Group Variable Selection via a Hierarchical Lasso and Its Oracle Property

Nengfeng Zhou
Consumer Credit Risk Solutions
Bank of America
Charlotte, NC 28255

Ji Zhu*
Department of Statistics
University of Michigan
Ann Arbor, MI 48109
August 10, 2010

*Address for correspondence: Ji Zhu, 439 West Hall, 1085 South University Ave, Ann Arbor, MI 48109-1107. E-mail: jizhu@umich.edu.
Abstract

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 such as Antoniadis and Fan (2001), Yuan and Lin (2006) and Zhao, Rocha and Yu (2009) 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. 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 as in Fan and Li (2001) and Fan and Peng (2004).

Keywords: Group selection; Lasso; Oracle property; Regularization; Variable selection
1 Introduction

Consider the usual regression situation: we have training data, \((x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)\), where \(x_i = (x_{i1}, \ldots, x_{ip})\) 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:

\[
y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon,
\]

where \(\varepsilon\) 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 \(k\)th group contains \(p_k\) variables. Specifically, the linear model (1) is now written as

\[
y_i = \beta_0 + \sum_{k=1}^{K} \sum_{j=1}^{p_k} \beta_{kj} x_{i,kj} + \varepsilon_i.
\]

And we are interested in finding out which variables, especially which “groups,” have an important effect on the response. For example, \((x_{11}, \ldots, x_{1p_1}), (x_{21}, \ldots, x_{2p_2}), \ldots, (x_{K1}, \ldots, x_{Kp_K})\) 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, for example, see George and McCulloch (1993), Breiman (1995), Tibshirani (1996), George and Foster (2000), Fan and Li (2001), Zou and Hastie (2005), Lin and Zhang (2006) and Wu, Boos and Stefanski (2007). In particular, lasso (Tibshirani 1996) has gained much attention in recent years. The lasso criterion penalizes the $L_1$-norm of the regression coefficients to achieve a sparse model:

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

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 sub-optimal. For example, suppose the $k$th group is unimportant, then lasso tends to make individual estimated coefficients in the $k$th 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.

Antoniadis and Fan (2001), Yuan and Lin (2006) and Zhao, Rocha and Yu (2009) have addressed the group variable selection problem in the literature. Antoniadis and Fan (2001) 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.

Yuan and Lin (2006) and Zhao, Rocha and Yu (2009) extended the lasso model (3) for group variable selection. Yuan and Lin (2006) chose to penalize the $L_2$-norm of the
coefficients within each group, i.e., \( \sum_{k=1}^{K} \| \beta_k \|_2 \), where
\[
\| \beta_k \|_2 = \sqrt{\beta_{k1}^2 + \ldots + \beta_{kp_k}^2}.
\] (4)

Due to the singularity of \( \| \beta_k \|_2 \) at \( \beta_k = 0 \), appropriately tuning \( \lambda \) can set the whole coefficient vector \( \beta_k = 0 \), hence the \( k \)th group is removed from the fitted model. We note that in the setting of wavelet analysis, this method reduces to Antoniadis and Fan (2001).

Instead of using the \( L_2 \)-norm penalty, Zhao, Rocha and Yu (2009) suggested using the \( L_\infty \)-norm penalty, i.e., \( \sum_{k=1}^{K} \| \beta_k \|_\infty \), where
\[
\| \beta_k \|_\infty = \max(|\beta_{k1}|, |\beta_{k2}|, \ldots, |\beta_{kp_k}|).
\] (5)

Similar to the \( L_2 \)-norm, the \( L_\infty \)-norm of \( \beta_k \) is also singular when \( \beta_k = 0 \); hence when \( \lambda \) is appropriately tuned, the \( L_\infty \)-norm can also effectively remove unimportant groups.

However, there are some possible limitations with these methods: Both the \( L_2 \)-norm penalty and the \( L_\infty \)-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 \( \| \beta_k \|_2 \) and \( \| \beta_k \|_\infty \) are singular only when the whole vector \( \beta_k = 0 \). Once a component of \( \beta_k \) 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 \( L_\infty \)-norm \( (5) \), if the “\( \max(\cdot) \)” 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 $L_\infty$-norm penalty, another possible limitation is that the estimated coefficients within a group tend to have the same magnitude, i.e. $|\beta_{k1}| = |\beta_{k2}| = \ldots = |\beta_{kp_k}|$; 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 (Fan and Li 2001, Fan and Peng 2004), 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 the within 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 the within group level.

We reparameterize $\beta_{kj}$ as

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

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 $k$th group, by treating each $\beta_{kj}$ hierarchically. $d_k$ is at the first level of the hierarchy, controlling $\beta_{kj}, j = 1, \ldots, p_k$, as a group; $\alpha_{kj}$’s are at the second level of the hierarchy, reflecting differences within the $k$th group.
For the purpose of variable selection, we consider the following penalized least squares criterion:

\[
\max_{d_k, \alpha_{kj}} \left(-\frac{1}{2} \sum_{i=1}^{n} \left( y_i - \sum_{k=1}^{K} d_k \sum_{j=1}^{p_k} \alpha_{kj} x_{i,kj} \right)^2 - \lambda_1 \cdot \sum_{k=1}^{K} d_k - \lambda_2 \cdot \sum_{k=1}^{K} \sum_{j=1}^{p_k} |\alpha_{kj}| \right) \tag{7}
\]

subject to \( d_k \geq 0, \ k = 1, \ldots, K \),

where \( \lambda_1 \geq 0 \) and \( \lambda_2 \geq 0 \) are tuning parameters. \( \lambda_1 \) controls the estimates at the group level, and it can effectively remove unimportant groups: if \( d_k \) is shrunken to zero, all \( \beta_{kj} \) in the \( k \)th group will be equal to zero. \( \lambda_2 \) controls the estimates at the variable-specific level: if \( d_k \) is not equal to zero, some of the \( \alpha_{kj} \) 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 \( \lambda_1 \) and \( \lambda_2 \) in (7) can be simplified into one. Specifically, let \( \lambda = \lambda_1 \cdot \lambda_2 \), we can show that (7) is equivalent to

\[
\max_{d_k, \alpha_{kj}} \left(-\frac{1}{2} \sum_{i=1}^{n} \left( y_i - \sum_{k=1}^{K} d_k \sum_{j=1}^{p_k} \alpha_{kj} x_{i,kj} \right)^2 - \sum_{k=1}^{K} d_k - \lambda \sum_{k=1}^{K} \sum_{j=1}^{p_k} |\alpha_{kj}| \right) \tag{8}
\]

subject to \( d_k \geq 0, \ k = 1, \ldots, K \).

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

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 \(\alpha_{kj}\). This also implies that in practice, we do not need to tune \(\lambda_1\) and \(\lambda_2\) 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 $\alpha_{kj}$ in (8), we can use an iterative approach, i.e., we first fix $d_k$ and estimate $\alpha_{kj}$, then we fix $\alpha_{kj}$ 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 (Efron, Hastie, Johnstone and Tibshirani 2004) or a quadratic programming package to efficiently solve for $\alpha_{kj}$. When $\alpha_{kj}$ 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 $y$. Center and normalize $x_{kj}$.

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 $y$ on each of the $x_{kj}$ for $\alpha_{kj}^{(0)}$. Let $\beta_{kj}^{(0)} = d_k^{(0)} \alpha_{kj}^{(0)}$ and $m = 1$.

3. (Update $\alpha_{kj}$) Let

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

then

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

4. (Update $d_k$) Let

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

then

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

5. (Update $\beta_{kj}$) 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 \(x_{kj}\) 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 \(x_{kj}^T x_{kj} = 1\) and \(x_{kj}^T x_{k'j'} = 0\) if \(k \neq k'\) or \(j \neq j'\), then Step 3 and Step 4 in the above algorithm have closed form solutions.

Let \(\hat{\beta}_{kj}^{\text{ols}} = x_{kj}^T y\) be the ordinary least squares solution when \(x_{kj}\) are orthonormal to each other.

