfloor 24U_{ij} \rfloor$, so the number of levels $K_j = 24$.

Simulation setting

- The error ϵ_i were independently distributed as $N(0, \sigma^2)$.
- Model performance is measured using mean squared prediction error (MSPE):

$$\text{MSPE} := \mathbb{E}_x\{g(x) - \hat{g}(x)\}^2,$$

  where g is the true regression function, $\hat{g}$ an estimate.
- 10^5 new observations generated as the training data and the whole process was repeated 500 times.
- various settings of $(n, p, \rho, \boldsymbol{\theta}^0, \sigma^2)$ below with low- and high-dimensional scenarios.

Low-dimensional experiments

Three setting is given below.

1. $\boldsymbol{\theta}_j^0 = (\overbrace{-3,\ldots,-3}^{10\text{ times}}, \overbrace{0,\ldots,0}^{4\text{ times}}, \overbrace{3,\ldots,3}^{10\text{ times}})$ for $j = 1, 2, 3$ and $\boldsymbol{\theta}_j^0 = 0$ otherwise; $\rho = 0$
2. $\boldsymbol{\theta}_j^0 = (\overbrace{-3,\ldots,-3}^{8\text{ times}}, \overbrace{0,\ldots,0}^{8\text{ times}}, \overbrace{3,\ldots,3}^{8\text{ times}})$ for $j = 1, 2, 3$ and $\boldsymbol{\theta}_j^0 = 0$ otherwise; $\rho = 0$
3. As setting 1, but with $\rho = 0.8$,

with $n = 500, p = 10$ and noise variance $\sigma^2 = 1, 6.25, 25, 100$.

Low-dimensional experiments

	Setting 1			
σ^2:	**1**	**6.25**	**25**	**100**
SNR:	**4.7**	**1.9**	**0.95**	**0.47**
SCOPE-8	**0.014** (0.0)	0.450 (0.5)	4.571 (1.0)	12.936 (2.8)
SCOPE-32	0.018 (0.0)	0.878 (0.6)	4.151 (0.9)	**12.356** (2.1)
SCOPE-CV	0.015 (0.0)	**0.407** (0.4)	**4.120** (0.9)	12.513 (2.5)
Linear regression	0.851 (0.1)	5.317 (0.7)	21.503 (2.7)	86.745 (10.7)
Oracle least squares	0.014 (0.0)	0.091 (0.1)	0.333 (0.2)	1.405 (0.8)
CAS-ANOVA	0.617 (0.3)	1.602 (0.3)	5.448 (1.0)	14.814 (2.2)
Adaptive CAS-ANOVA	0.135 (0.1)	0.880 (0.4)	5.076 (1.2)	22.896 (4.7)
DMR	**0.014** (0.0)	0.448 (0.4)	4.884 (1.4)	18.394 (3.6)
BEF	0.020 (0.0)	2.209 (1.1)	6.297 (1.8)	21.927 (2.3)
CART	3.844 (0.4)	5.099 (0.9)	13.219 (2.1)	22.431 (1.2)
RF	9.621 (0.5)	10.944 (0.5)	13.217 (0.7)	16.344 (0.9)

Figure 3: Mean squared prediction errors (and standard deviations thereof) of various methods on the settings described.

Low-dimensional experiments

	Setting 2			
σ^2:	**1**	**6.25**	**25**	**100**
SNR:	**4.2**	**1.7**	**0.85**	**0.42**
SCOPE-8	**0.015** (0.0)	**0.285** (0.3)	6.775 (0.9)	12.697 (2.3)
SCOPE-32	0.019 (0.0)	0.655 (0.4)	5.026 (1.0)	**12.037** (2.0)
SCOPE-CV	0.016 (0.0)	0.292 (0.3)	**5.005** (1.1)	12.444 (2.5)
Linear regression	0.869 (0.1)	5.406 (0.7)	21.216 (2.5)	85.439 (10.9)
Oracle least squares	0.014 (0.0)	0.088 (0.0)	0.336 (0.2)	1.532 (0.8)
CAS-ANOVA	1.483 (0.4)	1.626 (0.3)	5.466 (1.0)	13.421 (2.2)
Adaptive CAS-ANOVA	0.134 (0.1)	0.912 (0.3)	5.535 (1.2)	22.213 (4.9)
DMR	0.016 (0.0)	0.409 (0.4)	6.430 (1.4)	17.457 (2.1)
BEF	0.019 (0.0)	1.055 (0.9)	8.183 (2.0)	18.236 (1.5)
CART	5.530 (0.6)	7.457 (0.9)	13.280 (1.8)	18.198 (0.7)
RF	8.947 (0.3)	9.747 (0.4)	11.249 (0.6)	13.646 (0.8)

Figure 4: Mean squared prediction errors (and standard deviations thereof) of various methods on the settings described.

