Supporting Information for "A Hierarchical Integrative Group LASSO (HiGLASSO) framework for analyzing environmental mixtures" by Boss et al.

Web Appendix A: HiGLASSO algorithm

A.1. *Objective Function*

The HiGLASSO objective function is:

$$\begin{aligned}\arg\min_{\boldsymbol{\beta}_j, \boldsymbol{\eta}_{jj'}} \frac{1}{2}\Big\|\boldsymbol{y} - \sum_{j=1}^{S} \boldsymbol{X}_j \boldsymbol{\beta}_j - \sum_{1 \le j < j' \le S} \boldsymbol{X}_{jj'}[\boldsymbol{\eta}_{jj'} \odot (\boldsymbol{\beta}_j \otimes \boldsymbol{\beta}_{j'})]\Big\|_2^2 \\ + \lambda_1 \sum_{j=1}^{S} w_j ||\boldsymbol{\beta}_j||_2 + \lambda_2 \sum_{1 \le j < j' \le S} w_{jj'} ||\boldsymbol{\eta}_{jj'}||_2,\end{aligned} \tag{1}$$

$$w_j \equiv \exp\Big\{ - \frac{||\boldsymbol{\beta}_j||_\infty}{\sigma} \Big\} \text{ for } j = 1, \cdots, S, \tag{2}$$

$$w_{jj'} \equiv \exp\Big\{ - \frac{||\boldsymbol{\eta}_{jj'}||_\infty}{\sigma} \Big\} \text{ for } 1 \le j < j' \le S, \tag{3}$$

A.2. *Updating main effect coefficients*

By substituting our weight function (2) into (1), given the current $\hat{\boldsymbol{\beta}}_{j'}$'s with $j' \neq j$ and $\hat{\boldsymbol{\eta}}_{jj'}$'s, the objective function can be written as

$$\arg\min_{\boldsymbol{\beta}_j} \frac{1}{2}\big\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}_j \boldsymbol{\beta}_j\big\|_2^2 + \lambda_1 \exp\Big\{ - \frac{||\boldsymbol{\beta}_j||_\infty}{\sigma} \Big\} ||\boldsymbol{\beta}_j||_2, \tag{4}$$

such that

$$\tilde{\boldsymbol{y}} = \boldsymbol{y} - \sum_{k \neq j} \boldsymbol{X}_k \hat{\boldsymbol{\beta}}_k - \sum_{k,l \neq j} \boldsymbol{X}_{kl}[\hat{\boldsymbol{\eta}}_{kl} \odot (\hat{\boldsymbol{\beta}}_k \otimes \hat{\boldsymbol{\beta}}_l)],$$

$$\tilde{\boldsymbol{X}}_j = \boldsymbol{X}_j + \sum_{k<j} \boldsymbol{X}_{kj} \cdot \text{diag}(\hat{\boldsymbol{\eta}}_{kj})(\hat{\boldsymbol{\beta}}_k \otimes \boldsymbol{I}_{p_j}) + \sum_{l>j} \boldsymbol{X}_{jl} \cdot \text{diag}(\hat{\boldsymbol{\eta}}_{jl})(\boldsymbol{I}_{p_j} \otimes \hat{\boldsymbol{\beta}}_l),$$

where $\boldsymbol{I}_{p_j}$ is p_j dimensional identity matrix. $\tilde{\boldsymbol{X}}_j$ and $\tilde{\boldsymbol{y}}$ represent the design matrix and response vector at current step. (4) can be directly solved using gradient descent or the Newton-Raphson algorithm (Bauer and Cai, 2009).

Alternatively, we obtain updating algorithm for $\boldsymbol{\beta}_j$ in closed form using local quadratic approximation (LQA) (Fan and Li, 2001). Let $\mathbf{Pen}_1(\boldsymbol{\beta}_j)$ denote the penalty term in (4). We approximate $\mathbf{Pen}_1(\boldsymbol{\beta}_j)$ by

$$\mathbf{Pen}_1(\boldsymbol{\beta}_j) \approx \mathbf{Pen}_1\left(\hat{\boldsymbol{\beta}}_j^{(m)}\right) + \frac{1}{2}\sum_{k=1}^{p_j} d_{jk}^{(m)}\left[\beta_{jk}^2 - \left(\hat{\beta}_{jk}^{(m)}\right)^2\right]$$

where β_{jk} is the k^{th} element of $\boldsymbol{\beta}_j$, $\hat{\boldsymbol{\beta}}_j^{(m)}$ is the estimate of $\boldsymbol{\beta}_j$ from m^{th} iteration, and d_{jk} is defined through

$$\frac{\partial \mathbf{Pen}_1(\boldsymbol{\beta}_j)}{\partial \beta_{jk}} = d_{jk}\beta_{jk}.$$

By calculating the derivative of $\mathbf{Pen}_1(\boldsymbol{\beta}_j)$, we have

$$d_{jk} = \begin{cases} \exp\Big\{-\frac{||\boldsymbol{\beta}_j||_\infty}{\sigma}\Big\}\big(||\boldsymbol{\beta}_j||_2\big)^{-1}, & \text{if } |\beta_{jk}| \neq ||\boldsymbol{\beta}_j||_\infty \\ \exp\Big\{-\frac{||\boldsymbol{\beta}_j||_\infty}{\sigma}\Big\}\Big[\big(||\boldsymbol{\beta}_j||_2\big)^{-1} - ||\boldsymbol{\beta}_j||_2\big(|\beta_{jk}|\sigma\big)^{-1}\Big], & \text{if } |\beta_{jk}| = ||\boldsymbol{\beta}_j||_\infty. \end{cases} \tag{5}$$

The problem with LQA is that d_{jk}, which represents the second-degree derivative of $\mathbf{Pen}_1(\boldsymbol{\beta}_j)$, might be negative when $|\beta_{jk}| = ||\boldsymbol{\beta}_j||_\infty$. Therefore, it is not guaranteed that the approximated $\mathbf{Pen}_j(\boldsymbol{\beta}_j)$ will be convex.

Pan and Zhao proposed generalized local quadratic approximation (GLQA) to employ convex quadratic approximation to the penalty function (Pan and Zhao, 2016). Let $\mathcal{P}_1(\boldsymbol{\beta}_j)$ denote GLQA of $\mathbf{Pen}_1(\boldsymbol{\beta}_j)$ that satisfies the following three properties

1. $\mathcal{P}_1(\boldsymbol{\beta}_j)$ is convex,
2. $\mathcal{P}_1\left(\hat{\boldsymbol{\beta}}_j^{(m)}\right) = \mathbf{Pen}_1\left(\hat{\boldsymbol{\beta}}_j^{(m)}\right)$,
3. $\frac{\partial \mathcal{P}_1(\boldsymbol{\beta}_j)}{\partial \beta_{jk}}\Big|_{\beta_{jk}=\hat{\beta}_{jk}^{(m)}} = \frac{\partial \mathbf{Pen}_1(\boldsymbol{\beta}_j)}{\partial \beta_{jk}}\Big|_{\beta_{jk}=\hat{\beta}_{jk}^{(m)}}\ \forall\ k.$

A simple choice takes the form of

$$\mathcal{P}_1(\boldsymbol{\beta}_j) = \mathbf{Pen}_1\left(\hat{\boldsymbol{\beta}}_j^{(m)}\right) + \frac{1}{2}\sum_{k=1}^{p_j}\left|d_{jk}^{(m)}\right|\left[(\beta_{jk}^2 + c_1)^2 + c_2\right].$$

Solving c_1 and c_2 according to the second and third conditions gives

$$\mathcal{P}_1(\boldsymbol{\beta}_j) = \mathbf{Pen}_1\left(\hat{\boldsymbol{\beta}}_j^{(m)}\right) + \frac{1}{2}\sum_{k=1}^{p_j}\left|d_{jk}^{(m)}\right|\left[\left(\beta_{jk}^2 - \left(1 - \frac{d_{jk}^{(m)}}{|d_{jk}^{(m)}|}\right)\hat{\beta}_{jk}^{(m)}\right)^2 - \left(\hat{\beta}_{jk}^{(m)}\right)^2\right].$$

Rewriting the $\mathcal{P}_1(\boldsymbol{\beta}_j)$ in matrix form, (4) can be approximated as

$$\frac{1}{2}||\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}_j\boldsymbol{\beta}_j||_2^2 + \frac{1}{2}\lambda_1\boldsymbol{\beta}_j^\top \boldsymbol{D}_j^{(m)}\boldsymbol{\beta}_j - \lambda_1\boldsymbol{c}^{(m)\top}\boldsymbol{\beta}_j + \text{Constant}$$

where

$$\boldsymbol{D}_j^{(m)} = \text{diag}\Big[\Big(d_{j1}^{(m)}, \cdots, d_{jp_j}^{(m)}\Big)\Big] \text{ and}$$

$$\boldsymbol{c}^{(m)} = \Big\{\Big(|d_{j1}^{(m)}| - d_{j1}^{(m)}\Big)\hat{\beta}_{j1}^{(m)}, \cdots, \Big(|d_{jp_j}^{(m)}| - d_{jp_j}^{(m)}\Big)\hat{\beta}_{jp_j}^{(m)}\Big\}^\top.$$

$\boldsymbol{\beta}_j$ can be updated in closed-form as

$$\hat{\boldsymbol{\beta}}_j = \Big(\tilde{\boldsymbol{X}}_j^\top\tilde{\boldsymbol{X}}_j + n\lambda_1\boldsymbol{D}_j^{(m)}\Big)^{-1}\Big(\tilde{\boldsymbol{X}}_j^\top\tilde{\boldsymbol{y}} + \lambda_1 \cdot \boldsymbol{c}^{(m)}\Big). \tag{6}$$

A.3. *Updating scalar terms associated with interactions*

By substituting the specified weight function (3) into (1), given $\hat{\boldsymbol{\beta}}_j$'s, the objective function can be expressed as

$$\underset{\boldsymbol{\eta}_{jj'}}{\arg\min}\ \frac{1}{2}\Big\|\tilde{\boldsymbol{y}} - \sum_{j<j'}\tilde{\boldsymbol{X}}_{jj'}\boldsymbol{\eta}_{jj'}\Big\|_2^2 + \lambda_2\sum_{j<j'}\exp\Big\{-\frac{||\boldsymbol{\eta}_{jj'}||_\infty}{\sigma}\Big\}||\boldsymbol{\eta}_{jj'}||_2 \tag{7}$$

where

$$\tilde{\boldsymbol{y}} = \boldsymbol{y} - \sum_{k=1}^{S}\boldsymbol{X}_k\hat{\boldsymbol{\beta}}_k$$

and

$$\tilde{\boldsymbol{X}}_{jj'} = \boldsymbol{X}_{jj'}\text{diag}\big[(\hat{\boldsymbol{\beta}}_j \otimes \hat{\boldsymbol{\beta}}_{j'})\big] \text{ for } 1 \le j < j' \le S.$$