**Step 3.** When \(d_k\) is fixed,

\[
\alpha_{kj}^{(m)} = \mathbb{I}(d_{k}^{(m-1)} > 0) \cdot \text{sgn}(\hat{\beta}_{kj}^{\text{ols}}) \cdot \left( \frac{|\hat{\beta}_{kj}^{\text{ols}}|}{d_{k}^{(m-1)}} - \frac{\lambda}{(d_{k}^{(m-1)})^2} \right) + .
\]

(Equation 9)

**Step 4.** When \(\alpha_{kj}\) is fixed,

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

(Equation 10)

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 \(\alpha_{kj}\) to be zero because of the penalty on \(\alpha_{kj}\). 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 \(\beta_{kj}\), we obtain \(\hat{\alpha}_{kj} = \hat{\beta}_{kj}^{\text{ols}} / d_{k}^{(m-1)}\). Equation (9) shrinks \(\hat{\alpha}_{kj}\), 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 \(k\)th group is important, the amount of shrinkage is small, while when \(d_{k}^{(m-1)}\) is small, which indicates the \(k\)th group is less important, the amount of shrinkage is large.

7
Now considering $d_k^{(m)}$ in equation (10). If all $\alpha_{kj}^{(m)}$ are zero, it is natural to estimate $d_k$ also to be zero because of the penalty on $d_k$. If not all $\alpha_{kj}^{(m)}$ are zero, say $\alpha_{kj1}^{(m)}, \ldots, \alpha_{kjr}^{(m)}$ are not zero, then we have $d_k = \beta_{kjs}/\alpha_{kjs}^{(m)}, 1 \leq s \leq r$. Again, plugging in $\hat{\beta}_{kjs}^{\text{ols}}$ for $\beta_{kjs}$, we obtain $r$ estimates for $d_k$: $\tilde{d}_k = \hat{\beta}_{kjs}^{\text{ols}}/\alpha_{kjs}^{(m)}, 1 \leq s \leq r$. 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 $\alpha_{kj}$. We first show that it can also be written in an equivalent form using the original regression coefficients $\beta_{kj}$.

**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

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

(11)

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 $\hat{d}_k = \sqrt{\lambda \|\hat{\beta}_k\|_1^{1}}, \hat{\alpha}_k = \frac{\hat{\beta}_k}{\sqrt{\lambda \|\hat{\beta}_k\|_1^{1}}}$ 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 $\beta_k$, rather than the sum of squares. However, unlike the $L_2$-norm, which is singular only at the point $\beta_k = 0$, (i.e., the whole vector is equal to 0), the square root of the $L_1$-norm is singular at $\beta_{kj} = 0$ no matter what are the values of other $\beta_{kj}$'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 in Zhao, Rocha and Yu (2009).

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 \( \infty \), 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, \( \beta \), \( y_i \) and \( x_{i,kj} \) are also changed to \( \beta_n \), \( y_{ni} \) and \( x_{ni,kj} \). We write \( 2\sqrt{\lambda} \) in (11) as \( n\lambda_n \), and the criterion (11) is re-written as

\[
\max_{\beta_n,kj} \left\{ -\frac{1}{2} \sum_{i=1}^{n} \left( y_{ni} - \sum_{k=1}^{K_n} \sum_{j=1}^{p_{nk}} x_{ni,kj} \beta_n,kj \right)^2 \right. \\
\left. -n\lambda_n \sum_{k=1}^{K_n} \sqrt{\beta_{n,k1}^2 + \beta_{n,k2}^2 + \ldots + \beta_{n,kp_{nk}}^2} \right\}
\]  

(12)

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

Let \( \beta_n^0 = (\beta_{An}^0, \beta_{Bn}^0, \beta_{Cn}^0)^T \) be the underlying true parameters, where

\[
\begin{align*}
A_n &= \{(k, j) : \beta_{n,kj}^0 \neq 0\}, \\
B_n &= \{(k, j) : \beta_{n,kj}^0 = 0, \beta_{nk}^0 \neq 0\}, \\
C_n &= \{(k, j) : \beta_{nk}^0 = 0\}, \\
D_n &= B_n \cup C_n.
\end{align*}
\]  

(13)

Note that \( A_n \) contains the indices of coefficients which are truly non-zero, \( C_n \) contains the indices where the whole “groups” are truly zero, and \( B_n \) contains the indices of zero coefficients, but they belong to some non-zero groups. So \( A_n, B_n \) and \( C_n \) 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 \( \hat{\beta}_n = (\hat{\beta}_{An}, \hat{\beta}_{Bn}, \hat{\beta}_{Cn})^T \) of (12), and if also \( P_n n^{-3/4}/\lambda_n \to 0 \) as \( n \to \infty \), then \( \Pr(\hat{\beta}_{Cn} = 0) \to 1 \).

Theorem 2 implies that the hierarchical lasso method can effectively remove unimportant groups. For the above root-\((n/P_n)\) consistent estimate, however, if \( B_n \neq \emptyset \) (empty set), then
Pr(\(\hat{\beta}_{B_n} = 0\)) \rightarrow 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 which has been used in Breiman (1995), Wang, Li and Tsai (2006), Zhang and Lu (2007), and Zou (2006), i.e., to penalize different coefficients differently. Specifically, we consider

\[
\max_{\beta_{n,kj}} \left( -\frac{1}{2} \sum_{i=1}^{n} \left( y_{ni} - \sum_{k=1}^{K_n} \sum_{j=1}^{p_k} x_{ni,kj} \beta_{n,kj} \right)^2 - n\lambda_n \sum_{k=1}^{K_n} \sqrt{w_{n,k1}|\beta_{n,k1}| + w_{n,k2}|\beta_{n,k2}| + \ldots + w_{n,kp_k}|\beta_{n,kp_k}|} \right),
\]

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,

\[
w_{n,kj} = \frac{1}{|\hat{\beta}_{n,kj}^{\text{ols}}|^\gamma} \quad \text{or} \quad w_{n,kj} = \frac{1}{|\hat{\beta}_{n,kj}^{\text{ridge}}|^\gamma},
\]

where \(\gamma\) 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 $V_{ni} = (X_{ni}, Y_{ni})$, $i = 1, \ldots, n$ are independent and identically distributed for every $n$. Conditioning on $X_{ni} = x_{ni}$, $Y_{ni}$ has a density $f_n(g_n(x^T_{ni} \beta_n), Y_{ni})$, where $g_n(\cdot)$ is a known link function. We maximize the penalized log-likelihood

$$
\max_{\beta_{n,kj}} Q_n(\beta_n) = L_n(\beta_n) - J_n(\beta_n)
\hspace{1cm} (16)
$$

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

$$
p_{\lambda_n, w_n}(\beta_{nk}) = \lambda_n \sqrt{w_{n,k1}|\beta_{n,k1}| + \cdots + w_{n,kp}|\beta_{n,kp}|}.
$$

Note that under the normal distribution, $L_n(g_n(x^T_{ni} \beta_n), y_{ni}) = -\frac{(y_{ni} - x^T_{ni} \beta_n)^2}{2c_1} + C_2$, 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 Fan and Li (2001) and Fan and Peng (2004), but due to the grouping structure and the adaptive weights, they are non-trivial extensions of Fan and Li (2001) and Fan and Peng (2004).

To control the adaptive weights, we define:

$$
a_n = \max \{w_{n,kj} : \beta^0_{n,kj} \neq 0\},
$$

$$
b_n = \min \{w_{n,kj} : \beta^0_{n,kj} = 0\}.
$$

We assume that

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

Then we have the following results.

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

Then we have the following results.

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

11
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 $\{V_{ni}, i = 1, 2, \ldots, n\}$ are independent and identically distributed, each with a density $f_n(V_{n1}, \beta_n)$ that satisfies conditions (A1)-(A3) in the Appendix. If $\frac{p_n}{n} \to 0$, $\lambda_n \sqrt{a_n} = O_p(n^{-1/2})$ and $\frac{p_n^2}{\lambda_n b_n} = o_p(n)$, then there exists a root-$(n/P_n)$ consistent local maximizer $\hat{\beta}_n$ such that:

(a) Sparsity: $\Pr(\hat{\beta}_{n,D_n} = 0) \to 1$, where $D_n = B_n \cup C_n$.

