Supplementary materials for "`cmenet` – a new method for bi-level variable selection of conditional main effects"

Simon Mak and C. F. Jeff Wu

Georgia Institute of Technology

March 1, 2018

Contents

1 Proofs of technical results

1.1 Proof of Theorem 1

The proof of this requires a simple lemma on normal orthant probabilities:

Lemma 1. *(Stuart and Ord, 1994) Let $(X_1, \cdots, X_p)$ follow the equicorrelated normal distribution, with $\mathbb{E}(X_j) = 0$, $\mathbb{E}(X_j^2) = 1$ and $\mathbb{E}(X_j X_k) = \rho$ for all $j \neq k$, and let $p_m = \mathbb{P}(X_1 > 0, \cdots, X_m > 0)$. Then:*

$$p_2 = \frac{\sin^{-1}\rho}{2\pi} + \frac{1}{4} \quad \textit{and} \quad p_3 = \frac{3\sin^{-1}\rho}{4\pi} + \frac{1}{8}.$$

For the main proof, note that each row of the latent matrix $\mathbf{Z}$ is i.i.d., so it suffices to fix $n = 1$ and explore the correlation amongst the scalar ME quantities $\tilde{x}_{1,A}$ and CME quantities $\tilde{x}_{1,A|B+}$. We denote these as $\tilde{x}_A$ and $\tilde{x}_{A|B+}$ for brevity. Under the latent equicorrelated distribution $\mathcal{N}\{\mathbf{0}, \rho\mathbf{J} + (1-\rho)\mathbf{I}\}$, it is easy to show that $\mathbb{E}[\tilde{x}_A] = 0$ and $\text{Var}[\tilde{x}_A] = 1$. Moreover, the CME $\tilde{x}_{A|B+}$ can be conditionally decomposed as $\tilde{x}_{A|B+} \stackrel{d}{=} R[2p_2]$ if $\tilde{x}_B = +1$, and 0 if $\tilde{x}_B = -1$, where $R[q]$ is the Rademacher random variable taking on $+1$ w.p. $q \in [0,1]$ and -1 otherwise. From this, we get:

$$\begin{aligned}
\mu_c &\equiv \mathbb{E}[\tilde{x}_{A|B+}] = \mathbb{E}[\mathbb{E}[\tilde{x}_{A|B+}|\tilde{x}_B]] = \frac{1}{2}(4p_2 - 1),\\
\sigma_c^2 &\equiv \text{Var}[\tilde{x}_{A|B+}] = \text{Var}[\mathbb{E}[\tilde{x}_{A|B+}|\tilde{x}_B]] + \mathbb{E}[\text{Var}[\tilde{x}_{A|B+}|\tilde{x}_B]] = \frac{1}{2} - \left(\frac{\sin^{-1}\rho}{\pi}\right)^2.
\end{aligned}$$

Consider the correlation between the MEs $\tilde{x}_A$ and $\tilde{x}_B$. Note that $\tilde{x}_A\tilde{x}_B$ equals $+1$ when $\tilde{x}_A$ and $\tilde{x}_B$ have the same sign, and equals -1 otherwise. Letting $\mathbb{P}(++)$ be the probability of $(\tilde{x}_A, \tilde{x}_B) = (+1, +1)$ (with similar notation for $+-$, $-+$ and $--$), Lemma 1 then gives:

$$\text{Corr}(\tilde{x}_A, \tilde{x}_B) = [\mathbb{P}(++) + \mathbb{P}(++)] - [\mathbb{P}(+-) + \mathbb{P}(-+)] = 2p_2 - 2[1/2 - p_2] = \frac{2\sin^{-1}\rho}{\pi}.$$

Next, consider the two sibling CMEs $\tilde{x}_{A|B+}$ and $\tilde{x}_{A|C+}$. Note that $\tilde{x}_{A|B+}\tilde{x}_{A|C+}$ equals +1 when both $\tilde{x}_B = +1$ and $\tilde{x}_C = +1$, and equals 0 otherwise. It follows that:

$$\text{Corr}(\tilde{x}_{A|B+}, \tilde{x}_{A|C+}) = \frac{1}{\sigma_c^2}[\mathbb{P}(++) - \mu_c^2] = \frac{1}{\sigma_c^2}[p_2 - \mu_c^2] = \frac{1}{\sigma_c^2}\left\{-\left(\frac{\sin^{-1}\rho}{\pi}\right)^2 + \frac{\sin^{-1}\rho}{2\pi} + \frac{1}{4}\right\}.$$

The correlation for parent-child pairs can be proved in an analogous way.

