Group variable selection via a hierarchical lasso and its oracle property

NENGFENG ZHOU AND JI ZHU*

In many engineering and scientific applications, prediction variables are grouped, for example, in biological applications where assayed genes or proteins can be grouped by biological roles or biological pathways. Common statistical analysis methods such as ANOVA, factor analysis, and functional modeling with basis sets also exhibit natural variable groupings. Existing successful group variable selection methods have the limitation of selecting variables in an "all-in-all-out" fashion, i.e., when one variable in a group is selected, all other variables in the same group are also selected [1, 23, 25]. In many real problems, however, we may want to keep the flexibility of selecting variables within a group, such as in gene-set selection. In this paper, we develop a new group variable selection method that not only removes unimportant groups effectively, but also keeps the flexibility of selecting variables within a group. We also show that the new method offers the potential for achieving the theoretical "oracle" property [6, 7].

KEYWORDS AND PHRASES: Group selection, Lasso, Oracle property, Regularization, Variable selection.

1. INTRODUCTION

Consider the usual regression situation: we have training data, $(\boldsymbol{x}_1, y_1)$, ..., $(\boldsymbol{x}_i, y_i)$, ..., $(\boldsymbol{x}_n, y_n)$, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ are the predictors and y_i is the response. To model the response y in terms of the predictors $x_1, \ldots, x_p$, one may consider the linear model:

$$(1) \qquad y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon,$$

where ε is the error term. In many important practical problems, however, prediction variables are "grouped." For example, in ANOVA factor analysis, a factor may have several levels and can be expressed via several dummy variables, then the dummy variables corresponding to the same factor form a natural "group." Similarly, in additive models, each original prediction variable may be expanded into different order polynomials or a set of basis functions, then these polynomials (or basis functions) corresponding to the same

original prediction variable form a natural "group." Another example is in biological applications, where assayed genes or proteins can be grouped by biological roles (or biological pathways).

For the rest of the paper, we assume that the prediction variables can be divided into K groups and the kth group contains p_k variables. Specifically, the linear model (1) is now written as

$$(2) \qquad y_i = \beta_0 + \sum_{k=1}^{K} \sum_{j=1}^{p_k} \beta_{kj} x_{i,kj} + \varepsilon_i.$$

And we are interested in finding out which variables, especially which "groups," have an important effect on the response. For example, $(x_{11}, \ldots, x_{1p_1})$, $(x_{21}, \ldots, x_{2p_2})$, ..., $(x_{K1}, \ldots, x_{Kp_K})$ may represent different biological pathways, y may represent a certain phenotype and we are interested in deciphering which and how these biological pathways "work together" to affect the phenotype.

There are two important challenges in this problem: prediction accuracy and interpretation. We would like our model to accurately predict on future data. Prediction accuracy can often be improved by shrinking the regression coefficients. Shrinkage sacrifices some bias to reduce the variance of the predicted value and hence may improve the overall prediction accuracy. Interpretability is often realized via variable selection. With a large number of prediction variables, we often would like to determine a smaller subset that exhibits the strongest effects.

Variable selection has been studied extensively in the literature [2, 6, 8, 9, 12, 18, 22, 27]. In particular, lasso [18] has gained much attention in recent years. The lasso criterion penalizes the L_1-norm of the regression coefficients to achieve a sparse model:

$$(3)$$
$$\max_{\beta_0, \beta_{kj}} -\frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{k=1}^{K} \sum_{j=1}^{p_k} \beta_{kj} x_{i,kj} \right)^2 - \lambda \sum_{k=1}^{K} \sum_{j=1}^{p_k} |\beta_{kj}|,$$

where $\lambda \geq 0$ is a tuning parameter. Note that by location transformation, we can always assume that the predictors and the response have mean 0, so we can ignore the intercept in equation (3).

Due to the singularity at $\beta_{kj} = 0$, the L_1-norm penalty can shrink some of the fitted coefficients to be *exact* zero

when making the tuning parameter sufficiently large. However, lasso and other methods above are for the case when the candidate variables can be treated individually or "flatly." When variables are grouped, ignoring the group structure and directly applying lasso as in (3) may be suboptimal. For example, suppose the kth group is unimportant, then lasso tends to make individual estimated coefficients in the kth group to be zero, rather than the whole group to be zero, i.e., lasso tends to make selection based on the strength of individual variables rather than the strength of the group, often resulting in selecting more groups than necessary.

Group variable selection has been addressed in some literature [1, 23, 25]. In [1], the authors proposed to use a blockwise additive penalty in the setting of wavelet approximations. To increase the estimation precision, empirical wavelet coefficients were thresholded or shrunken in blocks (or groups) rather than individually.

In [23] and [25], Lasso model (3) is extended for group variable selection. In [23], the authors chose to penalize the L_2-norm of the coefficients within each group, i.e., $\sum_{k=1}^{K} \|\boldsymbol{\beta}_k\|_2$, where

(4)
$$\|\boldsymbol{\beta}_k\|_2 = \sqrt{\beta_{k1}^2 + \cdots + \beta_{kp_k}^2}.$$

Due to the singularity of $\|\boldsymbol{\beta}_k\|_2$ at $\boldsymbol{\beta}_k = \mathbf{0}$, appropriately tuning λ can set the whole coefficient vector $\boldsymbol{\beta}_k = \mathbf{0}$, hence the kth group is removed from the fitted model. We note that in the setting of wavelet analysis, this method reduces to that in [1].

Instead of using the L_2-norm penalty, in [25], the authors suggested using the L_∞-norm penalty, i.e., $\sum_{k=1}^{K} \|\boldsymbol{\beta}_k\|_\infty$, where

(5)
$$\|\boldsymbol{\beta}_k\|_\infty = \max(|\beta_{k1}|, |\beta_{k2}|, \ldots, |\beta_{kp_k}|).$$

Similar to the L_2-norm, the L_∞-norm of $\boldsymbol{\beta}_k$ is also singular when $\boldsymbol{\beta}_k = \mathbf{0}$; hence when λ is appropriately tuned, the L_∞-norm can also effectively remove unimportant groups.

However, there are some possible limitations with these methods: Both the L_2-norm penalty and the L_∞-norm penalty select variables in an "all-in-all-out" fashion, i.e., when one variable in a group is selected, *all* other variables in the same group are also selected. The reason is that both $\|\boldsymbol{\beta}_k\|_2$ and $\|\boldsymbol{\beta}_k\|_\infty$ are singular only when the whole vector $\boldsymbol{\beta}_k = \mathbf{0}$. Once a component of $\boldsymbol{\beta}_k$ is non-zero, the two norm functions are no longer singular. This can also be heuristically understood as the following: for the L_2-norm (4), it is the ridge penalty that is under the square root; since the ridge penalty can not do variable selection (as in ridge regression), once the L_2-norm is non-zero (or the corresponding group is selected), all components will be non-zero. For the L_∞-norm (5), if the "max(·)" is non-zero, there is no increase in the penalty for letting all the individual components move away from zero. Hence if one variable in a group is selected, all other variables are also automatically selected.

In many important real problems, however, we may want to keep the flexibility of selecting variables *within* a group. For example, in the gene-set selection problem, a biological pathway may be related to a certain biological process, but it does not necessarily mean all the genes in the pathway are all related to the biological process. We may want to not only remove unimportant pathways effectively, but also identify important genes within important pathways.

For the L_∞-norm penalty, another possible limitation is that the estimated coefficients within a group tend to have the same magnitude, i.e. $|\beta_{k1}| = |\beta_{k2}| = \cdots = |\beta_{kp_k}|$; and this may cause some serious bias, which jeopardizes the prediction accuracy.

In this paper, we propose an extension of lasso for group variable selection, which we call hierarchical lasso (HLasso). Our method not only removes unimportant groups effectively, but also keeps the flexibility of selecting variables within a group. Furthermore, asymptotic studies motivate us to improve our model and show that when the tuning parameter is appropriately chosen, the improved model has the *oracle* property [6, 7], i.e., it performs as well as if the correct underlying model were given in advance. Such a theoretical property has not been previously studied for group variable selection at both the group level and within the group level.

The rest of the paper is organized as follows. In Section 2, we introduce our new method: the hierarchical lasso. We propose an algorithm to compute the solution for the hierarchical lasso in Section 3. In Sections 4 and 5, we study the asymptotic behavior of the hierarchical lasso and propose an improvement for the hierarchical lasso. Numerical results are in Sections 6 and 7, and we conclude the paper with Section 8.

2. HIERARCHICAL LASSO

In this section, we extend the lasso method for group variable selection so that we can effectively remove unimportant variables at both the group level and within the group level.

We reparameterize β_{kj} as

(6)
$$\beta_{kj} = d_k \alpha_{kj}, \quad k = 1, \ldots, K; j = 1, \ldots, p_k,$$

where $d_k \geq 0$ (for identifiability reasons). This decomposition reflects the information that $\beta_{kj}, j = 1, \ldots, p_k$, all belong to the kh group, by treating each β_{kj} hierarchically. d_k is at the first level of the hierarchy, controlling $\beta_{kj}, j = 1, \ldots, p_k$, as a group; α_{kj}'s are at the second level of the hierarchy, reflecting differences within the kth group.

For the purpose of variable selection, we consider the following penalized least squares criterion:

$$(7) \qquad \max_{d_k, \alpha_{kj}} -\frac{1}{2} \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{K} d_k \sum_{j=1}^{p_k} \alpha_{kj} x_{i,kj} \right)^2$$
$$-\lambda_1 \cdot \sum_{k=1}^{K} d_k - \lambda_2 \cdot \sum_{k=1}^{K} \sum_{j=1}^{p_k} |\alpha_{kj}|$$
$$\text{subject to} \quad d_k \geq 0, \ k = 1, \ldots, K,$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are tuning parameters. λ_1 controls the estimates at the group level, and it can effectively remove unimportant groups: if d_k is shrunken to zero, all β_{kj} in the kth group will be equal to zero. λ_2 controls the estimates at the variable-specific level: if d_k is not equal to zero, some of the α_{kj} hence some of the $\beta_{kj}, j = 1, \ldots, p_k$, still have the possibility of being zero; in this sense, the hierarchical penalty keeps the flexibility of the L_1-norm penalty.

One may complain that such a hierarchical penalty may be more complicated to tune in practice, however, it turns out that the two tuning parameters λ_1 and λ_2 in (7) can be simplified into one. Specifically, let $\lambda = \lambda_1 \cdot \lambda_2$, we can show that (7) is equivalent to

$$(8) \qquad \max_{d_k, \alpha_{kj}} -\frac{1}{2} \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{K} d_k \sum_{j=1}^{p_k} \alpha_{kj} x_{i,kj} \right)^2$$
$$-\sum_{k=1}^{K} d_k - \lambda \sum_{k=1}^{K} \sum_{j=1}^{p_k} |\alpha_{kj}|$$
$$\text{subject to} \qquad d_k \geq 0, k = 1, \ldots, K.$$

Lemma 1. *Let $(\hat{\boldsymbol{d}}^*, \hat{\boldsymbol{\alpha}}^*)$ be a local maximizer of (7), then there exists a local maximizer $(\hat{\boldsymbol{d}}^\star, \hat{\boldsymbol{\alpha}}^\star)$ of (8) such that $\hat{d}_k^* \hat{\alpha}_{kj}^* = \hat{d}_k^\star \hat{\alpha}_{kj}^\star$. Similarly, if $(\hat{\boldsymbol{d}}^\star, \hat{\boldsymbol{\alpha}}^\star)$ is a local maximizer of (8), there exists a local maximizer $(\hat{\boldsymbol{d}}^*, \hat{\boldsymbol{\alpha}}^*)$ of (7) such that $\hat{d}_k^* \hat{\alpha}_{kj}^* = \hat{d}_k^\star \hat{\alpha}_{kj}^\star$.*

The proof is in the Appendix. This lemma indicates that the final fitted models from (7) and (8) are the same, although they may provide different d_k and α_{kj}. This also implies that in practice, we do not need to tune λ_1 and λ_2 separately; we only need to tune one parameter $\lambda = \lambda_1 \cdot \lambda_2$ as in (8).

3. ALGORITHM

To estimate the d_k and α_{kj} in (8), we can use an iterative approach, i.e., we first fix d_k and estimate α_{kj}, then we fix α_{kj} and estimate d_k, and we iterate between these two steps until the solution converges. Since at each step, the value of the objective function (8) decreases, the solution is guaranteed to converge.

When d_k is fixed, (8) becomes a lasso problem, hence we can use either the LAR/LASSO algorithm [4] or a quadratic

programming package to efficiently solve for α_{kj}. When α_{kj} is fixed, (8) becomes a non-negative garrote problem. Again, we can use either an efficient solution path algorithm or a quadratic programming package to solve for d_k. In summary, the algorithm proceeds as follows:

1. (Standardization) Center $\boldsymbol{y}$. Center and normalize $\boldsymbol{x}_{kj}$.
2. (Initialization) Initialize $d_k^{(0)}$ and $\alpha_{kj}^{(0)}$ with some plausible values. For example, we can set $d_k^{(0)} = 1$ and use the least squares estimates or the simple regression estimates by regressing the response $\boldsymbol{y}$ on each of the $\boldsymbol{x}_{kj}$ for $\alpha_{kj}^{(0)}$. Let $\beta_{kj}^{(0)} = d_k^{(0)} \alpha_{kj}^{(0)}$ and $m = 1$.
3. (Update α_{kj}) Let

$$\tilde{x}_{i,kj} = d_k^{(m-1)} x_{i,kj}, \ k = 1, \ldots, K; \ j = 1, \ldots, p_k,$$

then

$$\alpha_{kj}^{(m)} = \arg\max_{\alpha_{kj}} \left\{ -\frac{1}{2} \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{K} \sum_{j=1}^{p_k} \alpha_{kj} \tilde{x}_{i,kj} \right)^2 -\lambda \sum_{k=1}^{K} \sum_{j=1}^{p_k} |\alpha_{kj}| \right\}.$$

4. (Update d_k) Let

$$\tilde{x}_{i,k} = \sum_{j=1}^{p_k} \alpha_{kj}^{(m)} x_{i,kj}, \quad k = 1, \ldots, K,$$

then

$$d_k^{(m)} = \arg\max_{d_k \geq 0} \left\{ -\frac{1}{2} \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{K} d_k \tilde{x}_{i,k} \right)^2 - \sum_{k=1}^{K} d_k \right\}.$$

5. (Update β_{kj}) Let

$$\beta_{kj}^{(m)} = d_k^{(m)} \alpha_{kj}^{(m)}.$$

6. If $\|\beta_{kj}^{(m)} - \beta_{kj}^{(m-1)}\|$ is small enough, stop the algorithm. Otherwise, let $m \leftarrow m + 1$ and go back to Step 3.

3.1 Orthogonal case

To gain more insight into the hierarchical penalty, we have also studied the algorithm in the orthogonal design case. This can be useful, for example, in the wavelet setting, where each $\boldsymbol{x}_{kj}$ corresponds to a wavelet basis function, different k may correspond to different "frequency" scales, and different j with the same k correspond to different "time" locations. Specifically, suppose $\boldsymbol{x}_{kj}^\mathsf{T} \boldsymbol{x}_{kj} = 1$ and $\boldsymbol{x}_{kj}^\mathsf{T} \boldsymbol{x}_{k'j'} = 0$ if $k \neq k'$ or $j \neq j'$, then Step 3 and Step 4 in the above algorithm have closed form solutions.

Let $\hat{\beta}_{kj}^{\text{ols}} = \boldsymbol{x}_{kj}^\mathsf{T} \boldsymbol{y}$ be the ordinary least squares solution when $\boldsymbol{x}_{kj}$ are orthonormal to each other.

3. When d_k is fixed,

$$\text{(9)} \qquad \alpha_{kj}^{(m)} = \mathbb{I}(d_k^{(m-1)} > 0) \cdot \text{sgn}(\hat{\beta}_{kj}^{\text{ols}})$$
$$\cdot \left(\frac{|\hat{\beta}_{kj}^{\text{ols}}|}{d_k^{(m-1)}} - \frac{\lambda}{(d_k^{(m-1)})^2} \right)_+ .$$

4. When α_{kj} is fixed,

(10)

$$d_k^{(m)} = \mathbb{I}(\exists j, \alpha_{kj}^{(m)} \neq 0)$$
$$\cdot \left(\sum_{j=1}^{p_k} \frac{(\alpha_{kj}^{(m)})^2}{\sum_{j=1}^{p_k} (\alpha_{kj}^{(m)})^2} \frac{\hat{\beta}_{kj}^{\text{ols}}}{\alpha_{kj}^{(m)}} - \frac{1}{\sum_{j=1}^{p_k} (\alpha_{kj}^{(m)})^2} \right)_+ .$$

Equations (9) and (10) show that both $d_k^{(m)}$ and $\alpha_{kj}^{(m)}$ are soft-thresholding estimates. Here we provide some intuitive explanation.

