Electronic Journal of Statistics
Vol. 10 (2016) 3775–3806
ISSN: 1935-7524
DOI: 10.1214/16-EJS1217

Model selection of hierarchically structured covariates using elastic net*

Wenqian Qiao

Department of Statistics
Rutgers University
Piscataway, New Jersey 08854
USA
e-mail: joe.wenqian@gmail.com

Heng Lian

School of Mathematics and Statistics
University of New South Wales
Sydney, Australia, 2052
e-mail: heng.lian@unsw.edu.au

and

Min-ge Xie

Department of Statistics
Rutgers University
Piscataway, New Jersey 08854
USA
e-mail: mxie@stat.rutgers.edu

Abstract: Hierarchically associated covariates are common in many fields, and it is often of interest to incorporate their information in statistical inference. This paper proposes a novel way to explicitly integrate the information of a given hierarchical tree of covariates in high-dimensional model selection. Specifically, a set of *hierarchical scores* is introduced to quantify the hierarchical positions of the terminal nodes of the given hierarchical tree, where a terminal node represents either a single covariate or a group of covariates. These scores are then used to weight the corresponding penalty terms in a model selection approach. We show that the proposed estimation approach has a *hierarchical grouping property*, namely, two highly correlated covariates that are close to each other in the hierarchical tree will be more likely included or excluded together in the model than those which are far away. We also prove *model selection consistency* of the proposed estimator both between and within groups. The theoretical results are illustrated by simulation and also a real data analysis on the Systemic Lupus Erythematosus (SLE) dataset.

*This research is partly supported by research grants from NSF (DMS-1107012 and DMS-1513483). It is developed from a chapter of the first author's thesis. The authors wish to thank the editor, the associate editor and two reviewers for their constructive suggestions that helped improve the quality of the paper. They also wish to thank Dr. Jianqing Fan for his discussion and suggestions on the development.

Keywords and phrases: Grouping property, hierarchical elastic net, hierarchical covariate tree, variable selection.

Received September 2015.

1. Introduction

Consider the ordinary linear regression model with n observations and p covariates:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \tag{1.1}$$

where $\mathbf{y}$ is an n-dimensional response vector, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ is an $n \times p$ deterministic design matrix, $\beta = (\beta_1, \ldots, \beta_p)^T$ is the corresponding regression coefficients and ε is the vector of independent random errors. We assume that p is very large and β is sparse, under which setting there is a large body of literature on developing consistent model selection approaches in the past two decades. Examples include Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001; Fan and Lv, 2011), SIS (Sure independence screening) and two-scale method (Fan and Lv, 2008), MCP (Zhang, 2010) and many others. In many cases, there are high correlations among the covariates $\mathbf{x}_i$'s. A number of publications have shown that the correlation should be taken into account to produce a stable and consistent result; see, e.g., Elastic Net (Enet) method (Zou and Hastie, 2005), the OSCAR (octagonal shrinkage and clustering algorithm for regression) approach (Bondell and Reich, 2008), the Mnet method (Huang et al, 2010), the SLS method (Huang et al, 2011), among others. Indeed, structured information in covariates is important in covariate selection. For instance, Yuan and Lin (2006) studied a Group-Lasso method that introduced the concept of group sparsity where individual covariates are grouped for selection. The benefit of grouping is greater regularization and improvements in the prediction power due to the stability in the presence of highly correlated covariates. Similarly, there are publications on the use of hierarchical information in covariate selection. In those studies, the purpose of hierarchical structure is to impose the prioritization of the covariates and their ancestors being selected. Examples include composite absolute penalty (CAP) (Zhao et al, 2009), structured covariate selection with sparsity-inducing norms (Jenatton et al, 2011), structured covariate selection and estimation (Yuan et al, 2009), A LASSO for hierarchical interactions (Bien et al, 2013) and learning with structured sparsity (Huang et al, 2011).

