Supplement A of "Selection and Fusion of Categorical Predictors with L_0-Type Penalties"

Margret-Ruth Oelker[*,†], Wolfgang Pößnecker[†] & Gerhard Tutz[†]

The supplement is organized as follows: first, the proof of Proposition 1 is presented. Then, it is shown how pairwise fusion penalties can be represented as weighted sums of adjacent parameter differences.

1 Proof of Proposition 1

Lemma 1. *Consider the estimate* $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \cdot P(\boldsymbol{\beta})$ *of a penalized linear model with orthonormal design* $\boldsymbol{X}^T\boldsymbol{X} = \mathbb{I}_{(k+1)\times(k+1)}$ *and the general penalty* $P(\boldsymbol{\beta}) = \sum_{r\in\mathcal{I}_1, s\in\mathcal{I}_2} g(|\beta_r - \beta_s|)$*, where* $\mathcal{I}_1$*,* $\mathcal{I}_2$ *denote nonempty sets of indices, and where* $g : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ *denotes a monotonically increasing function. Then it holds that* $\sum_{r=0}^k \hat{\beta}_r = \sum_{r=0}^k \hat{\beta}_r^{ML}$ *and thus,* $\bar{\hat{\boldsymbol{\beta}}} = \bar{\boldsymbol{\beta}}^{ML}$.

Proof. Consider

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{(k+1)}} \left(\mathcal{M}(\boldsymbol{\beta}) := ||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}||_2^2 + \lambda P(\boldsymbol{\beta})\right), \tag{1}$$

for any input vector $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^{(k+1)}$, for any $\lambda \geq 0$ and for the penalty $P(\boldsymbol{\beta})$ that is defined in Lemma 1. The penalty P and thus the objective function $\mathcal{M}$ can be non-convex such that $\boldsymbol{\beta}^*$ is not unique. By definition, P and thus $\mathcal{M}$ are bounded by 0 such that $\mathcal{M}$ has a unique minimum nonetheless. The proof relies only on the uniqueness of this minimum and can be applied to all solutions of (1).

[*]Corresponding author: `margret.oelker@stat.uni-muenchen.de`
[†]Department of Statistics, Ludwig-Maximilians-Universität München, Germany

Let $m \in \mathbb{R}$ be a scalar and let $\mathbb{1}_{k+1}$ denote a vector of ones of length $k+1$. Consider the point $\boldsymbol{u} := \boldsymbol{\beta}^* - m \cdot \mathbb{1}_{k+1}$ and compare $\mathcal{M}(\boldsymbol{\beta}^*)$ with $\mathcal{M}(\boldsymbol{u})$.

First of all, note that, for any $m \in \mathbb{R}$,

$$\begin{aligned} P(\boldsymbol{u}) = P(\boldsymbol{\beta}^* - m \cdot \mathbb{1}_{k+1}) &= \sum_{r \in \mathcal{I}_1} \sum_{s \in \mathcal{I}_2} g\Big(\big|(\beta_r^* - m) - (\beta_s^* - m)\big|\Big) \\ &= \sum_{r \in \mathcal{I}_1} \sum_{s \in \mathcal{I}_2} g\big(|\beta_r^* - \beta_s^*|\big) \\ &= P(\boldsymbol{\beta}^*). \end{aligned}$$

Hence, the penalty is irrelevant for the comparison of $\mathcal{M}(\boldsymbol{\beta}^*)$ and $\mathcal{M}(\boldsymbol{u})$.

Differentiation of the L_2^2-term in $\mathcal{M}(\boldsymbol{u})$ with respect to m shows that

$$m^* = \arg\min_{m \in \mathbb{R}} ||\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}} - m \cdot \mathbb{1}_{k+1}||_2^2 = \frac{1}{k+1} \sum_{r=0}^{k} (\beta_r^* - \tilde{\beta}_r).$$

For $\boldsymbol{u}^* = \boldsymbol{\beta}^* - m^* \cdot \mathbb{1}_{k+1}$, it holds that

$$\begin{aligned} \mathcal{M}(\boldsymbol{u}^*) - \mathcal{M}(\boldsymbol{\beta}^*) &= \Big(||\boldsymbol{u}^* - \tilde{\boldsymbol{\beta}}||_2^2 + \lambda P(\boldsymbol{u}^*)\Big) - \Big(||\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}||_2^2 + \lambda P(\boldsymbol{\beta}^*)\Big) \\ &= ||\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}} - m^* \cdot \mathbb{1}_{k+1}||_2^2 - ||\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}||_2^2 \\ &\leq 0 \\ \Leftrightarrow \mathcal{M}(\boldsymbol{u}^*) &\leq \mathcal{M}(\boldsymbol{\beta}^*). \end{aligned}$$