(b) Asymptotic normality: If $\lambda_n \sqrt{a_n} = o_p((nP_n)^{-1/2})$ and $\frac{p_n^2}{\lambda_n b_n} \to 0$ as $n \to \infty$, then we also have:

$$\sqrt{n} A_n I_n^{1/2}(\beta_0^0, A_n)(\hat{\beta}_n, A_n - \beta_0^0, A_n) \to N(0, G),$$

where $A_n$ is a $q \times |A_n|$ matrix such that $A_n A_n^T \to I_q$ for a fixed $q$. This problem includes testing simultaneously the significance of several covariate variables.

The above requirements $\lambda_n \sqrt{a_n} = o_p((nP_n)^{-1/2})$ and $\frac{p_n^2}{\lambda_n b_n} = o_p(n)$ as $n \to \infty$ can be satisfied by selecting $\lambda_n$ and $w_{n,kj}$ appropriately. For example, we may let $\lambda_n = \frac{(nP_n)^{-1/2}}{\log n}$ and $w_{n,kj} = \frac{1}{|\hat{\beta}_0^0_{n,kj}|^2}$, where $\hat{\beta}_0^0_{n,kj}$ is the un-penalized likelihood estimate of $\beta_0^0_{n,kj}$, 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 $\frac{p_n^2}{\lambda_n b_n} = o_p(n)$ are satisfied when $\frac{p_n}{n} \to 0$.

### 5.2 Likelihood Ratio Test

Similarly as in Fan and Peng (2004), we develop a likelihood ratio test for testing linear hypotheses:

$$H_0 : A_n \beta_0^0, A_n = 0 \quad \text{vs.} \quad H_1 : A_n \beta_0^0, A_n \neq 0,$$

where $A_n$ is a $q \times |A_n|$ matrix and $A_n A_n^T \to I_q$ 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.

\[ T_n = 2 \left\{ \sup_{\Omega_n} Q_n(\beta_n | V) - \sup_{\Omega_n, A_n, A_n = 0} Q_n(\beta_n | V) \right\}, \]

where \( \Omega_n \) is the parameter space for \( \beta_n \). 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

\[ T_n \to \chi^2_q, \quad \text{as } n \to \infty. \]

## 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 \( L_\infty \)-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 Yuan and Lin (2006). 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”**

\[
Y = \left[ X_3 + 0.5X_3^2 + 0.1X_3^3 + 0.1X_3^4 \right] + \left[ X_6 - 0.5X_6^2 + 0.15X_6^3 + 0.1X_6^4 \right] + [\mathbb{I}(X_9 = 0) + \mathbb{I}(X_9 = 1) + \mathbb{I}(X_9 = 2)] + \varepsilon.
\]
Case 2. “Not all-in-all-out”

\[ Y = (X_3 + X_3^2) + (2X_6 - 1.5X_6^2) + [\mathbb{1}(X_9 = 0) + 2 \mathbb{1}(X_9 = 1)] + \varepsilon. \]

For all the simulations above, the error term \( \varepsilon \) follows a normal distribution \( N(0, \sigma^2) \), where \( \sigma^2 \) was set such that each of the signal to noise ratios, \( \text{Var}(\mathbf{X}^T \mathbf{\beta})/\text{Var}(\varepsilon) \), 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 \( \lambda \)’s that minimized the validation error. Using these selected \( \lambda \)’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 \( L_\infty \)-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 \( L_\infty \)-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
Table 1: Comparison of several group variable selection methods, including the $L_2$-norm group lasso, the $L_\infty$-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 | $L_\infty$ | $L_2$ | 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 | $L_\infty$ | $L_2$ | 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%) |

the grouping structure information. As a benchmark, we also computed the Oracle OLS results, i.e., OLS using only the important variables. The results 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 \( n = 200, 400, 800, 1600 \) and 3200. 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.

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%)  |
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.
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 (Olivier et al. 2002).

Instead of single-gene analysis, gene-set information has recently been used to analyze gene expression data. For example, Subramanian et al. (2005) developed the Gene Set Enrichment Analysis (GSEA), which is found to be more stable and more powerful than single-gene analysis. Efron and Tibshirani (2007) improved the GSEA method 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 (Subramanian et al. 2005). 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 gene-sets, we thus extend the hierarchical lasso to the case of overlapping groups. Suppose there are $K$ groups and $J$ variables. Let $G_k$ denote the set of indices of the variables in the $k$th group. One way to model the overlapping situation is to extend the criterion (8) as the following:

$$\max_{d_k, \alpha_j} \sum_{i=1}^{n} \ell \left( \sum_{k=1}^{K} d_k \sum_{j \in G_k} \alpha_j x_{i,j}, y_i \right)$$

$$- \sum_{k=1}^{K} d_k - \lambda \sum_{j=1}^{J} |\alpha_j|$$

subject to $d_k \geq 0$, $k = 1, \ldots, K$,

where $\alpha_j$ 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 $d_k$. 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 (Meier, van der Geer and Buhlmann 2008) 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%) 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 (Loh, Moritz, Contente and Dobbelstein 2003).

We also compared the gene-sets selected by the logistic hierarchical lasso with those
selected by the GSEA of Subramanian et al. (2005) and the GSA of Efron and Tibshirani (2007). 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 (Kubbutat, Jones and Vousden 1997, Moll and Petrenko 2003). 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
Figure 2: The number of samples vs the frequency that a sample was correctly classified on 100 random splits of the p53 data.

problem to illustrate the point (Breiman and Friedman 1997, Turlach, Venables and Wright 2005). Suppose we observe \((x_1, y_1), \ldots, (x_n, y_n)\), where each \(y_i = (y_{i1}, \ldots, y_{iK})\) 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, \(f_k(x) = \beta_{k1}x_1 + \cdots + \beta_{kp}x_p, k = 1, \ldots, K\). 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 \(j\)th
prediction variable will be removed from all $K$ predictions. If $d_j$ is not equal to zero, then some of the $\alpha_{kj}$ and hence some of the $\beta_{kj}$, $k = 1, \ldots, K$, still have the possibility of being zero. Therefore, the $j$th variable may be predictive for some responses but non-predictive for others.

One referee pointed out the work by Huang, Ma, Xie and Zhang (2009), which we were not aware of when our manuscript was first completed and submitted in 2007. We acknowledge that the work by Huang, Ma, Xie and Zhang (2009) 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 Huang, Ma, Xie and Zhang (2009) considered the oracle property only for group selection.

- Our theory applies to the generalized maximum likelihood estimate, while Huang, Ma, Xie and Zhang (2009) considered the penalized least squares estimate.

- Handling overlapping groups. It is not unusual for a variable to be a member of several groups. The gene expression data 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 Huang, Ma, Xie and Zhang (2009), 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

Zhou and Zhu were partially supported by grants DMS-0505432, DMS-0705532 and DMS-0748389 from the National Science Foundation.

References

Antoniadis, A., and Fan, J. (2001), “Regularization of Wavelet Approximations,” *Journal of the American Statistical Association*, 96, 939–967.

Breiman, L. (1995), “Better Subset Regression Using the Nonnegative Garrote,” *Technometrics*, 37, 373–384.

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.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), “Least Angle Regression,” *The Annals of Statistics*, 32, 407–499.

Efron, B., and Tibshirani, R. (2007), “On Testing the Significance of Sets of Genes,” *The Annals of Applied Statistics*, 1, 107–129.

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.

Fan, J., and Peng, H. (2004), “Nonconcave Penalized Likelihood with a Diverging Number of Parameters,” *The Annals of Statistics*, 32(3), 928–961.

George, E., and Foster, D. (2000), “Calibration and Empirical Bayes Variable Selection,” *Biometrika*, 87, 731–747.

George, E., and McCulloch, R. (1993), “Variable Selection via Gibbs Sampling,” *Journal of the American Statistical Association*, 88, 881–889.
Huang, J., Ma, S., Xie, H., and Zhang, C. (2009), “A Group Bridge Approach for Variable Selection,” *Biometrika*, 96, 339–355.

Kubbutat, M., Jones, S., and Vousden, K. (1997), “Regulation of p53 Stability by MDM2,” *Nature*, 387, 299–303.

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.

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.

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.

Moll, U., and Petrenko, O. (2003), “The MDM2-p53 Interaction,” *Molecular Cancer Research*, 1, 1001–1008.

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.

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.