In this work, we focus on a different type of hierarchical structure which is particularly common in hierarchical clustering analysis: instead of representing the priority of the covariates, the hierarchy contains the information on the split (or merge if agglomerative) sequences of clusters (i.e. groups of covariates). To avoid ambiguity, throughout the rest of this paper, "hierarchical structure" refers to this type of information. For instance, in cell biology and genetic studies, genes are often organized into groups based on their biological characteristics or genetic functions, and in many studies the groups are organized in hierarchical layers; see, e.g., (Nei, 1973; Beibbarth and Speed, 2004). See

Subsection 1.1 for additional examples. It is prudent to utilize this type of information in our analysis to seek important gene predictors. However, to the best of our knowledge, the question on how to utilize the hierarchy information in model selection problems is still open, even though hierarchical structures are common in many applications. Specifically, we introduce a novel hierarchical scoring system to quantify the hierarchical positions and, based on it, develop a new high-dimensional model selection approach to incorporate the information of given covariate hierarchical trees. The advantage of our approach is that the hierarchical information can influence the final outcome of the analysis, leading to better scientific and statistical interpretations which are fully consistent with the given hierarchical structure.

Our motivation comes from the peripheral-blood mononuclear cell (PBMC) study reported in Chaussabel et al. (2008). The objective of the study is to eliminate the trivial genes and identify the important genes which can be used to predict the Systemic Lupus Erythematosus disease-activity index (SLEDAI) among 4779 potential candidates with 47 individual samples. According to Chaussabel et al. (2008), those 4779 genes are distributed among 28 modules (groups). The transcripts within each module are highly correlated. On top of these 28 modules, we can also obtain a hierarchical structure with each terminal node representing a single module; see Figure 1. The challenge is how to take advantage of the information presented in the hierarchical structure and use it in a model selection procedure which selects covariates at both group and individual levels.

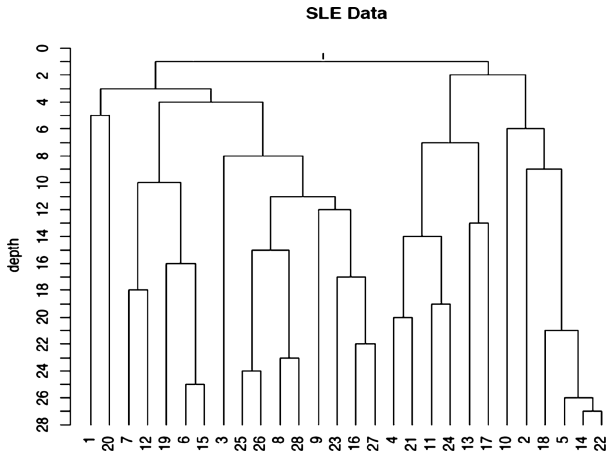

FIG 1. *Hierarchical structure for the SLE dataset. Each of the 28 terminal nodes contains a group (module) of genes.*

We use an Enet-type penalty throughout to illustrate our developments. This simplifies our presentation and keeps our focus on the main goal of incorporating the information contained in a covariate hierarchical tree. An Enet estimator is

$$\hat{\beta} = \operatorname{argmin}_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \beta^T \beta; \tag{1.2}$$

cf. Zou and Hastie (2005). The Enet approach encourages sparsity and grouping simultaneously. It also consistently selects the true model under certain condition; c.f., Jia and Yu (2010). Here and throughout the paper, we use l_1, l_2 and l_∞ norms. In particular, for a vector $\mathbf{a} = (a_1, \ldots, a_p)^T$, the l_1, l_2 and l_∞ norms are denoted by $\|\mathbf{a}\|_1 = \sum_i |a_i|$, $\|\mathbf{a}\|_2 = (\sum_i a_i^2)^{1/2}$ and $\|\mathbf{a}\|_\infty = \max_i |a_i|$, respectively. We also denote $\|\mathbf{A}\|_2 = \sup_{\|\mathbf{a}\|_2=1} \|\mathbf{A}\mathbf{a}\|_2$ and $\|\mathbf{A}\|_\infty = \sup_{\|\mathbf{a}\|_\infty=1} \|\mathbf{A}\mathbf{a}\|_\infty$ for a matrix $\mathbf{A}$.