As the the L_2^2-terms are strictly convex, $\mathcal{M}(\boldsymbol{u}^*) = \mathcal{M}(\boldsymbol{\beta}^*)$ holds if and only if $\boldsymbol{u}^* = \boldsymbol{\beta}^*$.

Hence, as $\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^{(k+1)}} \mathcal{M}(\boldsymbol{\beta})$, any $\boldsymbol{u}^* \neq \boldsymbol{\beta}^*$ is a contradiction. Thus, it holds that

$$\begin{aligned} \boldsymbol{u}^* &= \boldsymbol{\beta}^* - m^* \cdot \mathbb{1}_{k+1} \\ &= \boldsymbol{\beta}^* \\ \Leftrightarrow m^* &= \frac{1}{k+1} \sum_{r=0}^{k} (\beta_r^* - \tilde{\beta}_r) \\ &= 0. \end{aligned}$$

As $\boldsymbol{X}^T \boldsymbol{X} = \mathbb{I}_{(k+1)\times(k+1)}$, $\hat{\boldsymbol{\beta}}^{ML} = \boldsymbol{X}^T \boldsymbol{y}$.

According to Fan and Li (2001), in this case, the objective can be rewritten as

$$\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda P(\boldsymbol{\beta}) \quad = \quad \left\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{ML}\right\|_2^2+\lambda P(\boldsymbol{\beta})+\text{const.}$$

Hence, the results obtained above can be applied to the assumed orthonormal setting with $\tilde{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}^{ML}$; thus, Lemma 1 holds. □

Proposition 1. *Assume a penalized linear model with orthonormal design; that is* $\boldsymbol{X}^T\boldsymbol{X}=\mathbb{I}_{(k+1)\times(k+1)}$ *where* $\boldsymbol{X}\in\mathbb{R}^{(k+1)\times(k+1)}$ *denotes the design matrix without an intercept and where* $\mathbb{I}$ *denotes the identity matrix. Let the ML estimates be ordered* $\hat{\beta}_0^{ML}<\ldots<\hat{\beta}_k^{ML}$ *and employ penalty (2.3) with a fixed penalty parameter* λ, $\lambda\geq 0$. *Then for* j, $\hat{\beta}_j^{ML}<\bar{\beta}^{ML}$, $\bar{\beta}^{ML}=\frac{1}{k+1}\sum_{j=0}^{k}\hat{\beta}_j^{ML}$, *one obtains*

$$\hat{\beta}_j=\min\left\{\bar{\beta}^{ML},\ \max\{\hat{\beta}_l^{ML},\hat{\beta}_j^{ML}\}+\frac{(\lambda-\lambda_l)I_{(l\geq j)}}{2(l+1)}\right\},$$

where $l=\max_{l=0,\ldots,k}(\lambda_l<\lambda)$, $\lambda_l=\sum_{u=1}^{l}2u\left|\hat{\beta}_u^{ML}-\hat{\beta}_{u-1}^{ML}\right|$, *and with indicator function* I. *For* $\hat{\beta}_j^{ML}\geq\bar{\beta}^{ML}$, *one obtains analogously*

$$\hat{\beta}_j=\max\left\{\bar{\beta}^{ML},\ \min\{\hat{\beta}_l^{ML},\hat{\beta}_j^{ML}\}-\frac{(\lambda-\lambda_l)I_{(k-l\geq j)}}{2(l+1)}\right\},$$

with $\lambda_l=\sum_{u=l}^{k-1}2(k-u)\left|\hat{\beta}_{u+1}^{ML}-\hat{\beta}_u^{ML}\right|$ *and* l *as before.*