Consider now the two cousin CMEs $\tilde{x}_{B|A+}$ and $\tilde{x}_{C|A+}$. Note that $\tilde{x}_{B|A+}\tilde{x}_{C|A+}$ equals +1 when $\tilde{x}_A = +1$ and $\tilde{x}_B = \tilde{x}_C$, $\tilde{x}_{B|A+}\tilde{x}_{C|A+}$ equals -1 when $\tilde{x}_A = +1$ and $\tilde{x}_B \neq \tilde{x}_C$, and equals 0 otherwise. We then have:

$$\begin{aligned}
\text{Corr}(\tilde{x}_{B|A+}, \tilde{x}_{C|A+}) &= \frac{1}{\sigma_c^2}\left[\{\mathbb{P}(+++) + \mathbb{P}(+--)\} - \{\mathbb{P}(++-) + \mathbb{P}(++-)\} - \mu_c^2\right] \\
&= \frac{1}{\sigma_c^2}\left[\{\mathbb{P}(+++) + (\mathbb{P}(--) - \mathbb{P}(---))\} - 2\{\mathbb{P}(++) - \mathbb{P}(+++)\} - \mu_c^2\right] \\
&= \frac{1}{\sigma_c^2}[2p_3 - p_2 - \mu_c^2] = \frac{1}{\sigma_c^2}\left\{-\left(\frac{\sin^{-1}\rho}{\pi}\right)^2 + \frac{\sin^{-1}\rho}{\pi}\right\}.
\end{aligned}$$

1.2 Proof of Theorem 2

Let $\mathbf{X} \in \mathbb{R}^{n\times p'}$ be the normalized model matrix consisting of all main effects and CMEs, where $p' = p + 4\binom{p}{2}$. By the strong law of large numbers, the sample covariance matrix $\mathbf{C}_n = \mathbf{X}^T\mathbf{X}/n$ converges elementwise to some matrix $\mathbf{C} \in \mathbb{R}^{p'\times p'}$ with unit diagonal entries and off-diagonal entries given in Theorem 1. Consider the following block partition of $\mathbf{C} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix}$, where $\mathbf{C}_{11}$ is the block for the active set $\mathcal{A}$, and $\mathbf{C}_{22}$ the block for the remaining variables. Zhao and Yu (2006) proved that the LASSO is sign-selection consistent only when the (weak) *irrepresentability condition* holds: $\forall\boldsymbol{\zeta} \in \{-1,+1\}^{p'}$, $|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\boldsymbol{\zeta}| < \mathbf{1}$ (this is a slight simplification of the original condition under the current i.i.d. setting). Hence, sign-selection inconsistency can be proven if $\exists\boldsymbol{\zeta} \in \{-1,+1\}^{p'}$ and an inactive effect

j satisfying:

$$|\mathbf{C}_{21,j}\mathbf{C}_{11}^{-1}\boldsymbol{\zeta}| \geq 1, \quad \text{where} \quad \mathbf{C}_{21,j} \text{ is the row corresponding to effect } j. \tag{1}$$

Consider first a model with only $q \geq 3$ active siblings of the form $A|B+$, $A|C-$, ..., $A|R-$. Using the same principles as in Theorem 1, $\mathbf{C}_{11}$ can be shown to be a $q \times q$ matrix with unit diagonal, $[(1/2 - p_2) - \mu_c^2]/\sigma_c^2$ for off-diagonal entries in the first row and column, and $\psi_{sib}(\rho)$ for all other off-diagonal entries [1]. Letting A be the inactive effect, we have $\mathbf{C}_{21,A} = \psi_{pc}(\rho)\mathbf{1}_q^T$, and letting $\boldsymbol{\zeta} = \mathbf{1}_q$, it follows that $|\mathbf{C}_{21,A}\mathbf{C}_{11}^{-1}\boldsymbol{\zeta}| \geq 1$ for $\rho \geq 0$. By (1), part (a) is proven.

Next, consider a model with only $q = 2$ active main effects, say, A and $-B$. From Theorem 1, $\mathbf{C}_{11}$ is a $q \times q$ matrix with unit diagonal and $-\psi_{me}(\rho)$ on the off-diagonals. Let $A|B-$ be the inactive effect, so $\mathbf{C}_{21,A|B-} = (\psi_{pc}(\rho), \tilde{\psi}(\rho))$. Taking $\boldsymbol{\zeta} = (1,1)^T$, $|\mathbf{C}_{21,A|B-}\mathbf{C}_{11}^{-1}\boldsymbol{\zeta}| \geq 1$ for $\rho \geq 0.27$, thereby proving selection inconsistency.