In order to integrate information of the covariate hierarchical tree, we propose a scoring system to quantify the positions of the terminal nodes in the given hierarchical tree. For ease of presentation and also avoiding other complications, we first start with a simple case in which each terminal node of the hierarchical tree represents only one covariate. In this setting, each covariate will be assigned a score which is derived from the hierarchical structure. The set of scores quantify the hierarchical information among the covariates, and we integrate them into the Enet penalty function. It can be shown that the resulting procedure not only performs model selection and estimation simultaneously, but also enjoys a desired feature, called *hierarchical grouping property*, which can be generally described as follows:

- Two highly correlated covariates which are "close" to each other in the hierarchical tree will more likely be included or dropped together from the model than those that are "far away".

A formal definition of this hierarchical grouping property will be provided in Section 2.1.

In practice, the terminal nodes of many hierarchical trees contain multiple covariates; see, e.g., Breiman et al. (1984) and also our motivating example of SLE study mentioned above. We extend our development for the simply setting in the first part to the more complicated case in which each terminal node can contain potentially multiple covariates (thus a *terminal group* of covariates). Given a hierarchical structure, we assign each terminal group a score and prove that with an appropriate choice of group penalty function, the resulting procedure still retain the *hierarchical grouping property* for the terminal nodes. That is, the procedure will include or exclude together those highly correlated terminal groups (nodes) which are "close" in the hierarchical structure. Furthermore, we prove that the proposed estimator has model selection consistency at both levels, i.e., between-terminal-groups and within-terminal-groups.

The rest of the paper is organized as follows. The remaining of this section (Section 1.1) introduces several terminologies and notations to be used throughout the paper. In Section 2, we define the "hierarchical grouping property" and construct the "hierarchical score" along with the corresponding "hierarchical Elastic Net" estimator for the simplified case in which each terminal node represents only one covariate. We also prove that the proposed estimator has the "hierarchical grouping property" and can provide consistent result in model selection. Section 3 extends the results to the general case in which each terminal node corresponds to a group of covariates. Computational algorithm is presented in Section 4. Numerical studies including simulations and a real data analysis

are carried out in Section 5. Further remarks, including potential extensions of the method to other types of penalty functions, are provided in Section 6. We relegate technical proofs to the Appendix.

1.1. Terminologies and notations

Consider a hierarchical structure represented by an upside-down tree; for example, Figure 2 (a) or (b). The top of the tree is the *root* and we refer to the nodes at the bottom of the upside-down tree as *terminal nodes*. The nodes (splits) in between are the *internal nodes* which, viewed from the bottom to the top, show how individual covariates are grouped and how the groups are merged into supergroups. The concept *depth* is defined for each node (split) as its position in the sequence of splitting when viewed from the top to the bottom. The root node, which represents the first split, has depth 1. Then the node representing the second split has depth 2, etc. See the trees in Figure 2 (a) and (b), in which we have marked the depths of their splits on the vertical coordinate. We assume in this paper that the hierarchical tree structure is known and all internal nodes have distinct depths, i.e. the split or merge is sequential. In real applications, the depth values can be obtained from different ways. For instance, in a biological evolutionary tree, each splitting of lineages of species can be arranged in chronological order. In this case, the depth values can be derived from the time of origination of each new species. Also, classical divisive hierarchical clustering algorithms recursively divide one of the existing clusters into two sub-clusters at each iteration (cf., e.g., Breiman et al. (1984)). In this case, the iteration order can be used to define the depths. How to handle the uncertainty of the tree structure (especially when the tree is generated from a computer algorithm) is not a topic of the current paper; Section 6 contains further discussions.