Low-dimensional experiments

	Setting 3			
σ^2:	**1**	**6.25**	**25**	**100**
SNR:	**7.3**	**2.9**	**1.5**	**0.73**
SCOPE-8	**0.015** (0.0)	0.967 (0.7)	5.060 (1.3)	14.555 (2.9)
SCOPE-32	0.018 (0.0)	0.713 (0.4)	3.580 (0.8)	**9.721** (1.9)
SCOPE-CV	0.022 (0.1)	0.582 (0.3)	**3.368** (0.9)	10.168 (2.6)
Linear regression	0.879 (0.1)	5.485 (0.7)	21.987 (2.7)	87.820 (11.9)
Oracle least squares	0.014 (0.0)	0.092 (0.0)	0.362 (0.2)	1.488 (1.0)
CAS-ANOVA	0.710 (0.2)	1.601 (0.3)	4.732 (0.9)	12.708 (2.1)
Adaptive CAS-ANOVA	0.189 (0.2)	0.701 (0.3)	3.705 (1.0)	16.186 (3.6)
DMR	**0.015** (0.0)	**0.553** (0.5)	5.730 (1.9)	18.594 (4.5)
BEF	0.019 (0.0)	1.716 (0.9)	8.143 (2.6)	26.923 (7.0)
CART	4.336 (0.6)	5.685 (1.0)	9.910 (1.7)	18.543 (2.2)
RF	4.039 (0.3)	5.673 (0.5)	9.157 (0.9)	13.766 (1.7)

Figure 5: Mean squared prediction errors (and standard deviations thereof) of various methods on the settings described.

High-dimensional experiments

Eight setting is given below.

1. $\boldsymbol{\theta}_j^0 = (\overbrace{-2,\ldots,-2}^{8\text{ times}}, \overbrace{0,\ldots,0}^{8\text{ times}}, \overbrace{2,\ldots,2}^{8\text{ times}})$ for $j = 1, 2, 3$ and $\boldsymbol{\theta}_j^0 = (\overbrace{-2,\ldots,-2}^{10\text{ times}}, \overbrace{0,\ldots,0}^{4\text{ times}}, \overbrace{2,\ldots,2}^{10\text{ times}})$ for $j = 4, 5, 6$ and $\boldsymbol{\theta}_j^0 = 0$ otherwise; $\rho = 0$ and $\sigma^2 = 50$.
2. As setting 1, but with $\rho = 0.5$.
3. $\boldsymbol{\theta}_j^0 = (\overbrace{-2,\ldots,-2}^{8\text{ times}}, \overbrace{0,\ldots,0}^{8\text{ times}}, \overbrace{2,\ldots,2}^{8\text{ times}})$ for $j = 1, 2, 3$ and $\boldsymbol{\theta}_j^0 = (\overbrace{-2,\ldots,-2}^{16\text{ times}}, \overbrace{3,\ldots,3}^{8\text{ times}})$ for $j = 4, 5, 6$ and $\boldsymbol{\theta}_j^0 = 0$ otherwise; $\rho = 0.5$ and $\sigma^2 = 100$.
4. $\boldsymbol{\theta}_j^0 = (\overbrace{-2,\ldots,-2}^{5\text{ times}}, \overbrace{-1,\ldots,-1}^{5\text{ times}}, \overbrace{0,\ldots,0}^{4\text{ times}}, \overbrace{1,\ldots,1}^{5\text{ times}}, \overbrace{2,\ldots,2}^{5\text{ times}})$ for $j = 1, \ldots, 5$ and $\boldsymbol{\theta}_j^0 = 0$ otherwise; $\rho = 0.5$ and $\sigma^2 = 25$.