We first look at $\alpha_{kj}^{(m)}$ in equation (9). If $d_k^{(m-1)} = 0$, it is natural to estimate all α_{kj} to be zero because of the penalty on α_{kj}. If $d_k^{(m-1)} > 0$, then from our reparametrization, we have $\alpha_{kj} = \beta_{kj}/d_k^{(m-1)}$, $j = 1, \ldots, p_k$. Plugging in $\hat{\beta}_{kj}^{\text{ols}}$ for β_{kj}, we obtain $\tilde{\alpha}_{kj} = \hat{\beta}_{kj}^{\text{ols}}/d_k^{(m-1)}$. Equation (9) shrinks $\tilde{\alpha}_{kj}$, and the amount of shrinkage is inversely proportional to $(d_k^{(m-1)})^2$. So when $d_k^{(m-1)}$ is large, which indicates the kth group is important, the amount of shrinkage is small, while when $d_k^{(m-1)}$ is small, which indicates the kth group is less important, the amount of shrinkage is large.

Now considering $d_k^{(m)}$ in equation (10). If all $\alpha_{kj}^{(m)}$ are zero, it is natural to estimate $d_k^{(m)}$ also to be zero because of the penalty on d_k. If not all $\alpha_{kj}^{(m)}$ are 0, say $\alpha_{kj_1}^{(m)}, \ldots, \alpha_{kj_r}^{(m)}$ are not zero, then we have $d_k = \beta_{kj_s}/\alpha_{kj_s}^{(m)}, 1 \leq s \leq r$. Again, plugging in $\hat{\beta}_{kj_s}^{\text{ols}}$ for β_{kj_s}, we obtain r estimates for d_k: $\tilde{d}_k = \hat{\beta}_{kj_s}^{\text{ols}}/\alpha_{kj_s}^{(m)}, 1 \leq s \leq r$. A natural estimate for d_k is then a weighted average of the $\tilde{d}_k$, and equation (10) provides such a (shrunken) average, with weights proportional to $(\alpha_{kj}^{(m)})^2$.

4. ASYMPTOTIC THEORY

In this section, we explore the asymptotic behavior of the hierarchical lasso method.

The hierarchical lasso criterion (8) uses d_k and α_{kj}. We first show that it can also be written in an equivalent form using the original regression coefficients β_{kj}.

Theorem 1. *If $(\hat{d}, \hat{\alpha})$ is a local maximizer of* (8), *then $\hat{\beta}$, where $\hat{\beta}_{kj} = \hat{d}_k \hat{\alpha}_{kj}$, is a local maximizer of*

$$\text{(11)} \qquad \max_{\beta_{kj}} \left\{ -\frac{1}{2} \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{K} \sum_{j=1}^{p_k} x_{i,kj} \beta_{kj} \right)^2 \right.$$
$$\left. - 2\sqrt{\lambda} \cdot \sum_{k=1}^{K} \sqrt{|\beta_{k1}| + |\beta_{k2}| + \cdots + |\beta_{kp_k}|} \right\}.$$

On the other hand, if $\hat{\beta}$ is a local maximizer of (11), *then we define $(\hat{d}, \hat{\alpha})$, where $\hat{d}_k = 0, \hat{\alpha}_k = 0$ if $\|\hat{\beta}_k\|_1 = 0$, and $\hat{d}_k = \sqrt{\lambda \|\hat{\beta}_k\|_1}, \hat{\alpha}_k = \frac{\hat{\beta}_k}{\sqrt{\lambda \|\hat{\beta}_k\|_1}}$ if $\|\hat{\beta}_k\|_1 \neq 0$. Then the so-defined $(\hat{d}, \hat{\alpha})$ is a local maximizer of* (8).

Note that the penalty term in (11) is similar to the L_2-norm penalty (4), except that under each square root, we now penalize the L_1-norm of β_k, rather than the sum of squares. However, unlike the L_2-norm, which is singular only at the point $\beta_k = \mathbf{0}$, (i.e., the whole vector is equal to $\mathbf{0}$), the square root of the L_1-norm is singular at $\beta_{kj} = 0$ no matter what are the values of other β_{kj}'s. This explains, from a different perspective, why the hierarchical lasso can remove not only groups, but also variables within a group even when the group is selected. Equation (11) also implies that the hierarchical lasso belongs to the "CAP" family [25].

We study the asymptotic properties allowing the total number of variables P_n, as well as the number of groups K_n and the number of variables within each group p_{nk}, to go to ∞, where $P_n = \sum_{k=1}^{K_n} p_{nk}$. Note that we add a subscript "n" to K and p_k to denote that these quantities can change with n. Accordingly, β, y_i and $x_{i,kj}$ are also changed to β_n, y_{ni} and $x_{ni,kj}$. We write $2\sqrt{\lambda}$ in (11) as $n\lambda_n$, and the criterion (11) is re-written as

$$\text{(12)} \qquad \max_{\beta_{n,kj}} \left\{ -\frac{1}{2} \sum_{i=1}^{n} \left(y_{ni} - \sum_{k=1}^{K_n} \sum_{j=1}^{p_{nk}} x_{ni,kj} \beta_{n,kj} \right)^2 \right.$$
$$\left. - n\lambda_n \cdot \sum_{k=1}^{K_n} \sqrt{|\beta_{n,k1}| + \cdots + |\beta_{n,kp_{nk}}|} \right\}.$$

Our asymptotic analysis in this section is based on the criterion (12).

Let $\beta_n^0 = (\beta_{\mathcal{A}_n}^0, \beta_{\mathcal{B}_n}^0, \beta_{\mathcal{C}_n}^0)^\top$ be the underlying true parameters, where

$$\text{(13)} \qquad \begin{aligned} \mathcal{A}_n &= \{(k,j) : \beta_{n,kj}^0 \neq 0\}, \\ \mathcal{B}_n &= \{(k,j) : \beta_{n,kj}^0 = 0, \beta_{nk}^0 \neq \mathbf{0}\}, \\ \mathcal{C}_n &= \{(k,j) : \beta_{nk}^0 = \mathbf{0}\}, \\ \mathcal{D}_n &= \mathcal{B}_n \cup \mathcal{C}_n. \end{aligned}$$

Note that $\mathcal{A}_n$ contains the indices of coefficients which are truly non-zero, $\mathcal{C}_n$ contains the indices where the whole "groups" are truly zero, and $\mathcal{B}_n$ contains the indices of zero coefficients, but they belong to some non-zero groups. So $\mathcal{A}_n$, $\mathcal{B}_n$ and $\mathcal{C}_n$ are disjoint and they partition all the indices. We have the following theorem.

Theorem 2. *If $\sqrt{n}\lambda_n = O(1)$, then there exists a root-(n/P_n) consistent local maximizer $\hat{\beta}_n = (\hat{\beta}_{\mathcal{A}_n}, \hat{\beta}_{\mathcal{B}_n}, \hat{\beta}_{\mathcal{C}_n})^\top$ of* (12), *and if also $P_n n^{-3/4}/\lambda_n \to 0$ as $n \to \infty$, then $\Pr(\hat{\beta}_{\mathcal{C}_n} = 0) \to 1$.*

Theorem 2 implies that the hierarchical lasso method can effectively remove unimportant *groups*. For the above root-(n/P_n) consistent estimate, however, if $\mathcal{B}_n \neq \emptyset$ (empty set), then $\Pr(\hat{\boldsymbol{\beta}}_{\mathcal{B}_n} = 0) \to 1$ is not always true. This means that although the hierarchical lasso method can effectively remove *all* unimportant *groups* and *some* of the unimportant *variables* within important groups, it cannot effectively remove *all* unimportant *variables* within important groups.

In the next section, we improve the hierarchical lasso method to tackle this limitation.

5. ADAPTIVE HIERARCHICAL LASSO

To improve the hierarchical lasso method, we apply the adaptive idea [2, 21, 24, 26], i.e., to penalize different coefficients differently. Specifically, we consider

$$(14) \quad \max_{\beta_{n,kj}} \left\{ -\frac{1}{2} \sum_{i=1}^{n} \left(y_{ni} - \sum_{k=1}^{K_n} \sum_{j=1}^{p_k} x_{ni,kj} \beta_{n,kj} \right)^2 \right.$$
$$\left. - n\lambda_n \cdot \sum_{k=1}^{K_n} \sqrt{w_{n,k1}|\beta_{n,k1}| + \cdots + w_{n,kp_{nk}}|\beta_{n,kp_{nk}}|} \right\},$$

where $w_{n,kj}$ are pre-specified weights. The intuition is that if the effect of a variable is strong, we would like the corresponding weight to be small, hence the corresponding coefficient is lightly penalized. If the effect of a variable is not strong, we would like the corresponding weight to be large, hence the corresponding coefficient is heavily penalized. In practice, we may consider using the ordinary least squares estimates or the ridge regression estimates to help us compute the weights, for example,

$$(15) \qquad w_{n,kj} = \frac{1}{|\hat{\beta}_{n,kj}^{\text{ols}}|^\gamma} \quad \text{or} \quad w_{n,kj} = \frac{1}{|\hat{\beta}_{n,kj}^{\text{ridge}}|^\gamma},$$

where γ is a positive constant.

5.1 Oracle property

Problem setup

Since the theoretical results we develop for (14) are not restricted to the squared error loss, for the rest of the section, we consider the generalized linear model. For generalized linear models, statistical inferences are based on underlying likelihood functions. We assume that the data $\boldsymbol{V}_{ni} = (\boldsymbol{X}_{ni}, Y_{ni})$, $i = 1, \ldots, n$ are independent and identically distributed for every n. Conditioning on $\boldsymbol{X}_{ni} = \boldsymbol{x}_{ni}$, Y_{ni} has a density $f_n(g_n(\boldsymbol{x}_{ni}^\top \boldsymbol{\beta}_n), Y_{ni})$, where $g_n(\cdot)$ is a known link function. We maximize the penalized log-likelihood

$$(16)$$

$$\max_{\beta_{n,kj}} Q_n(\boldsymbol{\beta}_n) = L_n(\boldsymbol{\beta}_n) - J_n(\boldsymbol{\beta}_n)$$

$$= \sum_{i=1}^{n} \ell_n(g_n(\boldsymbol{x}_{ni}^\top \boldsymbol{\beta}_n), y_{ni}) - n \sum_{k=1}^{K} p_{\lambda_n, \boldsymbol{w}_n}(\boldsymbol{\beta}_{nk}),$$

where $\ell_n(\cdot, \cdot) = \log f_n(\cdot, \cdot)$ denotes the conditional log-likelihood of Y, and

$$p_{\lambda_n, \boldsymbol{w}_n}(\boldsymbol{\beta}_{nk}) = \lambda_n \sqrt{w_{n,k1}|\beta_{n,k1}| + \cdots + w_{n,kp_k}|\beta_{n,kp_{nk}}|}.$$

Note that under the normal distribution, $\ell_n(g_n(\boldsymbol{x}_{ni}^\top \boldsymbol{\beta}_n), y_{ni}) = -\frac{(y_{ni} - \boldsymbol{x}_{ni}^\top \boldsymbol{\beta}_n)^2}{2C_1} + C_2$, hence (16) reduces to (14).

The asymptotic properties of (16) are described in the following theorems, and the proofs are in the Appendix. We note that the proofs follow the spirit of previous work [6, 7], but due to the grouping structure and the adaptive weights, they are non-trivial extensions of this work.

To control the adaptive weights, we define:

$$a_n = \max\{w_{n,kj} : \beta_{n,kj}^0 \neq 0\},$$
$$b_n = \min\{w_{n,kj} : \beta_{n,kj}^0 = 0\}.$$

We assume that

$$0 < c_1 < \min\{\beta_{n,kj}^0 : \beta_{n,kj}^0 \neq 0\}$$
$$< \max\{\beta_{n,kj}^0 : \beta_{n,kj}^0 \neq 0\} < c_2 < \infty.$$

Then we have the following results.

Theorem 3. *For every n, the observations $\{\boldsymbol{V}_{ni}, i = 1, 2, \ldots, n\}$ are independent and identically distributed, each with a density $f_n(\boldsymbol{V}_{n1}, \boldsymbol{\beta}_n)$ that satisfies conditions (A1)–(A3) in the Appendix. If $\frac{P_n^4}{n} \to 0$ and $P_n^2 \lambda_n \sqrt{a_n} = o_p(1)$, then there exists a local maximizer $\hat{\boldsymbol{\beta}}_n$ of $Q_n(\boldsymbol{\beta}_n)$ such that $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^0\| = O_p(\sqrt{P_n}(n^{-1/2} + \lambda_n \sqrt{a_n}))$.*

Hence by choosing $\lambda_n \sqrt{a_n} = O_p(n^{-1/2})$, there exists a root-$(n/P_n)$ consistent penalized likelihood estimate.

Theorem 4. *For every n, the observations $\{\boldsymbol{V}_{ni}, i = 1, 2, \ldots, n\}$ are independent and identically distributed, each with a density $f_n(\boldsymbol{V}_{n1}, \boldsymbol{\beta}_n)$ that satisfies conditions (A1)–(A3) in the Appendix. If $\frac{P_n^4}{n} \to 0$, $\lambda_n \sqrt{a_n} = O_p(n^{-1/2})$ and $\frac{P_n^2}{\lambda_n^2 b_n} = o_p(n)$, then there exists a root-(n/P_n) consistent local maximizer $\hat{\boldsymbol{\beta}}_n$ such that:*

(a) *Sparsity:* $\Pr(\hat{\boldsymbol{\beta}}_{n,\mathcal{D}_n} = 0) \to 1$, *where $\mathcal{D}_n = \mathcal{B}_n \cup \mathcal{C}_n$.*
(b) *Asymptotic normality: If $\lambda_n \sqrt{a_n} = o_p((nP_n)^{-1/2})$ and $\frac{P_n^5}{n} \to 0$ as $n \to \infty$, then we also have:*

$$\sqrt{n} \boldsymbol{A}_n \boldsymbol{I}_n^{1/2}(\boldsymbol{\beta}_{n,\mathcal{A}_n}^0)(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n} - \boldsymbol{\beta}_{n,\mathcal{A}_n}^0) \to \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}),$$

where $\boldsymbol{A}_n$ is a $q \times |\mathcal{A}_n|$ matrix such that $\boldsymbol{A}_n \boldsymbol{A}_n^\top \to \boldsymbol{G}$ and $\boldsymbol{G}$ is a $q \times q$ nonnegative symmetric matrix. $\boldsymbol{I}_n(\boldsymbol{\beta}_{n,\mathcal{A}_n}^0)$ is the Fisher information matrix knowing $\boldsymbol{\beta}_{\mathcal{D}_n}^0 = 0$.

The above requirements $\lambda_n \sqrt{a_n} = o_p((nP_n)^{-1/2})$ and $\frac{P_n^2}{\lambda_n^2 b_n} = o_p(n)$ as $n \to \infty$ can be satisfied by selecting λ_n and

$w_{n,kj}$ appropriately. For example, we may let $\lambda_n = \frac{(nP_n)^{-1/2}}{\log n}$ and $w_{n,kj} = \frac{1}{|\hat{\beta}_{n,kj}^0|^2}$, where $\hat{\beta}_{n,kj}^0$ is the un-penalized likelihood estimate of $\beta_{n,kj}^0$, which is root-(n/P_n) consistent. Then we have $a_n = O_p(1)$ and $\frac{1}{b_n} = O_p(\frac{P_n}{n})$. Hence $\lambda_n \sqrt{a_n} = o_p((nP_n)^{-1/2})$ and $\frac{P_n^2}{\lambda_n^2 b_n} = o_p(n)$ are satisfied when $\frac{P_n^5}{n} \to 0$.