Lastly, consider a model with only $q \geq 6$ active cousins of the form $B|A+$, $C|A-$, ..., $R|A-$. Using the same principles as in Theorem 1, $\mathbf{C}_{11}$ is a $q \times q$ matrix with unit diagonal, $-\mu_c^2/\sigma_c^2$ for the off-diagonal entries in the first row and column, and $\psi_{cou}(\rho)$ for all other off-diagonal entries. Let B be the inactive effect with $\mathbf{C}_{21,B} = (\psi_{sib}(\rho), \tilde{\psi}(\rho)\mathbf{1}_{q-1})$. Taking $\boldsymbol{\zeta} = \mathbf{1}_q$, $|\mathbf{C}_{21,B}\mathbf{C}_{11}^{-1}\boldsymbol{\zeta}| \geq 1$ for $\rho \geq 0.29$, which proves inconsistency.

1.3 Proof of Proposition 1

As a note, since the objective $Q(\boldsymbol{\beta})$ is non-differentiable at $\boldsymbol{\beta} = \mathbf{0}$, what we mean by strict convexity here is that $\nabla_{\mathbf{u}}^2 Q(\boldsymbol{\beta})$, the directional Hessian of $Q(\boldsymbol{\beta})$ in direction $\mathbf{u}$, is positive-definite for all $\boldsymbol{\beta}$ and all $\|\mathbf{u}\| = 1$. We follow a similar approach as Proposition 1 of Breheny (2015). Note that $\nabla^2 \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = 2\mathbf{X}^T\mathbf{X}$. Moreover, with $\eta'_{\lambda,\tau}(\theta) = \lambda \exp(-\theta\tau/\lambda)$ and

[1] $\psi_{me}(\rho)$, $\psi_{sib}(\rho)$, $\psi_{pc}(\rho)$ and $\psi_{cou}(\rho)$ are the pairwise correlations in Theorem 1 for main effects, siblings, parent-child pairs and cousins, respectively. $\tilde{\psi}(\rho) = \sin^{-1}(\rho)/(\pi\sigma_c)$ is the pairwise correlation between a CME and its conditioned effect.

$\eta''_{\lambda,\tau}(\theta) = -\tau \exp(-\theta\tau/\lambda)$, one can show that $\nabla^2_{\mathbf{u}} P_s(\boldsymbol{\beta}) \geq -\tau(1) + \lambda(-1/(\lambda\gamma)) = -\tau - 1/\gamma$ and similarly $\nabla^2_{\mathbf{u}} P_c(\boldsymbol{\beta}) \geq -\tau - 1/\gamma$, for all $\mathbf{u}$ and $\boldsymbol{\beta}$. Hence:

$$\nabla^2_{\mathbf{u}} Q(\boldsymbol{\beta}) = \nabla^2_{\mathbf{u}} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + P_s(\boldsymbol{\beta}) + P_c(\boldsymbol{\beta}) \right\} \geq \frac{\lambda_{min}(\mathbf{X}^T\mathbf{X})}{n} - 2\left(\tau + \frac{1}{\gamma}\right) \text{ for all } \mathbf{u} \text{ and } \boldsymbol{\beta},$$

which is strictly positive when $\tau + 1/\gamma < \lambda_{min}(\mathbf{X}^T\mathbf{X})/(2n)$. The second part of the claim follows by replacing $\mathbf{X}$ with $\mathbf{x}_j$ in the argument above, and using the fact that $\|\mathbf{x}_j\|_2^2 = n$.

1.4 Proofs of Theorem 3 and Corollary 1

The majorization claim a) follows from a first-order Taylor expansion of the outer penalty: $\eta_{\lambda,\tau}(\|\boldsymbol{\beta}_g\|_{\lambda,\gamma}) \geq \eta_{\lambda,\tau}(\|\tilde{\boldsymbol{\beta}}_g\|_{\lambda,\gamma}) + \tilde{\Delta}_g \left\{ \|\boldsymbol{\beta}_g\|_{\lambda,\gamma} - \|\tilde{\boldsymbol{\beta}}_g\|_{\lambda,\gamma} \right\}$, where the inequality holds due to the concavity of η. See Lemma 1 in Breheny (2015) for details.