Tibshirani, R. (1996), “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Turlach, B., Venables, W., and Wright, S. (2005), “Simultaneous Variable Selection,” *Technometrics*, 47(3), 349–363.
Appendix

Proof of Lemma 1

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

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

We immediately have

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

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

\[
\frac{\lambda_1 \|d'' - \hat{d}^*\| + \frac{1}{\lambda_1} \|\lambda_1 \alpha'' - \hat{\alpha}^*\|}{\min(\lambda_1, \frac{1}{\lambda_1})} \leq \frac{\|d'' - \hat{d}^*\| + \|\alpha'' - \hat{\alpha}^*\|}{\min(\lambda_1, \frac{1}{\lambda_1})} \leq \frac{\delta'}{\min(\lambda_1, \frac{1}{\lambda_1})} < \delta.
\]

Hence

\[
Q^*(\lambda, \hat{d}'', \hat{\alpha}'') = Q^*(\lambda_1, \lambda_2, \hat{d}'', \lambda_1 \hat{\alpha}'') \leq Q^*(\lambda_1, \lambda_2, \hat{d}, \hat{\alpha}^*) = Q^*(\lambda, \hat{d}, \hat{\alpha}^*).
\]

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

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

**Lemma 2** Suppose \((\hat{d}, \hat{\alpha})\) is a local maximizer of (8). Let \(\hat{\beta}\) be the Hierarchical Lasso estimate related to \((\hat{d}, \hat{\alpha})\), i.e., \(\hat{\beta}_{kj} = \hat{d}_k \hat{\alpha}_{kj}\). If \(\hat{d}_k = 0\), then \(\hat{\alpha}_k = 0\); if \(\hat{d}_k \neq 0\), then \(\|\hat{\beta}_k\|_1 \neq 0\) and \(\hat{d}_k = \sqrt{\lambda \|\hat{\beta}_k\|_1}, \hat{\alpha}_k = \frac{\hat{\beta}_k}{\sqrt{\lambda \|\hat{\beta}_k\|_1}}\).
We prove $\hat{d}_k = \sqrt{\lambda \| \hat{\beta}_k \|_1}, \hat{\alpha}_k = \frac{\hat{\beta}_k}{\sqrt{\lambda \| \hat{\beta}_k \|_1}}$ for $\hat{d}_k \neq 0$ by contradiction. Suppose $\exists k'$ such that $\hat{d}_{k'} \neq 0$ and $\hat{d}_{k'}' \neq \sqrt{\lambda \| \hat{\beta}_{k'} \|_1}$. Let $\sqrt{\lambda \| \hat{d}_{k'} \|_1} = c$. Then $\hat{\alpha}_k = c \frac{\hat{\beta}_k}{\sqrt{\lambda \| \hat{\beta}_k \|_1}}$. Suppose $c > 1$.

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

Then we have

$$Q^*(\lambda, \tilde{d}, \tilde{\alpha}) - Q^*(\lambda, \hat{d}, \hat{\alpha}) = -\delta' |\tilde{d}_{k'}| - \frac{1}{\delta'} \lambda \| \tilde{\alpha}_{k'} \|_1 + |\hat{d}_{k'}| + \lambda \| \hat{\alpha}_{k'} \|_1$$

$$= \left( -\frac{\delta'}{c} - \frac{c}{\delta'} + 1 + c \right) \sqrt{\lambda \| \hat{\beta}_{k'} \|_1}$$

$$= \frac{1}{c} (\delta' - 1) \left( \frac{c^2}{\delta'} - 1 \right) \sqrt{\lambda \| \hat{\beta}_{k'} \|_1}$$

$$> 0.$$ 

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

Similarly for the case when $c < 1$. Hence, we have the result that if $\hat{d}_k \neq 0$, then

$$\hat{d}_k = \sqrt{\lambda \| \hat{\beta}_k \|_1}, \hat{\alpha}_k = \frac{\hat{\beta}_k}{\sqrt{\lambda \| \hat{\beta}_k \|_1}}.$$

**Proof of Theorem 1**

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

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

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

Now we show $Q(\lambda, \hat{\beta} + \Delta \beta^{(1)}) \leq Q(\lambda, \hat{\beta})$ if $\delta'$ is small enough. By Lemma 2, we have $\hat{d}_k = \sqrt{\lambda \| \hat{\beta}_k \|_1}, \hat{\alpha}_k = \frac{\hat{\beta}_k}{\sqrt{\lambda \| \hat{\beta}_k \|_1}}$ if $\|\hat{d}_k\|_1 \neq 0$ and $\hat{\alpha}_k = 0$ if $\|\hat{d}_k\|_1 = 0$. Furthermore, let $\hat{d}_k' = \sqrt{\lambda \| \hat{\beta}_k + \Delta \beta^{(1)} \|_1}, \hat{\alpha}_k' = \frac{\hat{\beta}_k + \Delta \beta^{(1)}_k}{\sqrt{\lambda \| \hat{\beta}_k + \Delta \beta^{(1)}_k \|_1}}$ if $\|\hat{d}_k\|_1 \neq 0$. Let $\hat{d}_k' = 0, \hat{\alpha}_k' = 0$ if $\|\hat{d}_k\|_1 = 0$.

Then we have $Q^*(\lambda, \hat{d}', \hat{\alpha}') = Q(\lambda, \hat{\beta} + \Delta \beta^{(1)})$ and $Q^*(\lambda, \hat{d}, \hat{\alpha}) = Q(\lambda, \hat{\beta})$. Hence we only need
to show that $Q^*(\lambda, \hat{d}', \alpha') \leq Q^*(\lambda, \hat{d}, \hat{\alpha})$. Note that $(\hat{d}, \hat{\alpha})$ is a local maximizer of $Q^*(\lambda, d, \alpha)$. Therefore there exists a $\delta$ such that for any $d', \alpha'$ satisfying $\|d' - \hat{d}\|_1 + \|\alpha' - \hat{\alpha}\|_1 < \delta$, we have $Q^*(\lambda, d', \alpha') \leq Q^*(\lambda, \hat{d}, \hat{\alpha})$.

Now since

$$|d_k' - d_k| = \left| \sqrt{\lambda \|\hat{\beta}_k + \Delta \beta_k^{(1)}\|_1} - \sqrt{\lambda \|\hat{\beta}_k\|_1} \right|$$

$$\leq \left| \sqrt{\lambda \|\hat{\beta}_k\|_1} - \lambda \|\Delta \beta_k^{(1)}\|_1 - \sqrt{\lambda \|\hat{\beta}_k\|_1} \right|$$

$$\leq \frac{1}{\sqrt{\lambda \|\hat{\beta}_k\|_1}} \|\Delta \beta_k^{(1)}\|_1$$

$$\leq \frac{1}{2} \sqrt{\lambda a - \lambda \delta'}$$

$$\leq \frac{1}{2} \sqrt{\lambda a/2},$$

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

Furthermore

$$\|\hat{\alpha}'_k - \alpha_k\|_1 = \left\| \frac{\hat{\beta}_k + \Delta \beta_k^{(1)}}{\sqrt{\lambda \|\hat{\beta}_k + \Delta \beta_k^{(1)}\|_1}} - \frac{\hat{\beta}_k}{\sqrt{\lambda \|\hat{\beta}_k\|_1}} \right\|_1$$

$$\leq \left\| \frac{\hat{\beta}_k + \Delta \beta_k^{(1)}}{\sqrt{\lambda \|\hat{\beta}_k + \Delta \beta_k^{(1)}\|_1}} - \frac{\hat{\beta}_k}{\sqrt{\lambda \|\hat{\beta}_k\|_1}} \right\|_1$$

$$+ \left\| \frac{\hat{\beta}_k}{\sqrt{\lambda \|\hat{\beta}_k\|_1}} - \frac{\hat{\beta}_k}{\sqrt{\lambda \|\hat{\beta}_k\|_1}} \right\|_1$$

$$\leq \frac{\|\Delta \beta_k^{(1)}\|_1}{\sqrt{\lambda a/2}}$$

$$+ \frac{\|\hat{\beta}_k\|_1}{\sqrt{\lambda \|\hat{\beta}_k\|_1}} \left( \frac{\lambda \|\Delta \beta_k^{(1)}\|_1}{\sqrt{\lambda a/2}} - \sqrt{\lambda \|\hat{\beta}_k\|_1} \right)$$

$$\leq \frac{\|\Delta \beta_k^{(1)}\|_1}{\sqrt{\lambda a/2}} + \frac{b}{\sqrt{\lambda a/2} \sqrt{\lambda a}} \left( \frac{\frac{1}{2} \lambda \|\Delta \beta_k^{(1)}\|_1}{\sqrt{\lambda a/2}} \right)$$

$$\leq \frac{\|\Delta \beta_k^{(1)}\|_1}{\sqrt{\lambda a/2}} + \frac{b}{\frac{1}{2} \sqrt{\lambda a}}.$$
where \( b = \max\{\|\hat{\beta}_k\|_1 : \|\hat{\beta}_k\|_1 \neq 0\} \).