Let $\mathbf{Pen}_2(\boldsymbol{\eta}_{jj'})$ denote the individual penalty term in (7) and let $\mathcal{P}_2(\boldsymbol{\beta}_{jj'})$ denote GLQA of $\mathbf{Pen}_2(\boldsymbol{\eta}_{jj'})$. We have

$$\mathcal{P}_2(\boldsymbol{\eta}_{jj'}) = \mathbf{Pen}_1\Big(\hat{\boldsymbol{\eta}}_{jj'}^{(m)}\Big) + \frac{1}{2}\sum_{k=1}^{p_jp_{j'}}|d_{jj'k}^{(m)}|\left[\left(\eta_{jj'k}^2 - \left(1 - \frac{d_{jj'k}^{(m)}}{|d_{jj'k}^{(m)}|}\right)\hat{\eta}_{jj'k}^{(m)}\right)^2 - \Big(\hat{\eta}_{jj'k}^{(m)}\Big)^2\right]$$

where $\eta_{jj'k}$ is the k^{th} element of $(p_jp_{j'})-$vector of $\boldsymbol{\eta}_{jj'}$ and $d_{jj'k}$ is similarly defined through

$$\frac{\partial\mathbf{Pen}_2(\boldsymbol{\eta}_{jj'})}{\partial\eta_{jj'k}} = d_{jj'k}\eta_{jj'k}$$

as (5). (7) can be approximated as

$$\frac{1}{2}||\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\eta}||_2^2 + \frac{1}{2}\lambda_2\boldsymbol{\eta}^\top\boldsymbol{D}^{(m)}\boldsymbol{\eta} - \lambda_2\boldsymbol{C}^{(m)\top}\boldsymbol{\eta} + \text{Constant}$$

where $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{X}}_{12}, \cdots, \tilde{\boldsymbol{X}}_{S-1,S}]$, $\boldsymbol{\eta} = \big(\boldsymbol{\eta}_{12}^\top, \cdots, \boldsymbol{\eta}_{S-1,S}^\top\big)^\top$,

$$\boldsymbol{D}^{(m)} = \text{diag}\Big[d_{121}^{(m)}, \cdots, d_{12(p_1p_2)}^{(m)}, \cdots, d_{(S-1)S(p_{S-1}p_S)}^{(m)}\Big],$$

and $\boldsymbol{C}^{(m)}$ is a $[S(S-1)/2] \times [\sum_{j<j'} p_j p_{j'}]$ block column vector such that the block corresponding to the interaction between group j and group j' is defined as a vector of length $p_j p_{j'}$ with the k^{th} element equal to $\left(|d_{jj'k}^{(m)}| - d_{jj'k}^{(m)}\right)\hat{\eta}_{jj'k}^{(m)}$. $\boldsymbol{\eta}_{jj'}$s can then be updated in closed form as

$$\hat{\boldsymbol{\eta}} = \left(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + n\lambda_2 \boldsymbol{D}^{(m)}\right)^{-1} \left(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{y}} + \lambda_2 \cdot \boldsymbol{C}^{(m)}\right). \quad (8)$$

A.4. *Algorithm*

We describe the full algorithm for estimating $\boldsymbol{\beta}_j$'s and $\boldsymbol{\eta}_{jj'}$'s in (1). We first fix $\boldsymbol{\eta}_{jj'}$ to estimate $\boldsymbol{\beta}_j$, then fix $\boldsymbol{\beta}_j$ to estimate $\boldsymbol{\eta}_{jj'}$, and iterate the two steps until convergence. The algorithm can be summarized as follows:

1. Obtain basis-expanded main effect matrices for each covariate, denoted by $\boldsymbol{X}_j$ for $j = 1, \ldots, S$. Normalize $\boldsymbol{X}_j$. Calculate interaction design matrices $\boldsymbol{X}_{jj'}$ from the normalized $\boldsymbol{X}_j$ for $1 \le j \le j' \le S$. Normalize $\boldsymbol{X}_{jj'}$. Orthogonalize $\boldsymbol{X}_j$ and $\boldsymbol{X}_{jj'}$ using QR decomposition and center the response vector $\boldsymbol{y}$. Scale $\boldsymbol{X}_j$ and $\boldsymbol{X}_{jj'}$ to have unit variance.

2. Initialize $\hat{\boldsymbol{\beta}}_j^{(0)}$ for $j = 1, \cdots, S$ and $\hat{\boldsymbol{\eta}}_{jj'}^{(0)}$ for $1 \le j < j' \le S$. Set $m = 1$. A feasible choice for the initialization $\hat{\boldsymbol{\beta}}_j^{(0)}$ and $\hat{\boldsymbol{\eta}}_{jj'}^{(0)}$ can be obtained using the adaptive elastic-net estimator. We use this as the initialization in our implementation.

3. For each j in $1, \cdots, S$, update $\hat{\boldsymbol{\beta}}_j^{(m)}$ via closed-form formula in (6), given $\hat{\boldsymbol{\eta}}_{kj}^{(m-1)}$ and $\hat{\boldsymbol{\beta}}_k^{(m)}$ for $k < j$, and $\hat{\boldsymbol{\eta}}_{jl}^{(m-1)}$ and $\hat{\boldsymbol{\beta}}_l^{(m-1)}$ for $l > j$. A backtracking line search algorithm is followed to guarantee that $\hat{\boldsymbol{\beta}}_j^{(m)}$ leads to a lower value of the objective function (4) than $\hat{\boldsymbol{\beta}}_j^{(m)}$.

4. Given $\hat{\boldsymbol{\beta}}_j^{(m)}$ for $j = 1, \cdots, S$, update the $\hat{\boldsymbol{\eta}}_{jj'}^{(m)}$'s via the closed-form formula in (8). A backtracking line search algorithm is followed to guarantee that the $\hat{\boldsymbol{\eta}}_{jj'}^{(m)}$'s lead to a lower value of the objective function in (7) compared to the $\hat{\boldsymbol{\eta}}_{jj'}^{(m-1)}$'s.

5. Stop if change in the penalized likelihood is less than a pre-specified margin δ, namely
$$|P_n^{(m-1)} - P_n^{(m)}| < \delta.$$
where $P_n^{(m)}$ is the value of (1) evaluated at the $\hat{\boldsymbol{\beta}}_j^{(m)}$'s and $\hat{\boldsymbol{\eta}}_{jj'}^{(m)}$'s.

Remark 2: We note that there is no guarantee that each of the $S+1$ updates decreases the value of penalized least squares criterion since we utilize approximations to the original penalty. We therefore employ a backtracking line search algorithm (Dennis and Schnabel, 1996) to ensure that the penalized least squares criterion monotonically decreases throughout the entire procedure. The maximum amount to move along a given search direction is determined by the Armijo-Goldstein condition (Armijo, 1966).

Remark 3: Steps (3) and (4) in the HiGLASSO algorithm could be easily modified to accommodate objective functions without the least squares criterion. However, closed-form updates may not be avilable, thus requiring one-step gradient descent.

Web Appendix B: Sparsistency proof

B.1. *Notation*

Let $\boldsymbol{X} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_S, \boldsymbol{X}_{12}, \cdots, \boldsymbol{X}_{S,S-1}]$ be the design matrix containing main effect and interaction terms. Without loss of generality, we rearrange the group indices so that the first $s_0 \le S$ groups of predictors have nonzero main effects. Suppose there are i_0 nonzero two-way interaction terms out of at most $s_0(s_0 - 1)/2$ possible pairs under strong heredity constraints.

The HiGLASSO estimator is defined as:

$$\begin{aligned}
&\underset{\boldsymbol{\beta}_j, \boldsymbol{\eta}_{jj'}}{\arg\min} \frac{1}{2}\Big\|\boldsymbol{y} - \sum_{j=1}^{S} \boldsymbol{X}_j\boldsymbol{\beta}_j - \sum_{1\le j<j'\le S} \boldsymbol{X}_{jj'}[\boldsymbol{\eta}_{jj'} \odot (\boldsymbol{\beta}_j \otimes \boldsymbol{\beta}_{j'})]\Big\|_2^2 \\
&+ \lambda_1(n)\sum_{j=1}^{S} w_j(\boldsymbol{\beta}_j)||\boldsymbol{\beta}_j||_2 + \lambda_2(n)\sum_{1\le j<j'\le S} w_{jj'}(\boldsymbol{\eta}_{jj'})||\boldsymbol{\eta}_{jj'}||_2. \qquad (9)
\end{aligned}$$

B.2. *Directional Derivatives of HiGLASSO Objective Function*

Consider the following function

$$f(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_S, \boldsymbol{\eta}_{12}, ..., \boldsymbol{\eta}_{S-1,S}) = \frac{1}{2}\Big\|\boldsymbol{y} - \sum_{j=1}^{S} \boldsymbol{X}_j\boldsymbol{\beta}_j - \sum_{1\le j<j'\le S} \boldsymbol{X}_{jj'}[\boldsymbol{\eta}_{jj'} \odot (\boldsymbol{\beta}_j \otimes \boldsymbol{\beta}_{j'})]\Big\|_2^2$$