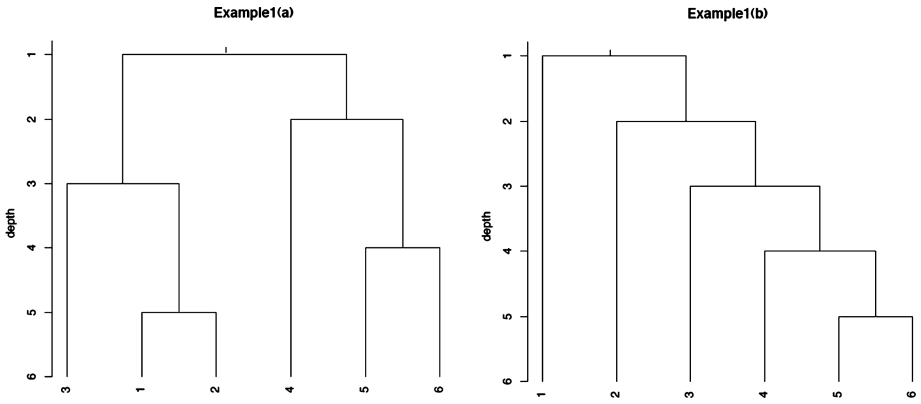

FIG 2. *Hierarchical trees used to simulate data in Example 1 of Section 4. Each of the six terminal nodes in (a) and (b) has only one covariate, i.e.,* $\mathbf{x}_1, ..., \mathbf{x}_6$*, respectively.*

In addition, following Zou and Hastie (2005) and others, we assume that each

covariate has been standardized to have l_2-norm n before data analysis:

$$\sum_{i=1}^{n} x_{ij} = 0 \quad \text{and} \quad \|\mathbf{x}_j\|_2^2 = \sum_{i=1}^{n} x_{ij}^2 = n, \quad \text{for } j = 1, \ldots, p. \tag{1.3}$$

2. Hierarchical covariate selection when a terminal node contains only a single predictor

This section considers the simple case in which each terminal node of the given hierarchical tree contains only one covariate. We design in Subsections 2.1 and 2.2 a set of positive *hierarchical scores* s_j for covariates $\mathbf{x}_j$, for $j = 1, \ldots, p$, to reflect the structure of a given tree. We use these hierarchical scores as a set of weights on the Enet penalty terms (1.2), and propose the following *Hierarchical Enet (HEnet)* estimator:

$$\hat{\beta} = \text{argmin}_\beta \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_1 \|S^{-1}\beta\|_1 + \lambda_2 \beta^T S^{-1} \beta. \tag{2.1}$$

Here, (λ_1, λ_2) are the tuning parameters and $S = \text{diag}\{s_1, \ldots, s_p\}$. In the special case when all $s_j \equiv 1$, (2.1) reduces back to the conventional Enet estimator (1.2). This Hierarchical Enet (HEnet) estimator has several desirable properties.

2.1. Hierarchical grouping property

We define below the notion of *ancestors*, which will be used to determine the *closeness* of a pair of covariates in a given tree.

Definition 2.1 (**Ancestors & closeness**). For a covariate $\mathbf{x}_j$ in a given hierarchical structure, we define *ancestors* of the covariate $\mathbf{x}_j$, denoted by $\mathcal{A}_{\mathbf{x}_j}$, as the set of depths associated with the splits in the hierarchical tree that lead to the terminal node x_i. We also define *ancestors* of a set of covariates $B = \{\mathbf{x}_i, i \in I\}$, denoted by $\mathcal{A}_B$, as the common ancestors of $\mathbf{x}_i$ over $i \in I$, i.e., $\mathcal{A}_B = \cap_{i \in I} \mathcal{A}_{\mathbf{x}_i}$. In addition, we call $\mathbf{x}_j$ is *closer* to $\mathbf{x}_i$ than to $\mathbf{x}_k$ in the hierarchy tree, if $\mathbf{x}_i$ and $\mathbf{x}_j$ share more ancestors than $\mathbf{x}_i$ and $\mathbf{x}_k$, i.e., $\mathcal{A}_{\{\mathbf{x}_i, \mathbf{x}_j\}} \supset \mathcal{A}_{\{\mathbf{x}_i, \mathbf{x}_k\}}$.