High-dimensional experiments

5. $\boldsymbol{\theta}_j^0 = (\overbrace{-2, \ldots, -2}^{16 \text{ times}}, \overbrace{3, \ldots, 3}^{8 \text{ times}})$ for $j = 1, \ldots, 25$ and $\boldsymbol{\theta}_j^0 = 0$ otherwise; $\rho = 0$ and $\sigma^2 = 1$.
6. As setting 5, but with $\rho = 0.5$.
7. $\boldsymbol{\theta}_j^0 = (\overbrace{-2, \ldots, -2}^{4 \text{ times}}, \overbrace{0, \ldots, 0}^{12 \text{ times}}, \overbrace{2, \ldots, 2}^{8 \text{ times}})$ for $j = 1, \ldots, 10$ and $\boldsymbol{\theta}_j^0 = 0$ otherwise; $\rho = 0$ and $\sigma^2 = 25$.
8. $\boldsymbol{\theta}_j^0 = (\overbrace{-3, \ldots, -3}^{6 \text{ times}}, \overbrace{-1, \ldots, -1}^{6 \text{ times}}, \overbrace{1, \ldots, 1}^{6 \text{ times}}, \overbrace{3, \ldots, 3}^{6 \text{ times}})$ for $j = 1, \ldots, 5$ and $\boldsymbol{\theta}_j^0 = 0$ otherwise; $\rho = 0$ and $\sigma^2 = 25$.

High-dimensional experiments

Setting:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
SNR:	**0.6**	**1.0**	**1.0**	**0.64**	**12**	**36**	**0.87**	**1.0**
SCOPE-8	14.319 (2.0)	15.445 (2.9)	30.597 (5.6)	7.254 (1.2)	96.538 (25.0)	7.960 (23.2)	15.867 (1.4)	11.028 (1.6)
SCOPE-32	**14.009** (1.6)	**10.780** (1.6)	**21.841** (3.4)	7.256 (0.9)	65.344 (13.4)	0.107 (0.0)	14.867 (1.2)	11.218 (1.4)
SCOPE-CV	14.026 (1.7)	10.843 (1.8)	22.004 (3.9)	**7.191** (1.0)	**54.030** (19.2)	**0.084** (0.0)	**14.865** (1.3)	**10.941** (1.5)
Oracle LSE	5.044 (0.6)	5.130 (0.6)	2.664 (1.0)	1.09 (0.3)	0.054 (0.0)	0.055 (0.0)	1.087 (0.3)	0.799 (0.3)
DMR	18.199 (1.4)	22.627 (4.4)	42.979 (9.2)	9.645 (1.2)	139.095 (4.3)	213.691 (35.7)	19.298 (0.8)	11.737 (2.4)
CART	18.146 (0.5)	31.235 (3.6)	58.73 (6.6)	10.466 (0.3)	139.35 (2.1)	614.739 (42.8)	19.021 (0.4)	23.775 (1.5)
RF	16.181 (0.6)	16.345 (1.4)	31.561 (2.6)	9.053 (0.4)	128.618 (2.2)	264.374 (14.4)	17.224 (0.4)	19.783 (0.7)
Lasso	18.136 (0.5)	24.839 (1.3)	48.162 (2.5)	10.473 (0.4)	135.375 (5.0)	154.656 (7.8)	18.886 (0.6)	23.813 (1.6)

Figure 6: Mean squared prediction errors (and standard deviations thereof) of various methods on the settings described.

High-dimensional experiments

Setting:	1	2	3	4	5	6	7	8
SCOPE-8	0.02/0.35	0.04/0.23	0.04/0.25	0.02/0.15	0.02/0.23	0.02/0.01	0.02/0.35	0.01/0.00
SCOPE-32	0.14/0.15	0.30/0.02	0.30/0.02	0.15/0.04	0.52/0.00	0.00/0.00	0.21/0.08	0.21/0.00
SCOPE-CV	0.12/0.20	0.30/0.02	0.29/0.03	0.12/0.07	0.59/0.00	0.00/0.00	0.21/0.11	0.09/0.00
DMR	0.00/0.86	0.00/0.44	0.00/0.47	0.00/0.62	0.00/0.91	0.03/0.60	0.00/0.88	0.00/0.02
Lasso	0.01/0.88	0.00/1.00	0.00/1.00	0.01/0.83	0.00/0.98	0.00/1.00	0.00/0.91	0.00/0.90

Figure 7: (False positive rate)/(False negative rate) of linear modelling methods considered in the high-dimensional settings

4. Adult dataset analysis

- 45,222 observations based on information from the 1994 US.
- The response is binary variable, 0 is the individual earns at most $50,000 a year, and 1 otherwise.
- 93 predictive variable.