First we will calculate the directional derivative in the $\boldsymbol{u}$ direction with respect to $\boldsymbol{\beta}_k$. By definition the directional derivative is given by:

$$\lim_{t\to 0^+} \frac{f(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_{k-1}, \boldsymbol{\beta}_k + t\boldsymbol{u}, \boldsymbol{\beta}_{k+1}, ..., \boldsymbol{\beta}_S, \boldsymbol{\eta}_{12}, ..., \boldsymbol{\eta}_{S-1,S}) - f(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_S, \boldsymbol{\eta}_{12}, ..., \boldsymbol{\eta}_{S-1,S})}{t}$$

$$\begin{aligned}
&f(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_{k-1}, \boldsymbol{\beta}_k + t\boldsymbol{u}, \boldsymbol{\beta}_{k+1}, ..., \boldsymbol{\beta}_S, \boldsymbol{\eta}_{12}, ..., \boldsymbol{\eta}_{S-1,S}) \\
&= \frac{1}{2}\Big\|\boldsymbol{y} - \boldsymbol{X}_k(\boldsymbol{\beta}_k + t\boldsymbol{u}) - \sum_{j\ne k} \boldsymbol{X}_j\boldsymbol{\beta}_j - \sum_{1\le k<j'\le S} \boldsymbol{X}_{kj'}[\boldsymbol{\eta}_{kj'} \odot (\boldsymbol{\beta}_k + t\boldsymbol{u} \otimes \boldsymbol{\beta}_{j'})] \\
&- \sum_{1\le j<k\le S} \boldsymbol{X}_{jk}[\boldsymbol{\eta}_{jk} \odot (\boldsymbol{\beta}_j \otimes \boldsymbol{\beta}_k + t\boldsymbol{u})] - \sum_{1\le j<j'\le S: j,j'\ne k} \boldsymbol{X}_{jj'}[\boldsymbol{\eta}_{jj'} \odot (\boldsymbol{\beta}_j \otimes \boldsymbol{\beta}_{j'})]\Big\|_2^2
\end{aligned}$$

Note that

$$\boldsymbol{\eta}_{jk} \odot (\boldsymbol{\beta}_j \otimes \boldsymbol{\beta}_k + t\boldsymbol{u}) = \boldsymbol{\eta}_{jk} \odot (\boldsymbol{\beta}_j \otimes \boldsymbol{\beta}_k) + t(\boldsymbol{\eta}_{jk} \odot (\boldsymbol{\beta}_j \otimes \boldsymbol{u}))$$

and

$$\boldsymbol{\eta}_{kj'} \odot (\boldsymbol{\beta}_k + t\boldsymbol{u} \otimes \boldsymbol{\beta}_{j'}) = \boldsymbol{\eta}_{kj'} \odot (\boldsymbol{\beta}_k \otimes \boldsymbol{\beta}_{j'}) + t(\boldsymbol{\eta}_{kj'} \odot (\boldsymbol{u} \otimes \boldsymbol{\beta}_{j'}))$$

Thus, the expression becomes

$$\frac{1}{2}\Big\| \boldsymbol{y} - t\boldsymbol{X}_k\boldsymbol{u} - \sum_{j=1}^{S}\boldsymbol{X}_j\boldsymbol{\beta}_j - t\sum_{1\le k<j'\le S}\boldsymbol{X}_{kj'}[\boldsymbol{\eta}_{kj'}\odot(\boldsymbol{u}\otimes\boldsymbol{\beta}_{j'})]$$

$$-t\sum_{1\le j<k\le S}\boldsymbol{X}_{jk}[\boldsymbol{\eta}_{jk}\odot(\boldsymbol{\beta}_j\otimes\boldsymbol{u})] - \sum_{1\le j<j'\le S}\boldsymbol{X}_{jj'}[\boldsymbol{\eta}_{jj'}\odot(\boldsymbol{\beta}_j\otimes\boldsymbol{\beta}_{j'})]\Big\|_2^2$$

Observe that as we take the limit to 0 we get that the terms with a t^2 term go to 0 and the terms without a t cancel with $f(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_S, \boldsymbol{\eta}_{12}, ..., \boldsymbol{\eta}_{S-1,S})$. Therefore, we only need to keep track of the terms that are linear in t. To simplify notation, let

$$\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{y} - \sum_{j=1}^{S}\boldsymbol{X}_j\boldsymbol{\beta}_j - \sum_{1\le j<j'\le S}\boldsymbol{X}_{jj'}[\boldsymbol{\eta}_{jj'}\odot(\boldsymbol{\beta}_j\otimes\boldsymbol{\beta}_{j'})]$$

Then, the expression becomes

$$\frac{1}{2}\Big\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}-t\boldsymbol{X}_k\boldsymbol{u}-t\sum_{1\le k<j'\le S}\boldsymbol{X}_{kj'}[\boldsymbol{\eta}_{kj'}\odot(\boldsymbol{u}\otimes\boldsymbol{\beta}_{j'})]-t\sum_{1\le j<k\le S}\boldsymbol{X}_{jk}[\boldsymbol{\eta}_{jk}\odot(\boldsymbol{\beta}_j\otimes\boldsymbol{u})]\Big\|_2^2$$

Therefore, the directional derivative is:

$$-\left(\boldsymbol{X}_k\boldsymbol{u}+\sum_{1\le k<j'\le S}\boldsymbol{X}_{kj'}[\boldsymbol{\eta}_{kj'}\odot(\boldsymbol{u}\otimes\boldsymbol{\beta}_{j'})]+\sum_{1\le j<k\le S}\boldsymbol{X}_{jk}[\boldsymbol{\eta}_{jk}\odot(\boldsymbol{\beta}_j\otimes\boldsymbol{u})]\right)^{\top}\Big(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\Big)$$

Lastly, from the proof of Theorem 1 in (Pan and Zhao, 2016), we have that the directional derivative of $\lambda_1(n)w_k(\boldsymbol{\beta}_k)\|\boldsymbol{\beta}_k\|_2$ in the $\boldsymbol{u}$ direction evaluated at zero is $\lambda_1(n)$.

Next we will calculate the directional derivative in the $\boldsymbol{u}$ direction with respect to $\boldsymbol{\eta}_{kk'}$. By definition the directional derivative is given by:

$$\lim_{t\to 0^+}\frac{f(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_S, \boldsymbol{\eta}_{12}, ..., \boldsymbol{\eta}_{kk'}+t\boldsymbol{u}, ..., \boldsymbol{\eta}_{S-1,S}) - f(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_S, \boldsymbol{\eta}_{12}, ..., \boldsymbol{\eta}_{S-1,S})}{t},$$

$$f(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_S, \boldsymbol{\eta}_{12}, ..., \boldsymbol{\eta}_{kk'}+t\boldsymbol{u}, ..., \boldsymbol{\eta}_{S-1,S}) = \frac{1}{2}\Big\|\boldsymbol{y}-\sum_{j=1}^{S}\boldsymbol{X}_j\boldsymbol{\beta}_j-\boldsymbol{X}_{kk'}[(\boldsymbol{\eta}_{kk'}+t\boldsymbol{u})\odot(\boldsymbol{\beta}_k\otimes\boldsymbol{\beta}_{k'})]$$

$$-\sum_{1\le j<j'\le S:(j,j')\ne(k,k')}\boldsymbol{X}_{jj'}[\boldsymbol{\eta}_{jj'}\odot(\boldsymbol{\beta}_j\otimes\boldsymbol{\beta}_{j'})]\Big\|_2^2$$

Again, note that

$$(\boldsymbol{\eta}_{kk'}+t\boldsymbol{u})\odot(\boldsymbol{\beta}_k\otimes\boldsymbol{\beta}_{k'}) = \boldsymbol{\eta}_{kk'}\odot(\boldsymbol{\beta}_k\otimes\boldsymbol{\beta}_{k'}) + t(\boldsymbol{u}\odot(\boldsymbol{\beta}_k\otimes\boldsymbol{\beta}_{k'}))$$

Thus the expression becomes

$$\frac{1}{2}\Big\|\boldsymbol{y} - \sum_{j=1}^{S}\boldsymbol{X}_j\boldsymbol{\beta}_j - \sum_{1\le j<j'\le S}\boldsymbol{X}_{jj'}[\boldsymbol{\eta}_{jj'} \odot (\boldsymbol{\beta}_j \otimes \boldsymbol{\beta}_{j'})] - t\boldsymbol{X}_{kk'}[\boldsymbol{u} \odot (\boldsymbol{\beta}_k \otimes \boldsymbol{\beta}_{k'})]\Big\|_2^2$$

$$= \frac{1}{2}\Big\|\boldsymbol{y} - \boldsymbol{X\theta} - t\boldsymbol{X}_{kk'}[\boldsymbol{u} \odot (\boldsymbol{\beta}_k \otimes \boldsymbol{\beta}_{k'})]\Big\|_2^2$$

Following the same argument as above the directional derivative of $\boldsymbol{\beta}_k$, as we take the limit to 0 we get that the terms with a t^2 term go to 0 and the terms without a t cancel with $f(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_S, \boldsymbol{\eta}_{12}, ..., \boldsymbol{\eta}_{S-1,S})$. Therefore, we only need to keep track of the terms that are linear in t. That is, the directional derivative is,

$$-\Big(\boldsymbol{X}_{kk'}[\boldsymbol{u} \odot (\boldsymbol{\beta}_k \otimes \boldsymbol{\beta}_{k'})]\Big)^{\top}\Big(\boldsymbol{y} - \boldsymbol{X\theta}\Big)$$

Again, from the proof of Theorem 1 in (Pan and Zhao, 2016), we have that the directional derivative of $\lambda_2(n)w_{kk'}(\boldsymbol{\eta}_{kk'})\|\boldsymbol{\eta}_{kk'}\|_2$ in the $\boldsymbol{u}$ direction evaluated at zero is $\lambda_2(n)$.