5.2 Likelihood ratio test

Similarly as in [7], we develop a likelihood ratio test for testing linear hypotheses:

$$H_0 : \boldsymbol{A}_n \boldsymbol{\beta}_{n,\mathcal{A}_n}^0 = 0 \quad \text{vs.} \quad H_1 : \boldsymbol{A}_n \boldsymbol{\beta}_{n,\mathcal{A}_n}^0 \neq 0,$$

where $\boldsymbol{A}_n$ is a $q \times |\mathcal{A}_n|$ matrix and $\boldsymbol{A}_n \boldsymbol{A}_n^\mathsf{T} \to \boldsymbol{I}_q$ for a fixed q. This problem includes testing simultaneously the significance of several covariate variables.

We introduce a natural likelihood ratio test statistic, i.e.

$$T_n = 2\Big\{ \sup_{\Omega_n} Q_n(\boldsymbol{\beta}_n | \boldsymbol{V}) - \sup_{\Omega_n, \boldsymbol{A}_n \boldsymbol{\beta}_{n,\mathcal{A}_n}=0} Q_n(\boldsymbol{\beta}_n | \boldsymbol{V}) \Big\},$$

where Ω_n is the parameter space for $\boldsymbol{\beta}_n$. Then we can obtain the following theorem regarding the asymptotic null distribution of the test statistic.

Theorem 5. *When conditions in* (b) *of Theorem 4 are satisfied, under H_0 we have*

$$T_n \to \chi_q^2, \quad \text{as } n \to \infty.$$

6. SIMULATION STUDY

In this section, we use simulations to demonstrate the hierarchical lasso (HLasso) method, and compare the results with those of some existing methods.

Specifically, we first compare hierarchical lasso with some other group variable selection methods, i.e., the L_2-norm group lasso (4) and the L_∞-norm group lasso (5). Then we compare the adaptive hierarchical lasso with some other "oracle" (but non-group variable selection) methods, i.e., the SCAD and the adaptive lasso.

We extended the simulations in [23]. We considered a model which had both categorical and continuous prediction variables. We first generated seventeen independent standard normal variables, $Z_1, \ldots, Z_{16}$ and W. The covariates were then defined as $X_j = (Z_j + W)/\sqrt{2}$. Then the last eight covariates $X_9, \ldots, X_{16}$ were discretized to 0, 1, 2, and 3 by $\Phi^{-1}(1/4)$, $\Phi^{-1}(1/2)$ and $\Phi^{-1}(3/4)$. Each of $X_1, \ldots, X_8$ was expanded through a fourth-order polynomial, and only the main effects of $X_9, \ldots, X_{16}$ were considered. This gave us a total of eight continuous groups with four variables in each group and eight categorical groups with three variables in each group. We considered two cases.

Case 1. "All-in-all-out"

$$\begin{aligned} Y = &[X_3 + 0.5X_3^2 + 0.1X_3^3 + 0.1X_3^4] \\ &+ [X_6 - 0.5X_6^2 + 0.15X_6^3 + 0.1X_6^4] \\ &+ [\mathbb{I}(X_9 = 0) + \mathbb{I}(X_9 = 1) + \mathbb{I}(X_9 = 2)] + \varepsilon. \end{aligned}$$

Case 2. "Not all-in-all-out"

$$\begin{aligned} Y = &(X_3 + X_3^2) + (2X_6 - 1.5X_6^2) \\ &+ [\mathbb{I}(X_9 = 0) + 2\,\mathbb{I}(X_9 = 1)] + \varepsilon. \end{aligned}$$

For all the simulations above, the error term ε follows a normal distribution $\mathrm{N}(0, \sigma^2)$, where σ^2 was set such that each of the signal to noise ratios, $\mathrm{Var}(\boldsymbol{X}^\mathsf{T} \boldsymbol{\beta})/\mathrm{Var}(\epsilon)$, was equal to 3. We generated $n = 400$ training observations from each of the above models, along with 200 validation observations and 10,000 test observations. The validation set was used to select the tuning parameters λ's that minimized the validation error. Using these selected λ's, we calculated the mean squared error (MSE) with the test set. We repeated this 200 times and computed the average MSEs and their corresponding standard errors. We also recorded how frequently the important variables were selected and how frequently the unimportant variables were removed. The results are summarized in Table 1.

As we can see, all shrinkage methods perform much better than OLS; this illustrates that some regularization is crucial for prediction accuracy. In terms of prediction accuracy, we can also see that when variables in a group follow the "all-in-all-out" pattern, the L_2-norm (group lasso) method performs slightly better than the hierarchical lasso method (Case 1 of Table 1). When variables in a group do not follow the "all-in-all-out" pattern, however, the hierarchical lasso method performs slightly better than the L_2-norm method (Case 2 of Table 1). For variable selection, we can see that in terms of identifying important variables, the four shrinkage methods, the lasso, the L_∞-norm, the L_2-norm, and the hierarchical lasso all perform similarly ("Non-zero Var." of Table 1). However, the L_2-norm method and the hierarchical lasso method are more effective at removing unimportant variables than lasso and the L_∞-norm method ("Zero Var." of Table 1).

In the above analysis, we used the criterion (8) or (11) for the hierarchical lasso, i.e., we did not use the adaptive weights w_{kj} to penalize different coefficients differently. To assess the improved version of the hierarchical lasso, i.e. criterion (14), we also considered using adaptive weights. Specifically, we applied the OLS weights in (15) to (14) with $\gamma = 1$. We compared the results with those of SCAD and the adaptive lasso, which also enjoy the oracle property. However, we note that SCAD and the adaptive lasso do not take advantage of the grouping structure information. As a benchmark, we also computed the Oracle OLS results, i.e., OLS using only the important variables. The results

Table 1. Comparison of several group variable selection methods, including the L_2-norm group lasso, the L_∞-norm group lasso and the hierarchical lasso. The OLS and the regular lasso are used as benchmarks. The upper part is for Case 1, and the lower part is for Case 2. "MSE" is the mean squared error on the test set. "Zero Var." is the percentage of correctly removed unimportant variables. "Non-zero Var." is the percentage of correctly identified important variables. All the numbers outside parentheses are means over 200 repetitions, and the numbers in the parentheses are the corresponding standard errors

Case 1: "All-in-all-out"					
	OLS	Lasso	L_∞	L_2	HLasso
MSE	0.92 (0.018)	0.47 (0.011)	0.31 (0.008)	0.18 (0.009)	0.24 (0.008)
Zero Var.	–	57% (1.6%)	29% (1.4%)	96% (0.8%)	94% (0.7%)
Non-Zero Var.	–	92% (0.6%)	100% (0%)	100% (0%)	98% (0.3%)
Case 2: "Not all-in-all-out"					
	OLS	Lasso	L_∞	L_2	HLasso
MSE	0.91 (0.018)	0.26 (0.008)	0.46 (0.012)	0.21 (0.01)	0.15 (0.006)
Zero Var.	–	70% (1%)	17% (1.2%)	87% (0.8%)	91% (0.5%)
Non-zero Var.	–	99% (0.3%)	100% (0%)	100% (0.2%)	100% (0.1%)

Table 2. Comparison of several "oracle" methods, including the adaptive hierarchical lasso, SCAD and the adaptive lasso. SCAD and adaptive lasso do not take advantage of the grouping structure information. The Oracle OLS uses only important variables. Descriptions for the rows are the same as those in the caption of Table 1

Case 1: "All-in-all-out"				
	Oracle OLS	Ada Lasso	SCAD	Ada HLasso
MSE	0.16 (0.006)	0.37 (0.011)	0.35 (0.011)	0.24 (0.009)
Zero Var.	–	77% (0.7%)	79% (1.1%)	98% (0.3%)
Non-Zero Var.	–	94% (0.5%)	91% (0.6%)	96% (0.5%)
Case 2: "Not all-in-all-out"				
	Oracle OLS	Ada Lasso	SCAD	Ada HLasso
MSE	0.07 (0.003)	0.13 (0.005)	0.11 (0.004)	0.10 (0.005)
Zero Var.	–	91% (0.3%)	91% (0.4%)	98% (0.1%)
Non-zero Var.	–	98% (0.4%)	99% (0.3%)	99% (0.3%)

are summarized in Table 2. We can see that in the "all-in-all-out" case, the adaptive hierarchical lasso removes unimportant variables more effectively than SCAD and adaptive lasso, and consequently, the adaptive hierarchical lasso outperforms SCAD and adaptive lasso by a significant margin in terms of prediction accuracy. In the "not all-in-all-out" case, the advantage of knowing the grouping structure information is reduced, however, the adaptive hierarchical lasso still performs slightly better than SCAD and adaptive lasso, especially in terms of removing unimportant variables.

To assess how the sample size affects different "oracle" methods, we also considered $n = 200, 400, 800, 1,600$ and 3,200. The results are summarized in Figure 1, where the first row corresponds to the "all-in-all-out" case, and the second row corresponds to the "not all-in-all-out" case. Not surprisingly, as the sample size increases, the performances of different methods all improve: in terms of prediction accuracy, the MSE's all decrease (at about the same rate) and get closer to that of the Oracle OLS; in terms of variable selection, the probabilities of identifying the correct model all increase and approach one. However, overall, the adaptive hierarchical lasso always performs the best among the three "oracle" methods, and the gap is especially prominent in terms of removing unimportant variables when the sample size is moderate.

7. REAL DATA ANALYSIS

In this section, we use a gene expression dataset from the NCI-60 collection of cancer cell lines to further illustrate the hierarchical lasso method. We sought to use this dataset to identify targets of the transcription factor p53, which regulates gene expression in response to various signals of cellular stress. The mutational status of the p53 gene has been reported for 50 of the NCI-60 cell lines, with 17 being classified as normal and 33 as carrying mutations [16].

Instead of single-gene analysis, gene-set information has recently been used to analyze gene expression data. For example, a Gene Set Enrichment Analysis (GSEA) is developed [17], which is found to be more stable and more powerful than single-gene analysis. In [5], the GSEA method is improved by using new statistics for summarizing gene-sets. Both methods are based on hypothesis testing. In this analysis, we consider using the hierarchical lasso method for gene-set selection. The gene-sets used here are the cytogenetic gene-sets and the functionals gene-sets from the GSEA package [17]. We considered 391 overlapping gene-sets with the size of each set greater than 15.

Since the response here is binary (normal vs mutation), following the result in Section 5.1, we use the logistic hierarchical lasso regression, instead of the least square hierarchical lasso. Note that a gene may belong to multiple
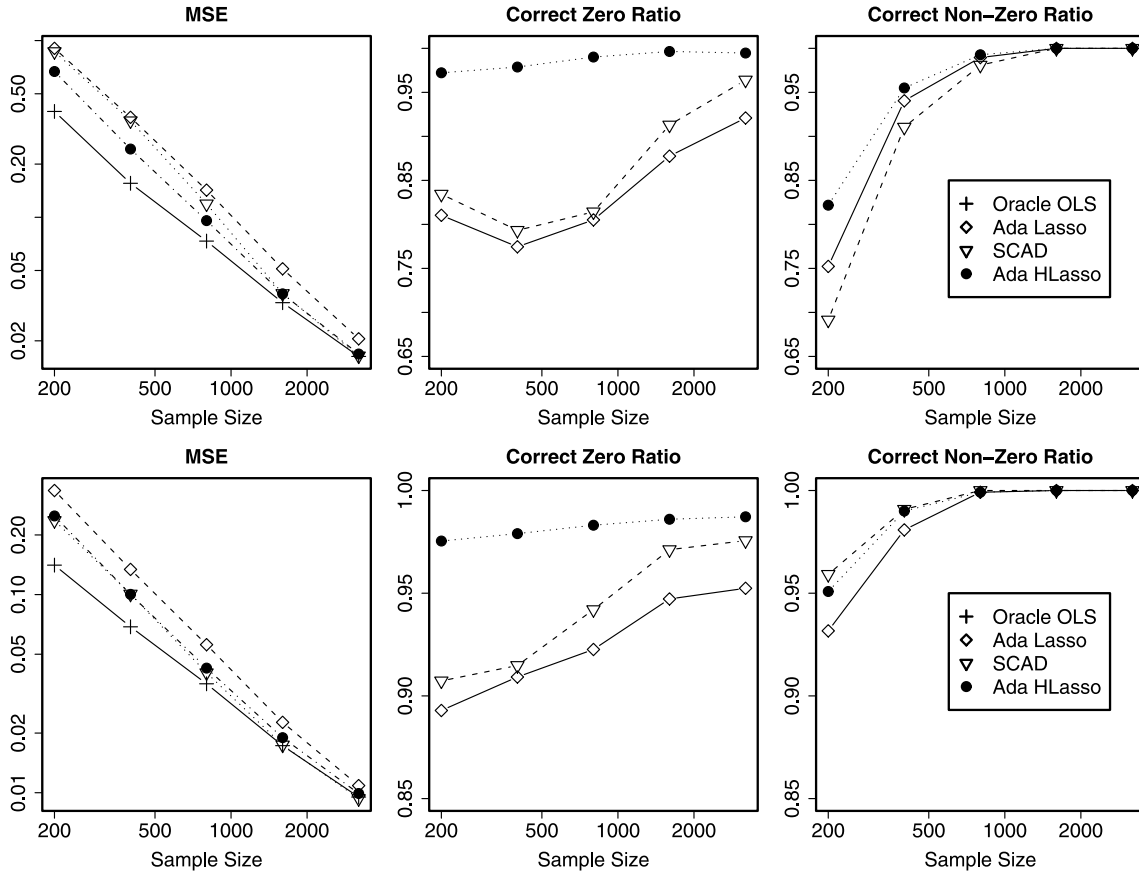
Figure 1. Comparison of several oracle methods, including the SCAD, the adaptive lasso and the adaptive hierarchical lasso. SCAD and adaptive lasso do not take advantage of the grouping structure information. The Oracle OLS uses only important variables. The first row corresponds to the "all-in-all-out" case, and the second row corresponds to the "not all-in-all-out" case. "Correct zero ratio" records the percentage of correctly removed unimportant variables. "Correct non-zero ratio" records the percentage of correctly identified important variables.

gene-sets, we thus extend the hierarchical lasso to the case of overlapping groups. Suppose there are K groups and J variables. Let $\mathcal{G}_k$ denote the set of indices of the variables in the kth group. One way to model the overlapping situation is to extend the criterion (8) as the following:

$$(17) \quad \max_{d_k, \alpha_j} \sum_{i=1}^n \ell\left(\sum_{k=1}^K d_k \sum_{j:j \in \mathcal{G}_k} \alpha_j x_{i,j}, \ y_i \right)$$
$$- \sum_{k=1}^K \lambda_k d_k - \lambda \cdot \sum_{j=1}^J |\alpha_j|$$
$$\text{subject to} \quad d_k \geq 0, \ k = 1, \dots, K,$$

where α_j can be considered as the "intrinsic" effect of a variable (no matter which group it belongs to), and different group effects are represented via different d_k. Since the number of genes in each gene-set may differ significantly from set to set, we use λ_k to allow different gene-sets to be penalized differently. Specifically, we have considered $\lambda_k = 1$,

$\lambda_k = \sqrt{p_k}$ and $\lambda_k = p_k$; we will report the results based on $\lambda_k = \sqrt{p_k}$ as it echoes the square root in Section 4. Similar adjustments can also be applied to the methods that we have compared with. In this section, $\ell(\eta_i, y_i) = y_i \eta_i - \log(1 + e^{\eta_i})$ is the logistic log-likelihood function with y_i being a 0/1 response. Also notice that if each variable belongs to only one group, the model reduces to the non-overlapping criterion (8).

We randomly split the 50 samples into the training and test sets 100 times; for each split, 33 samples (22 carrying mutations and 11 being normal) were used for training and the rest 17 samples (11 carrying mutations and 6 being normal) were for testing. For each split, we applied three methods, the logistic lasso, the logistic L_2-norm group lasso [14] and the logistic hierarchical lasso. Tuning parameters were chosen using five-fold cross-validation.