4. Adult dataset analysis

Variable	Coefficient	Levels
Intercept	−3.048	—
Age	0.027	—
Hours per week	0.029	—
Work class	0.378	Federal government, Self-employed (incorporated)
	0.058	Private
	−0.143	Local government
	−0.434	Self-employed (not incorporated), State government, Without pay
Education level	1.691	Doctorate, Professional school
	1.023	Master's
	0.646	Bachelor's
	−0.132	Associate's (academic), Associate's (vocational), Some college (non-graduate)
	−0.546	12th, High school grad
	−1.539	Preschool, 1st–4th, 5th–6th, 7th–8th, 9th, 10th, 11th
Marital status	0.059	Divorced, Married (armed forces spouse), Married (civilian spouse), Married (absent spouse), Separated, Widowed
	−0.476	Never married
Occupation	0.560	Executive/Managerial
	0.311	Professional/Specialty, Protective service, Tech support
	−0.003	Armed forces, Sales
	−0.168	Admin/Clerical, Craft/Repair
	−0.443	Machine operative/inspector, Transport
	−1.107	Farming/Fishing, Handler/Cleaner, Other service, Private house servant

Figure 8: Coefficients of SCOPE model trained on the full dataset. Here, = 100 and was selected by fivefold cross-validation (with cross-validation error of 16.82%

4. Adult dataset analysis

Relationship*	1.498	Wife
	0.332	Husband
	−1.220	Not in family
	−1.482	Unmarried, Other relative
	−2.144	Own child
Race	0.013	White
	0.008	Asian/Pacific islander, Other
	−0.182	Native-American/Inuit, Black
Sex	0.139	Male
	−0.619	Female
Native country	0.018	KH, CA, CU, ENG, FR, DE, GR, HT, HN, HK, HU, IN, IR, IE, IT, JM, JP, PH, PL, PT, PR, TW, US, YU
	−0.882	CN, CO, DO, EC, SV, GT, NL, LA, MX, NI, GU-VI-etc, PE, SCT, ZA, TH, TT, VN

Figure 9: Coefficients of SCOPE model trained on the full dataset. Here, = 100 and was selected by fivefold cross-validation (with cross-validation error of 16.82%

4. Insurance data example

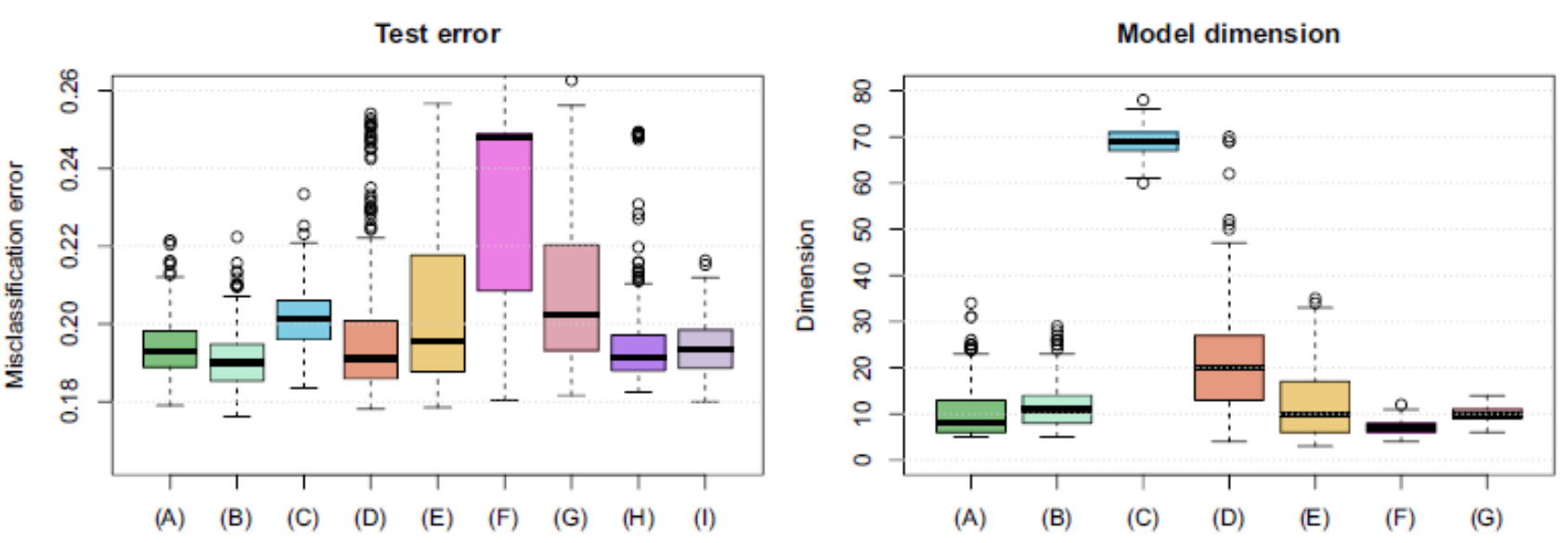

Figure 10: Prediction performance and fitted model dimension (respectively) of various methods on the Adult dataset: (A) SCOPE-100; (B) SCOPE-250; (C) Logistic regression; (D) CAS-ANOVA; (E) Adaptive CAS-ANOVA; (F) DMR; (G) BEF; (H) CART; (I) RF