B.3. *Derivative of HiGLASSO Objective Function*

First we calculate the derivative with respect to $\boldsymbol{\beta}_k$:

$$\frac{\partial}{\partial\boldsymbol{\beta}_k}f(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_S, \boldsymbol{\eta}_{12}, ..., \boldsymbol{\eta}_{S-1,S}) = \frac{\partial}{\partial\boldsymbol{\beta}_k}\left[\frac{1}{2}\Big\|\boldsymbol{y} - \sum_{j=1}^{S}\boldsymbol{X}_j\boldsymbol{\beta}_j - \sum_{1\le j<j'\le S}\boldsymbol{X}_{jj'}[\boldsymbol{\eta}_{jj'} \odot (\boldsymbol{\beta}_j \otimes \boldsymbol{\beta}_{j'})]\Big\|_2^2\right]$$

$$= \left(\frac{\partial}{\partial\boldsymbol{\beta}_k}\left[\boldsymbol{y} - \sum_{j=1}^{S}\boldsymbol{X}_j\boldsymbol{\beta}_j - \sum_{1\le j<j'\le S}\boldsymbol{X}_{jj'}[\boldsymbol{\eta}_{jj'} \odot (\boldsymbol{\beta}_j \otimes \boldsymbol{\beta}_{j'})]\right]\right)^{\top}\Big(\boldsymbol{y} - \boldsymbol{X\theta}\Big)$$

$$= \left(\frac{\partial}{\partial\boldsymbol{\beta}_k}\left[-\boldsymbol{X}_k\boldsymbol{\beta}_k - \sum_{1\le k<j'\le S}\boldsymbol{X}_{kj'}[\boldsymbol{\eta}_{kj'} \odot (\boldsymbol{\beta}_k \otimes \boldsymbol{\beta}_{j'})] - \sum_{1\le j<k\le S}\boldsymbol{X}_{jk}[\boldsymbol{\eta}_{jk} \odot (\boldsymbol{\beta}_j \otimes \boldsymbol{\beta}_k)]\right]\right)^{\top}\Big(\boldsymbol{y} - \boldsymbol{X\theta}\Big)$$

$$= -\left[\boldsymbol{X}_k + \sum_{1\le k<j'\le S}\boldsymbol{X}_{kj'}\frac{\partial}{\partial\boldsymbol{\beta}_k}[\boldsymbol{\eta}_{kj'} \odot (\boldsymbol{\beta}_k \otimes \boldsymbol{\beta}_{j'})] + \sum_{1\le j<k\le S}\boldsymbol{X}_{jk}\frac{\partial}{\partial\boldsymbol{\beta}_k}[\boldsymbol{\eta}_{jk} \odot (\boldsymbol{\beta}_j \otimes \boldsymbol{\beta}_k)]\right]^{\top}\Big(\boldsymbol{y} - \boldsymbol{X\theta}\Big)$$

$$= -\left[\boldsymbol{X}_k + \sum_{1\le k<j'\le S}\boldsymbol{X}_{kj'}\left[\text{diag}\big(\boldsymbol{\eta}_{kj'}\big)\frac{\partial}{\partial\boldsymbol{\beta}_k}(\boldsymbol{\beta}_k \otimes \boldsymbol{\beta}_{j'})\right]\right.$$

$$+\sum_{1\le j<k\le S} \boldsymbol{X}_{jk}\Big[\mathrm{diag}\big(\boldsymbol{\eta}_{jk}\big)\frac{\partial}{\partial \boldsymbol{\beta}_k}(\boldsymbol{\beta}_j\otimes\boldsymbol{\beta}_k)\Big]\Big]^\top\Big(\boldsymbol{y}-\boldsymbol{X\theta}\Big)$$

$$=-\Big[\boldsymbol{X}_k+\sum_{1\le k<j'\le S}\boldsymbol{X}_{kj'}\Big[\mathrm{diag}\big(\boldsymbol{\eta}_{kj'}\big)(\boldsymbol{I}\otimes\boldsymbol{\beta}_{j'})\Big]+\sum_{1\le j<k\le S}\boldsymbol{X}_{jk}\Big[\mathrm{diag}\big(\boldsymbol{\eta}_{jk}\big)(\boldsymbol{\beta}_j\otimes\boldsymbol{I})\Big]\Big]^\top\Big(\boldsymbol{y}-\boldsymbol{X\theta}\Big)$$

The derivative of the penalty function is:

$$\frac{\partial}{\partial\boldsymbol{\beta}_k}w_k(\boldsymbol{\beta}_k)||\boldsymbol{\beta}_k||_2=\frac{\partial}{\partial\boldsymbol{\beta}_k}\exp\bigg(-\frac{||\boldsymbol{\beta}_k||_\infty}{\sigma(n)}\bigg)||\boldsymbol{\beta}_k||_2$$

$$=||\boldsymbol{\beta}_k||_2\frac{\partial}{\partial\boldsymbol{\beta}_k}\exp\bigg(-\frac{||\boldsymbol{\beta}_k||_\infty}{\sigma(n)}\bigg)+\exp\bigg(-\frac{||\boldsymbol{\beta}_k||_\infty}{\sigma(n)}\bigg)\frac{\partial}{\partial\boldsymbol{\beta}_k}||\boldsymbol{\beta}_k||_2$$

$$=||\boldsymbol{\beta}_k||_2\bigg(-\frac{1}{\sigma(n)}\exp\bigg(-\frac{||\boldsymbol{\beta}_k||_\infty}{\sigma(n)}\bigg)\sum_{l=1}^{p_k}\mathrm{sign}(\beta_{kl})\vec{\boldsymbol{e}}_l I\Big(\beta_{kl}=||\boldsymbol{\beta}_k||_\infty\Big)\bigg)$$

$$+\exp\bigg(-\frac{||\boldsymbol{\beta}_k||_\infty}{\sigma(n)}\bigg)\Big(||\boldsymbol{\beta}_k||_2\Big)^{-1}\boldsymbol{\beta}_k,$$

where $\vec{\boldsymbol{e}}_l$ is the standard basis vector of dimension p_k such that the l-th component is equal to 1.

Next we calculate the derivative with respect to $\boldsymbol{\eta}_{kk'}$:

$$\frac{\partial}{\partial\boldsymbol{\eta}_{kk'}}f(\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_S,\boldsymbol{\eta}_{12},...,\boldsymbol{\eta}_{S-1,S})$$

$$=\frac{\partial}{\partial\boldsymbol{\eta}_{kk'}}\bigg[\frac{1}{2}\Big\|\boldsymbol{y}-\sum_{j=1}^{S}\boldsymbol{X}_j\boldsymbol{\beta}_j-\sum_{1\le j<j'\le S}\boldsymbol{X}_{jj'}[\boldsymbol{\eta}_{jj'}\odot(\boldsymbol{\beta}_j\otimes\boldsymbol{\beta}_{j'})]\Big\|_2^2\bigg]$$

$$=\bigg(\frac{\partial}{\partial\boldsymbol{\eta}_{kk'}}\bigg[\boldsymbol{y}-\sum_{j=1}^{S}\boldsymbol{X}_j\boldsymbol{\beta}_j-\sum_{1\le j<j'\le S}\boldsymbol{X}_{jj'}[\boldsymbol{\eta}_{jj'}\odot(\boldsymbol{\beta}_j\otimes\boldsymbol{\beta}_{j'})]\bigg]\bigg)^\top\Big(\boldsymbol{y}-\boldsymbol{X\theta}\Big)$$

$$=-\bigg(\frac{\partial}{\partial\boldsymbol{\eta}_{kk'}}\Big[\boldsymbol{X}_{kk'}[\boldsymbol{\eta}_{kk'}\odot(\boldsymbol{\beta}_k\otimes\boldsymbol{\beta}_{k'})]\Big]\bigg)^\top\Big(\boldsymbol{y}-\boldsymbol{X\theta}\Big)$$

$$=-\bigg[\boldsymbol{X}_{kk'}\Big[\frac{\partial}{\partial\boldsymbol{\eta}_{kk'}}\boldsymbol{\eta}_{kk'}\odot\mathrm{diag}(\boldsymbol{\beta}_k\otimes\boldsymbol{\beta}_{k'})\Big]\bigg]^\top\Big(\boldsymbol{y}-\boldsymbol{X\theta}\Big)$$

$$=-\bigg(\boldsymbol{X}_{kk'}\Big[\boldsymbol{I}\odot\mathrm{diag}(\boldsymbol{\beta}_k\otimes\boldsymbol{\beta}_{k'})\Big]\bigg)^\top\Big(\boldsymbol{y}-\boldsymbol{X\theta}\Big)$$

$$= -\left(\boldsymbol{X}_{kk'}\left[\text{diag}(\boldsymbol{\beta}_k \otimes \boldsymbol{\beta}_{k'})\right]\right)^{\top}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\right)$$

The derivative of the penalty function is:

$$\frac{\partial}{\partial \boldsymbol{\eta}_{kk'}} w_{kk'}(\boldsymbol{\eta}_{kk'})||\boldsymbol{\eta}_{kk'}||_2 = \frac{\partial}{\partial \boldsymbol{\eta}_{kk'}} \exp\left(-\frac{||\boldsymbol{\eta}_{kk'}||_\infty}{\sigma(n)}\right)||\boldsymbol{\eta}_{kk'}||_2$$

$$= ||\boldsymbol{\eta}_{kk'}||_2 \frac{\partial}{\partial \boldsymbol{\eta}_{kk'}} \exp\left(-\frac{||\boldsymbol{\eta}_{kk'}||_\infty}{\sigma(n)}\right) + \exp\left(-\frac{||\boldsymbol{\eta}_{kk'}||_\infty}{\sigma(n)}\right)\frac{\partial}{\partial \boldsymbol{\eta}_{kk'}}||\boldsymbol{\eta}_{kk'}||_2$$

$$= ||\boldsymbol{\eta}_{kk'}||_2\left(-\frac{1}{\sigma(n)}\exp\left(-\frac{||\boldsymbol{\eta}_{kk'}||_\infty}{\sigma(n)}\right)\sum_{l=1}^{p_k p_{k'}} \text{sign}(\eta_{kk'l})\vec{\boldsymbol{e}}_l I\left(\eta_{kk'l} = ||\boldsymbol{\eta}_{kk'}||_\infty\right)\right)$$

$$+ \exp\left(-\frac{||\boldsymbol{\eta}_{kk'}||_\infty}{\sigma(n)}\right)\left(||\boldsymbol{\eta}_{kk'}||_2\right)^{-1}\boldsymbol{\eta}_{kk'},$$

where $\vec{\boldsymbol{e}}_l$ is the standard basis vector of dimension $p_k p_{k'}$ such that the l-th component is equal to 1.