We first compare the prediction accuracy of the three methods. Over the 100 random splits, the logistic hierarchical lasso has an average misclassification rate of 14% with the standard error 1.8%, which is smaller than 23% (1.7%)
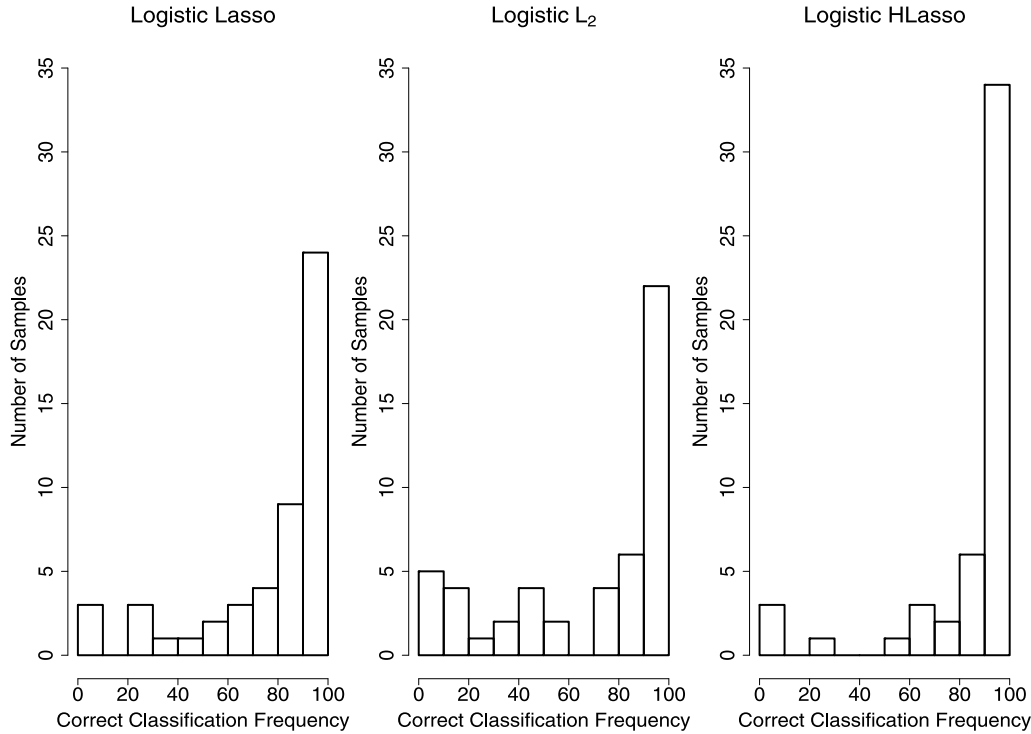
Figure 2. The number of samples vs the frequency that a sample was correctly classified on 100 random splits of the p53 data.

of the logistic lasso and 32% (1.2%) of the logistic group lasso. To assess the stability of the prediction, we recorded the frequency in which each sample, as a test observation, was correctly classified. For example, if a sample appeared in 40 test sets among the 100 random splits, and out of the 40 predictions, the sample was correctly classified 36 times, we recorded 36/40 for this sample. The results are shown in Figure 2. As we can see, for most samples, the logistic hierarchical lasso classified them correctly for most of the random splits, and the predictions seemed to be slightly more stable than the logistic lasso and the logistic L_2-norm group lasso.

Next, we compare gene-set selection of these three methods. The most notable difference is that both logistic lasso and the logistic hierarchical lasso selected gene CDKN1A most frequently out of the 100 random split, while the logistic L_2-norm group lasso rarely selected it. CDKN1A is also named as wild-type p53 activated fragment-1 (p21), and it is known that the expression of gene CDKN1A is tightly controlled by the tumor suppressor protein p53, through which this protein mediates the p53-dependent cell cycle G1 phase arrest in response to a variety of stress stimuli [13].

We also compared the gene-sets selected by the logistic hierarchical lasso with those selected by the GSEA [17] and the GSA [5]. The two most frequently selected gene-sets by the hierarchical lasso are *atm pathway* and *radiation sensitivity*. The most frequently selected genes in *atm pathway* by the logistic hierarchical lasso are CDKN1A, MDM2 and RELA, and the most frequently selected genes in *radiation*

sensitivity are CDKN1A, MDM2 and BCL2. It is known that MDM2, the second commonly selected gene, is a target gene of the transcription factor tumor protein p53, and the encoded protein in MDM2 is a nuclear phosphoprotein that binds and inhibits transactivation by tumor protein p53, as part of an autoregulatory negative feedback loop [11, 15]. Note that the gene-set *radiation sensitivity* was also selected by GSEA and GSA. Though the gene-set *atm pathway* was not selected by GSEA and GSA, it shares 7, 8, 6, and 3 genes with gene-sets *radiation sensitivity*, *p53 signalling*, *p53 hypoxia pathway* and *p53 Up* respectively, which were all selected by GSEA and GSA. We also note that GSEA and GSA are based on the *marginal* strength of each gene-set, while the logistic hierarchical lasso fits an "additive" model and uses the *joint* strengths of gene-sets.

8. DISCUSSION

In this paper, we have proposed a hierarchical lasso method for group variable selection. Different variable selection methods have their own advantages in different scenarios. The hierarchical lasso method not only effectively removes unimportant groups, but also keeps the flexibility of selecting variables within a group. We show that the improved hierarchical lasso method enjoys an oracle property, i.e., it performs as well as if the true sub-model were given in advance. Numerical results indicate that our method works well, especially when variables in a group are associated with the response in a "not all-in-all-out" fashion.

The grouping idea is also applicable to other regression and classification settings, for example, the multi-response regression and multi-class classification problems. In these problems, a grouping structure may not exist among the prediction variables, but instead, natural grouping structures exist among *parameters*. We use the multi-response regression problem to illustrate the point [3, 19]. Suppose we observe $(\boldsymbol{x}_1, \boldsymbol{y}_1)$, ..., $(\boldsymbol{x}_n, \boldsymbol{y}_n)$, where each $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iK})$ is a vector containing K responses, and we are interested in selecting a subset of the prediction variables that predict well for all of the multiple responses. Standard techniques estimate K prediction functions, one for each of the K responses, $f_k(\boldsymbol{x}) = \beta_{k1}x_1 + \cdots + \beta_{kp}x_p, k = 1, \ldots, K$. The prediction variables $(x_1, \ldots, x_p)$ may not have a grouping structure, however, we may consider the coefficients corresponding to the same prediction variable form a natural group, i.e., $(\beta_{1j}, \beta_{2j}, \ldots, \beta_{Kj})$. Using our hierarchical lasso idea, we reparameterize $\beta_{kj} = d_j\alpha_{kj}, d_j \geq 0$, and we consider

$$\max_{d_j \geq 0, \alpha_{kj}} - \frac{1}{2} \sum_{k=1}^{K} \sum_{i=1}^{n} \left(y_{ik} - \sum_{j=1}^{p} d_j\alpha_{kj}x_{ij} \right)^2$$
$$- \lambda_1 \cdot \sum_{j=1}^{p} d_j - \lambda_2 \cdot \sum_{j=1}^{p} \sum_{k=1}^{K} |\alpha_{kj}|.$$

Note that if d_j is shrunk to zero, all $\beta_{kj}, k = 1, \ldots, K$ will be equal to zero, hence the jth prediction variable will be removed from all K predictions. If d_j is not equal to zero, then some of the α_{kj} and hence some of the $\beta_{kj}, k = 1, \ldots, K$, still have the possibility of being zero. Therefore, the jth variable may be predictive for some responses but non-predictive for others.

One referee mentioned the paper [20], for which we are the co-authors. Most of this work was finished before the start of [20], which focuses on survival analysis. Furthermore, in [20], the number of prediction variables and the number of groups are fixed, while in this work, we allow both $p_{nk} \to \infty$ and $K_n \to \infty$ as $n \to \infty$.

Another referee pointed out the work in [10], which we were not aware of when our manuscript was first completed and submitted in 2007. It is true that if the penalty in (8) is modified to be $\sum_{k=1}^{K} d_k^{1/\gamma} + \lambda \sum_{k=1}^{K} \sum_{j=1}^{p_k} |\alpha_{kj}|$, then using the same argument as in Theorem 1, one can show that the hierarchical model is equivalent to a penalized linear regression model with the bridge penalty $\sum_{k=1}^{K} (|\beta_{k1}| + |\beta_{k2}| + \cdots + |\beta_{kp_k}|)^{1/(1+\gamma)}$ as in [10]. Thus we acknowledge that the work in [10] is closely related with ours, but there are also differences. For example:

- We proved the oracle property for both group selection and within group selection, while the oracle property is considered only for group selection in [10].
- Our theory applies to the generalized maximum likelihood estimate, while the penalized least squares estimate is considered in [10].

- Handling overlapping groups. It is not unusual for a variable to be a member of several groups. The gene expression date we analyzed in Section 7 is such an example: given a plethora of biologically defined gene-sets, not surprisingly, there will be considerable overlap among these sets. In [10], a prediction variable that appears in more than one group gets penalized more heavily than variables appearing in only one group. Therefore, a prediction variable belonging to multiple groups is more likely to be removed than a variable belonging to only one group. We are not sure whether this is an appealing property. In our approach, as shown in (17), if a prediction variable belongs to multiple groups, it does not get penalized more heavily than other variables that belong to only one group.

ACKNOWLEDGMENTS

APPENDIX

Proof of Lemma 1

Let $Q^*(\lambda_1, \lambda_2, \boldsymbol{d}, \boldsymbol{\alpha})$ be the criterion that we would like to maximize in equation (7) and let $Q^\star(\lambda, \boldsymbol{d}, \boldsymbol{\alpha})$ be the corresponding criterion in equation (8).

Let $(\hat{\boldsymbol{d}}^*, \hat{\boldsymbol{\alpha}}^*)$ be a local maximizer of $Q^*(\lambda_1, \lambda_2, \boldsymbol{d}, \boldsymbol{\alpha})$. We would like to prove $(\hat{\boldsymbol{d}}^\star = \lambda_1\hat{\boldsymbol{d}}^*, \hat{\boldsymbol{\alpha}}^\star = \hat{\boldsymbol{\alpha}}^*/\lambda_1)$ is a local maximizer of $Q^\star(\lambda, \boldsymbol{d}, \boldsymbol{\alpha})$.

We immediately have

$$Q^*(\lambda_1, \lambda_2, \boldsymbol{d}, \boldsymbol{\alpha}) = Q^\star(\lambda, \lambda_1\boldsymbol{d}, \boldsymbol{\alpha}/\lambda_1).$$

Since $(\hat{\boldsymbol{d}}^*, \hat{\boldsymbol{\alpha}}^*)$ is a local maximizer of $Q^*(\lambda_1, \lambda_2, \boldsymbol{d}, \boldsymbol{\alpha})$, there exists $\delta > 0$ such that if $\boldsymbol{d}', \boldsymbol{\alpha}'$ satisfy $\|\boldsymbol{d}' - \hat{\boldsymbol{d}}^*\| + \|\boldsymbol{\alpha}' - \hat{\boldsymbol{\alpha}}^*\| < \delta$ then $Q^*(\lambda_1, \lambda_2, \boldsymbol{d}', \boldsymbol{\alpha}') \leq Q^*(\lambda_1, \lambda_2, \hat{\boldsymbol{d}}^*, \hat{\boldsymbol{\alpha}}^*)$.

Choose δ' such that $\frac{\delta'}{\min(\lambda_1, \frac{1}{\lambda_1})} \leq \delta$, for any $(\boldsymbol{d}'', \boldsymbol{\alpha}'')$ satisfying $\|\boldsymbol{d}'' - \hat{\boldsymbol{d}}^\star\| + \|\boldsymbol{\alpha}'' - \hat{\boldsymbol{\alpha}}^\star\| < \delta'$ we have

$$\left\| \frac{\boldsymbol{d}''}{\lambda_1} - \hat{\boldsymbol{d}}^* \right\| + \|\lambda_1\boldsymbol{\alpha}'' - \hat{\boldsymbol{\alpha}}^*\|$$
$$\leq \frac{\lambda_1\|\frac{\boldsymbol{d}''}{\lambda_1} - \hat{\boldsymbol{d}}^*\| + \frac{1}{\lambda_1}\|\lambda_1\boldsymbol{\alpha}'' - \hat{\boldsymbol{\alpha}}^*\|}{\min(\lambda_1, \frac{1}{\lambda_1})}$$
$$= \frac{\|\boldsymbol{d}'' - \hat{\boldsymbol{d}}^\star\| + \|\boldsymbol{\alpha}'' - \hat{\boldsymbol{\alpha}}^\star\|}{\min(\lambda_1, \frac{1}{\lambda_1})} < \frac{\delta'}{\min(\lambda_1, \frac{1}{\lambda_1})} < \delta.$$

Hence

$$Q^\star(\lambda, \hat{\boldsymbol{d}}'', \hat{\boldsymbol{\alpha}}'') = Q^*(\lambda_1, \lambda_2, \boldsymbol{d}''/\lambda_1, \lambda_1\hat{\boldsymbol{\alpha}}'')$$
$$\leq Q^*(\lambda_1, \lambda_2, \hat{\boldsymbol{d}}^*, \hat{\boldsymbol{\alpha}}^*) = Q^\star(\lambda, \hat{\boldsymbol{d}}^\star, \hat{\boldsymbol{\alpha}}^\star).$$

Therefore, $(\hat{\boldsymbol{d}}^{\star} = \lambda_1 \hat{\boldsymbol{d}}^{*}, \hat{\boldsymbol{\alpha}}^{\star} = \hat{\boldsymbol{\alpha}}^{*}/\lambda_1)$ is a local maximizer of $Q^{\star}(\lambda, \boldsymbol{d}, \boldsymbol{\alpha})$.

Similarly we can prove that for any local maximizer $(\hat{\boldsymbol{d}}^{\star}, \hat{\boldsymbol{\alpha}}^{\star})$ of $Q^{\star}(\lambda, \boldsymbol{d}, \boldsymbol{\alpha})$, there is a corresponding local maximizer $(\hat{\boldsymbol{d}}^{*}, \hat{\boldsymbol{\alpha}}^{*})$ of $Q^{*}(\lambda_1, \lambda_2, \boldsymbol{d}, \boldsymbol{\alpha})$ such that $\hat{d}_k^{*}\hat{\alpha}_{kj}^{*} = \hat{d}_k^{\star}\hat{\alpha}_{kj}^{\star}$.

Lemma 2. *Suppose $(\hat{\boldsymbol{d}}, \hat{\boldsymbol{\alpha}})$ is a local maximizer of* (8). *Let $\hat{\boldsymbol{\beta}}$ be the Hierarchical Lasso estimate related to $(\hat{\boldsymbol{d}}, \hat{\boldsymbol{\alpha}})$, i.e., $\hat{\beta}_{kj} = \hat{d}_k\hat{\alpha}_{kj}$. If $\hat{d}_k = 0$, then $\hat{\boldsymbol{\alpha}}_k = 0$; if $\hat{d}_k \neq 0$, then $\|\hat{\boldsymbol{\beta}}_k\|_1 \neq 0$ and $\hat{d}_k = \sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k\|_1}, \hat{\boldsymbol{\alpha}}_k = \frac{\hat{\boldsymbol{\beta}}_k}{\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k\|_1}}$.*

Proof of Lemma 2

If $\hat{d}_k = 0$, then $\hat{\boldsymbol{\alpha}}_k = 0$ is quite obvious. Similarly, if $\hat{\boldsymbol{\alpha}}_k = 0$, then $\hat{d}_k = 0$. Therefore, if $\hat{d}_k \neq 0$, then $\hat{\boldsymbol{\alpha}}_k \neq 0$ and $\|\hat{\boldsymbol{\beta}}_k\|_1 \neq 0$.

We prove $\hat{d}_k = \sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k\|_1}, \hat{\boldsymbol{\alpha}}_k = \frac{\hat{\boldsymbol{\beta}}_k}{\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k\|_1}}$ for $\hat{d}_k \neq 0$ by contradiction. Suppose $\exists k'$ such that $\hat{d}_{k'} \neq 0$ and $\hat{d}_{k'} \neq \sqrt{\lambda\|\hat{\boldsymbol{\beta}}_{k'}\|_1}$. Let $\frac{\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_{k'}\|_1}}{\hat{d}_{k'}} = c$. Then $\hat{\boldsymbol{\alpha}}_k = c\frac{\hat{\boldsymbol{\beta}}_k}{\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k\|_1}}$. Suppose $c > 1$.