For examples, the ancestors of covariate $\mathbf{x}_1$ (corresponding to the terminal node 1) in Figure 2 (a) are the set of nodes (splits) with depths $1, 3, 5$, so $\mathcal{A}_{\mathbf{x}_1} = \{1, 3, 5\}$. The ancestors of covariate $\mathbf{x}_3$ (corresponding to the terminal node 3) are $\mathcal{A}_{\mathbf{x}_3} = \{1, 3\}$. The ancestors of the set of covariates $\{\mathbf{x}_1, \mathbf{x}_3\}$ are $\mathcal{A}_{\{\mathbf{x}_1, \mathbf{x}_3\}} = \{1, 3, 5\} \cap \{1, 3\} = \{1, 3\}$. Similarly, in Figure 2 (b), the ancestors of covariate $\mathbf{x}_1$, $\mathbf{x}_3$ and covariate set $\{\mathbf{x}_1, \mathbf{x}_3\}$ are $\mathcal{A}_{\mathbf{x}_1} = \{1\}$, $\mathcal{A}_{\mathbf{x}_3} = \{1, 2, 3\}$ and $\mathcal{A}_{\{\mathbf{x}_1, \mathbf{x}_3\}} = \{1\} \cap \{1, 2, 3\} = \{1\}$, respectively. Also, in Figure 2 (a), $\mathbf{x}_1$ is *closer* to $\mathbf{x}_3$ than to $\mathbf{x}_4$, because $\mathcal{A}_{\{\mathbf{x}_1, \mathbf{x}_3\}} = \{1, 3\} \supset \mathcal{A}_{\{\mathbf{x}_1, \mathbf{x}_4\}} = \{1\}$. In Figure 2 (b), $\mathbf{x}_5$ is *closer* to $\mathbf{x}_3$ than to $\mathbf{x}_1$, because $\mathcal{A}_{\{\mathbf{x}_3, \mathbf{x}_5\}} = \{1, 2, 3\} \supset \mathcal{A}_{\{\mathbf{x}_1, \mathbf{x}_5\}} = \{1\}$, etc.

Now, we formally state the *hierarchical grouping property* as follows.

Definition 2.2 (**Hierarchical grouping property**). We call predictor covariates *grouped*, if they are selected or dropped together by a model selection procedure. A *hierarchical grouping property* refers to:

P1. For two covariates that share the same ancestors, if they are also highly correlated, then they tend to selected or dropped together by the model selection procedure.
P2. For any three covariates, say $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k$, with $cor(\mathbf{x}_i, \mathbf{x}_j) = cor(\mathbf{x}_i, \mathbf{x}_k) > 0$, if $\mathbf{x}_j$ is closer to $\mathbf{x}_i$ than to $\mathbf{x}_k$ in a hierarchy tree, then $\mathbf{x}_i$ and $\mathbf{x}_j$ tend to be grouped together with a higher chance than that of $\mathbf{x}_i$ and $\mathbf{x}_k$.

The P1 property is similar to the conventional *grouping property* discussed in Zou and Hastie (2005) and Bondell and Reich (2008): highly correlated covariates tend to be selected or dropped together from the estimated model. The P2 property is in compliance with hierarchical structure: a pair of covariates that are closer (i.e., share more ancestors) in the hierarchy will be more likely grouped in the estimated model than those that are farther away (i.e., share fewer ancestors), provided that both pairs have the same correlation. Here, the interpretation of "more likely" is that $|\hat{\beta}_i - \hat{\beta}_j|$ has a smaller upper bound than $|\hat{\beta}_i - \hat{\beta}_k|$'s.

To achieve the goal of hierarchical grouping property, the proposed hierarchical scores (i.e., s_i's) used in our proposed approach (2.1) need to satisfy certain conditions, as discussed in the following:

Recall that the conventional Enet estimator in (1.2) has group properties. In particular, by taking derivatives of the right hand side of (1.2) with respect to β_j and β_k, respectively, we have

$$-2\mathbf{x}_j^T\{\mathbf{y} - \mathbf{X}\hat{\beta}\} + \lambda_1 \mathrm{sgn}\{\hat{\beta}_j\} + 2\lambda_2\hat{\beta}_j = 0$$
$$-2\mathbf{x}_k^T\{\mathbf{y} - \mathbf{X}\hat{\beta}\} + \lambda_1 \mathrm{sgn}\{\hat{\beta}_k\} + 2\lambda_2\hat{\beta}_k = 0$$

assuming $\hat{\beta}_j\hat{\beta}_k \neq 0$. Under the condition $\hat{\beta}_j\hat{\beta}_k > 0$, subtracting above two equations and combining with the fact that $\|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2 \leq \|\mathbf{y}\|_2$, we have

$$|\hat{\beta}_j - \hat{\beta}_k| \leq \frac{\|\mathbf{y}\|_2}{\lambda_2}\|\mathbf{x}_j - \mathbf{x}_k\|_2 = \frac{\|\mathbf{y}\|_2}{\lambda_2}\sqrt{2n(1-\phi_{jk})} \tag{2.2}$$

where $\phi_{jk} = \mathrm{cor}(\mathbf{x}_j, \mathbf{x}_k)$. The inequality (2.2) indicates that as $\phi_{jk} \to 1$, $|\hat{\beta}_j - \hat{\beta}_k| \to 0$; thus, the grouping property in Enet is realized.