Therefore, there exists a small enough \( \delta' \), if \( \|\triangle\beta\|_1 < \delta' \) we have \( \|\hat{d}' - \hat{d}\|_1 + \|\hat{\alpha}' - \hat{\alpha}\|_1 < \delta \). Hence \( Q^*(\lambda, \hat{d}', \hat{\alpha}') \leq Q^*(\lambda, \hat{d}, \hat{\alpha}) \) (due to local maximality) and \( Q(\lambda, \hat{\beta} + \triangle\beta(1)) \leq Q(\lambda, \hat{\beta}) \).

Next we show \( Q(\lambda, \hat{\beta} + \triangle\beta(1) + \triangle\beta(2)) \leq Q(\lambda, \hat{\beta} + \triangle\beta(1)) \). Note that
\[
\frac{Q(\lambda, \hat{\beta} + \triangle\beta(1) + \triangle\beta(2)) - Q(\lambda, \hat{\beta} + \triangle\beta(1))}{\lambda \|\triangle\beta(2)\|_1} = \sum_{k=1}^{K} \sqrt{\lambda \|\triangle\beta(2)\|_1},
\]
where \( \beta^* \) is a vector between \( \hat{\beta} + \triangle\beta(1) + \triangle\beta(2) \) and \( \hat{\beta} + \triangle\beta(1) \). Since \( \|\triangle\beta(2)\|_1 < \delta' \) is small enough, the second term dominates the first term, hence we have \( Q(\lambda, \hat{\beta} + \triangle\beta(1) + \triangle\beta(2)) \leq Q(\lambda, \hat{\beta} + \triangle\beta(1)) \).

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

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

**Regularity Conditions**

Let \( S_n \) be the number of non-zero groups, i.e., \( \|\beta^0_{nk}\| \neq 0 \). Without loss of generality, we assume
\[
\|\beta^0_{nk}\| \neq 0, \text{ for } k = 1, \ldots, S_n, \\
\|\beta^0_{nk}\| = 0, \text{ for } k = S_n + 1, \ldots, K_n.
\]

Let \( s_{nk} \) be the number of non-zero coefficients in group \( k, 1 \leq k \leq S_n \); again, without loss of generality, we assume
\[
\beta^0_{n,kj} \neq 0, \text{ for } k = 1, \ldots, S_n; j = 1, \ldots, s_{nk}, \\
\beta^0_{n,kj} = 0, \text{ for } k = 1, \ldots, S_n; j = s_{nk} + 1, \ldots, p_{nk}.
\]

For simplicity, we write \( \beta_{n,kj}, p_{nk} \) and \( s_{nk} \) as \( \beta_{kj}, p_k \) and \( s_k \) in the following.
For every $n$, there exists an open subset $\omega$ estimate under finite parameters (Lehmann and Casella 1998). 

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

$$
E_{\beta_n} \left[ \frac{\partial \log f_n(V_{n1}, \beta_n)}{\partial \beta_k} \right] = 0, \quad \text{for } k = 1, \ldots, K_n; \quad j = 1, \ldots, p_k
$$

$$
I_{k_1j_1k_2j_2}(\beta_n) = E_{\beta_n} \left[ \frac{\partial}{\partial \beta_{k_1j_1}} \log f_n(V_{n1}, \beta_n) \frac{\partial}{\partial \beta_{k_2j_2}} \log f_n(V_{n1}, \beta_n) \right]
$$

$$
= E_{\beta_n} \left[ -\frac{\partial^2}{\partial \beta_{k_1j_1} \partial \beta_{k_2j_2}} \log f_n(V_{n1}, \beta_n) \right].
$$

(A2) The Fisher information matrix

$$
I(\beta_n) = E_{\beta_n} \left[ \frac{\partial}{\partial \beta_n} \log f_n(V_{n1}, \beta_n) \frac{\partial^T}{\partial \beta_n} \log f_n(V_{n1}, \beta_n) \right]
$$

satisfies the condition

$$
0 < C_1 < \lambda_{\min}\{I(\beta_n)\} \leq \lambda_{\max}\{I(\beta_n)\} < C_2 < \infty,
$$

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

$$
E_{\beta_n} \left[ \frac{\partial}{\partial \beta_{k_1j_1}} \log f_n(V_{n1}, \beta_n) \frac{\partial}{\partial \beta_{k_2j_2}} \log f_n(V_{n1}, \beta_n) \right]^2 < C_3 < \infty,
$$

$$
E_{\beta_n} \left[ -\frac{\partial^2}{\partial \beta_{k_1j_1} \partial \beta_{k_2j_2}} \log f_n(V_{n1}, \beta_n) \right]^2 < C_4 < \infty.
$$

(A3) There exists an open subset $\omega_n$ of $\Omega_n \in R^{p_n}$ that contains the true parameter point $\beta_n^0$ such that for almost all $V_{n1}$, the density $f_n(V_{n1}, \beta_n)$ admits all third derivatives $\partial^3 f_n(V_{n1}, \beta_n)/(\partial \beta_{k_1j_1} \partial \beta_{k_2j_2} \partial \beta_{k_3j_3})$ for all $\beta_n \in \omega_n$. Furthermore, there exist functions $M_{nk_1j_1k_2j_2k_3j_3}$ such that

$$
\left| \frac{\partial^3}{\partial \beta_{k_1j_1} \partial \beta_{k_2j_2} \partial \beta_{k_3j_3}} \log f_n(V_{n1}, \beta_n) \right| \leq M_{nk_1j_1k_2j_2k_3j_3}(V_{n1}) \quad \text{for all } \beta_n \in \omega_n,
$$

and $E_{\beta_n}[M_{nk_1j_1k_2j_2k_3j_3}^2(V_{n1})] < C_5 < \infty.$
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

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

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

Using $p_{\lambda_n, w_n}(0) = 0$, we have

$$D_n(u) = Q_n (\beta_n^0 + \alpha_n u) - Q_n (\beta_n^0)$$

$$\leq L_n (\beta_n^0 + \alpha_n u) - L_n (\beta_n^0)$$

$$- n \sum_{k=1}^{S_n} \left( p_{\lambda_n, w_n} (\beta_n^0 + \alpha_n u_k) - p_{\lambda_n, w_n} (\beta_n^0 + \alpha_n u_k) \right)$$

$$\triangleq (I) + (II).$$

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

$$(I) = \alpha_n u^T \nabla L_n (\beta_n^0) + \frac{1}{2} u^T \nabla^2 L_n (\beta_n^0) u \alpha_n^2 + \frac{1}{6} u^T \nabla \{ u^T \nabla^2 L_n (\beta_n^*) u \} \alpha_n^3$$

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

where $\beta_n^*$ lies between $\beta_n^0$ and $\beta_n^0 + \alpha_n u$. Using the same argument as in the proof of Theorem 1 of Fan and Peng (2004), we have

$$|I_1| = O_p(\alpha_n^2 n \|u\|);$$

$$I_2 = - \frac{n \alpha_n^2}{2} u^T I_n (\beta_n^0) u + o_p(1)n \alpha_n^2 \|u\|^2,$$
\[ |I_3| = \left| \frac{1}{6} \sum_{k_1=1}^n \sum_{j_1=1}^n \sum_{k_2=1}^n \sum_{j_2=1}^n \sum_{k_3=1}^n \sum_{j_3=1}^n \frac{\partial^3 L_n(\beta^*_n)}{\partial \beta_{k_1j_1} \partial \beta_{k_2j_2} \partial \beta_{k_3j_3}} u_{k_1j_1} u_{k_2j_2} u_{k_3j_3} \alpha^3_n \right|. \]

\[
\leq \frac{1}{6} \sum_{i=1}^n \left\{ \sum_{k_1=1}^n \sum_{j_1=1}^n \sum_{k_2=1}^n \sum_{j_2=1}^n \sum_{k_3=1}^n \sum_{j_3=1}^n \sum_{p_1=1}^n \sum_{p_2=1}^n \sum_{p_3=1}^n \sum_{p_4=1}^n M^2_{nk_1j_1k_2j_2k_3j_3} (V_{ni}) \right\}^{1/2} \|u\|^2 \alpha^3_n.
\]