Proof. According to Fan and Li (2001), the objective and the estimate are defined by

$$\begin{aligned}\mathcal{M}_{pen}(\boldsymbol{\beta}) &= \left\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{ML}\right\|_2^2+\lambda\left\|\boldsymbol{R}\boldsymbol{\beta}\right\|_1, \\ \hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}}\mathcal{M}_{pen}(\boldsymbol{\beta}),\end{aligned} \tag{2}$$

where λ denotes the tuning parameter of the penalty, and where $\boldsymbol{R}\boldsymbol{\beta}$ with

$$\boldsymbol{R}=\begin{pmatrix}-1 & 1 & 0 & & \ldots & 0\\ 0 & -1 & 1 & 0 & & \vdots\\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots\\ \vdots & & 0 & -1 & 1 & 0\\ 0 & \ldots & & 0 & -1 & 1\end{pmatrix}\in\mathbb{R}^{k\times(k+1)},$$

builds the adjacent differences of coefficients.

As the objective (2) is convex, the Karush-Kuhn-Tucker conditions (KKT; Boyd and Vandenberghe, 2004, p. 243-244) are sufficient for a solution. The necessary background on subdifferential calculus for the following proof can be found in Hiriart-Urruty and Lemaréchal, 2001. With $\nabla\mathcal{M}_{pen}$ denoting the subdifferential or, depending on context, the gradient of $\mathcal{M}_{pen}$, each solution $\hat{\boldsymbol{\beta}}$ is characterized by the condition

$$0 \in \nabla\mathcal{M}_{pen}(\hat{\boldsymbol{\beta}}).$$

Hence, $\hat{\boldsymbol{\beta}}$ is obtained by solving the following equation for $\boldsymbol{\beta}$:

$$\begin{aligned} 0 &\in \nabla\mathcal{M}_{pen}(\boldsymbol{\beta}) = 2(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{ML}) + \lambda \cdot \nabla \left\|\boldsymbol{R}\boldsymbol{\beta}\right\|_1 \\ \Leftrightarrow \quad \hat{\boldsymbol{\beta}}^{ML} - \boldsymbol{\beta} &\in \tfrac{\lambda}{2}\nabla \left\|\boldsymbol{R}\boldsymbol{\beta}\right\|_1, \end{aligned} \tag{3}$$

In order to obtain $\hat{\beta}_j$, start with $\lambda = 0$ and increase λ gradually. For $\lambda = 0$, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{ML}$. For $\lambda > 0$, let λ_1 denote the value of λ for which the first pair of coefficients is fused. That is, for $0 < \lambda < \lambda_1$, all differences in $\boldsymbol{R}\boldsymbol{\beta}$ are unequal zero; the penalty term is differentiable:

$$[\nabla \left\|\boldsymbol{R}\boldsymbol{\beta}\right\|_1]_j = \begin{cases} \frac{\partial}{\partial\beta_j}\left(|\beta_j - \beta_{j-1}| + |\beta_{j+1} - \beta_j|\right) = 1 - 1 = 0 & \text{for } 0 < j < k, \\ \frac{\partial}{\partial\beta_j}\left(|\beta_{j+1} - \beta_j|\right) = -1 & \text{for } j = 0, \\ \frac{\partial}{\partial\beta_j}\left(|\beta_j - \beta_{j-1}|\right) = 1 & \text{for } j = k. \end{cases} \tag{4}$$

Hence, for $\lambda > 0$, a distinction of cases is helpful. As the ML estimate is assumed to be ordered and due to Lemma 1, distinguish coefficients with $\hat{\beta}_j^{ML} < \bar{\beta}^{ML}$ and with $\hat{\beta}_j^{ML} \geq \bar{\beta}^{ML}$.

- **Case 1:** β_j with $\hat{\beta}_j^{ML} < \bar{\beta}^{ML}$

  Due to (4), for $0 < \lambda \leq \lambda_1$, shrinkage only affects β_0. There is no shrinkage for $j > 0$; the first fusion of coefficients at $\lambda = \lambda_1$ must affect β_0, β_1. If the coefficients are fused, it holds that $|\beta_1 - \beta_0| = 0$. Therefore, define the subdifferential v of $|\xi|$:

$$v \begin{cases} \in [-1, 1] & \text{for } \xi = 0, \\ = \text{sign}(\xi) & \text{else wise.} \end{cases}$$

Thus, for $0 < \lambda \leq \lambda_1$,

$$\begin{aligned} [\nabla \|\boldsymbol{R}\boldsymbol{\beta}\|_1]_0 &= \frac{\partial}{\partial \beta_0} |\beta_1 - \beta_0| \\ &= -v. \end{aligned}$$