Similarly, for the HEnet approach (2.1), by taking derivative on the right hand side, we have after a simple calculation,

$$-2s_j\mathbf{x}_j^T\{\mathbf{y} - \mathbf{X}\hat{\beta}\} + \lambda_1 \mathrm{sgn}\{\hat{\beta}_j\} + 2\lambda_2\hat{\beta}_j = 0.$$

Thus, parallel to (2.2), we have

$$|\hat{\beta}_j - \hat{\beta}_k| \leq \frac{\|\mathbf{y}\|_2}{\lambda_2}\|s_j\mathbf{x}_j - s_k\mathbf{x}_k\|_2 \leq \frac{\|\mathbf{y}\|_2 s^{(p)}}{\lambda_2}\sqrt{2n(1-\varphi_{jk}\phi_{jk})}, \tag{2.3}$$

where $s^{(p)} = \max_{1\le i\le p} s_i$ and

$$\varphi_{jk} = \frac{2s_j s_k}{s_j^2 + s_k^2} = 1 - \frac{(s_j - s_k)^2}{s_j^2 + s_k^2}. \tag{2.4}$$

By definition (2.4), we always have $0 \le \varphi_{jk} \le 1$. When $s^{(p)}$ is upper bounded, the only difference in the upper bound between (2.2) and (2.3) is the scaling term φ_{jk} which, under a careful design of s_i's to be specified below, could lead to desired hierarchical grouping property defined above. The inequality (2.3) is key step of our development, and we will formally state the inequality (2.3) in Theorem 2.2 in Subsection 2.3.

Based on (2.3) and (2.4), the hierarchical grouping property P1 and P2 can be related to the following two (sufficient) conditions on the hierarchical scores s_i, respectively:

C1. For any given pair of predictors $\mathbf{x}_j$ and $\mathbf{x}_k$, $s_j = s_k$ if and only if they have exactly the same ancestors.

C2. For any given predictors $\mathbf{x}_i$, $\mathbf{x}_j$ and $\mathbf{x}_k$, if the ancestors of $\mathbf{x}_i$ and $\mathbf{x}_k$ are a subset of the ancestors of $\mathbf{x}_i$ and $\mathbf{x}_j$, then $\min(s_i/s_j, s_j/s_i) > \min(s_i/s_k, s_k/s_i)$.

Formally, we have the following lemma. A proof is provided in the Appendix.

Lemma 2.1. *Under the setting described above, we have (i) Condition C1 ensures hierarchical grouping property P1; and (ii) Condition C2 ensures hierarchical grouping property P2.*

The remaining question is how to construct a set of hierarchical scores s_i's that satisfies conditions C1 and C2. We provide a method to construct such scores next in Section 2.2.

2.2. Construction of hierarchical scores

For a given terminal node with covariate $\mathbf{x}_i$, we define a binary vector $\mathbf{v}_i \in \mathbb{R}^{p-1}$ such that, for $l = 1, \ldots, p-1$, the lth element of $\mathbf{v}_i$ is:

$$\mathbf{v}_i(l) = \begin{cases} 1 & \text{if } l \in \mathcal{A}_{\mathbf{x}_i}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

Here, $\mathcal{A}_{\mathbf{x}_i}$ is the set of ancestors of $\mathbf{x}_i$ defined in Section 2.1. For example, corresponding to the terminal node $\mathbf{x}_1$ in Figure 2 (a), $\mathcal{A}_{\mathbf{x}_1} = \{1, 3, 5\}$. Thus, by (2.5), the binary vector $\mathbf{v}_1 = (1, 0, 1, 0, 1)$. Similarly, corresponding to the predictor $\mathbf{x}_3$ in Figure 2 (a), $\mathcal{A}_{\mathbf{x}_3} = \{1, 3\}$ and, by (2.5), the binary vector $\mathbf{v}_3 = (1, 0, 1, 0, 0)$.