Let $\tilde{d}_k = \hat{d}_k$ and $\tilde{\boldsymbol{\alpha}}_k = \hat{\boldsymbol{\alpha}}_k$ for $k \neq k'$ and $\tilde{d}_{k'} = \delta'\hat{d}_{k'}$ and $\tilde{\boldsymbol{\alpha}}_{k'} = \hat{\boldsymbol{\alpha}}_{k'}\frac{1}{\delta'}$, where δ' satisfies $c > \delta' > 1$ and is very close to 1 such that $\|\tilde{d}_{k'} - \hat{d}_{k'}\|_1 + \|\tilde{\boldsymbol{\alpha}}_{k'} - \hat{\boldsymbol{\alpha}}_{k'}\|_1 < \delta$ for some $\delta > 0$.

Then we have

$$Q^{\star}(\lambda, \tilde{\boldsymbol{d}}, \tilde{\boldsymbol{\alpha}}) - Q^{\star}(\lambda, \hat{\boldsymbol{d}}, \hat{\boldsymbol{\alpha}})$$
$$= -\delta'|\hat{d}_{k'}| - \frac{1}{\delta'}\lambda\|\hat{\boldsymbol{\alpha}}_{k'}\|_1 + |\hat{d}_{k'}| + \lambda\|\hat{\boldsymbol{\alpha}}_{k'}\|_1$$
$$= \left(-\frac{\delta'}{c} - \frac{c}{\delta'} + \frac{1}{c} + c\right)\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_{k'}\|_1}$$
$$= \frac{1}{c}(\delta' - 1)\left(\frac{c^2}{\delta'} - 1\right)\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_{k'}\|_1} > 0.$$

Therefore, for any $\delta > 0$, we can find $\tilde{\boldsymbol{d}}, \tilde{\boldsymbol{\alpha}}$ such that $\|\tilde{\boldsymbol{d}} - \hat{\boldsymbol{d}}\|_1 + \|\tilde{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}\|_1 < \delta$ and $Q^{\star}(\lambda, \tilde{\boldsymbol{d}}, \tilde{\boldsymbol{\alpha}}) > Q^{\star}(\lambda, \hat{\boldsymbol{d}}, \hat{\boldsymbol{\alpha}})$. These contradict with $(\hat{\boldsymbol{d}}, \hat{\boldsymbol{\alpha}})$ being a local maximizer.

Similarly for the case when $c < 1$. Hence, we have the result that if $\hat{d}_k \neq 0$, then $\hat{d}_k = \sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k\|_1}, \hat{\boldsymbol{\alpha}}_k = \frac{\hat{\boldsymbol{\beta}}_k}{\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k\|_1}}$.

Proof of Theorem 1

Let $Q(\lambda, \boldsymbol{\beta})$ be the corresponding criterion in equation (11).

Suppose $(\hat{\boldsymbol{d}}, \hat{\boldsymbol{\alpha}})$ is a local maximizer of $Q^{\star}(\lambda, \boldsymbol{d}, \boldsymbol{\alpha})$, we first show that $\hat{\boldsymbol{\beta}}$, where $\hat{\beta}_{kj} = \hat{d}_k\hat{\alpha}_{kj}$, is a local maximizer of $Q(\lambda, \boldsymbol{\beta})$, i.e. there exists a δ' such that if $\|\triangle\boldsymbol{\beta}\|_1 < \delta'$ then $Q(\lambda, \hat{\boldsymbol{\beta}} + \triangle\boldsymbol{\beta}) \leq Q(\lambda, \hat{\boldsymbol{\beta}})$.

We denote $\triangle\boldsymbol{\beta} = \triangle\boldsymbol{\beta}^{(1)} + \triangle\boldsymbol{\beta}^{(2)}$, where $\triangle\boldsymbol{\beta}_k^{(1)} = 0$ if $\|\hat{\boldsymbol{\beta}}_k\|_1 = 0$ and $\triangle\boldsymbol{\beta}_k^{(2)} = 0$ if $\|\hat{\boldsymbol{\beta}}_k\|_1 \neq 0$. We have $\|\triangle\boldsymbol{\beta}\|_1 = \|\triangle\boldsymbol{\beta}^{(1)}\|_1 + \|\triangle\boldsymbol{\beta}^{(2)}\|_1$.

Now we show $Q(\lambda, \hat{\boldsymbol{\beta}} + \triangle\boldsymbol{\beta}^{(1)}) \leq Q(\lambda, \hat{\boldsymbol{\beta}})$ if δ' is small enough. By Lemma 2, we have $\hat{d}_k = \sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k\|_1}, \hat{\boldsymbol{\alpha}}_k = \frac{\hat{\boldsymbol{\beta}}_k}{\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k\|_1}}$ if $\|\hat{d}_k\|_1 \neq 0$ and $\hat{\boldsymbol{\alpha}}_k = \boldsymbol{0}$ if $\|\hat{d}_k\|_1 = 0$. Furthermore, let $\hat{d}_k' = \sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k + \triangle\boldsymbol{\beta}_k^{(1)}\|_1}, \hat{\boldsymbol{\alpha}}_k' = \frac{\hat{\boldsymbol{\beta}}_k + \triangle\boldsymbol{\beta}_k^{(1)}}{\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k + \triangle\boldsymbol{\beta}_k^{(1)}\|_1}}$ if $\|\hat{d}_k\|_1 \neq 0$. Let $\hat{d}_k' = 0, \hat{\boldsymbol{\alpha}}_k' = \boldsymbol{0}$ if $\|\hat{d}_k\|_1 = 0$. Then we have $Q^{\star}(\lambda, \hat{\boldsymbol{d}}', \hat{\boldsymbol{\alpha}}') = Q(\lambda, \hat{\boldsymbol{\beta}} + \triangle\boldsymbol{\beta}^{(1)})$ and $Q^{\star}(\lambda, \hat{\boldsymbol{d}}, \hat{\boldsymbol{\alpha}}) = Q(\lambda, \hat{\boldsymbol{\beta}})$. Hence we only need to show that $Q^{\star}(\lambda, \hat{\boldsymbol{d}}', \hat{\boldsymbol{\alpha}}') \leq Q^{\star}(\lambda, \hat{\boldsymbol{d}}, \hat{\boldsymbol{\alpha}})$. Note that $(\hat{\boldsymbol{d}}, \hat{\boldsymbol{\alpha}})$ ia a local maximizer of $Q^{\star}(\lambda, \boldsymbol{d}, \boldsymbol{\alpha})$. Therefore there exists a δ such that for any $\boldsymbol{d}', \boldsymbol{\alpha}'$ satisfying $\|\boldsymbol{d}' - \hat{\boldsymbol{d}}\|_1 + \|\boldsymbol{\alpha}' - \hat{\boldsymbol{\alpha}}\|_1 < \delta$, we have $Q^{\star}(\lambda, \boldsymbol{d}', \boldsymbol{\alpha}') \leq Q^{\star}(\lambda, \hat{\boldsymbol{d}}, \hat{\boldsymbol{\alpha}})$.

Now since

$$|\hat{d}_k' - \hat{d}_k| = |\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k + \triangle\boldsymbol{\beta}_k^{(1)}\|_1} - \sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k\|_1}|$$
$$\leq |\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k\|_1 - \lambda\|\triangle\boldsymbol{\beta}_k^{(1)}\|_1} - \sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k\|_1}|$$
$$\leq \frac{1}{2}\frac{\lambda\|\triangle\boldsymbol{\beta}_k^{(1)}\|_1}{\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k\|_1 - \lambda\|\triangle\boldsymbol{\beta}_k^{(1)}\|_1}}$$
$$\leq \frac{1}{2}\frac{\lambda\|\triangle\boldsymbol{\beta}_k^{(1)}\|_1}{\sqrt{\lambda a - \lambda\delta'}} \leq \frac{1}{2}\frac{\lambda\|\triangle\boldsymbol{\beta}_k^{(1)}\|_1}{\sqrt{\lambda a/2}},$$

where $a = \min\{\|\hat{\boldsymbol{\beta}}_k\|_1 : \|\hat{\boldsymbol{\beta}}_k\|_1 \neq 0\}$ and $\delta' < a/2$.

Furthermore

$$\|\hat{\boldsymbol{\alpha}}_k' - \hat{\boldsymbol{\alpha}}_k\|_1 = \left\|\frac{\hat{\boldsymbol{\beta}}_k + \triangle\boldsymbol{\beta}_k^{(1)}}{\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k + \triangle\boldsymbol{\beta}_k^{(1)}\|_1}} - \frac{\hat{\boldsymbol{\beta}}_k}{\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k\|_1}}\right\|_1$$
$$\leq \left\|\frac{\hat{\boldsymbol{\beta}}_k + \triangle\boldsymbol{\beta}_k^{(1)}}{\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k + \triangle\boldsymbol{\beta}_k^{(1)}\|_1}} - \frac{\hat{\boldsymbol{\beta}}_k}{\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k + \triangle\boldsymbol{\beta}_k^{(1)}\|_1}}\right\|_1$$
$$+ \left\|\frac{\hat{\boldsymbol{\beta}}_k}{\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k + \triangle\boldsymbol{\beta}_k^{(1)}\|_1}} - \frac{\hat{\boldsymbol{\beta}}_k}{\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k\|_1}}\right\|_1$$
$$\leq \frac{\|\triangle\boldsymbol{\beta}_k^{(1)}\|_1}{\sqrt{\lambda a/2}}$$
$$+ \frac{\|\hat{\boldsymbol{\beta}}_k\|_1|\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k + \triangle\boldsymbol{\beta}_k^{(1)}\|_1} - \sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k\|_1}|}{\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k + \triangle\boldsymbol{\beta}_k^{(1)}\|_1}\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k\|_1}}$$
$$\leq \frac{\|\triangle\boldsymbol{\beta}_k^{(1)}\|_1}{\sqrt{\lambda a/2}} + \frac{b}{\sqrt{\lambda a/2}\sqrt{\lambda a}}\left(\frac{1}{2}\frac{\lambda\|\triangle\boldsymbol{\beta}_k^{(1)}\|_1}{\sqrt{\lambda a/2}}\right)$$
$$\leq \|\triangle\boldsymbol{\beta}_k^{(1)}\|_1\left(\frac{1}{\sqrt{\lambda a/2}} + \frac{b}{a\sqrt{\lambda a}}\right),$$

where $b = \max\{\|\hat{\boldsymbol{\beta}}_k\|_1 : \|\hat{\boldsymbol{\beta}}_k\|_1 \neq 0\}$.

Therefore, there exists a small enough δ', if $\|\triangle\boldsymbol{\beta}^{(1)}\|_1 < \delta'$ we have $\|\hat{\boldsymbol{d}}' - \hat{\boldsymbol{d}}\|_1 + \|\hat{\boldsymbol{\alpha}}' - \hat{\boldsymbol{\alpha}}\|_1 < \delta$. Hence $Q^{\star}(\lambda, \hat{\boldsymbol{d}}', \hat{\boldsymbol{\alpha}}') \leq$

$Q^\star(\lambda, \hat{\boldsymbol{d}}, \hat{\boldsymbol{\alpha}})$ (due to local maximality) and $Q(\lambda, \hat{\boldsymbol{\beta}} + \triangle\boldsymbol{\beta}^{(1)}) \le Q(\lambda, \hat{\boldsymbol{\beta}})$.

Next we show $Q(\lambda, \hat{\boldsymbol{\beta}} + \triangle\boldsymbol{\beta}^{(1)} + \triangle\boldsymbol{\beta}^{(2)}) \le Q(\lambda, \hat{\boldsymbol{\beta}} + \triangle\boldsymbol{\beta}^{(1)})$. Note that

$$Q(\lambda, \hat{\boldsymbol{\beta}} + \triangle\boldsymbol{\beta}^{(1)} + \triangle\boldsymbol{\beta}^{(2)}) - Q(\lambda, \hat{\boldsymbol{\beta}} + \triangle\boldsymbol{\beta}^{(1)})$$
$$= \triangle\boldsymbol{\beta}^{(2)\top}\nabla L(\hat{\boldsymbol{\beta}}^*) - \sum_{k=1}^{K}\sqrt{\lambda\|\triangle\boldsymbol{\beta}^{(2)}\|_1},$$

where $\boldsymbol{\beta}^*$ is a vector between $\hat{\boldsymbol{\beta}} + \triangle\boldsymbol{\beta}^{(1)} + \triangle\boldsymbol{\beta}^{(2)}$ and $\hat{\boldsymbol{\beta}} + \triangle\boldsymbol{\beta}^{(1)}$. Since $\|\triangle\boldsymbol{\beta}^{(2)}\|_1 < \delta'$ is small enough, the second term dominates the first term, hence we have $Q(\lambda, \hat{\boldsymbol{\beta}} + \triangle\boldsymbol{\beta}^{(1)} + \triangle\boldsymbol{\beta}^{(2)}) \le Q(\lambda, \hat{\boldsymbol{\beta}} + \triangle\boldsymbol{\beta}^{(1)})$.

Overall, we have that there exists a small enough δ', if $\|\triangle\boldsymbol{\beta}\|_1 < \delta'$, then $Q(\lambda, \hat{\boldsymbol{\beta}} + \triangle\boldsymbol{\beta}) \le Q(\lambda, \hat{\boldsymbol{\beta}})$, which implies that $\hat{\boldsymbol{\beta}}$ is a local maximizer of $Q(\lambda, \boldsymbol{\beta})$.

Similarly, we can prove that if $\hat{\boldsymbol{\beta}}$ is a local maximizer of $Q(\lambda, \boldsymbol{\beta})$, and if we let $\hat{d}_k = \sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k\|_1}, \hat{\boldsymbol{\alpha}}_k = \frac{\hat{\boldsymbol{\beta}}_k}{\sqrt{\lambda\|\hat{\boldsymbol{\beta}}_k\|_1}}$ for $\|\hat{\boldsymbol{\beta}}_k\|_1 \ne 0$ and let $\hat{d}_k = 0, \hat{\boldsymbol{\alpha}}_k = \boldsymbol{0}$ for $\|\hat{\boldsymbol{\beta}}_k\|_1 = 0$, then $(\hat{\boldsymbol{d}}, \hat{\boldsymbol{\alpha}})$ is a local maximizer of $Q^\star(\lambda, \boldsymbol{d}, \boldsymbol{\alpha})$.

Regularity conditions

Let S_n be the number of non-zero groups, i.e., $\|\boldsymbol{\beta}_{nk}^0\| \ne 0$. Without loss of generality, we assume

$$\|\boldsymbol{\beta}_{nk}^0\| \ne 0, \quad \text{for } k = 1, \ldots, S_n,$$
$$\|\boldsymbol{\beta}_{nk}^0\| = 0, \quad \text{for } k = S_n + 1, \ldots, K_n.$$

Let s_{nk} be the number of non-zero coefficients in group $k, 1 \le k \le S_n$; again, without loss of generality, we assume

$$\beta_{n,kj}^0 \ne 0, \quad \text{for } k = 1, \ldots, S_n; \; j = 1, \ldots, s_{nk},$$
$$\beta_{n,kj}^0 = 0, \quad \text{for } k = 1, \ldots, S_n; \; j = s_{nk} + 1, \ldots, p_{nk}.$$

For simplicity, we write $\beta_{n,kj}$, p_{nk} and s_{nk} as β_{kj}, p_k and s_k in the following.

Since we have a diverging number of parameters, to keep the uniform properties of the likelihood function, we need some conditions on the higher-order moment of the likelihood function, as compared to the usual condition in the asymptotic theory of the likelihood estimate under finite parameters (Lehmann and Casella 1998).