To derive the threshold function in b), take the following optimization problem:

$$\hat{\beta}_j = \underset{\beta_j}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\mathbf{r} - \mathbf{x}_j \beta_j\|_2^2 + \Delta_1 g_{\lambda_1,\gamma}(\beta_j) + \Delta_2 g_{\lambda_2,\gamma}(\beta_j) \right\}. \tag{2}$$

The KKT condition for (2) is:

$$0 \in -\frac{1}{n}\mathbf{x}_j^T \mathbf{r} + \hat{\beta}_j + \Delta_1 \partial_{\lambda_1,\gamma} \hat{\beta}_j + \Delta_2 \partial_{\lambda_2,\gamma} \hat{\beta}_j, \quad \partial_{\lambda,\gamma}\beta_j = \begin{cases} \operatorname{sgn}(\beta_j)\left(1 - \frac{|\beta_j|}{\lambda\gamma}\right)_+ & \text{if } |\beta_j| > 0, \\ [-1, 1] & \text{if } \beta_j = 0. \end{cases} \tag{3}$$

Without loss of generality, assume $z \equiv \mathbf{x}_j^T \mathbf{r}/n > 0$. Consider the same four cases for z as presented in equation (9) in the paper:

1. $z \geq \lambda_{(1)}\gamma$: Suppose $\hat{\beta}_j = z$. Then the KKT condition (3) becomes $0 \in -z + \hat{\beta}_j$, which is satisfied. Since (2) is strictly convex, $\hat{\beta}_j = z$ must be its unique solution.

2. $c_2 \le z < \lambda_{(1)}\gamma$ (see equation (9) in the paper for c_2): Suppose $\hat{\beta}_j = (z - \Delta_{(1)}) / \left(1 - \frac{\Delta_{(1)}}{\lambda_{(1)}\gamma}\right)$. Since $\lambda_{(2)}\gamma \le \hat{\beta}_j < \lambda_{(1)}\gamma$, the KKT condition (3) becomes $0 \in -z + \hat{\beta}_j + \Delta_{(1)} \left(1 - \frac{\hat{\beta}_j}{\lambda_{(1)}\gamma}\right)$, which is satisfied. Hence, $\hat{\beta}_j$ is the unique solution to (2).

3. $\Delta_{(1)} + \Delta_{(2)} \le z < c_2$ (see equation (9) in the paper for c_3): Suppose $\hat{\beta}_j = (z - \Delta_{(1)} - \Delta_{(2)}) / \left(1 - \frac{\Delta_{(1)}}{\lambda_{(1)}\gamma} - \frac{\Delta_{(2)}}{\lambda_{(2)}\gamma}\right)$. Since $0 < \hat{\beta}_j < \lambda_{(2)}\gamma$, the KKT condition (3) becomes $0 \in -z + \hat{\beta}_j + \Delta_{(1)} \left(1 - \frac{\hat{\beta}_j}{\lambda_{(1)}\gamma}\right) + \Delta_{(2)} \left(1 - \frac{\hat{\beta}_j}{\lambda_{(2)}\gamma}\right)$, which is satisfied. Hence, $\hat{\beta}_j$ is the unique solution to (2).

4. $0 \le z < \Delta_{(1)} + \Delta_{(2)}$: Suppose $\hat{\beta}_j = 0$. The KKT condition then becomes $0 \in -z + (\Delta_{(1)} + \Delta_{(2)})[-1, 1]$, which is satisfied, so $\hat{\beta}_j$ is the unique solution to (2).

From this, Corollary 1 can be proved in a similar way as Proposition 3 of Breheny (2015).

1.5 Proof of Proposition 2

Since $Q(\boldsymbol{\beta})$ is strictly convex, it must have at most one minimizer $\boldsymbol{\beta}$. By definition, $\boldsymbol{\beta}$ must satisfy the KKT condition:

$$0 \in -\frac{1}{n}\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \Delta_{\mathcal{S}}(\boldsymbol{\beta})\partial_{\lambda_s,\gamma}\beta_j + \Delta_{\mathcal{C}}(\boldsymbol{\beta})\partial_{\lambda_c,\gamma}\beta_j, \quad j = 1, \cdots, p', \tag{4}$$

where $\partial_{\lambda,\gamma}\beta_j$ is the subgradient defined in (3), and $\Delta_{\mathcal{S}}(\boldsymbol{\beta})$ and $\Delta_{\mathcal{C}}(\boldsymbol{\beta})$ are the linearized slopes for the sibling and cousin groups of effect j (see equation (5) of the paper). Setting $\beta = \mathbf{0}$, the right side of (4) becomes:

$$-\frac{1}{n}\mathbf{x}_j^T\mathbf{y} + \lambda_s[-1, 1] + \lambda_c[-1, 1] = -\frac{1}{n}\mathbf{x}_j^T\mathbf{y} + [-\lambda_s - \lambda_c, \lambda_s + \lambda_c],$$

which contains 0 when $\lambda_s + \lambda_c \ge |\mathbf{x}_j^T\mathbf{y}|/n$. Hence, when $\lambda_s + \lambda_c \ge \max_{j=1,\cdots,p'} |\mathbf{x}_j^T\mathbf{y}|/n$, one can invoke the strict convexity of $Q(\boldsymbol{\beta})$ to show that the trivial solution $\boldsymbol{\beta} = \mathbf{0}$ is indeed the unique minimizer.

2 Algorithm statement for `cv.cmenet`

Algorithm 1 `cv.cmenet`: A cross-validation algorithm for tuning `cmenet`

function CV.CMENET$(\mathbf{X}, \mathbf{y}, K)$
- Initialize grid of potential parameters $\max_{j=1,\cdots,p'} |\mathbf{x}_j^T \mathbf{y}|/n > \lambda_s^1 > \cdots > \lambda_s^L > 0$, $\max_{j=1,\cdots,p'} |\mathbf{x}_j^T \mathbf{y}|/n > \lambda_c^1 > \cdots > \lambda_c^M > 0$, $\gamma^1 < \cdots < \gamma^G$ and $\tau^1 < \cdots < \tau^T$ (satisfying $\tau + 1/\gamma < 1/2$).
- Obtain the tuned MC+ parameters (λ^*, γ^*) using `cv.sparsenet` in the R package SPARSENET, and set $\lambda_s^*, \lambda_c^* \leftarrow \lambda^*/2$ as an initial estimate.
- Randomly partition the data $\mathcal{D} = (\mathbf{X}, y)$ into K equal pieces $\{\mathcal{D}_1, \cdots, \mathcal{D}_K\}$.

for $k = 1, \cdots, K$ **do** ▷ K-fold CV for tuning γ and τ
  for $\gamma \in \{\gamma_1, \cdots, \gamma_G\}$ **do** ▷ For each γ...
    • $\boldsymbol{\beta}_{prev} \leftarrow \mathbf{0}_{p'}$ ▷ Reset warm start solution
    for $\tau \in \{\tau_1, \cdots, \tau_T\}$ **do** ▷ For each τ...
      • $\boldsymbol{\beta}_{\lambda_s^*,\lambda_c^*}(\gamma, \tau; k) \leftarrow$ `cmenet`$(\mathbf{X}_{-k}, \mathbf{y}_{-k}, \lambda_s^*, \lambda_c^*, \gamma, \tau, \boldsymbol{\beta}_{prev})$ ▷ Train w/o part k
      • $\boldsymbol{\beta}_{prev} \leftarrow \boldsymbol{\beta}_{\lambda_s^*,\lambda_c^*}(\gamma, \tau; k)$ ▷ Update warm start solution

• $(\gamma^*, \tau^*) \leftarrow \operatorname{argmin}_{\gamma,\tau} \sum_{k=1}^{K} \|\mathbf{y}_k - \mathbf{X}_k \boldsymbol{\beta}_{\lambda_s^*,\lambda_c^*}(\gamma, \tau; k)\|_2^2$ ▷ Estimate optimal γ and τ

for $k = 1, \cdots, K$ **do** ▷ K-fold CV for tuning λ_s and λ_c
  for $\lambda_c \in \{\lambda_c^1, \cdots, \lambda_c^M\}$ **do** ▷ For each λ_c...
    • $\boldsymbol{\beta}_{prev} \leftarrow \mathbf{0}_{p'}$
    for $\lambda_s \in \{\lambda_s^1, \cdots, \lambda_s^L\}$ **do** ▷ For each λ_s...
      if $\lambda_c + \lambda_s < \max_{j=1,\cdots,p'} |\mathbf{x}_j^T \mathbf{y}|/n$ **then**
        • Screen using the three strong rules in Section 4.3.
        • $\boldsymbol{\beta}_{\lambda_s,\lambda_c}(\gamma^*, \tau^*; k) \leftarrow$ `cmenet`$(\mathbf{X}_{-k}, \mathbf{y}_{-k}, \lambda_s, \lambda_c, \gamma^*, \tau^*, \boldsymbol{\beta}_{prev})$, using only screened effects.
        • Check KKT conditions on converged solution $\boldsymbol{\beta}_{\lambda_s,\lambda_c}(\gamma^*, \tau^*; k)$.
        • $\boldsymbol{\beta}_{prev} \leftarrow \boldsymbol{\beta}_{\lambda_s,\lambda_c}(\gamma^*, \tau^*; k)$