B.4. *Sparsistency Proof*

The proof closely follows the proof of Theorem 1 in (Pan and Zhao, 2016). Define the HiGLASSO estimator of a re-parameterized version of (9) such that only the covariates corresponding to the non-zero coefficient set are included:

$$\arg\min_{\boldsymbol{\theta}_{\mathcal{P}}}\Bigg\{||\boldsymbol{y} - \boldsymbol{X}_{\mathcal{P}}\boldsymbol{\theta}_{\mathcal{P}}||_2^2 + \lambda_1(n)\sum_{j\in\mathcal{P}_1} w_j(\boldsymbol{\theta}_j)||\boldsymbol{\theta}_j||_2 + \lambda_2(n)\sum_{(j,j')\in\mathcal{P}_2} w_{jj'}(\boldsymbol{\eta}_{jj'})||\boldsymbol{\eta}_{jj'}||_2\Bigg\}. \quad (10)$$

Let $\tilde{\boldsymbol{\theta}}_{\mathcal{P}}$ be the solution to (10). From the assumptions of the Theorem we have that

$$\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{y} = \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}_{\mathcal{P}}\boldsymbol{\theta}_{\mathcal{P}} + \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\epsilon}$$
$$= \left[E\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}_{\mathcal{P}}\right) + O_p\left(n^{-1/2}\right)\right]\boldsymbol{\theta}_{\mathcal{P}} + O_p\left(n^{-1/2}\right)$$
$$= E\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}_{\mathcal{P}}\right)\boldsymbol{\theta}_{\mathcal{P}} + O_p\left(n^{-1/2}\right),$$

which implies that

$$\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{y} - \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}_{\mathcal{P}}\tilde{\boldsymbol{\theta}}_{\mathcal{P}} = E\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}_{\mathcal{P}}\right)(\boldsymbol{\theta}_{\mathcal{P}} - \tilde{\boldsymbol{\theta}}_{\mathcal{P}}) + O_p\left(n^{-1/2}\right). \quad (11)$$

(11) can be decomposed as

$$\frac{1}{n}\boldsymbol{X}_{\mathcal{P}}^{\top}\boldsymbol{y} - \frac{1}{n}\boldsymbol{X}_{\mathcal{P}}^{\top}\boldsymbol{X}_{\mathcal{P}}\tilde{\boldsymbol{\theta}}_{\mathcal{P}} = E\left(\frac{1}{n}\boldsymbol{X}_{\mathcal{P}}^{\top}\boldsymbol{X}_{\mathcal{P}}\right)(\boldsymbol{\theta}_{\mathcal{P}} - \tilde{\boldsymbol{\theta}}_{\mathcal{P}}) + O_p(n^{-1/2}) \tag{12}$$

$$\frac{1}{n}\boldsymbol{X}_{\mathcal{P}^{\mathsf{c}}}^{\top}\boldsymbol{y} - \frac{1}{n}\boldsymbol{X}_{\mathcal{P}^{\mathsf{c}}}^{\top}\boldsymbol{X}_{\mathcal{P}}\tilde{\boldsymbol{\theta}}_{\mathcal{P}} = E\left(\frac{1}{n}\boldsymbol{X}_{\mathcal{P}^{\mathsf{c}}}^{\top}\boldsymbol{X}_{\mathcal{P}}\right)(\boldsymbol{\theta}_{\mathcal{P}} - \tilde{\boldsymbol{\theta}}_{\mathcal{P}}) + O_p(n^{-1/2}) \tag{13}$$

From (12) we get

$$\boldsymbol{\theta}_{\mathcal{P}} - \tilde{\boldsymbol{\theta}}_{\mathcal{P}} = E^{-1}\left(\frac{1}{n}\boldsymbol{X}_{\mathcal{P}}^{\top}\boldsymbol{X}_{\mathcal{P}}\right)\frac{1}{n}\boldsymbol{X}_{\mathcal{P}}^{\top}(\boldsymbol{y} - \boldsymbol{X}_{\mathcal{P}}\tilde{\boldsymbol{\theta}}_{\mathcal{P}}) + O_p(n^{-1/2})$$

and substituting into (13) we obtain

$$\begin{aligned}&\frac{1}{n}\boldsymbol{X}_{\mathcal{P}^{\mathsf{c}}}^{\top}(\boldsymbol{y} - \boldsymbol{X}_{\mathcal{P}}\tilde{\boldsymbol{\theta}}_{\mathcal{P}})\\ &= E\left(\frac{1}{n}\boldsymbol{X}_{\mathcal{P}^{\mathsf{c}}}^{\top}\boldsymbol{X}_{\mathcal{P}}\right)E^{-1}\left(\frac{1}{n}\boldsymbol{X}_{\mathcal{P}}^{\top}\boldsymbol{X}_{\mathcal{P}}\right)\frac{1}{n}\boldsymbol{X}_{\mathcal{P}}^{\top}(\boldsymbol{y} - \boldsymbol{X}_{\mathcal{P}}\tilde{\boldsymbol{\theta}}_{\mathcal{P}}) + O_p(n^{-1/2}).\end{aligned}$$

Multiplying both sides by n/a_n, we get

$$\begin{aligned}&\frac{n}{a_n}\left(\frac{1}{n}\boldsymbol{X}_{\mathcal{P}^{\mathsf{c}}}^{\top}(\boldsymbol{y} - \boldsymbol{X}_{\mathcal{P}}\tilde{\boldsymbol{\theta}}_{\mathcal{P}})\right)\\ &= E\left(\frac{1}{n}\boldsymbol{X}_{\mathcal{P}^{\mathsf{c}}}^{\top}\boldsymbol{X}_{\mathcal{P}}\right)E^{-1}\left(\frac{1}{n}\boldsymbol{X}_{\mathcal{P}}^{\top}\boldsymbol{X}_{\mathcal{P}}\right)\frac{1}{a_n}\boldsymbol{X}_{\mathcal{P}}^{\top}(\boldsymbol{y} - \boldsymbol{X}_{\mathcal{P}}\tilde{\boldsymbol{\theta}}_{\mathcal{P}}) + O_p\left(\frac{\sqrt{n}}{a_n}\right).\end{aligned}$$

Therefore, when $b_n \to 0$, $a_n/\sqrt{n} \to \infty$, and $a_n/n \to 0$ we have

$$\frac{n}{a_n}\left\|\frac{1}{n}\boldsymbol{X}_{\mathcal{P}^{\mathsf{c}}}^{\top}(\boldsymbol{y} - \boldsymbol{X}_{\mathcal{P}}\tilde{\boldsymbol{\theta}}_{\mathcal{P}})\right\|_2 \to_p 0,$$

which also implies that

$$\frac{n}{\lambda_1(n)}\left\|\frac{1}{n}\boldsymbol{X}_{[k]}^{\top}(\boldsymbol{y} - \boldsymbol{X}_{\mathcal{P}}\tilde{\boldsymbol{\theta}}_{\mathcal{P}})\right\|_2 \to_p 0, \;\; \forall k \in \mathcal{P}_1^{\mathsf{c}}$$

$$\frac{n}{\lambda_2(n)}\left\|\frac{1}{n}\boldsymbol{X}_{kk'}^{\top}(\boldsymbol{y} - \boldsymbol{X}_{\mathcal{P}}\tilde{\boldsymbol{\theta}}_{\mathcal{P}})\right\|_2 \to_p 0, \;\; \forall (k,k') \in \mathcal{P}_2^{\mathsf{c}}$$

where

$$\boldsymbol{X}_{[k]} = \begin{pmatrix}\boldsymbol{X}_k, & \boldsymbol{X}_{k,k+1}, & \cdots & \boldsymbol{X}_{k,S}, & \boldsymbol{X}_{1,k}, & \cdots & \boldsymbol{X}_{k-1,k}\end{pmatrix}$$

is the submatrix of the design matrix corresponding to the kth covariate. These two convergence in probability statements imply that

$$P\left(\forall k \in \mathcal{P}_1^{\mathsf{c}}, \frac{nB_1}{\lambda_1(n)}\left\|\frac{1}{n}\boldsymbol{X}_{[k]}^{\top}(\boldsymbol{y} - \boldsymbol{X}_{\mathcal{P}}\tilde{\boldsymbol{\theta}}_{\mathcal{P}})\right\|_2 \leq 1\right) \to 1 \tag{14}$$

$$P\left(\forall(k,k')\in\mathcal{P}_2^{\mathsf{c}}, \frac{nB_2}{\lambda_2(n)}\left\|\frac{1}{n}\boldsymbol{X}_{kk'}^\top\big(\boldsymbol{y}-\boldsymbol{X}_{\mathcal{P}}\tilde{\boldsymbol{\theta}}_{\mathcal{P}}\big)\right\|_2\le 1\right)\to 1 \tag{15}$$

for any finite constants B_1 and B_2.

Define $\tilde{\boldsymbol{\theta}}$ as the concatenation of $\tilde{\boldsymbol{\theta}}_{\mathcal{P}_1}$, a vector of zeros with length equal to the number of columns in $\boldsymbol{X}$ corresponding to $\mathcal{P}_1^{\mathsf{c}}$, $\tilde{\boldsymbol{\theta}}_{\mathcal{P}_2}$, and a vector of zeros with length equal to the number of columns in $\boldsymbol{X}$ corresponding to $\mathcal{P}_2^{\mathsf{c}}$. The assumption that the L_2 norm of the HiGLASSO estimator is uniformly bounded for all n coupled with (14) and (15) imply that with probability approaching one

$$\frac{1}{n}\tilde{\boldsymbol{C}}_{[k]}^\top\boldsymbol{X}_{[k]}^\top\big(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\theta}}\big)=\frac{\lambda_1(n)}{n}D_k(\tilde{\boldsymbol{\beta}}_k),\ \ \forall k\in\mathcal{P}_1 \tag{16}$$

$$\frac{1}{n}\mathrm{diag}(\tilde{\boldsymbol{\beta}}_k\otimes\tilde{\boldsymbol{\beta}}_{k'})\boldsymbol{X}_{kk'}^\top\big(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\theta}}\big)=\frac{\lambda_2(n)}{n}D_{kk'}(\tilde{\boldsymbol{\eta}}_{kk'}),\ \ \forall(k,k')\in\mathcal{P}_2 \tag{17}$$

$$\left\|\frac{1}{n}\tilde{\boldsymbol{C}}_{[k]}^\top\boldsymbol{X}_{[k]}^\top\big(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\theta}}\big)\right\|_2\le\frac{\lambda_1(n)}{n},\ \ \forall k\in\mathcal{P}_1^{\mathsf{c}} \tag{18}$$

$$\left\|\frac{1}{n}\mathrm{diag}(\tilde{\boldsymbol{\beta}}_k\otimes\tilde{\boldsymbol{\beta}}_{k'})\boldsymbol{X}_{kk'}^\top\big(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\theta}}\big)\right\|_2\le\frac{\lambda_2(n)}{n},\ \ \forall(k,k')\in\mathcal{P}_2^{\mathsf{c}} \tag{19}$$

where

$$\tilde{\boldsymbol{C}}_{[k]}=\begin{pmatrix}\boldsymbol{I}_{p_k\times p_k}\\ \mathrm{diag}(\tilde{\boldsymbol{\eta}}_{k,k+1})(\boldsymbol{I}_{p_k\times p_k}\otimes\tilde{\boldsymbol{\beta}}_{k+1})\\ \vdots\\ \mathrm{diag}(\tilde{\boldsymbol{\eta}}_{k,S})(\boldsymbol{I}_{p_k\times p_k}\otimes\tilde{\boldsymbol{\beta}}_S)\\ \mathrm{diag}(\tilde{\boldsymbol{\eta}}_{1,k})(\tilde{\boldsymbol{\beta}}_1\otimes\boldsymbol{I}_{p_k\times p_k})\\ \vdots\\ \mathrm{diag}(\tilde{\boldsymbol{\eta}}_{k-1,k})(\tilde{\boldsymbol{\beta}}_{k-1}\otimes\boldsymbol{I}_{p_k\times p_k})\end{pmatrix}$$

$$D_k(\tilde{\boldsymbol{\beta}}_k)=\frac{\partial}{\partial\boldsymbol{\beta}_k}w_k(\boldsymbol{\beta}_k)||\boldsymbol{\beta}_k||_2\bigg|_{\boldsymbol{\beta}_k=\tilde{\boldsymbol{\beta}}_k}$$

$$D_{kk'}(\tilde{\boldsymbol{\eta}}_{kk'})=\frac{\partial}{\partial\boldsymbol{\eta}_{kk'}}w_{kk'}(\boldsymbol{\eta}_{kk'})||\boldsymbol{\eta}_{kk'}||_2\bigg|_{\boldsymbol{\eta}_{kk'}=\tilde{\boldsymbol{\eta}}_{kk'}}$$