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

\[ |I_3| = o_p(n \alpha^3_n) \|u\|^2. \] (23)

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

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 \( \|u\| \leq C \) we have

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

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

\[
p_{\lambda_n,w_n}(\beta_{nk}^0 + \alpha_n u_k) - p_{\lambda_n,w_n}(\beta_{nk}^0)
= \lambda_n(\sqrt{w_{nk,1} |\beta_{nk}^0 + \alpha_n u_{k1}| + \ldots + w_{nk,p_k} |\beta_{nk}^0 + \alpha_n u_{kp_k}|} - w_{nk,1} |\beta_{nk}^0| + \ldots + w_{nk,p_k} |\beta_{nk}^0|)
\geq \lambda_n(\sqrt{w_{nk,1} |\beta_{nk}^0 + \alpha_n u_{k1}| + \ldots + w_{nk,k_{sh}} |\beta_{nk}^0 + \alpha_n u_{k_{sh}}|} - w_{nk,1} |\beta_{nk}^0| + \ldots + w_{nk,k_{sh}} |\beta_{nk}^0|)
\geq \lambda_n(\sqrt{w_{nk,1} |\beta_{nk}^0| + \ldots + w_{nk,k_{sh}} |\beta_{nk}^0|} - \alpha_n (w_{nk,1} |u_{k1}| + \ldots + w_{nk,k_{sh}} |u_{k_{sh}}|))
= \lambda_n \sqrt{w_{nk,1} |\beta_{nk}^0| + \ldots + w_{nk,k_{sh}} |\beta_{nk}^0|} \left( \sqrt{1 - \gamma_{nk}} - 1 \right),
\]

where \( \gamma_{nk} \) is defined as \( \gamma_{nk} = \frac{\alpha_n (w_{nk,1} |u_{k1}| + \ldots + w_{nk,k_{sh}} |u_{k_{sh}}|)}{w_{nk,1} |\beta_{nk}^0| + \ldots + w_{nk,k_{sh}} |\beta_{nk}^0|} \). For \( n \) large enough, we have \( 0 \leq \gamma_{nk} < 1 \) and \( \gamma_{nk} \leq \frac{\alpha_n |u_k| (w_{nk,1} + \ldots + w_{nk,k_{sh}})}{c_1 (w_{nk,1} + \ldots + w_{nk,k_{sh}})} = \frac{\alpha_n |u_k|}{c_1} \leq \frac{\alpha_n C}{c_1} \to 0 \) with probability tending to 0 as \( n \to \infty \).
Therefore,
\[
p_{\lambda_n} w_n (\beta_{nk}^0 + \alpha_n u_k) - p_{\lambda_n} w_n (\beta_{nk}^0) \\
\geq \lambda_n \sqrt{w_{n,k_1} |\beta_{k_1}^0| + \ldots + w_{n,k_{s_k}} |\beta_{k_{s_k}}^0|} (\sqrt{1 - \gamma_{nk}} - 1) \\
\geq \lambda_n \sqrt{w_{n,k_1} |\beta_{k_1}^0| + \ldots + w_{n,k_{s_k}} |\beta_{k_{s_k}}^0|} \left( \frac{1 + |o_p(1)|}{2} (-\gamma_{nk}) \right) \\
\text{(Using } \gamma_{nk} = o_p(1) \text{ and Taylor expansion)} \\
\geq -\lambda_n \frac{\alpha_n (w_{n,k_1} |u_{k_1}| + \ldots + w_{n,k_{s_k}} |u_{k_{s_k}}|)}{\sqrt{w_{n,k_1} |\beta_{k_1}^0| + \ldots + w_{n,k_{s_k}} |\beta_{k_{s_k}}^0|}} \left( \frac{1 + |o_p(1)|}{2} \right) \\
\geq -\alpha_n \lambda_n \frac{\|u_k\| \sqrt{a_n s_k}}{2 \sqrt{c_1}} (1 + |o_p(1)|).
\]

Therefore, the term (II) in (19) is bounded by
\[
n\alpha_n \lambda_n \left( \sum_{k=1}^{s_n} \frac{\|u_k\| \sqrt{a_n s_k}}{2 \sqrt{c_1}} \right) (1 + |o_p(1)|),
\]
which is further bounded by
\[
n\alpha_n \lambda_n \sqrt{a_n} (\|u\| \cdot \frac{\sqrt{P_n}}{2 \sqrt{c_1}}) (1 + |o_p(1)|).
\]
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
\[
\|u\| n\alpha_n^2 (1 + |o_p(1)|).
\]
This term is also dominated by the first term of \( I_2 \) on \( \|u\| = C \) uniformly. Therefore, \( D_n(u) < 0 \) is satisfied uniformly on \( \|u\| = C \). 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 \( \hat{\beta}_n \). 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

\[
\frac{\partial Q_n(\beta_n)}{\partial \beta_{kj}} = \frac{\partial L_n(\beta_n)}{\partial \beta_{kj}} - n \frac{\partial p_{\lambda_n w_n}(\beta_{nk})}{\partial \beta_{kj}} \\
= \frac{\partial L_n(\beta_n^0)}{\partial \beta_{kj}} + \sum_{k_1=1}^{K_n} \sum_{j_1=1}^{p_{k_1}} \frac{\partial^2 L_n(\beta_n^0)}{\partial \beta_{kj} \partial \beta_{k_1 j_1}} (\beta_{k_1 j_1} - \beta_{k_1 j_1}^0) \\
+ \frac{1}{2} \sum_{k_1=1}^{K_n} \sum_{j_1=1}^{p_{k_1}} \sum_{k_2=1}^{p_{k_2}} \sum_{j_2=1}^{p_{k_2}} \frac{\partial^3 L_n(\beta_n^*)}{\partial \beta_{kj} \partial \beta_{k_1 j_1} \partial \beta_{k_2 j_2}} (\beta_{k_1 j_1} - \beta_{k_1 j_1}^0) (\beta_{k_2 j_2} - \beta_{k_2 j_2}^0) \\
- \frac{n \lambda_n w_{n,kj}}{2 \sqrt{w_{n,k1} |\beta_{k1}| + \ldots + w_{n,kp_k} |\beta_{kp_k}|}} \text{sgn}(\beta_{kj}) \\
\triangleq I_1 + I_2 + I_3 + I_4, (25)
\]

where \(\beta_n^*\) lies between \(\beta_n\) and \(\beta_n^0\).

Using the argument in the proof of Lemma 5 of Fan and Peng (2004), for any \(\beta_n\) satisfying

\[
\|\beta_n - \beta_n^0\| = O_p(\sqrt{P_n/n}),
\]
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 \(\hat{\beta}_n\) is a root-(\(n/P_n\)) consistent estimate maximizing \(Q_n(\beta_n)\), if \(\hat{\beta}_{kj} \neq 0\), we have

\[
\left. \frac{\partial Q_n(\beta_n)}{\partial \beta_{kj}} \right|_{\beta_n = \hat{\beta}_n} = O_p(\sqrt{nP_n}) - \frac{n \lambda_n w_{n,kj}}{2 \sqrt{w_{n,k1} |\hat{\beta}_{k1}| + \ldots + w_{n,kp_k} |\hat{\beta}_{kp_k}|}} \text{sgn}(\hat{\beta}_{kj}) \\
= 0. (26)
\]

Therefore,

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

This can be extended to

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

34
for any \( \hat{\beta}_{kj} \) with \( \hat{\beta}_{nk} \neq 0 \). If we sum this over all \( j \) in the \( k \)th group, we have

\[
\begin{align*}
n\lambda_n \sqrt{w_{n,k1}|\hat{\beta}_{k1}| + \ldots + w_{n,kp_k}|\hat{\beta}_{kp_k}|} = \sum_{j=1}^{p_k} |\hat{\beta}_{kj}|O_p(\sqrt{nP_n}).
\end{align*}
\] (27)