• $(\lambda_s^*, \lambda_c^*) \leftarrow \operatorname{argmin}_{\lambda_s,\lambda_c} \sum_{k=1}^{K} \|\mathbf{y}_k - \mathbf{X}_k \boldsymbol{\beta}_{\lambda_s,\lambda_c}(\gamma^*, \tau^*; k)\|_2^2$ ▷ Estimate optimal λ_s and λ_c

• $\hat{\boldsymbol{\beta}} \leftarrow$ `cmenet`$(\mathbf{X}, \mathbf{y}, \lambda_s^*, \lambda_c^*, \gamma^*, \tau^*, \mathbf{0}_{p'})$ ▷ Refit using optimal parameters
  return optimal coefficients $\hat{\boldsymbol{\beta}}$.

Some comments on the implementation of active set optimization within `cmenet`:

- The active set of variables is initialized by performing the full coordinate descent cycle for 25 iterations, then choosing the variables whose coefficients are non-zero.

- Repeat coordinate descent iterations over the active set until convergence.
- Perform a full coordinate descent cycle over all p' variables. If this cycle does not change the active set, `cmenet` is terminated; otherwise, the active set is updated, and the above steps repeated.

3 Theoretical derivation of CME screening rules

Fix γ and τ, and suppose $\hat{\beta}_j(\lambda_s, \lambda_c) \in (0, \min\{\Delta_{(1)} + \Delta_{(2)}, \lambda_{(2)}\gamma\})$. For brevity, we denote $\hat{\beta}_j(\lambda_s, \lambda_c)$ as $\hat{\beta}_j$ from here on. Using equation (9) in the paper, we know that $\hat{\beta}_j$ takes the form:

$$\begin{aligned}\hat{\beta}_j &= \text{sgn}(z_j)\left(|z_j| - \Delta_{(1)} - \Delta_{(2)}\right)_+ / \left(1 - \frac{\Delta_{(1)}}{\lambda_{(1)}\gamma} - \frac{\Delta_{(2)}}{\lambda_{(2)}\gamma}\right) \\ &= \text{sgn}(z_j)\left(|z_j| - \Delta_S - \Delta_C\right)_+ / \left(1 - \frac{\Delta_S}{\lambda_S\gamma} - \frac{\Delta_C}{\lambda_C\gamma}\right),\end{aligned} \tag{5}$$

where $z_j = \mathbf{x}_j^T \mathbf{r}_{-j}/n$ (see Theorem 3), and $\Delta_{\mathcal{S}}$ and $\Delta_{\mathcal{C}}$ are the linearized slopes for the current penalty setting (λ_s, λ_c). Plugging this expression into (4), the KKT condition for $\hat{\beta}_j$ can be simplified to:

$$\begin{aligned}0 &= -c_j(\lambda_s, \lambda_c) + \text{sgn}(\hat{\beta}_j)\Delta_{\mathcal{S}}\left\{1 - \frac{(|z_j| - \Delta_{\mathcal{S}} - \Delta_{\mathcal{C}})_+}{\lambda_s\left(\gamma - \frac{\Delta_{\mathcal{S}}}{\lambda_s} - \frac{\Delta_{\mathcal{C}}}{\lambda_c}\right)}\right\} + \text{sgn}(\hat{\beta}_j)\Delta_{\mathcal{C}}\left\{1 - \frac{(|z_j| - \Delta_{\mathcal{S}} - \Delta_{\mathcal{C}})_+}{\lambda_c\left(\gamma - \frac{\Delta_{\mathcal{S}}}{\lambda_s} - \frac{\Delta_{\mathcal{C}}}{\lambda_c}\right)}\right\} \\ \Leftrightarrow\ c_j(\lambda_s, \lambda_c) &= \text{sgn}(\hat{\beta}_j)\Delta_{\mathcal{S}}\left\{1 - \frac{(|z_j| - \Delta_{\mathcal{S}} - \Delta_{\mathcal{C}})_+}{\lambda_s\left(\gamma - \frac{\Delta_{\mathcal{S}}}{\lambda_s} - \frac{\Delta_{\mathcal{C}}}{\lambda_c}\right)}\right\} + \text{sgn}(\hat{\beta}_j)\Delta_{\mathcal{C}}\left\{1 - \frac{(|z_j| - \Delta_{\mathcal{S}} - \Delta_{\mathcal{C}})_+}{\lambda_c\left(\gamma - \frac{\Delta_{\mathcal{S}}}{\lambda_s} - \frac{\Delta_{\mathcal{C}}}{\lambda_c}\right)}\right\}.\end{aligned} \tag{6}$$