(A1) For every n, the observations $\{\boldsymbol{V}_{ni}, i = 1, 2, \ldots, n\}$ are independent and identically distributed, each with a density $f_n(\boldsymbol{V}_{n1}, \boldsymbol{\beta}_n)$. $f_n(\boldsymbol{V}_{n1}, \boldsymbol{\beta}_n)$ has a common support and the model is identifiable. Furthermore, the first and second logarithmic derivatives of f_n satisfy the equations

$$\mathrm{E}_{\boldsymbol{\beta}_n}\left[\frac{\partial \log f_n(\boldsymbol{V}_{n1}, \boldsymbol{\beta}_n)}{\partial \beta_{kj}}\right] = 0,$$
$$\text{for } k = 1, \ldots, K_n; \; j = 1, \ldots, p_k$$
$$\boldsymbol{I}_{k_1 j_1 k_2 j_2}(\boldsymbol{\beta}_n) = \mathrm{E}_{\boldsymbol{\beta}_n}\left[\frac{\partial}{\partial \beta_{k_1 j_1}}\log f_n(\boldsymbol{V}_{n1}, \boldsymbol{\beta}_n)\right.$$
$$\left.\cdot \frac{\partial}{\partial \beta_{k_2 j_2}}\log f_n(\boldsymbol{V}_{n1}, \boldsymbol{\beta}_n)\right]$$
$$= \mathrm{E}_{\boldsymbol{\beta}_n}\left[-\frac{\partial^2}{\partial \beta_{k_1 j_2}\partial \beta_{k_2 j_2}}\log f_n(\boldsymbol{V}_{n1}, \boldsymbol{\beta}_n)\right].$$

(A2) The Fisher information matrix

$$\boldsymbol{I}(\boldsymbol{\beta}_n) =$$
$$\mathrm{E}_{\boldsymbol{\beta}_n}\left[\frac{\partial}{\partial\boldsymbol{\beta}_n}\log f_n(\boldsymbol{V}_{n1}, \boldsymbol{\beta}_n)\frac{\partial^\top}{\partial\boldsymbol{\beta}_n}\log f_n(\boldsymbol{V}_{n1}, \boldsymbol{\beta}_n)\right]$$

satisfies the condition

$$0 < C_1 < \lambda_{\min}\{\boldsymbol{I}(\boldsymbol{\beta}_n)\} \le \lambda_{\max}\{\boldsymbol{I}(\boldsymbol{\beta}_n)\} < C_2 < \infty,$$

and for any k_1, j_1, k_2, j_2, we have

$$\mathrm{E}_{\boldsymbol{\beta}_n}\left[\frac{\partial}{\partial \beta_{k_1 j_1}}\log f_n(\boldsymbol{V}_{n1}, \boldsymbol{\beta}_n)\right.$$
$$\left.\cdot \frac{\partial}{\partial \beta_{k_2 j_2}}\log f_n(\boldsymbol{V}_{n1}, \boldsymbol{\beta}_n)\right]^2 < C_3 < \infty,$$
$$\mathrm{E}_{\boldsymbol{\beta}_n}\left[-\frac{\partial^2}{\partial \beta_{k_1 j_1}\partial \beta_{k_2 j_2}}\log f_n(\boldsymbol{V}_{n1}, \boldsymbol{\beta}_n)\right]^2 < C_4 < \infty.$$

(A3) There exists an open subset ω_n of $\Omega_n \in R^{P_n}$ that contains the true parameter point $\boldsymbol{\beta}_n^0$ such that for almost all $\boldsymbol{V}_{n1}$, the density $f_n(\boldsymbol{V}_{n1}, \boldsymbol{\beta}_n)$ admits all third derivatives $\partial^3 f_n(\boldsymbol{V}_{n1}, \boldsymbol{\beta}_n)/(\partial \beta_{k_1 j_1}\partial \beta_{k_2 j_2}\partial \beta_{k_3 j_3})$ for all $\boldsymbol{\beta}_n \in \omega_n$. Furthermore, there exist functions $M_{nk_1 j_1 k_2 j_2 k_3 j_3}$ such that

$$\left|\frac{\partial^3}{\partial \beta_{k_1 j_1}\partial \beta_{k_2 j_2}\partial \beta_{k_3 j_3}}\log f_n(\boldsymbol{V}_{n1}, \boldsymbol{\beta}_n)\right|$$
$$\le M_{nk_1 j_1 k_2 j_2 k_3 j_3}(\boldsymbol{V}_{n1}) \text{ for all } \boldsymbol{\beta}_n \in \omega_n,$$

and $\mathrm{E}_{\boldsymbol{\beta}_n}[M_{nk_1 j_1 k_2 j_2 k_3 j_3}^2(\boldsymbol{V}_{n1})] < C_5 < \infty.$

These regularity conditions guarantee the asymptotic normality of the ordinary maximum likelihood estimates for diverging number of parameters.

For expositional simplicity, we will first prove Theorem 3 and Theorem 4, then prove Theorem 2.

Proof of Theorem 3

We will show that for any given $\epsilon > 0$, there exists a constant C such that

$$(18) \qquad \Pr\left\{\sup_{\|\boldsymbol{u}\|=C} Q_n(\boldsymbol{\beta}_n^0 + \alpha_n\boldsymbol{u}) < Q_n(\boldsymbol{\beta}_n^0)\right\} \geq 1 - \epsilon,$$

where $\alpha_n = \sqrt{P_n}(n^{-1/2} + \lambda_n\sqrt{a_n}/2\sqrt{c_1})$. This implies that with probability at least $1-\epsilon$, that there exists a local maximum in the ball $\{\boldsymbol{\beta}_n^0 + \alpha_n\boldsymbol{u} : \|\boldsymbol{u}\| \leq C\}$. Hence, there exists a local maximizer such that $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^0\| = O_p(\alpha_n)$. Since $1/2\sqrt{c_1}$ is a constant, we have $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^0\| = O_p(\sqrt{P_n}(n^{-1/2} + \lambda_n\sqrt{a_n}))$.

Using $p_{\lambda_n,\boldsymbol{w}_n}(0) = 0$, we have

(19)

$$\begin{aligned}
D_n(\boldsymbol{u}) &= Q_n(\boldsymbol{\beta}_n^0 + \alpha_n\boldsymbol{u}) - Q_n(\boldsymbol{\beta}_n^0) \\
&\leq L_n(\boldsymbol{\beta}_n^0 + \alpha_n\boldsymbol{u}) - L_n(\boldsymbol{\beta}_n^0) \\
&\quad - n\sum_{k=1}^{S_n}\{p_{\lambda_n,\boldsymbol{w}_n}(\boldsymbol{\beta}_{nk}^0 + \alpha_n\boldsymbol{u}_k) - p_{\lambda_n,\boldsymbol{w}_n}(\boldsymbol{\beta}_{nk}^0)\} \\
&\triangleq (I) + (II).
\end{aligned}$$

Using the standard argument on the Taylor expansion of the likelihood function, we have

$$\begin{aligned}
(20) \qquad (I) &= \alpha_n\boldsymbol{u}^\mathsf{T}\nabla L_n(\boldsymbol{\beta}_n^0) + \frac{1}{2}\boldsymbol{u}^\mathsf{T}\nabla^2 L_n(\boldsymbol{\beta}_n^0)\boldsymbol{u}\alpha_n^2 \\
&\quad + \frac{1}{6}\boldsymbol{u}^\mathsf{T}\nabla\{\boldsymbol{u}^\mathsf{T}\nabla^2 L_n(\boldsymbol{\beta}_n^*)\boldsymbol{u}\}\alpha_n^3 \\
&\triangleq I_1 + I_2 + I_3,
\end{aligned}$$

where $\boldsymbol{\beta}_n^*$ lies between $\boldsymbol{\beta}_n^0$ and $\boldsymbol{\beta}_n^0 + \alpha_n\boldsymbol{u}$. Using the same argument as in the proof of Theorem 1 in [7], we have

$$(21) \qquad |I_1| = O_p(\alpha_n^2 n)\|\boldsymbol{u}\|,$$

$$(22) \qquad I_2 = -\frac{n\alpha_n^2}{2}\boldsymbol{u}^\mathsf{T}\boldsymbol{I}_n(\boldsymbol{\beta}_n^0)\boldsymbol{u} + o_p(1)n\alpha_n^2\|\boldsymbol{u}\|^2,$$

and

$$\begin{aligned}
|I_3| &= \left|\frac{1}{6}\sum_{k_1=1}^{K_n}\sum_{j_1=1}^{p_k}\sum_{k_2=1}^{K_n}\sum_{j_2=1}^{p_k}\sum_{k_3=1}^{K_n}\sum_{j_3=1}^{p_k}\right. \\
&\qquad \left.\frac{\partial^3 L_n(\boldsymbol{\beta}_n^*)}{\partial\beta_{k_1 j_1}\partial\beta_{k_2 j_2}\partial\beta_{k_3 j_3}}u_{k_1 j_1}u_{k_2 j_2}u_{k_3 j_3}\alpha_n^3\right| \\
&\leq \frac{1}{6}\sum_{i=1}^{n}\left\{\sum_{k_1=1}^{K_n}\sum_{j_1=1}^{p_k}\sum_{k_2=1}^{K_n}\sum_{j_2=1}^{p_k}\sum_{k_3=1}^{K_n}\sum_{j_3=1}^{p_k}\right. \\
&\qquad \left. M_{nk_1 j_1 k_2 j_2 k_3 j_3}^2(V_{ni})\right\}^{1/2}\|\boldsymbol{u}\|^3\alpha_n^3 \\
&= O_p(P_n^{3/2}\alpha_n)n\alpha_n^2\|\boldsymbol{u}\|^3.
\end{aligned}$$

Since $\frac{P_n^4}{n} \to 0$ and $P_n^2\lambda_n\sqrt{a_n} \to 0$ as $n \to \infty$, we have

$$(23) \qquad |I_3| = o_p(n\alpha_n^2)\|\boldsymbol{u}\|^3.$$

From (21)–(23), we can see that, by choosing a sufficiently large C, the first term in I_2 dominates I_1 uniformly on $\|\boldsymbol{u}\| = C$; when n is large enough, I_2 also dominates I_3 uniformly on $\|\boldsymbol{u}\| = C$.

Now we consider (II). Since $\alpha_n = \sqrt{P_n}(n^{-1/2} + \lambda_n\sqrt{a_n}/2\sqrt{c_1}) \to 0$, for $\|\boldsymbol{u}\| \leq C$ we have

$$(24) \qquad |\beta_{kj}^0 + \alpha_n u_{kj}| \geq |\beta_{kj}^0| - |\alpha_n u_{kj}| > 0$$

for n large enough and $\beta_{kj}^0 \neq 0$. Hence, we have

$$\begin{aligned}
&p_{\lambda_n,\boldsymbol{w}_n}(\boldsymbol{\beta}_{nk}^0 + \alpha_n\boldsymbol{u}_k) - p_{\lambda_n,\boldsymbol{w}_n}(\boldsymbol{\beta}_{nk}^0) \\
&= \left(\sqrt{w_{n,k1}|\beta_{k1}^0 + \alpha_n u_{k1}| + \cdots + w_{n,kp_k}|\beta_{kp_k}^0 + \alpha_n u_{kp_k}|}\right. \\
&\qquad \left. - \sqrt{w_{n,k1}|\beta_{k1}^0| + \cdots + w_{n,kp_k}|\beta_{kp_k}^0|}\right)\lambda_n \\
&\geq \left(\sqrt{w_{n,k1}|\beta_{k1}^0 + \alpha_n u_{k1}| + \cdots + w_{n,ks_k}|\beta_{ks_k}^0 + \alpha_n u_{ks_k}|}\right. \\
&\qquad \left. - \sqrt{w_{n,k1}|\beta_{k1}^0| + \cdots + w_{n,ks_k}|\beta_{ks_k}^0|}\right)\lambda_n \\
&\geq \left(\{w_{n,k1}|\beta_{k1}^0| + \cdots + w_{n,ks_k}|\beta_{ks_k}^0|\right. \\
&\qquad - \alpha_n(w_{n,k1}|u_{k1}| + \cdots + w_{n,ks_k}|u_{ks_k}|)\}^{1/2} \\
&\qquad \left. - \sqrt{w_{n,k1}|\beta_{k1}^0| + \cdots + w_{n,ks_k}|\beta_{ks_k}^0|}\right)\lambda_n \\
&\qquad (\text{for } n \text{ large enough, by (24)}) \\
&= \lambda_n\sqrt{w_{n,k1}|\beta_{k1}^0| + \cdots + w_{n,ks_k}|\beta_{ks_k}^0|}(\sqrt{1 - \gamma_{nk}} - 1),
\end{aligned}$$

where γ_{nk} is defined as $\gamma_{nk} = \frac{\alpha_n(w_{n,k1}|u_{k1}| + \cdots + w_{n,ks_k}|u_{ks_k}|)}{w_{n,k1}|\beta_{k1}^0| + \cdots + w_{n,ks_k}|\beta_{ks_k}^0|}$. For n large enough, we have $0 \leq \gamma_{nk} < 1$ and $\gamma_{nk} \leq \frac{\alpha_n\|\boldsymbol{u}_k\|(w_{n,k1} + \cdots + w_{n,ks_k})}{c_1(w_{n,k1} + \cdots + w_{n,ks_k})} = \frac{\alpha_n\|\boldsymbol{u}_k\|}{c_1} \leq \frac{\alpha_n C}{c_1} \to 0$ with probability tending to 1 as $n \to \infty$.

Therefore,

$$\begin{aligned}
&p_{\lambda_n,\boldsymbol{w}_n}(\boldsymbol{\beta}_{nk}^0 + \alpha_n\boldsymbol{u}_k) - p_{\lambda_n,\boldsymbol{w}_n}(\boldsymbol{\beta}_{nk}^0) \\
&\geq \lambda_n\sqrt{w_{n,k1}|\beta_{k1}^0| + \cdots + w_{n,ks_k}|\beta_{ks_k}^0|} \\
&\qquad \cdot (\sqrt{1 - \gamma_{nk}} - 1) \\
&\geq \lambda_n\sqrt{w_{n,k1}|\beta_{k1}^0| + \cdots + w_{n,ks_k}|\beta_{ks_k}^0|} \\
&\qquad \cdot \left(\frac{1 + |o_p(1)|}{2}(-\gamma_{nk})\right) \\
&\qquad (\text{Using } \gamma_{nk} = o_p(1) \text{ and Taylor expansion}) \\
&\geq -\lambda_n\frac{\alpha_n(w_{n,k1}|u_{k1}| + \cdots + w_{n,ks_k}|u_{ks_k}|)}{\sqrt{w_{n,k1}|\beta_{k1}^0| + \cdots + w_{n,ks_k}|\beta_{ks_k}^0|}} \\
&\qquad \cdot \left(\frac{1 + |o_p(1)|}{2}\right) \\
&\geq -\alpha_n\lambda_n\frac{\|\boldsymbol{u}_k\|\sqrt{a_n s_k}}{2\sqrt{c_1}}(1 + |o_p(1)|).
\end{aligned}$$

Therefore, the term (II) in (19) is bounded by

$$n\alpha_n\lambda_n\left(\sum_{k=1}^{S_n}\frac{\|\boldsymbol{u}_k\|\sqrt{a_n s_k}}{2\sqrt{c_1}}\right)(1+|o_p(1)|),$$

which is further bounded by

$$n\alpha_n\lambda_n\sqrt{a_n}\left(\|\boldsymbol{u}\|\cdot\frac{\sqrt{P_n}}{2\sqrt{c_1}}\right)(1+|o_p(1)|).$$

Note that $\alpha_n = \sqrt{P_n}(n^{-1/2}+\lambda_n\sqrt{a_n}/2\sqrt{c_1})$, hence the above expression is bounded by

$$\|\boldsymbol{u}\|n\alpha_n^2(1+|o_p(1)|).$$

This term is also dominated by the first term of I_2 on $\|\boldsymbol{u}\| = C$ uniformly. Therefore, $D_n(\boldsymbol{u}) < 0$ is satisfied uniformly on $\|\boldsymbol{u}\| = C$. This completes the proof of the theorem.

Proof of Theorem 4

We have proved that if $\lambda_n\sqrt{a_n} = O_p(n^{-1/2})$, there exists a root-$(n/P_n)$ consistent estimate $\hat{\boldsymbol{\beta}}_n$. Now we prove that this root-(n/P_n) consistent estimate has the oracle sparsity under the condition $\frac{P_n^2}{\lambda_n^2 b_n} = o_p(n)$, i.e., $\hat{\beta}_{kj} = 0$ with probability tending to 1 if $\beta_{kj}^0 = 0$.