The directional derivative with respect to $\boldsymbol{\beta}_k$ in the u direction of (9) is

$$-\left(\boldsymbol{X}_k\boldsymbol{u}+\sum_{1\le k<j'\le S}\boldsymbol{X}_{kj'}[\boldsymbol{\eta}_{kj'}\odot(\boldsymbol{u}\otimes\boldsymbol{\beta}_{j'})]+\sum_{1\le j<k\le S}\boldsymbol{X}_{jk}[\boldsymbol{\eta}_{jk}\odot(\boldsymbol{\beta}_j\otimes\boldsymbol{u})]\right)^\top\times$$

$$\big(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\big) + \lambda_1(n)$$

$$= -\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{\eta}_{k,k+1} \odot (\boldsymbol{u} \otimes \boldsymbol{\beta}_{k+1}) \\ \vdots \\ \boldsymbol{\eta}_{k,S} \odot (\boldsymbol{u} \otimes \boldsymbol{\beta}_S) \\ \boldsymbol{\eta}_{1,k} \odot (\boldsymbol{\beta}_1 \otimes \boldsymbol{u}) \\ \vdots \\ \boldsymbol{\eta}_{k-1,k} \odot (\boldsymbol{\beta}_{k-1} \otimes \boldsymbol{u}) \end{pmatrix}^\top \boldsymbol{X}_{[k]}^\top \big(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\big) + \lambda_1(n).$$

For $\tilde{\boldsymbol{\beta}}_j$ and $\tilde{\boldsymbol{\eta}}_{jj'}$ to be the minimizer's of (9), we need

$$-\begin{pmatrix} \boldsymbol{u} \\ \tilde{\boldsymbol{\eta}}_{k,k+1} \odot (\boldsymbol{u} \otimes \tilde{\boldsymbol{\beta}}_{k+1}) \\ \vdots \\ \tilde{\boldsymbol{\eta}}_{k,S} \odot (\boldsymbol{u} \otimes \tilde{\boldsymbol{\beta}}_S) \\ \tilde{\boldsymbol{\eta}}_{1,k} \odot (\tilde{\boldsymbol{\beta}}_1 \otimes \boldsymbol{u}) \\ \vdots \\ \tilde{\boldsymbol{\eta}}_{k-1,k} \odot (\tilde{\boldsymbol{\beta}}_{k-1} \otimes \boldsymbol{u}) \end{pmatrix}^\top \boldsymbol{X}_{[k]}^\top \big(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\big) + \lambda_1(n) \geq 0,$$

for all p_k dimensional unit vectors $\boldsymbol{u}$. To verify this we must substitute the negative normalized gradient in for $\boldsymbol{u}$, and see when the inequality holds. The negative normalized gradient is given by

$$\boldsymbol{u}^* = \frac{\big(\boldsymbol{X}_{[k]}\boldsymbol{C}_{[k]}\big)^\top \big(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\big)}{\Big\|\big(\boldsymbol{X}_{[k]}\boldsymbol{C}_{[k]}\big)^\top \big(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\big)\Big\|_2}.$$

Then we have that

$$\boldsymbol{u}^{*\top} \boldsymbol{X}_k^\top \big(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\big)$$

$$= \frac{\big(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\big)^\top \boldsymbol{X}_{[k]}\boldsymbol{C}_{[k]}\boldsymbol{I}_{p_k \times p_k}^\top \boldsymbol{X}_k^\top \big(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\big)}{\Big\|\big(\boldsymbol{X}_{[k]}\boldsymbol{C}_{[k]}\big)^\top \big(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\big)\Big\|_2}$$

$$\Big[\tilde{\boldsymbol{\eta}}_{kj'}^\top \odot \big(\boldsymbol{u}^{*\top} \otimes \tilde{\boldsymbol{\beta}}_{j'}^\top\big)\Big] \boldsymbol{X}_{kj'}^\top \big(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\big)$$

$$= \frac{\Big[\tilde{\boldsymbol{\eta}}_{kj'}^\top \odot \Big(\Big(\big(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\big)^\top \boldsymbol{X}_{[k]}\boldsymbol{C}_{[k]}\Big) \otimes \tilde{\boldsymbol{\beta}}_{j'}^\top\Big)\Big] \boldsymbol{X}_{kj'}^\top \big(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\big)}{\Big\|\big(\boldsymbol{X}_{[k]}\boldsymbol{C}_{[k]}\big)^\top \big(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\big)\Big\|_2}$$

$$= \frac{\big(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\big)^\top \boldsymbol{X}_{[k]}\boldsymbol{C}_{[k]}\Big(\boldsymbol{I}_{p_k \times p_k} \otimes \tilde{\boldsymbol{\beta}}_{j'}^\top\Big) \mathrm{diag}\big(\tilde{\boldsymbol{\eta}}_{kj'}\big) \boldsymbol{X}_{kj'}^\top \big(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\big)}{\Big\|\big(\boldsymbol{X}_{[k]}\boldsymbol{C}_{[k]}\big)^\top \big(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\big)\Big\|_2}$$

$$\left[\tilde{\boldsymbol{\eta}}_{jk}^{\top} \odot \left(\tilde{\boldsymbol{\beta}}_j^{\top} \otimes \boldsymbol{u}^{*\top}\right)\right] \boldsymbol{X}_{jk}^{\top}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})$$

$$= \frac{\left[\tilde{\boldsymbol{\eta}}_{jk}^{\top} \odot \left(\tilde{\boldsymbol{\beta}}_j^{\top} \otimes \left(\left(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\right)^{\top} \boldsymbol{X}_{[k]} \boldsymbol{C}_{[k]}\right)\right)\right] \boldsymbol{X}_{jk}^{\top}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})}{\left\|\left(\boldsymbol{X}_{[k]} \boldsymbol{C}_{[k]}\right)^{\top}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})\right\|_2}$$

$$= \frac{\left(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\right)^{\top} \boldsymbol{X}_{[k]} \boldsymbol{C}_{[k]} \left(\tilde{\boldsymbol{\beta}}_j^{\top} \otimes \boldsymbol{I}_{p_k \times p_k}\right) \operatorname{diag}(\tilde{\boldsymbol{\eta}}_{jk}) \boldsymbol{X}_{jk}^{\top}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})}{\left\|\left(\boldsymbol{X}_{[k]} \boldsymbol{C}_{[k]}\right)^{\top}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})\right\|_2}$$

Substituting this result in, we get:

$$- \begin{pmatrix} \boldsymbol{u}^* \\ \tilde{\boldsymbol{\eta}}_{k,k+1} \odot (\boldsymbol{u}^* \otimes \tilde{\boldsymbol{\beta}}_{k+1}) \\ \vdots \\ \tilde{\boldsymbol{\eta}}_{k,S} \odot (\boldsymbol{u}^* \otimes \tilde{\boldsymbol{\beta}}_S) \\ \tilde{\boldsymbol{\eta}}_{1,k} \odot (\tilde{\boldsymbol{\beta}}_1 \otimes \boldsymbol{u}^*) \\ \vdots \\ \tilde{\boldsymbol{\eta}}_{k-1,k} \odot (\tilde{\boldsymbol{\beta}}_{k-1} \otimes \boldsymbol{u}^*) \end{pmatrix}^{\top} \boldsymbol{X}_{[k]}^{\top}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})$$

$$= -\frac{\left(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\right)^{\top} \boldsymbol{X}_{[k]} \boldsymbol{C}_{[k]} \boldsymbol{I}_{p_k \times p_k}^{\top} \boldsymbol{X}_k^{\top}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})}{\left\|\left(\boldsymbol{X}_{[k]} \boldsymbol{C}_{[k]}\right)^{\top}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})\right\|_2}$$

$$- \sum_{j'>k} \left[\frac{\left(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\right)^{\top} \boldsymbol{X}_{[k]} \boldsymbol{C}_{[k]} \left(\boldsymbol{I}_{p_k \times p_k} \otimes \tilde{\boldsymbol{\beta}}_{j'}^{\top}\right) \operatorname{diag}(\tilde{\boldsymbol{\eta}}_{kj'}) \boldsymbol{X}_{kj'}^{\top}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})}{\left\|\left(\boldsymbol{X}_{[k]} \boldsymbol{C}_{[k]}\right)^{\top}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})\right\|_2}\right]$$

$$- \sum_{j<k} \left[\frac{\left(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\right)^{\top} \boldsymbol{X}_{[k]} \boldsymbol{C}_{[k]} \left(\tilde{\boldsymbol{\beta}}_j^{\top} \otimes \boldsymbol{I}_{p_k \times p_k}\right) \operatorname{diag}(\tilde{\boldsymbol{\eta}}_{jk}) \boldsymbol{X}_{jk}^{\top}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})}{\left\|\left(\boldsymbol{X}_{[k]} \boldsymbol{C}_{[k]}\right)^{\top}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})\right\|_2}\right]$$

$$= -\frac{\left(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}\right)^{\top} \boldsymbol{X}_{[k]} \boldsymbol{C}_{[k]} \boldsymbol{C}_{[k]}^{\top} \boldsymbol{X}_{[k]}^{\top}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})}{\left\|\left(\boldsymbol{X}_{[k]} \boldsymbol{C}_{[k]}\right)^{\top}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})\right\|_2}$$

$$= -\frac{\left\|\left(\boldsymbol{X}_{[k]} \boldsymbol{C}_{[k]}\right)^{\top}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})\right\|_2^2}{\left\|\left(\boldsymbol{X}_{[k]} \boldsymbol{C}_{[k]}\right)^{\top}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})\right\|_2}$$

$$= -\left\|\left(\boldsymbol{X}_{[k]} \boldsymbol{C}_{[k]}\right)^{\top}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})\right\|_2.$$