Suppose no effects are active in either the sibling group $\mathcal{S}$ or the cousin group $\mathcal{C}$, in which case $\Delta_{\mathcal{S}} = \lambda_s$ and $\Delta_{\mathcal{C}} = \lambda_c$. The KKT condition in (6) can then be rewritten as:

$$c_j(\lambda_s, \lambda_c) = \text{sgn}(\hat{\beta}_j)\left\{\lambda_s - \frac{(|z_j| - \lambda_s - \lambda_c)_+}{\gamma - 2}\right\} + \text{sgn}(\hat{\beta}_j)\left\{\lambda_c - \frac{(|z_j| - \lambda_s - \lambda_c)_+}{\gamma - 2}\right\}. \tag{7}$$

Taking the derivative with respect to λ_s (and assuming z_j is approximately constant in λ_s, following Lee and Breheny, 2015), we get:

$$\left|\frac{\partial}{\partial \lambda_s} c_j(\lambda_s, \lambda_c)\right| \lesssim 1 + \frac{1}{\gamma - 2} + \frac{1}{\gamma - 2} = \frac{\gamma}{\gamma - 2}. \tag{8}$$

A similar argument shows that this approximate upper bound also holds for $|(\partial/\partial\lambda_c)\, c_j(\lambda_s, \lambda_c)|$.

Now, suppose no effects are active in the sibling group $\mathcal{S}$ (but some in the cousin group $\mathcal{C}$), in which case $\Delta_{\mathcal{S}} = \lambda_s$. The KKT condition in (6) can then be rewritten as:

$$c_j(\lambda_s, \lambda_c) = \text{sgn}(\hat{\beta}_j) \left\{ \lambda_s - \frac{(|z_j| - \lambda_s - \Delta_{\mathcal{C}})_+}{\gamma - 1 - \frac{\Delta_{\mathcal{C}}}{\lambda_c}} \right\} + \text{sgn}(\hat{\beta}_j) \Delta_{\mathcal{C}} \left\{ 1 - \frac{(|z_j| - \lambda_s - \Delta_{\mathcal{C}})_+}{\lambda_c \left(\gamma - 1 - \frac{\Delta_{\mathcal{C}}}{\lambda_c} \right)} \right\}. \tag{9}$$

Taking the derivative on λ_s (and assuming z_j is approximately constant in λ_s), we get:

$$\left|\frac{\partial}{\partial \lambda_s} c_j(\lambda_s, \lambda_c)\right| \lesssim 1 + \frac{1}{\gamma - 1 - \frac{\Delta_{\mathcal{C}}}{\lambda_c}} + \frac{\frac{\Delta_{\mathcal{C}}}{\lambda_c}}{\gamma - 1 - \frac{\Delta_{\mathcal{C}}}{\lambda_c}} = \frac{\gamma}{\gamma - 1 - \frac{\Delta_{\mathcal{C}}}{\lambda_c}}. \tag{10}$$

Finally, suppose there are no active effects in the cousin group $\mathcal{C}$ (but some in sibling group $\mathcal{S}$). One can do a similar approximation and show that:

$$\left|\frac{\partial}{\partial \lambda_c} c_j(\lambda_s, \lambda_c)\right| \lesssim 1 + \frac{1}{\gamma - \frac{\Delta_{\mathcal{S}}}{\lambda_s} - 1} + \frac{\frac{\Delta_{\mathcal{S}}}{\lambda_s}}{\gamma - \frac{\Delta_{\mathcal{S}}}{\lambda_s} - 1} = \frac{\gamma}{\gamma - \frac{\Delta_{\mathcal{S}}}{\lambda_s} - 1}. \tag{11}$$

These upper bounds on the absolute derivatives of $c_j(\lambda_s, \lambda_c)$, along with the proposed strong rules in Section 4.3, can then be used to demonstrate the inactivity of effect j at penalty setting $(\lambda_s^l, \lambda_c^m)$:

1. Consider the first part of the first strong rule, which applies when no active effects are in $\mathcal{S}$ and $\mathcal{C}$ for setting $(\lambda_s^{l-1}, \lambda_c^m)$. This rule discards effect j at setting $(\lambda_s^l, \lambda_c^m)$ if:

$$|c_j(\lambda_s^{l-1}, \lambda_c^m)| < \lambda_s^l + \lambda_c^m + \frac{\gamma}{\gamma - 2}(\lambda_s^l - \lambda_s^{l-1}).$$