Using Taylor's expansion, we have

(25)
$$\frac{\partial Q_n(\boldsymbol{\beta}_n)}{\partial\beta_{kj}} = \frac{\partial L_n(\boldsymbol{\beta}_n)}{\partial\beta_{kj}} - n\frac{\partial p_{\lambda_n,\boldsymbol{w}_n}(\boldsymbol{\beta}_{nk})}{\partial\beta_{kj}}$$

$$= \frac{\partial L_n(\boldsymbol{\beta}_n^0)}{\partial\beta_{kj}} + \sum_{k_1=1}^{K_n}\sum_{j_1=1}^{p_{k_1}}\frac{\partial^2 L_n(\boldsymbol{\beta}^0)}{\partial\beta_{kj}\partial\beta_{k_2 j_2}}(\beta_{k_1 j_1} - \beta_{k_1 j_1}^0)$$

$$+ \frac{1}{2}\sum_{k_1=1}^{K_n}\sum_{j_1=1}^{p_{k_1}}\sum_{k_2=1}^{K_n}\sum_{j_2=1}^{p_{k_2}}$$

$$\frac{\partial^3 L_n(\boldsymbol{\beta}_n^*)}{\partial\beta_{kj}\partial\beta_{k_1 j_1}\partial\beta_{k_2 j_2}}(\beta_{k_1 j_1} - \beta_{k_1 j_1}^0)(\beta_{k_2 j_2} - \beta_{k_2 j_2}^0)$$

$$- \frac{n\lambda_n w_{n,kj}}{2\sqrt{w_{n,k1}|\beta_{k1}| + \cdots + w_{n,kp_k}|\beta_{kp_k}|}}\operatorname{sgn}(\beta_{kj})$$

$$\triangleq I_1 + I_2 + I_3 + I_4,$$

where $\boldsymbol{\beta}_n^*$ lies between $\boldsymbol{\beta}_n$ and $\boldsymbol{\beta}_n^0$.

Using the argument in the proof of Lemma 5 in [7], for any $\boldsymbol{\beta}_n$ satisfying $\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^0\| = O_p(\sqrt{P_n/n})$, we have

$$I_1 = O_p(\sqrt{n}) = O_p(\sqrt{nP_n}),$$
$$I_2 = O_p(\sqrt{nP_n}),$$
$$I_3 = o_p(\sqrt{nP_n}).$$

Then, since $\hat{\boldsymbol{\beta}}_n$ is a root-(n/P_n) consistent estimate maximizing $Q_n(\boldsymbol{\beta}_n)$, if $\hat{\beta}_{kj} \neq 0$, we have

(26)
$$\frac{\partial Q_n(\boldsymbol{\beta}_n)}{\partial\beta_{kj}}\bigg|_{\boldsymbol{\beta}_n=\hat{\boldsymbol{\beta}}_n}$$
$$= O_p(\sqrt{nP_n})$$
$$- \frac{n\lambda_n w_{n,kj}}{2\sqrt{w_{n,k1}|\hat{\beta}_{k1}| + \cdots + w_{n,kp_k}|\hat{\beta}_{kp_k}|}}\operatorname{sgn}(\hat{\beta}_{kj}) = 0.$$

Therefore,

$$\frac{n\lambda_n w_{n,kj}}{\sqrt{w_{n,k1}|\hat{\beta}_{k1}| + \cdots + w_{n,kp_k}|\hat{\beta}_{kp_k}|}} = O_p(\sqrt{nP_n}) \text{ for } \hat{\beta}_{kj} \neq 0.$$

This can be extended to

$$\frac{n\lambda_n w_{n,kj}|\hat{\beta}_{kj}|}{\sqrt{w_{n,k1}|\hat{\beta}_{k1}| + \cdots + w_{n,kp_k}|\hat{\beta}_{kp_k}|}} = |\hat{\beta}_{kj}|O_p(\sqrt{nP_n}),$$

for any $\hat{\beta}_{kj}$ with $\hat{\boldsymbol{\beta}}_{nk} \neq 0$. If we sum this over all j in the kth group, we have

(27)
$$n\lambda_n\sqrt{w_{n,k1}|\hat{\beta}_{k1}| + \cdots + w_{n,kp_k}|\hat{\beta}_{kp_k}|}$$
$$= \sum_{j=1}^{p_k}|\hat{\beta}_{kj}|O_p(\sqrt{nP_n}).$$

Since $\hat{\boldsymbol{\beta}}_n$ is a root-(n/P_n) consistent estimate of $\boldsymbol{\beta}_n^0$, we have $|\hat{\beta}_{kj}| = O_p(1)$ for $(k,j) \in \mathcal{A}_n$ and $|\hat{\beta}_{kj}| = O_p(\sqrt{P_n/n})$ for $(k,j) \in \mathcal{B}_n \cup \mathcal{C}_n$.

Now for any k and j satisfying $\beta_{kj}^0 = 0$ and $\hat{\beta}_{kj} \neq 0$, equation (26) can be written as:

(28)
$$\frac{\partial Q_n(\boldsymbol{\beta}_n)}{\partial\beta_{kj}}\bigg|_{\boldsymbol{\beta}_n=\hat{\boldsymbol{\beta}}_n}$$
$$= \frac{1}{2\lambda_n\sqrt{w_{n,k1}|\hat{\beta}_{k1}| + \cdots + w_{n,kp_k}|\hat{\beta}_{kp_k}|}}$$
$$(O_p(\sqrt{P_n/n})n\lambda_n\sqrt{w_{n,k1}|\hat{\beta}_{k1}| + \cdots + w_{n,kp_k}|\hat{\beta}_{kp_k}|}$$
$$- n\lambda_n^2 w_{n,kj}\operatorname{sgn}(\hat{\beta}_{kj})) = 0.$$

Denote

$$h_{nk} = O_p(\sqrt{P_n/n})n\lambda_n \cdot \sqrt{w_{n,k1}|\hat{\beta}_{k1}| + \cdots + w_{n,kp_k}|\hat{\beta}_{kp_k}|}.$$

Let $h_n = \sum_{k=1}^{K_n}h_{nk}$. By equation (27), we have $h_n = \sum_{k=1}^{K_n}O_p(\sqrt{P_n/n})\sum_{j=1}^{p_k}|\hat{\beta}_{kj}|O_p(\sqrt{nP_n}) = O_p(P_n^2)$. Since $\frac{P_n^2}{\lambda_n^2 b_n} = o_p(n)$ guarantees that $n\lambda_n^2 b_n$ dominates h_n with probability tending to 1 as $n \to \infty$, the first term in (28) is dominated by the second term as $n \to \infty$ uniformly for all k and j satisfying $\beta_{kj}^0 = 0$ since $w_{n,kj} \geq b_n$ and $h_n > h_{nk}$. Denote $g_{nk} = 2\lambda_n\sqrt{w_{n,k1}|\hat{\beta}_{k1}| + \cdots + w_{n,kp_k}|\hat{\beta}_{kp_k}|}/(n\lambda_n^2 b_n)$. Let $g_n = \sum_{k=1}^{K_n}g_{nk}$. By equation (27), we have

$g_n = 2\sum_{k=1}^{K_n}(1/n)\sum_{j=1}^{p_k}|\hat{\beta}_{kj}|O_p(\sqrt{nP_n})/(n\lambda_n^2 b_n) = o_p(1/\sqrt{nP_n})$. The absolute value of the second term in (28) is bounded below by $1/g_n$. So with probability uniformly converging to 1 the second term in the derivative $\frac{\partial Q(\beta)}{\partial\beta_{kj}}|_{\beta=\hat{\beta}_n}$ will go to ∞ as $n \to \infty$, which is a contradiction with equation (28). Therefore, for any k and j satisfying $\beta_{kj}^0 = 0$, we have $\hat{\beta}_{kj} = 0$ with a probability tending to 1 as $n \to \infty$. We have $\hat{\beta}_{\mathcal{D}_n} = 0$ with probability tending to 1 as well.

Now we prove the second part of Theorem 4. From the above proof, we know that there exists $(\hat{\beta}_{n,\mathcal{A}_n}, \mathbf{0})$ with probability tending to 1, which is a root-(n/P_n) consistent local maximizer of $Q(\beta_n)$. With a slight abuse of notation, let $Q_n(\beta_{n,\mathcal{A}_n}) = Q_n(\beta_{n,\mathcal{A}_n}, \mathbf{0})$. Using the Taylor expansion on $\nabla Q_n(\hat{\beta}_{n,\mathcal{A}_n})$ at point $\beta_{n,\mathcal{A}_n}^0$, we have

$$(29) \quad \frac{1}{n}(\nabla^2 L_n(\beta_{n,\mathcal{A}_n}^0)(\hat{\beta}_{n,\mathcal{A}_n} - \beta_{n,\mathcal{A}_n}^0) - \nabla J_n(\hat{\beta}_{n,\mathcal{A}_n}))$$
$$= -\frac{1}{n}\bigg(\nabla L_n(\beta_{n,\mathcal{A}_n}^0)$$
$$+ \frac{1}{2}(\hat{\beta}_{n,\mathcal{A}_n} - \beta_{n,\mathcal{A}_n}^0)^\top \nabla^2\{\nabla L_n(\beta_{n,\mathcal{A}_n}^*)\}$$
$$(\hat{\beta}_{n,\mathcal{A}_n} - \beta_{n,\mathcal{A}_n}^0)\bigg),$$

where $\beta_{n,\mathcal{A}_n}^*$ lies between $\hat{\beta}_{n,\mathcal{A}_n}$ and $\beta_{n,\mathcal{A}_n}^0$.

Now we define

$$\mathcal{C}_n \triangleq \frac{1}{2}(\hat{\beta}_{n,\mathcal{A}_n} - \beta_{n,\mathcal{A}_n}^0)^\top \nabla^2\{\nabla L_n(\beta_{n,\mathcal{A}_n}^*)\}(\hat{\beta}_{n,\mathcal{A}_n} - \beta_{n,\mathcal{A}_n}^0).$$

Using the Cauchy-Schwarz inequality, we have

$$(30) \quad \left\|\frac{1}{n}\mathcal{C}_n\right\|^2$$
$$\le \frac{1}{n^2}\sum_{i=1}^n n\|\hat{\beta}_{n,\mathcal{A}_n} - \beta_{n,\mathcal{A}_n}^0\|^4$$
$$\sum_{k_1=1}^{S_n}\sum_{j_1=1}^{p_k}\sum_{k_2=1}^{S_n}\sum_{j_2=1}^{p_k}\sum_{k_3=1}^{S_n}\sum_{j_3=1}^{p_k} M_{nk_1 j_1 k_2 j_2 k_3 j_2}^3(\mathbf{V}_{ni})$$
$$= O_p(P_n^2/n^2)O_p(P_n^3) = O_p(P_n^5/n^2) = o_p(1/n).$$

Since $\frac{P_n^5}{n} \to 0$ as $n \to \infty$, by Lemma 8 in [7], we have

$$\left\|\frac{1}{n}\nabla^2 L_n(\beta_{n,\mathcal{A}_n}^0) + \mathbf{I}_n(\beta_{n,\mathcal{A}_n}^0)\right\| = o_p(1/P_n)$$

and

$$(31) \quad \left\|\left(\frac{1}{n}\nabla^2 L_n(\beta_{n,\mathcal{A}_n}^0) + \mathbf{I}_n(\beta_{n,\mathcal{A}_n}^0)\right)(\hat{\beta}_{n,\mathcal{A}_n} - \beta_{n,\mathcal{A}_n}^0)\right\|$$
$$= o_p(1/\sqrt{nP_n}) = o_p(1/\sqrt{n}).$$

Since

$$\sqrt{w_{n,k1}|\hat{\beta}_{k1}| + \cdots + w_{n,ks_k}|\hat{\beta}_{ks_k}|}$$
$$= \{w_{n,k1}|\beta_{k1}^0|(1 + O_p(\sqrt{P_n/n})) + \cdots$$
$$+ w_{n,ks_k}|\beta_{ks_k}^0|(1 + O_p(\sqrt{P_n/n}))\}^{1/2}$$
$$= \sqrt{w_{n,k1}|\beta_{k1}^0| + \cdots + w_{n,ks_k}|\beta_{ks_k}^0|}$$
$$\cdot (1 + O_p(\sqrt{P_n/n})),$$

we have

$$\frac{\lambda_n w_{n,kj}}{\sqrt{w_{n,k1}|\hat{\beta}_{k1}| + \cdots + w_{n,ks_k}|\hat{\beta}_{ks_k}|}}$$
$$= \frac{\lambda_n w_{n,kj}}{\sqrt{w_{n,k1}|\beta_{k1}^0| + \cdots + w_{n,ks_k}|\beta_{ks_k}^0|}}(1 + O_p(\sqrt{P_n/n})).$$

Furthermore, since

$$\frac{\lambda_n w_{n,kj}}{\sqrt{w_{n,k1}|\beta_{k1}^0| + \cdots + w_{n,ks_k}|\beta_{ks_k}^0|}}$$
$$\le \frac{\lambda_n w_{n,kj}}{\sqrt{w_{n,kj}c_1}} \le \frac{\lambda_n\sqrt{a_n}}{\sqrt{c_1}} = o_p((nP_n)^{-1/2})$$

for $(k,j) \in \mathcal{A}_n$, we have

$$\left(\frac{1}{n}\nabla J_n(\hat{\beta}_{n,\mathcal{A}_n})\right)_{kj}$$
$$= \frac{\lambda_n w_{n,kj}}{2\sqrt{w_{n,k1}|\hat{\beta}_{k1}| + \cdots + w_{n,ks_k}|\hat{\beta}_{ks_k}|}}$$
$$= o_p((nP_n)^{-1/2})$$

and

$$(32) \quad \left\|\frac{1}{n}\nabla J_n(\hat{\beta}_{n,\mathcal{A}_n})\right\| \le \sqrt{P_n}o_p((nP_n)^{-1/2}) = o_p(1/\sqrt{n}).$$

Together with (30), (31) and (32), from (29) we have

$$\mathbf{I}_n(\beta_{n,\mathcal{A}_n}^0)(\hat{\beta}_{n,\mathcal{A}_n} - \beta_{n,\mathcal{A}_n}^0)$$
$$= \frac{1}{n}\nabla L_n(\beta_{n,\mathcal{A}_n}^0) + o_p(1/\sqrt{n}).$$

Now using the same argument as in the proof of Theorem 2 in [7], we have

$$\sqrt{n}\mathbf{A}_n\mathbf{I}_n^{1/2}(\beta_{n,\mathcal{A}_n}^0)(\hat{\beta}_{n,\mathcal{A}_n} - \beta_{n,\mathcal{A}_n}^0)$$
$$\to \sqrt{n}\mathbf{A}_n\mathbf{I}_n^{-1/2}(\beta_{n,\mathcal{A}_n}^0)\left(\frac{1}{n}\nabla L_n(\beta_{n,\mathcal{A}_n}^0)\right)$$
$$\to \mathcal{N}(\mathbf{0}, \mathbf{G}),$$

where $\mathbf{A}_n$ is a $q \times |\mathcal{A}_n|$ matrix such that $\mathbf{A}_n\mathbf{A}_n^\top \to \mathbf{G}$ and $\mathbf{G}$ is a $q \times q$ nonnegative symmetric matrix.

Proof of Theorem 2

Note that when $w_{n,kj} = 1$, we have $a_n = 1$ and $b_n = 1$. The conditions $\lambda_n \sqrt{a_n} = O_p(n^{-1/2})$ and $\frac{P_n^2}{\lambda_n^2 b_n} = o_p(n)$ in Theorem 4 become $\lambda_n \sqrt{n} = O_p(1)$ and $\frac{P_n}{\lambda_n \sqrt{n}} \to 0$. These two conditions cannot be satisfied simultaneously by adjusting λ_n, which implies that $\Pr(\hat{\boldsymbol{\beta}}_{\mathcal{D}} = 0) \to 1$ cannot be guaranteed.

We will prove that by choosing λ_n satisfying $\sqrt{n}\lambda_n = O_p(1)$ and $P_n n^{-3/4}/\lambda_n \to 0$ as $n \to \infty$, we can have a root-n consistent local maximizer $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}_{\mathcal{A}_n}, \hat{\boldsymbol{\beta}}_{\mathcal{B}_n}, \hat{\boldsymbol{\beta}}_{\mathcal{C}_n})^\top$ such that $\Pr(\hat{\boldsymbol{\beta}}_{\mathcal{C}_n} = 0) \to 1$.