Therefore, for $\tilde{\boldsymbol{\beta}}_j$ and $\tilde{\boldsymbol{\eta}}_{jj'}$ to be the minimizer's of (9), we need

$$-\left\|\left(\boldsymbol{X}_{[k]}\boldsymbol{C}_{[k]}\right)^\top\left(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\theta}}\right)\right\|_2+\lambda_1(n)\geq 0,$$

which implies that

$$\left\|\frac{1}{n}\boldsymbol{C}_{[k]}^\top\boldsymbol{X}_{[k]}^\top\left(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\theta}}\right)\right\|_2\leq\frac{\lambda_1(n)}{n}. \tag{20}$$

The directional derivative with respect to $\boldsymbol{\eta}_{kk'}$ in the u direction of (9) is

$$-\Big(\boldsymbol{X}_{kk'}[\boldsymbol{u}\odot(\boldsymbol{\beta}_k\otimes\boldsymbol{\beta}_{k'})]\Big)^\top\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\right)+\lambda_2(n)$$

For $\tilde{\boldsymbol{\beta}}_j$ and $\tilde{\boldsymbol{\eta}}_{jj'}$ to be the minimizer's of (9), we need

$$-\Big(\boldsymbol{X}_{kk'}[\boldsymbol{u}\odot(\tilde{\boldsymbol{\beta}}_k\otimes\tilde{\boldsymbol{\beta}}_{k'})]\Big)^\top\left(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\theta}}\right)+\lambda_2(n)\geq 0,$$

for all $p_kp_{k'}$ dimensional unit vectors $\boldsymbol{u}$. To verify this we must substitute the negative normalized gradient in for $\boldsymbol{u}$, and see when the inequality holds. The negative normalized gradient is given by

$$\boldsymbol{u}^*=\frac{\left(\boldsymbol{X}_{kk'}\left[\operatorname{diag}(\tilde{\boldsymbol{\beta}}_k\otimes\tilde{\boldsymbol{\beta}}_{k'})\right]\right)^\top\left(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\theta}}\right)}{\left\|\left(\boldsymbol{X}_{kk'}\left[\operatorname{diag}(\tilde{\boldsymbol{\beta}}_k\otimes\tilde{\boldsymbol{\beta}}_{k'})\right]\right)^\top\left(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\theta}}\right)\right\|_2}.$$

Substituting this into our expression for the we get

$$-\Big(\boldsymbol{X}_{kk'}[\boldsymbol{u}^*\odot(\tilde{\boldsymbol{\beta}}_k\otimes\tilde{\boldsymbol{\beta}}_{k'})]\Big)^\top\left(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\theta}}\right)$$

$$=-\frac{\left[\left(\operatorname{diag}(\tilde{\boldsymbol{\beta}}_k\otimes\tilde{\boldsymbol{\beta}}_{k'})\boldsymbol{X}_{kk'}^\top\left(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\theta}}\right)\right)\odot(\boldsymbol{\beta}_k\otimes\boldsymbol{\beta}_{k'})\right]^\top\boldsymbol{X}_{kk'}^\top\left(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\theta}}\right)}{\left\|\left(\boldsymbol{X}_{kk'}\left[\operatorname{diag}(\tilde{\boldsymbol{\beta}}_k\otimes\tilde{\boldsymbol{\beta}}_{k'})\right]\right)^\top\left(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\theta}}\right)\right\|_2}$$

$$=-\frac{\left(\operatorname{diag}(\tilde{\boldsymbol{\beta}}_k\otimes\tilde{\boldsymbol{\beta}}_{k'})\operatorname{diag}(\tilde{\boldsymbol{\beta}}_k\otimes\tilde{\boldsymbol{\beta}}_{k'})\boldsymbol{X}_{kk'}^\top\left(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\theta}}\right)\right)^\top\boldsymbol{X}_{kk'}^\top\left(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\theta}}\right)}{\left\|\left(\boldsymbol{X}_{kk'}\left[\operatorname{diag}(\tilde{\boldsymbol{\beta}}_k\otimes\tilde{\boldsymbol{\beta}}_{k'})\right]\right)^\top\left(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\theta}}\right)\right\|_2}$$

$$=-\frac{\left(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\theta}}\right)^\top\boldsymbol{X}_{kk'}\operatorname{diag}(\tilde{\boldsymbol{\beta}}_k\otimes\tilde{\boldsymbol{\beta}}_{k'})\operatorname{diag}(\tilde{\boldsymbol{\beta}}_k\otimes\tilde{\boldsymbol{\beta}}_{k'})\boldsymbol{X}_{kk'}^\top\left(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\theta}}\right)}{\left\|\left(\boldsymbol{X}_{kk'}\left[\operatorname{diag}(\tilde{\boldsymbol{\beta}}_k\otimes\tilde{\boldsymbol{\beta}}_{k'})\right]\right)^\top\left(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\theta}}\right)\right\|_2}$$

$$= -\frac{\left\|\left(\boldsymbol{X}_{kk'}\left[\operatorname{diag}(\tilde{\boldsymbol{\beta}}_k \otimes \tilde{\boldsymbol{\beta}}_{k'})\right]\right)^\top (\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})\right\|_2^2}{\left\|\left(\boldsymbol{X}_{kk'}\left[\operatorname{diag}(\tilde{\boldsymbol{\beta}}_k \otimes \tilde{\boldsymbol{\beta}}_{k'})\right]\right)^\top (\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})\right\|_2}$$

$$= -\left\|\left(\boldsymbol{X}_{kk'}\left[\operatorname{diag}(\tilde{\boldsymbol{\beta}}_k \otimes \tilde{\boldsymbol{\beta}}_{k'})\right]\right)^\top (\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})\right\|_2.$$

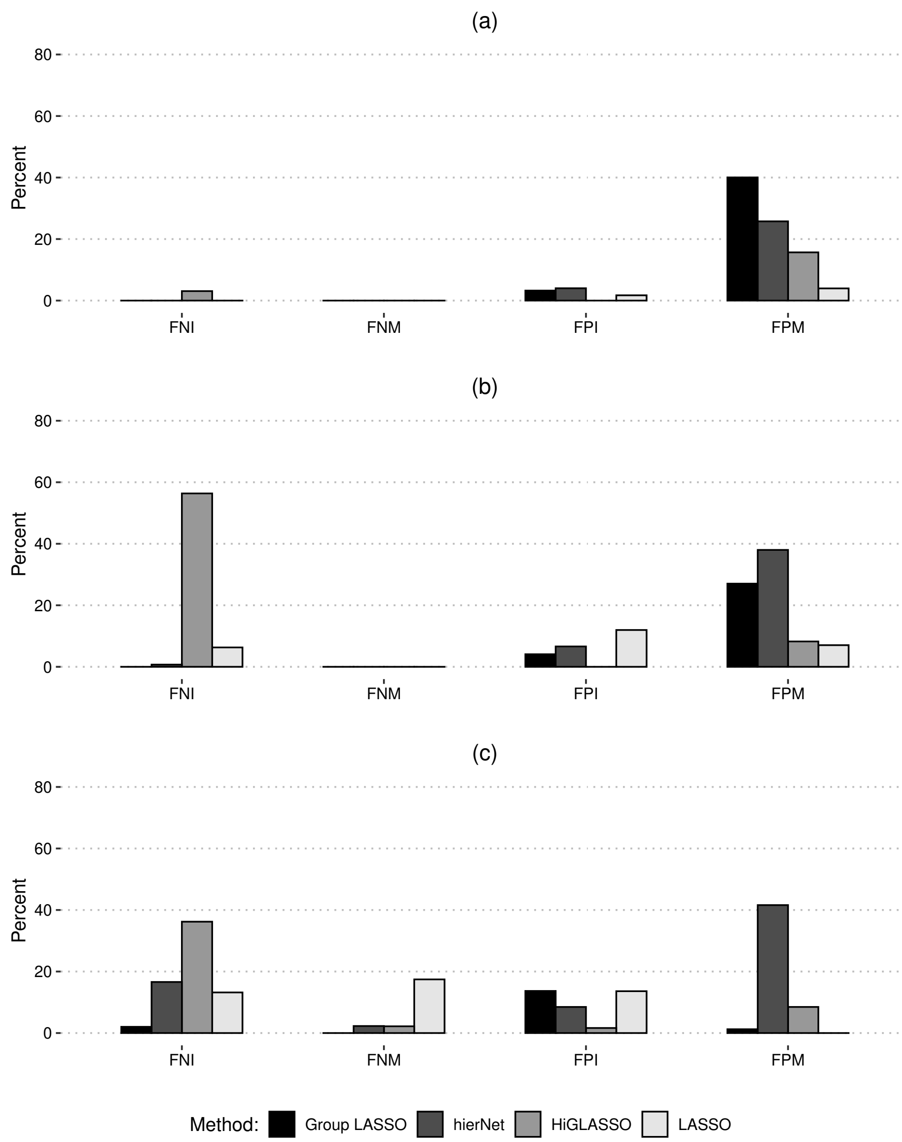
Therefore, for $\tilde{\boldsymbol{\beta}}_j$ and $\tilde{\boldsymbol{\eta}}_{jj'}$ to be the minimizer's of (9), we need

$$-\left\|\left(\boldsymbol{X}_{kk'}\left[\operatorname{diag}(\tilde{\boldsymbol{\beta}}_k \otimes \tilde{\boldsymbol{\beta}}_{k'})\right]\right)^\top (\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})\right\|_2 + \lambda_2(n) \geq 0,$$