Since \( \hat{\beta}_n \) is a root-\((n/P_n)\) consistent estimate of \( \beta^0_n \), 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 \( \beta^0_{kj} = 0 \) and \( \hat{\beta}_{kj} \neq 0 \), equation (26) can be written as:

\[
\begin{align*}
\frac{\partial Q_n(\beta_n)}{\partial \beta_{kj}} \bigg|_{\beta_n = \hat{\beta}_n} &= \frac{1}{2\lambda_n \sqrt{w_{n,k1}|\hat{\beta}_{k1}| + \ldots + w_{n,kp_k}|\hat{\beta}_{kp_k}|}}
\end{align*}
\]

\[
\begin{align*}
&\quad (O_p(\sqrt{P_n/n})n\lambda_n \sqrt{w_{n,k1}|\hat{\beta}_{k1}| + \ldots + w_{n,kp_k}|\hat{\beta}_{kp_k}|}
\end{align*}
\]

\[
\begin{align*}
&\quad -n\lambda^2_n w_{n,kj}\text{sgn}(\hat{\beta}_{kj}))
\end{align*}
\]

\[
= 0.
\]

Denote \( h_{nk} = O_p(\sqrt{P_n/n})n\lambda_n \sqrt{w_{n,k1}|\hat{\beta}_{k1}| + \ldots + w_{n,kp_k}|\hat{\beta}_{kp_k}|} \). Let \( h_n = \sum_{k=1}^{K_n} h_{nk} \). By equation (27), we have \( h_n = \sum_{k=1}^{K_n} O_p(\sqrt{P_n/n}) \sum_{j=1}^{p_k} |\hat{\beta}_{kj}|O_p(\sqrt{nP_n}) = O_p(P^2_n) \). Since \( \frac{P^2_n}{\lambda_n b_n} = o_p(n) \) guarantees that \( n\lambda^2_n b_n \) 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 \( \beta^0_{kj} = 0 \) since \( w_{n,kj} \geq b_n \) and \( h_n > h_{nk} \). Denote \( g_{nk} = 2\lambda_n \sqrt{w_{n,k1}|\hat{\beta}_{k1}| + \ldots + w_{n,kp_k}|\hat{\beta}_{kp_k}|}/(n\lambda^2_n b_n) \). Let \( g_n = \sum_{k=1}^{K_n} g_{nk} \). By equation (27), we have \( g_n = 2 \sum_{k=1}^{K_n} (1/n) \sum_{j=1}^{p_k} |\hat{\beta}_{kj}|O_p(\sqrt{nP_n})/(n\lambda^2_n b_n) = o_p(1/\sqrt{nP_n}) \). 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 \( \frac{\partial Q_n(\beta)}{\partial \beta_{kj}} \bigg|_{\beta = \hat{\beta}_n} \) will go to \( \infty \) as \( n \to \infty \), which is a contradiction with equation (28). Therefore, for any \( k \) and \( j \) satisfying \( \beta^0_{kj} = 0 \), we have \( \hat{\beta}_{kj} = 0 \) with a probability tending to 1 as \( n \to \infty \). We have \( \hat{\beta}_D_n = 0 \) 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 \((\hat{\beta}_{n,A_n}, 0)\) 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 \( Q_n(\beta_{n,A_n}) = Q_n(\beta_{n,A_n}, 0) \). Using
the Taylor expansion on $\nabla Q_n(\hat{\beta}_{n,A_n})$ at point $\beta_{n,A_n}^0$, we have

$$
\frac{1}{n}(\nabla^2 L_n(\beta_{n,A_n}^0) (\hat{\beta}_{n,A_n} - \beta_{n,A_n}^0) - \nabla J_n(\hat{\beta}_{n,A_n}))
\quad = \quad -\frac{1}{n} \left( \nabla L_n(\beta_{n,A_n}^0) + \frac{1}{2}(\hat{\beta}_{n,A_n} - \beta_{n,A_n}^0)^T \nabla^2 \{\nabla L_n(\beta_{n,A_n}^*)\} (\hat{\beta}_{n,A_n} - \beta_{n,A_n}^0) \right),
$$

(29)

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

Now we define

$$
C_n \triangleq \frac{1}{2}(\hat{\beta}_{n,A_n} - \beta_{n,A_n}^0)^T \nabla^2 \{\nabla L_n(\beta_{n,A_n}^*)\} (\hat{\beta}_{n,A_n} - \beta_{n,A_n}^0).
$$

Using the Cauchy-Schwarz inequality, we have

$$
\left\| \frac{1}{n} C_n \right\|^2 \leq \frac{1}{n^2} \sum_{i=1}^n \| \hat{\beta}_{n,A_n} - \beta_{n,A_n}^0 \|^2 \sum_{k=1}^S_p \sum_{j=1}^S_p M_{nk,jk}^2 (V_{ni})
\quad = \quad O_p(P_n^2/n^2)O_p(P_n^3) = O_p(P_n^5/n^2) = o_p(1/n).
$$

(30)

Since $\frac{P_n^5}{n} \to 0$ as $n \to \infty$, by Lemma 8 of Fan and Peng (2004), we have

$$
\left\| \frac{1}{n} \nabla^2 L_n(\beta_{n,A_n}^0) + I_n(\beta_{n,A_n}^0) \right\| = o_p(1/P_n)
$$

and

$$
\left\| \left( \frac{1}{n} \nabla^2 L_n(\beta_{n,A_n}^0) + I_n(\beta_{n,A_n}^0) \right) (\hat{\beta}_{n,A_n} - \beta_{n,A_n}^0) \right\| = o_p(1/\sqrt{nP_n}) = o_p(1/\sqrt{n}).
$$

(31)

Since

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

we have

$$
\frac{\lambda_n w_{n,kj}}{\sqrt{w_{n,k1}\beta_{k1} + \ldots + w_{n,ks_k}\beta_{ks_k}}}
\quad = \quad \frac{\lambda_n w_{n,kj}}{\sqrt{w_{n,k1}\beta_{k1}^0 + \ldots + w_{n,ks_k}\beta_{ks_k}^0}}(1 + O_p(\sqrt{P_n/n})).
$$

36
Furthermore, since
\[
\frac{\lambda_n w_{n,kj}}{\sqrt{\sum_{k,k_1} |\beta_{k,k_1}^0| + \ldots + \sum_{k,k_s} |\beta_{k,k_s}^0|}} \leq \frac{\lambda_n w_{n,kj}}{\sqrt{w_{n,kj}c_1}} \leq \frac{\lambda_n \sqrt{a_n}}{\sqrt{c_1}} = o_p((nP_n)^{-1/2})
\]
for \((k, j) \in A_n\), we have
\[
\left(\frac{1}{n} \nabla J_n(\hat{\beta}_{n,A_n})\right)_{kj} = \frac{\lambda_n w_{n,kj}}{2\sqrt{\sum_{k,k_1} |\beta_{k,k_1}^0| + \ldots + \sum_{k,k_s} |\beta_{k,k_s}^0|}} = o_p((nP_n)^{-1/2})
\]
and
\[
\left\|\frac{1}{n} \nabla J_n(\hat{\beta}_{n,A_n})\right\| \leq \sqrt{P_n o_p((nP_n)^{-1/2})} = o_p(1/\sqrt{n}).
\]
Together with (30), (31) and (32), from (29) we have
\[
I_n(\beta_{n,A_n}^0)(\hat{\beta}_{n,A_n} - \beta_{n,A_n}^0) = \frac{1}{n} \nabla L_n(\beta_{n,A_n}^0) + o_p(1/\sqrt{n}).
\]
Now using the same argument as in the proof of Theorem 2 of Fan and Peng (2004), we have
\[
\sqrt{n}A_n I_n^{1/2}(\beta_{n,A_n}^0)(\hat{\beta}_{n,A_n} - \beta_{n,A_n}^0) \to \sqrt{n}A_n I_n^{-1/2}(\beta_{n,A_n}^0) \left(\frac{1}{n} \nabla L_n(\beta_{n,A_n}^0)\right) \to \mathcal{N}(0, G),
\]
where \(A_n\) is a \(q \times |A_n|\) matrix such that \(A_n A_n^\top \to G\) and \(G\) 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 \(\frac{P_n}{\lambda_n b_n} = o_p(n)\) 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 \(\lambda_n\), which implies that \(\Pr(\hat{\beta}_D = 0) \to 1\) cannot be guaranteed.