With (3), it follows that

$$\begin{aligned} \hat{\beta}_j &= \hat{\beta}_j^{ML}, \quad j > 0 \\ \hat{\beta}_0 &= \begin{cases} \hat{\beta}_0^{ML} + \frac{1}{2}\lambda & \text{for } \lambda < 2(\hat{\beta}_1^{ML} - \hat{\beta}_0^{ML}), \\ \beta_1 & \text{for } \lambda = 2(\hat{\beta}_1^{ML} - \hat{\beta}_0^{ML}). \end{cases} \end{aligned}$$

That is, the first fusion takes place for $\lambda \geq \lambda_1 = 2(\hat{\beta}_1^{ML} - \hat{\beta}_0^{ML})$ so that the estimates of the coefficients β_0, β_1 are the same; for $\lambda = \lambda_1$, it holds that $\hat{\beta}_0 = \hat{\beta}_1 = \hat{\beta}_1^{ML}$. Let λ_2 denote the value of λ for which the second pair of coefficients is fused. Consider now the case $\lambda_1 = 2(\hat{\beta}_1^{ML} - \hat{\beta}_0^{ML}) < \lambda \leq \lambda_2$, where it holds that

$$\begin{aligned} [\nabla \|\boldsymbol{R}\boldsymbol{\beta}\|_1]_1 &= \frac{\partial}{\partial \beta_1} \left| \beta_2 - \frac{\beta_0 + \beta_1}{2} \right| \\ &= -\frac{v}{2}, \\ [\nabla \|\boldsymbol{R}\boldsymbol{\beta}\|_1]_2 &= 0. \end{aligned}$$

With the same arguments as above, we obtain

$$\hat{\beta}_1 = \begin{cases} \hat{\beta}_1^{ML} + \frac{1}{4}(\lambda - \lambda_1) & \text{for } \lambda < \lambda_1 + 4(\hat{\beta}_2^{ML} - \hat{\beta}_1^{ML}), \\ \beta_2 & \text{for } \lambda = \lambda_1 + 4(\hat{\beta}_2^{ML} - \hat{\beta}_1^{ML}). \end{cases}$$

That is, the estimates of β_0, β_1, β_2 are the same for $\lambda \geq \lambda_2 = \lambda_1 + 4(\hat{\beta}_2^{ML} - \hat{\beta}_1^{ML})$; and it holds that $\hat{\beta}_0 = \hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_2^{ML}$ for $\lambda = \lambda_2$. Recursive application gives

$$\hat{\beta}_j = \min\left\{\bar{\beta}^{ML},\ \max\{\hat{\beta}_l^{ML}, \hat{\beta}_j^{ML}\} + \frac{(\lambda - \lambda_l) I_{(l \geq j)}}{2(l+1)}\right\},$$

with $l = \max_{l=0,\ldots,k}(\lambda_l < \lambda)$, $\lambda_l = \sum_{u=1}^{l} 2u \left|\hat{\beta}_u^{ML} - \hat{\beta}_{u-1}^{ML}\right|$, and with indicator function I.

- **Case 2:** β_j with $\hat{\beta}_j^{ML} \geq \bar{\beta}^{ML}$

  Analogously, one obtains

$$\hat{\beta}_j = \max\left\{\bar{\beta}^{ML},\ \min\{\hat{\beta}_l^{ML}, \hat{\beta}_j^{ML}\} - \frac{(\lambda - \lambda_l) I_{(k-l \geq j)}}{2(l+1)}\right\},$$

  with $\lambda_l = \sum_{u=l}^{k-1} 2(k-u) \left|\hat{\beta}_{u+1}^{ML} - \hat{\beta}_u^{ML}\right|$ and l as before.

Note that, with λ_{max} denoting the minimal value of λ that effects maximal penalization, we have $\hat{\beta}_j = \bar{\beta}^{ML}$ for all j for $\lambda \geq \lambda_{max}$. Due to Lemma 1, for $\lambda = \lambda_{max}$, at least two (groups of) coefficients are fused with $\hat{\beta}_j \neq \bar{\beta}^{ML}$ for $\lambda < \lambda_{max}$. □

2 Representing Pairwise Fusion Penalties as Weighted Sum of Adjacent Differences

On page 7, it says: "Assume a fixed value of the tuning parameter λ and let $\beta_{(0)}, \beta_{(1)}, \ldots, \beta_{(k)}$ denote the (arbitrary) ordering of the solution. Then a short transformation (see Supplement A) shows that $\sum_{r>s} |\beta_{(r)} - \beta_{(s)}| = \sum_{r=1}^{k} w_{(r)} |\beta_{(r)} - \beta_{(r-1)}|$, where $w_{(r)} = r(k - r + 1)$."