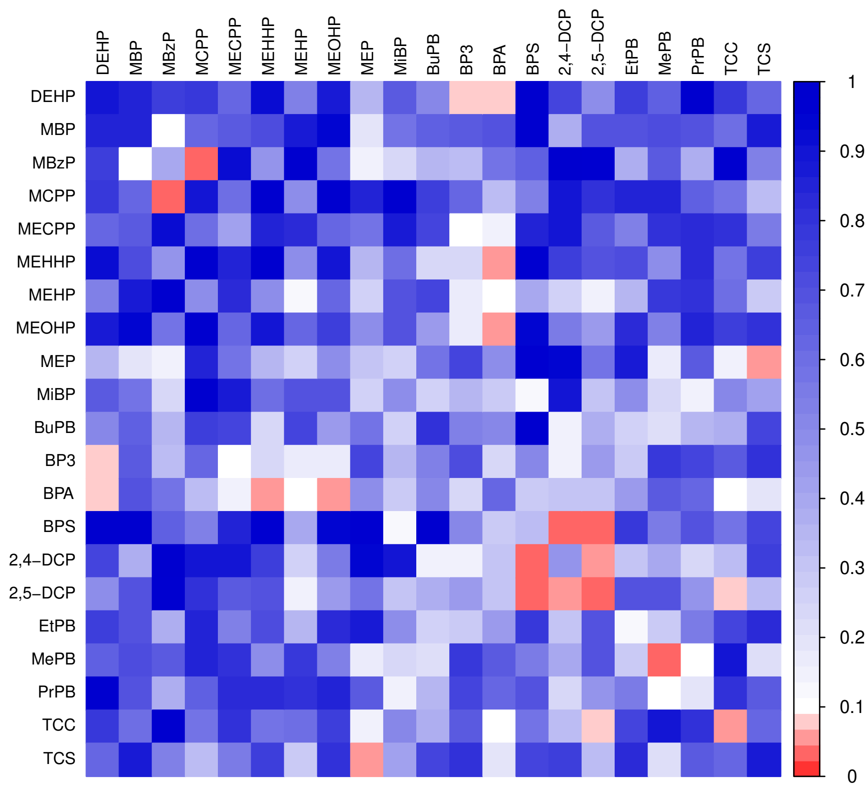
which implies that

$$\left\|\frac{1}{n}\operatorname{diag}(\tilde{\boldsymbol{\beta}}_k \otimes \tilde{\boldsymbol{\beta}}_{k'})\boldsymbol{X}_{kk'}^\top (\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}})\right\|_2 \leq \frac{\lambda_2(n)}{n}. \tag{21}$$

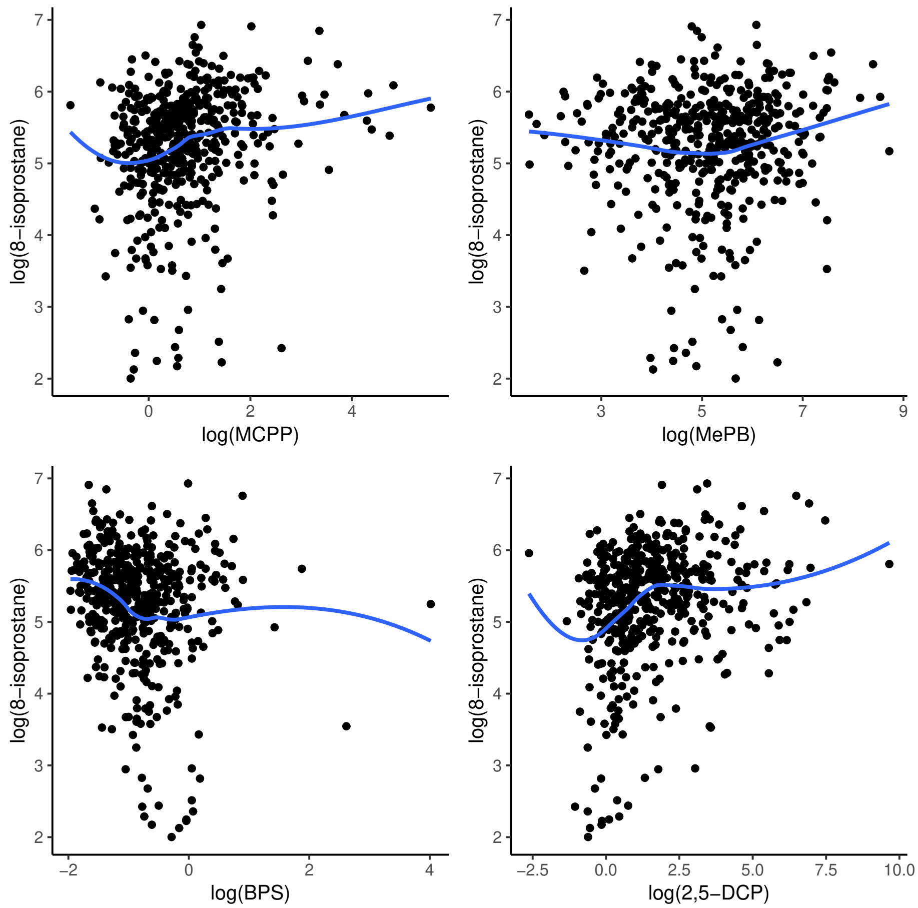
Since (20) is equivalent to (18) and (21) is equivalent to (19), this concludes the proof.



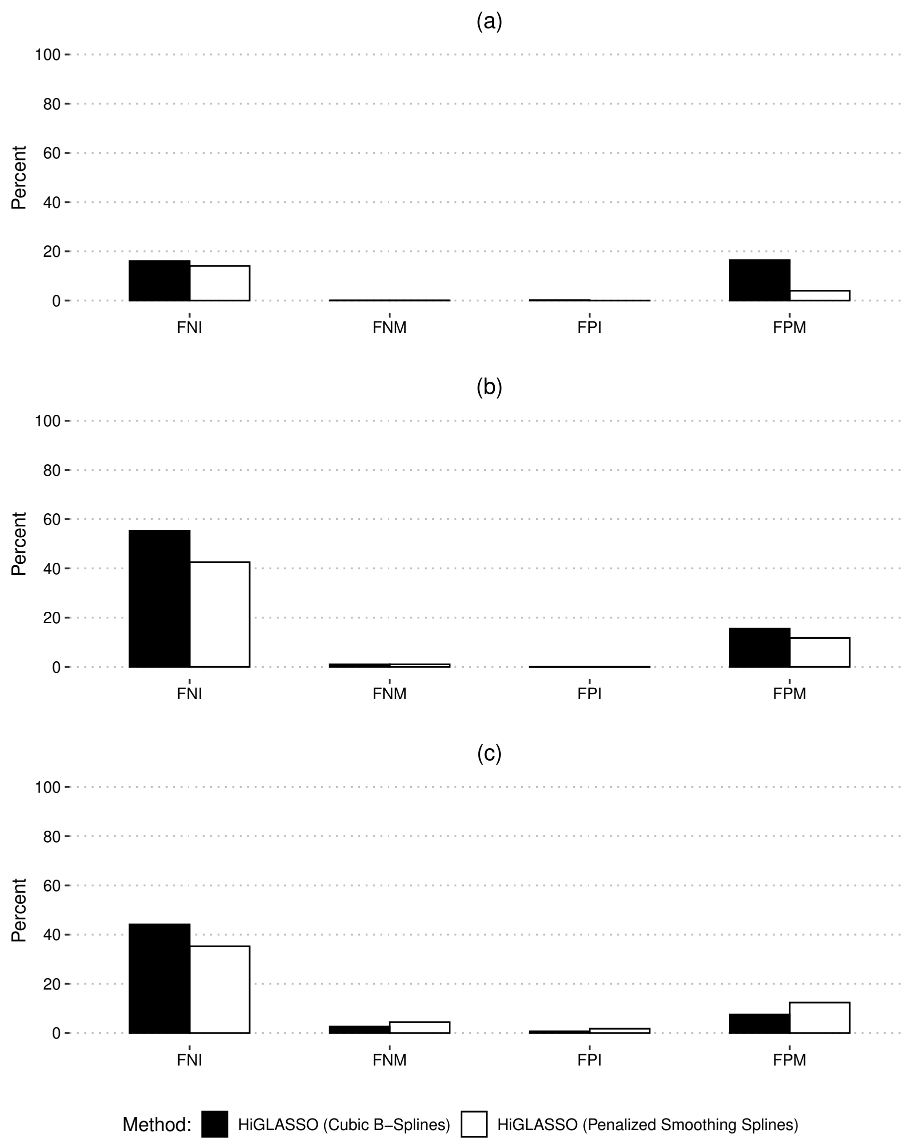
Web Figure 1: Simulation Results for the $n = 10000$ and $p = 10$ cases: (a) Linear main and interaction effects (b) Piecewise linear main and interaction effects (c) Nonlinear main and interaction effects. FNI, FNM, FPI, and FPM are defined in Section 4.2.



Web Figure 2: Heatmap for Wald test p-values corresponding to all pairwise linear interactions. Each p-value is obtained from a multiple regression model with 21 exposure main effect terms and a single pairwise linear interaction term. Diagonal elements indicate the addition of a squared term instead of an interaction.



Web Figure 3: Scatterplots between four exposures and 8-isoprostane superimposed with a Locally Weighted Scatterplot Smoothing (LOWESS) curve. The four exposures are mono(3-carboxypropyl) phthalate (MCPP), methyl paraben (MePB), Bisphenol S (BPS), and 2,5-Dichlorophenol (2,5-DCP).



Web Figure 4: Simulated comparison between HiGLASSO with cubic B-splines and HiGLASSO with penalized smoothing splines for the $n = 1000$ and $p = 10$ cases: (a) L10 (b) PL10 (c) NL10. FNI, FNM, FPI, and FPM are defined in Section 4.2 of the main text.

References

Armijo, L. (1966). Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics* **16,** 1–3.

Bauer, D. J. and Cai, L. (2009). Consequences of unmodeled nonlinear effects in multilevel models. *Journal of Educational and Behavioral Statistics* **34,** 97–114.

Dennis, J. E. and Schnabel, R. B. (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Classics in Applied Mathematics. Society for Industrial and Applied Mathematics.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96,** 1348–1360.

Pan, Q. and Zhao, Y. (2016). Integrative weighted group lasso and generalized local quadratic approximation. *Computational Statistics & Data Analysis* **104,** 66–78.