This can be justified as follows. Using the approximate upper bound in (8), the inner-product of effect j at setting $(\lambda_s^l, \lambda_c^m)$ can be approximately upper bounded as:

$$\begin{aligned}
|c_j(\lambda_s^l, \lambda_c^m)| &\leq |c_j(\lambda_s^l, \lambda_c^m) - c_j(\lambda_s^{l-1}, \lambda_c^m)| + |c_j(\lambda_s^{l-1}, \lambda_c^m)| \\
&\approx \left|\frac{\partial}{\partial \lambda_s} c_j(\lambda_s^{l-1}, \lambda_c^m)\right|(\lambda_s^{l-1} - \lambda_s^l) + |c_j(\lambda_s^{l-1}, \lambda_c^m)| \\
&< \frac{\gamma}{\gamma - 2}(\lambda_s^{l-1} - \lambda_s^l) + \left[\lambda_s^l + \lambda_c^m + \frac{\gamma}{\gamma - 2}(\lambda_s^l - \lambda_s^{l-1})\right] \\
&= \lambda_s^l + \lambda_c^m.
\end{aligned}$$

Assuming effect j is the first variable to potentially be selected in $\mathcal{S}$ or $\mathcal{C}$ at current setting $(\lambda_s^l, \lambda_c^m)$, the KKT conditions in (4) suggest that effect j is inactive, which justifies the screening rule. A similar argument can be used to derive the second part of this rule.

2. Consider next the second strong rule, which applies when no active effects are in $\mathcal{S}$ for setting $(\lambda_s^{l-1}, \lambda_c^m)$. This rule discards effect j at setting $(\lambda_s^l, \lambda_c^m)$ if:

$$|c_j(\lambda_s^{l-1}, \lambda_c^m)| < \lambda_s^l + \Delta'_{\mathcal{C}} + \frac{\gamma}{\gamma - (\Delta'_{\mathcal{C}}/\lambda_c^m + 1)}(\lambda_s^l - \lambda_s^{l-1}).$$

This can be justified as follows. Using the approximate upper bound in (10), the inner-product of effect j at setting $(\lambda_s^l, \lambda_c^m)$ can be approximately upper bounded as:

$$\begin{aligned}
|c_j(\lambda_s^l, \lambda_c^m)| &\leq |c_j(\lambda_s^l, \lambda_c^m) - c_j(\lambda_s^{l-1}, \lambda_c^m)| + |c_j(\lambda_s^{l-1}, \lambda_c^m)| \\
&\approx \left|\frac{\partial}{\partial \lambda_s} c_j(\lambda_s^{l-1}, \lambda_c^m)\right|(\lambda_s^{l-1} - \lambda_s^l) + |c_j(\lambda_s^{l-1}, \lambda_c^m)| \\
&< \frac{\gamma}{\gamma - (\Delta'_{\mathcal{C}}/\lambda_c^m + 1)}(\lambda_s^{l-1} - \lambda_s^l) + \left[\lambda_s^l + \Delta'_{\mathcal{C}} + \frac{\gamma}{\gamma - (\Delta'_{\mathcal{C}}/\lambda_c^m + 1)}(\lambda_s^l - \lambda_s^{l-1})\right] \\
&= \lambda_s^l + \Delta'_{\mathcal{C}}.
\end{aligned}$$

Assuming:

- Effect j is the first variable to potentially be selected in $\mathcal{S}$ at current setting $(\lambda_s^l, \lambda_c^m)$,
- The linearized slope $\Delta'_{\mathcal{C}}$ at previous setting $(\lambda_s^{l-1}, \lambda_c^m)$ is approximately the linearized slope $\Delta_{\mathcal{C}}$ at current setting $(\lambda_s^l, \lambda_c^m)$,

the KKT conditions in (4) suggest that effect j is inactive, which justifies the screening rule.

3. The third strong rule can be justified in a similar manner to the above two rules.

References

Breheny, P. (2015). The group exponential lasso for bi-level variable selection. *Biometrics*, 71(3):731–740.

Lee, S. and Breheny, P. (2015). Strong rules for nonconvex penalties and their implications for efficient algorithms in high-dimensional regression. *Journal of Computational and Graphical Statistics*, 24(4):1074–1091.

Stuart, A. and Ord, J. (1994). *Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory*. Arnold London.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.