4. Adult dataset analysis

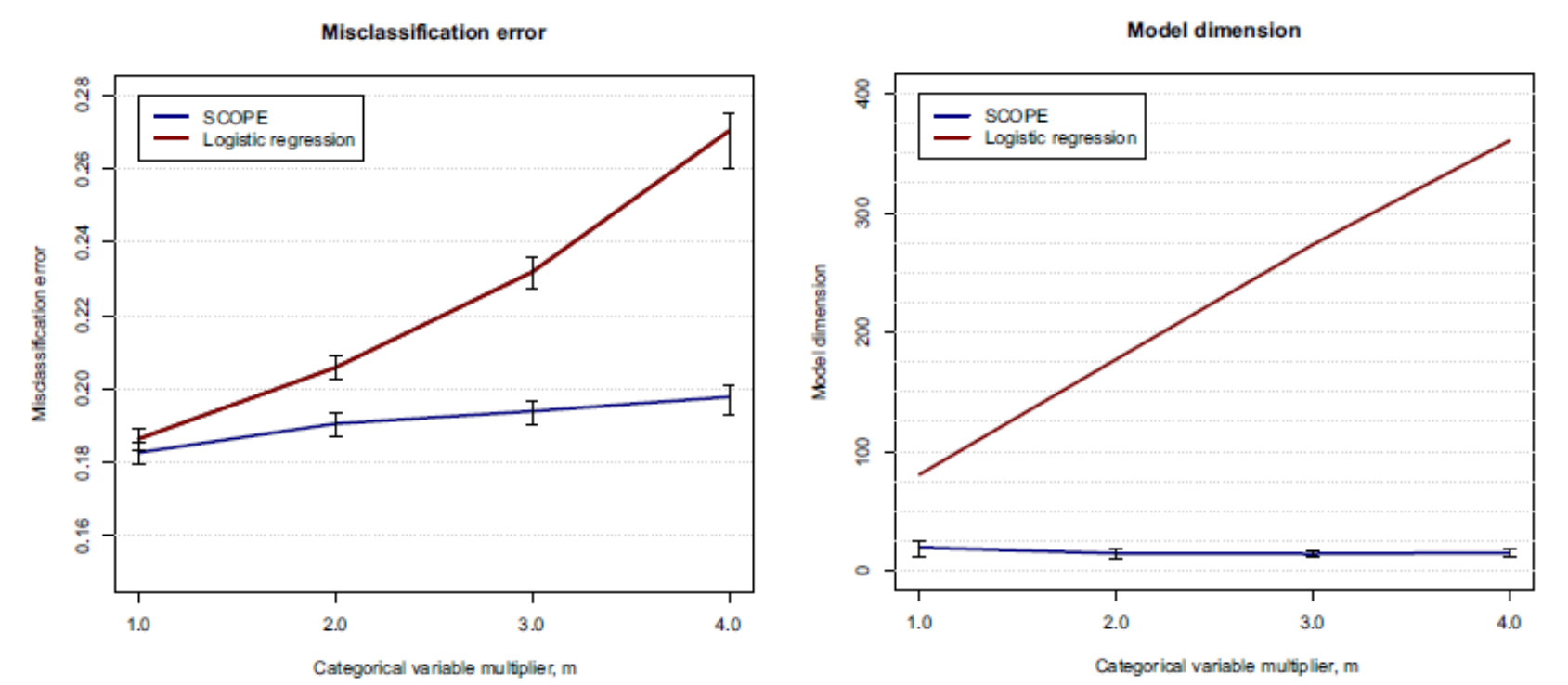

Figure 11: Misclassification error and dimensions of models fitted on a sample of the Adult dataset when levels have been artificially split m times

5. Discussion

5. Discussion

- Introducing a new penalty-based method for performing regression on categorical data.
- A penalty-based approach is attractive because it can be integrated easily with existing methods for regression with continuous data, such as the Lasso.
- The nonconvexity here is necessary in order to obtain sparse solutions, that is fusions of levels.
- Our dynamic programming algorithm can typically solve the resulting optimisation problem.