Similar as in the proof of Theorem 4, we let $h'_n = \sum_{k=S_n+1}^{K_n} h_{nk}$. By equation (27), we have $h'_n = \sum_{k=S_n+1}^{K_n} O_p(\sqrt{P_n/n}) \sum_{j=1}^{p_k} |\hat{\beta}_{kj}| O_p(\sqrt{nP_n}) = O_p(P_n^2/\sqrt{n})$. Since $P_n n^{-3/4}/\lambda_n \to 0$ guarantees that $n\lambda_n^2$ dominates h'_n with probability tending to 1 as $n \to \infty$, the first term in (28) is dominated by the second term as $n \to \infty$ uniformly for any k satisfying $\boldsymbol{\beta}_{nk}^0 = 0$ since $w_{n,kj} = 1$ and $h'_n > h_{nk}$. Similar as in the proof of Theorem 4, we have $\hat{\boldsymbol{\beta}}_{\mathcal{C}_n} = 0$ with probability tending to 1.

Proof of Theorem 5

Let $N_n = |\mathcal{A}_n|$ be the number of nonzero parameters. Let $\boldsymbol{B}_n$ be an $(N_n - q) \times N_n$ matrix which satisfies $\boldsymbol{B}_n \boldsymbol{B}_n^\top = \boldsymbol{I}_{N_n - q}$ and $\boldsymbol{A}_n \boldsymbol{B}_n^\top = 0$. As $\boldsymbol{\beta}_{n,\mathcal{A}_n}$ is in the orthogonal complement to the linear space that is spanned by the rows of $\boldsymbol{A}_n$ under the null hypothesis H_0, it follows that

$$\boldsymbol{\beta}_{n,\mathcal{A}_n} = \boldsymbol{B}_n^\top \boldsymbol{\gamma}_n,$$

where $\boldsymbol{\gamma}_n$ is an $(N_n - q) \times 1$ vector. Then, under H_0 the penalized likelihood estimator is also the local maximizer $\hat{\boldsymbol{\gamma}}_n$ of the problem

$$Q_n(\boldsymbol{\beta}_{n,\mathcal{A}_n}) = \max_{\boldsymbol{\gamma}_n} Q_n(\boldsymbol{B}_n^\top \boldsymbol{\gamma}_n).$$

To prove Theorem 5 we need the following two lemmas.

Lemma 3. *Under condition* (b) *of Theorem 4 and the null hypothesis H_0, we have*

$$
\begin{aligned}
&\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n} - \boldsymbol{\beta}_{n,\mathcal{A}_n}^0 \\
&\quad = \frac{1}{n} \boldsymbol{I}_n^{-1}(\boldsymbol{\beta}_{n,\mathcal{A}_n}^0) \nabla L_n(\boldsymbol{\beta}_{n,\mathcal{A}_n}^0) + o_p(n^{-1/2}), \\
&\boldsymbol{B}_n^\top(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_n^0) \\
&\quad = \frac{1}{n} \boldsymbol{B}_n^\top \{\boldsymbol{B}_n \boldsymbol{I}_n(\boldsymbol{\beta}_{n,\mathcal{A}_n}^0) \boldsymbol{B}_n^\top\}^{-1} \boldsymbol{B}_n \nabla L_n(\boldsymbol{\beta}_{n,\mathcal{A}_n}^0) \\
&\qquad + o_p(n^{-1/2}).
\end{aligned}
$$

Proof of Lemma 3

We need only prove the second equation. The first equation can be shown in the same manner. Following the proof of Theorem 4, it follows that under H_0,

$$
\begin{aligned}
&\boldsymbol{B}_n \boldsymbol{I}_n(\boldsymbol{\beta}_{n,\mathcal{A}_n}^0) \boldsymbol{B}_n^\top(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_n^0) \\
&\quad = \frac{1}{n} \boldsymbol{B}_n \nabla L_n(\boldsymbol{\beta}_{n,\mathcal{A}_n}^0) + o_p(n^{-1/2}).
\end{aligned}
$$

As the eigenvalue $\lambda_i(\boldsymbol{B}_n \boldsymbol{I}_n(\boldsymbol{\beta}_{n,\mathcal{A}_n}^0) \boldsymbol{B}_n^\top)$ is uniformly bounded away from 0 and infinity, we have

$$
\begin{aligned}
&\boldsymbol{B}_n^\top(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_n^0) \\
&\quad = \frac{1}{n} \boldsymbol{B}_n^\top \{\boldsymbol{B}_n \boldsymbol{I}_n(\boldsymbol{\beta}_{n,\mathcal{A}_n}^0) \boldsymbol{B}_n^\top\}^{-1} \boldsymbol{B}_n \nabla L_n(\boldsymbol{\beta}_{n,\mathcal{A}_n}^0) \\
&\qquad + o_p(n^{-1/2}).
\end{aligned}
$$

Lemma 4. *Under condition* (b) *of Theorem 4 and the null hypothesis H_0, we have*

$$
\begin{aligned}
(33) \quad &Q_n(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n}) - Q_n(\boldsymbol{B}_n^\top \hat{\boldsymbol{\gamma}}_n) \\
&= \frac{n}{2}(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n} - \boldsymbol{B}_n^\top \hat{\boldsymbol{\gamma}}_n)^\top \boldsymbol{I}_n(\boldsymbol{\beta}_{n,\mathcal{A}_n}^0)(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n} - \boldsymbol{B}_n^\top \hat{\boldsymbol{\gamma}}_n) \\
&\quad + o_p(1).
\end{aligned}
$$

Proof of Lemma 4

A Taylor's expansion of $Q_n(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n}) - Q_n(\boldsymbol{B}_n^\top \hat{\boldsymbol{\gamma}}_n)$ at the point $\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n}$ yields

$$Q_n(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n}) - Q_n(\boldsymbol{B}_n^\top \hat{\boldsymbol{\gamma}}_n) = T_1 + T_2 + T_3 + T_4,$$

where

$$
\begin{aligned}
T_1 &= \nabla^\top Q_n(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n})(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n} - \boldsymbol{B}_n^\top \hat{\boldsymbol{\gamma}}_n), \\
T_2 &= -\frac{1}{2}(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n} - \boldsymbol{B}_n^\top \hat{\boldsymbol{\gamma}}_n)^\top \nabla^2 L_n(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n}) \\
&\quad (\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n} - \boldsymbol{B}_n^\top \hat{\boldsymbol{\gamma}}_n), \\
T_3 &= \frac{1}{6}\nabla^\top \{(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n} - \boldsymbol{B}_n^\top \hat{\boldsymbol{\gamma}}_n)^\top \nabla^2 L_n(\boldsymbol{\beta}_{n,\mathcal{A}_n}^\star) \\
&\quad (\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n} - \boldsymbol{B}_n^\top \hat{\boldsymbol{\gamma}}_n)\}(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n} - \boldsymbol{B}_n^\top \hat{\boldsymbol{\gamma}}_n), \\
T_4 &= \frac{1}{2}(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n} - \boldsymbol{B}_n^\top \hat{\boldsymbol{\gamma}}_n)^\top \nabla^2 J_n(\boldsymbol{\beta}_{n,\mathcal{A}_n}^*) \\
&\quad (\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n} - \boldsymbol{B}_n^\top \hat{\boldsymbol{\gamma}}_n).
\end{aligned}
$$

We have $T_1 = 0$ as $\nabla^\top Q_n(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n}) = 0$.

Let $\boldsymbol{\Theta}_n = \boldsymbol{I}_n(\boldsymbol{\beta}_{n,\mathcal{A}_n}^0)$ and $\boldsymbol{\Phi}_n = \frac{1}{n}\nabla L_n(\boldsymbol{\beta}_{n,\mathcal{A}_n}^0)$. By Lemma 2 we have

$$
\begin{aligned}
&(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n} - \boldsymbol{B}_n^\top \hat{\boldsymbol{\gamma}}_n) \\
&= \boldsymbol{\Theta}_n^{-1/2}\{\boldsymbol{I}_n - \boldsymbol{\Theta}_n^{1/2} \boldsymbol{B}_n^\top (\boldsymbol{B}_n \boldsymbol{\Theta}_n \boldsymbol{B}_n^\top)^{-1} \boldsymbol{B}_n \boldsymbol{\Theta}_n^{1/2}\} \boldsymbol{\Theta}_n^{-1/2} \boldsymbol{\Phi}_n \\
&\quad + o_p(n^{-1/2}).
\end{aligned}
$$

$\boldsymbol{I}_n - \boldsymbol{\Theta}_n^{1/2} \boldsymbol{B}_n^\top (\boldsymbol{B}_n \boldsymbol{\Theta}_n \boldsymbol{B}_n^\top)^{-1} \boldsymbol{B}_n \boldsymbol{\Theta}_n^{1/2}$ is an idempotent matrix with rank q. Hence, by a standard argument and condition (A2),

$$(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n} - \boldsymbol{B}_n^\top \hat{\boldsymbol{\gamma}}_n) = O_p\left(\sqrt{\frac{q}{n}}\right).$$

We have

$$(34) \qquad \left(\frac{1}{n}\nabla^2 J_n(\boldsymbol{\beta}_{n,\mathcal{A}_n})\right)_{kjk_1j_1} = 0, \quad \text{for } k \neq k_1$$

and

$$(35) \qquad \left(\frac{1}{n}\nabla^2 J_n(\boldsymbol{\beta}_{n,\mathcal{A}_n}^*)\right)_{kjkj_1}$$

$$= \frac{\lambda_n w_{n,kj} w_{n,kj_1}}{4(w_{n,k1}|\beta_{k1}^*| + \cdots + w_{n,ks_k}|\beta_{ks_k}^*|)^{3/2}}$$

$$= \frac{\lambda_n w_{n,kj} w_{n,kj_1}}{4(w_{n,k1}|\beta_{k1}^0| + \cdots + w_{n,ks_k}|\beta_{ks_k}^0|)^{3/2}}(1 + o_p(1))$$

$$\leq \frac{\lambda_n \sqrt{a_n}}{4(c_1)^{3/2}}(1 + o_p(1))$$

$$= o_p((nP_n)^{-1/2}).$$

Combining (34), (35) and condition $q < P_n$, following the proof of I_3 in Theorem 3, we have

$$T_3 = O_p(nP_n^{3/2}n^{-3/2}q^{3/2}) = o_p(1)$$

and

$$T_4 \leq n\left\|\frac{1}{n}\nabla^2 J_n(\boldsymbol{\beta}_{n,\mathcal{A}_n}^*)\right\|\|\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n} - \boldsymbol{B}_n^\mathsf{T}\hat{\boldsymbol{\gamma}}_n\|^2$$

$$= nP_n o_p((nP_n)^{-1/2})O_p\left(\frac{q}{n}\right) = o_p(1).$$

Thus,

$$(36) \qquad Q_n(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n}) - Q_n(\boldsymbol{B}_n^\mathsf{T}\hat{\boldsymbol{\gamma}}_n) = T_2 + o_p(1).$$

It follows from Lemmas 8 and 9 in [7] that

$$\left\|\frac{1}{n}\nabla^2 L_n(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n}) + \boldsymbol{I}_n(\boldsymbol{\beta}_{n,\mathcal{A}_n}^0)\right\| = o_p\left(\frac{1}{\sqrt{P_n}}\right).$$

Hence, we have

$$(37) \quad \frac{1}{2}(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n} - \boldsymbol{B}_n^\mathsf{T}\hat{\boldsymbol{\gamma}}_n)^\mathsf{T}\{\nabla^2 L_n(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n}) + n\boldsymbol{I}_n(\boldsymbol{\beta}_{n,\mathcal{A}_n}^0)\}$$

$$(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_n} - \boldsymbol{B}_n^\mathsf{T}\hat{\boldsymbol{\gamma}}_n)$$

$$\leq o_p\left(n\frac{1}{\sqrt{P_n}}\right)O_p\left(\frac{q}{n}\right) = o_p(1).$$

The combination of (36) and (37) yields (33).

Proof of Theorem 5

Given Lemmas 3 and 4, the proof of the Theorem is similar to the proof of Theorem 4 in [7].

REFERENCES

[1] ANTONIADIS, A. and FAN, J. (2001). Regularization of Wavelet Approximations. *Journal of the American Statistical Association*, **96** 939–967. MR1946364

[2] BREIMAN, L. (1995). Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, **37** 373–384. MR1365720

[3] BREIMAN, L. and FRIEDMAN, J. (1997). Predicting Multivariate Responses in Multiple Linear Regression (with Discussion). *Journal of the Royal Statistical Society, Series B*, **59(1)** 3–54. MR1436554

[4] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least Angle Regression. *The Annals of Statistics*, **32** 407–499. MR2060166

[5] EFRON, B. and TIBSHIRANI, R. (2007). On Testing the Significance of Sets of Genes. *The Annals of Applied Statistics*, **1** 107–129. MR2393843

[6] 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. MR1946581

[7] FAN, J. and PENG, H. (2004). Nonconcave Penalized Likelihood with a Diverging Number of Parameters. *The Annals of Statistics*, **32(3)** 928–961. MR2065194

[8] GEORGE, E. and FOSTER, D. (2000). Calibration and Empirical Bayes Variable Selection. *Biometrika*, **87** 731–747. MR1813972

[9] GEORGE, E. and MCCULLOCH, R. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, **88** 881–889.

[10] HUANG, J., MA, S., XIE, H. and ZHANG, C. (2009). A Group Bridge Approach for Variable Selection. *Biometrika*, **96** 339–355.

[11] KUBBUTAT, M., JONES, S. and VOUSDEN, K. (1997). Regulation of p53 Stability by MDM2. *Nature*, **387** 299–303.

[12] LIN, Y. and ZHANG, H. (2006). Component Selection and Smoothing in Smoothing Spline Analysis of Variance Models. *The Annals of Statistics*, **34(5)** 2272–2297. MR2291500

[13] LOHR, K., MORITZ, C., CONTENTE, A. and DOBBELSTEIN, M. (2003). p21/CDKN1A mediates negative regulation of transcription by p53. *Journal of Biological Chemistry*, **278(35)** 32507–32516.

[14] MEIER, L., VAN DER GEER, S. and BUHLMANN, P. (2008). The Group Lasso for Logistic Regression. *Journal of the Royal Statistical Society: Series B*, **70(1)** 53–71. MR2412631

[15] MOLL, U. and PETRENKO, O. (2003). The MDM2-p53 Interaction. *Molecular Cancer Research*, **1** 1001–1008.

[16] OLIVIER, M., EELES, R., HOLLSTEIN, M., KHAN, M. A., HARRIS, C. C. and HAINAUT, P. (2002). The IARC TP53 Database: New Online Mutation Analysis and Recommendations to Users. *Human Mutation*, **19** 607–614.

[17] SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S. and MESIROV, J. P. (2005). Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *Proceedings of the National Academy of Sciences, USA*, **102** 15545–15550.

[18] TIBSHIRANI, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58** 267–288. MR1379242

[19] TURLACH, B., VENABLES, W. and WRIGHT, S. (2005). Simultaneous Variable Selection. *Technometrics*, **47(3)** 349–363. MR2164706

[20] WANG, S., NAN, B., ZHOU, N. and ZHU, J. (2009). Hierarchically Penalized Cox Regression with Grouped Variables. *Biometrika*, **96(2)**, 307–322. MR2507145

[21] WANG, H., LI, G. and TSAI, C. L. (2006). Regression Coefficient and Autoregressive Order Shrinkage and Selection via Lasso. *Journal of the Royal Statistical Society, Series B*, **69** 63–78. MR2301500

[22] Wu, Y., Boos, D. D. and Stefanski, L. A. (2007). Controlling Variable Selection by the Addition of Pseudovariables. *Journal of the American Statistical Association*, **102(477)** 235–243. MR2345541

[23] Yuan, M. and Lin, Y. (2006). Model Selection and Estimation in Regression with Grouped Variable. *Journal of the Royal Statistical Society, Series B*, **68(1)** 49–67. MR2212574

[24] Zhang, H. and Lu, W. (2007). Adaptive-Lasso for Cox's Proportional Hazards Model. *Biometrika*, **94(3)** 691–703. MR2410017

[25] Zhao, P., Rocha, G. and Yu, B. (2009). Grouped and Hierarchical Model Selection through Composite Absolute Penalties. *The Annals of Statistics*, **37(6)** 3468–3497. MR2549566

[26] Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101** 1418–1429. MR2279469

[27] Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, **67(2)** 301–320. MR2137327

Nengfeng Zhou
Consumer Credit Risk Solutions
Bank of America
Charlotte, NC 28255
USA
E-mail address: nengfeng.zhou@bankofamerica.com

Ji Zhu
Department of Statistics
University of Michigan
Ann Arbor, MI 48109
USA
E-mail address: jizhu@umich.edu