We will prove that by choosing \(\lambda_n\) 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 \(\hat{\beta}_n = (\hat{\beta}_{A_n}, \hat{\beta}_{B_n}, \hat{\beta}_{C_n})^\top\) such that \(\Pr(\hat{\beta}_{C_n} = 0) \to 1\).

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 \(h'_n = O_p(\sqrt{P_n/n} \sum_{j=1}^{p_k} |\hat{\beta}_{kj}|) = O_p(P_n^2/\sqrt{n})\). 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 \( \beta_{nk}^0 = 0 \) since \( w_{n,kj} = 1 \) and \( h_n' > h_{nk} \). Similar as in the proof of Theorem 4, we have \( \hat{\beta}_{c_n} = 0 \) with probability tending to 1.

**Proof of Theorem 5**

Let \( N_n = |A_n| \) be the number of nonzero parameters. Let \( B_n \) be an \((N_n - q) \times N_n\) matrix which satisfies \( B_nB_n^\top = I_{N_n-q} \) and \( A_nB_n^\top = 0 \). As \( \beta_{n,A_n} \) is in the orthogonal complement to the linear space that is spanned by the rows of \( A_n \) under the null hypothesis \( H_0 \), it follows that

\[
\beta_{n,A_n} = B_n^\top \gamma_n,
\]

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

\[
Q_n(\beta_{n,A_n}) = \max_{\gamma_n} Q_n(B_n^\top \gamma_n).
\]

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

\[
\hat{\beta}_{n,A_n} - \beta_{n,A_n}^0 = \frac{1}{n} \left[ I_n^{-1}(\beta_{n,A_n}^0) \nabla L_n(\beta_{n,A_n}^0) + o_p(n^{-1/2}) \right],
\]

\[
B_n^\top (\hat{\gamma}_n - \gamma_n^0) = \frac{1}{n} B_n^\top \{ B_n I_n(\beta_{n,A_n}^0) B_n^\top \}^{-1} B_n \nabla L_n(\beta_{n,A_n}^0) + o_p(n^{-1/2}).
\]

**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 \( H_0 \),

\[
B_n I_n(\beta_{n,A_n}^0) B_n^\top (\hat{\gamma}_n - \gamma_n^0) = \frac{1}{n} B_n \nabla L_n(\beta_{n,A_n}^0) + o_p(n^{-1/2}).
\]
As the eigenvalue $\lambda_i(B_n I_n(\beta_{n,A_n}^0) B_n^\top)$ is uniformly bounded away from 0 and infinity, we have

$$B_n^\top(\gamma_n - \gamma_n^0) = \frac{1}{n} B_n^\top \{ B_n I_n(\beta_{n,A_n}^0) B_n^\top \}^{-1} B_n \nabla L_n(\beta_{n,A_n}^0) + o_p(n^{-1/2}).$$

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

$$Q_n(\hat{\beta}_{n,A_n}) - Q_n(B_n^\top \hat{\gamma}_n) = \frac{n}{2} (\hat{\beta}_{n,A_n} - B_n^\top \hat{\gamma}_n)^\top I_n(\beta_{n,A_n}^0) (\hat{\beta}_{n,A_n} - B_n^\top \hat{\gamma}_n) + o_p(1).$$

**Proof of Lemma 4**

A Taylor’s expansion of $Q_n(\hat{\beta}_{n,A_n}) - Q_n(B_n^\top \hat{\gamma}_n)$ at the point $\hat{\beta}_{n,A_n}$ yields

$$Q_n(\hat{\beta}_{n,A_n}) - Q_n(B_n^\top \hat{\gamma}_n) = T_1 + T_2 + T_3 + T_4,$$

where

$$T_1 = \nabla^\top Q_n(\hat{\beta}_{n,A_n}) (\hat{\beta}_{n,A_n} - B_n^\top \hat{\gamma}_n),$$

$$T_2 = -\frac{1}{2} (\hat{\beta}_{n,A_n} - B_n^\top \hat{\gamma}_n)^\top \nabla^2 L_n(\hat{\beta}_{n,A_n}) (\hat{\beta}_{n,A_n} - B_n^\top \hat{\gamma}_n),$$

$$T_3 = \frac{1}{6} \nabla^\top \{ (\hat{\beta}_{n,A_n} - B_n^\top \hat{\gamma}_n)^\top \nabla^2 L_n(\beta_{n,A_n}^*) (\hat{\beta}_{n,A_n} - B_n^\top \hat{\gamma}_n) \} (\hat{\beta}_{n,A_n} - B_n^\top \hat{\gamma}_n),$$

$$T_4 = \frac{1}{2} (\hat{\beta}_{n,A_n} - B_n^\top \hat{\gamma}_n)^\top \nabla^2 I_n(\beta_{n,A_n}^*) (\hat{\beta}_{n,A_n} - B_n^\top \hat{\gamma}_n).$$

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

Let $\Theta_n = I_n(\beta_{n,A_n}^0)$ and $\Phi_n = \nabla L_n(\beta_{n,A_n}^0)$. By Lemma 2 we have

$$(\hat{\beta}_{n,A_n} - B_n^\top \hat{\gamma}_n)$$

$$= \Theta_n^{-1/2} (I_n - \Theta_n^{1/2} B_n^\top (B_n \Theta_n B_n^\top)^{-1} B_n \Theta_n^{1/2}) \Theta_n^{-1/2} \Phi_n$$

$$+ o_p(n^{-1/2}).$$

$I_n - \Theta_n^{1/2} B_n^\top (B_n \Theta_n B_n^\top)^{-1} B_n \Theta_n^{1/2}$ is an idempotent matrix with rank $q$. Hence, by a standard argument and condition (A2),

$$(\hat{\beta}_{n,A_n} - B_n^\top \hat{\gamma}_n) = O_p(\sqrt{\frac{q}{n}}).$$
We have
\[ \left( \frac{1}{n} \nabla^2 J_n(\beta_n, A_n) \right)_{kkj_1} = 0, \quad \text{for } k \neq k_1 \] (34)
and
\[
\left( \frac{1}{n} \nabla^2 J_n(\beta^*_n, A_n) \right)_{kkj_1} \nabla^2 L_n(\hat{\beta}_n, A_n) + I_n(\beta^0_n, A_n) \right)_{kkj_1} \]
\[
= \frac{\lambda_n w_{n,kj} w_{n,kj_1}}{4(w_{n,k1} |\beta^*_k| + \cdots + w_{n,ks} |\beta^*_s|)^{3/2}} (1 + o_p(1)) \]
\[
\leq \frac{\lambda_n \sqrt{a_n}}{4(c_1)^{3/2}} (1 + o_p(1)) \]
\[
= o_p((nP_n)^{-1/2}). \] (35)

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
\[
T_4 \leq n \left\| \frac{1}{n} \nabla^2 J_n(\beta^*_n, A_n) \right\| \left\| \hat{\beta}_n, A_n - B_n^T \hat{\gamma}_n \right\|^2 \]
\[
= nP_n o_p((nP_n)^{-1/2}) O_p(q_n) \]
\[
= o_p(1). \] (36)

Thus,
\[
Q_n(\hat{\beta}_n, A_n) - Q_n(B_n^T \hat{\gamma}_n) = T_2 + o_p(1). \]

It follows from Lemmas 8 and 9 of Fan and Peng (2004) that
\[
\left\| \frac{1}{n} \nabla^2 L_n(\hat{\beta}_n, A_n) + I_n(\beta^0_n, A_n) \right\| = o_p \left( \frac{1}{\sqrt{P_n}} \right). \]

Hence, we have
\[
\frac{1}{2}(\hat{\beta}_n, A_n - B_n^T \hat{\gamma}_n) \right\{ \nabla^2 L_n(\hat{\beta}_n, A_n) + nI_n(\beta^0_n, A_n) \right\}(\hat{\beta}_n, A_n - B_n^T \hat{\gamma}_n) \]
\[
\leq o_p \left( \frac{n}{\sqrt{P_n}} \right) O_p(q_n) = o_p(1). \] (37)

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 Fan and Peng (2004).