Proof. The ordering of the coefficients implies (for $r > s$) that

$$|\beta_{(r)} - \beta_{(s)}| = \sum_{l=s+1}^{r} |\beta_{(l)} - \beta_{(l-1)}|.$$

With

$$d_{(r)} \quad = \quad \left|\beta_{(r)} - \beta_{(r-1)}\right|,$$

one thus obtains

$$
\begin{aligned}
\sum_{r>s} |\beta_{(r)} - \beta_{(s)}| &= \sum_{s=1}^{k}\sum_{l=s}^{k}\sum_{r=s}^{l} d_{(r)}, \\
\sum_{r=1}^{k} w_{(r)} |\beta_{(r)} - \beta_{(r-1)}| &= \sum_{r=1}^{k} w_{(r)} d_{(r)}.
\end{aligned}
$$

Hence, it is to show that

$$
\begin{aligned}
\sum_{s=1}^{k}\sum_{l=s}^{k}\sum_{r=s}^{l} d_{(r)} &= \sum_{r=1}^{k} w_{(r)} d_{(r)}. \\
\sum_{s=1}^{k}\sum_{l=s}^{k}\sum_{r=s}^{l} d_{(r)} &= \underbrace{\sum_{s=1}^{k}}\quad \underbrace{\sum_{l=s}^{k}}\quad \underbrace{\sum_{r=s}^{l}}\quad d_{(r)}
\end{aligned}
$$

$$
\begin{array}{llllllllllll}
s=1 & l=1 & r=1 & d_{(1)} & & & & & & & & \\
 & l=2 & r=1,2 & d_{(1)} & + & d_{(2)} & & & & & & \\
 & l=3 & r=1,2,3 & d_{(1)} & + & d_{(2)} & + & d_{(3)} & & & & \\
 & \vdots & \vdots & \vdots & & \vdots & & \vdots & & & & \\
 & l=k & r=1,\ldots,k & d_{(1)} & + & d_{(2)} & + & d_{(3)} & + & \ldots & + & d_{(k)} \\
s=2 & l=2 & r=2 & & & d_{(2)} & & & & & & \\
 & l=3 & r=2,3 & & & d_{(2)} & + & d_{(3)} & & & & \\
 & \vdots & \vdots & & & \vdots & & \vdots & & & & \\
 & l=k & r=2,\ldots,k & & & d_{(2)} & + & d_{(3)} & + & \ldots & + & d_{(k)} \\
s=3 & l=3 & r=3 & & & & & d_{(3)} & & & & \\
 & \vdots & \vdots & & & & & \vdots & & & & \\
 & l=k & r=3,\ldots,k & & & & & d_{(3)} & + & \ldots & + & d_{(k)} \\
 & & & & & & & & & & & \\
\vdots & \vdots & \vdots & & & & & & & & & \\
s=k & l=k & r=k & & & & & & & & & d_{(k)} \\
 & & & \text{k terms} & & \text{2(k-1) terms} & & \text{3(k-2) terms} & & & & \text{k terms}
\end{array}
$$

$$
\begin{aligned}
&= k \cdot d_{(1)} + 2 \cdot (k-1) \cdot d_{(3)} + 3 \cdot (k-2) \cdot d_{(3)} + \ldots + k \cdot d_{(k)} \\
&= \sum_{r=1}^{k} r \cdot (k-r+1) \cdot d_{(r)} = \sum_{r=1}^{k} w_{(r)} d_{(r)}.
\end{aligned}
$$

If the ordering of the solution is not bijective as there are fused categories, the number of categories k has to be reduced accordingly and the procedure is the same as described above. $\square$

References

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc. 96*(456), 1348–1360.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. New York: Cambridge University Press.

Hiriart-Urruty, J.-B. and C. Lemaréchal (2001). *Fundamentals of Convex Analysis*. Berlin: Springer.