We define the *hierarchical score* s_i for the terminal node of the predictor $\mathbf{x}_i$ as:

$$s_i = \left\{ \left(\frac{1}{\tau}, \frac{1}{\tau^2}, \ldots, \frac{1}{\tau^{p-1}} \right) \mathbf{v}_i \right\}^{\alpha} = \left\{ \sum_{l=1}^{p-1} \tau^{-l} \mathbf{v}_i(l) \right\}^{\alpha}. \tag{2.6}$$

Here, τ and α are positive constants to be further explained. This set of scores s_i are bounded above by $\max_{1\le i\le p} s_i \le (\sum_{l=1}^{p-1}\tau^{-l})^{\alpha} = \{(\tau^{-1}-\tau^{-p})/(1-\tau^{-1})\}^{\alpha}$.

The following theorem states that the scores s_i defined in (2.6) satisfy Conditions C1 and C2 for any $\tau \ge 3$ and $\alpha > 0$. A proof of the theorem is given in the appendix. Note that, the requirement that $\tau \ge 3$ and $\alpha > 0$ also ensures that the mapping from $\mathbf{v}_i$ to s_i is one-to-one.

Theorem 2.1. *Suppose all the internal nodes have different depths in the given tree. If $\tau \ge 3$ and $\alpha > 0$, then the scores s_i defined in (2.6) satisfies Conditions C1 and C2.*

From (2.6), we can show that the absolute score difference of two different predictors (for example, in Figure 2 (a), $|s_1 - s_3| = \tau^{-5\alpha}$) is a decreasing function of τ. So the choice of $\tau = 3$ numerically maximizes the differentiation among the scores of the covariates. Based on this consideration, we fix $\tau = 3$ in all of our numerical studies.

For the parameter $\alpha > 0$, we treat it as a tuning parameter. In particular, the score s_i is decreasing in α and it also has the following properties:

- When $\alpha \to 0$, all scores $s_i \equiv 1$ and thus all scale weights $\varphi_{ij} \equiv 1$. In this case, the hierarchical tree structure is not taken into account in (2.3) and the HEnet estimator (2.1) is just the conventional Enet estimator (1.2).
- When $\alpha \to \infty$, only covariates sharing same ancestors have $\varphi_{ij} = 1$ and otherwise $\varphi_{ij} = 0$. So only the covariates with the same ancestors are considered for grouping and the hierarchical structure is strictly enforced.

Clearly, the tuning parameter α controls the extent to which the hierarchical structure impacts the regression parameter estimates. More details on how we choose α in our developments can be found in Sections 4 and 5.

2.3. Theoretical results

The proposed HEnet estimator in (2.1) can be re-expressed as

$$\hat{\beta} = \text{argmin}_{\beta}\left\{ \|\mathbf{y} - \sum_{j=1}^{p}\beta_j\mathbf{x}_j\|_2^2 + \lambda_1 \sum_{j=1}^{p}\frac{|\beta_j| + \delta\beta_j^2}{s_j} \right\}, \tag{2.7}$$

where $\delta = \lambda_2/\lambda_1$ is a new tuning parameter replacing λ_2. We formally state the result derived in Section 2.1 in the following theorem.

Theorem 2.2 (**Hierarchical grouping property**). *Let $\hat{\beta}$ be the estimator in (2.7). Suppose $\hat{\beta}_i\hat{\beta}_j > 0$, then*

$$|\hat{\beta}_i - \hat{\beta}_j| \le \frac{\|\mathbf{y}\|_2}{\lambda_1\delta}\|s_i\mathbf{x}_i - s_j\mathbf{x}_j\|_2 \le \frac{\sqrt{n}\|\mathbf{y}\|_2 s^{(p)}}{\lambda_1\delta}\sqrt{2(1-\varphi_{ij}\phi_{ij})}$$

where $s^{(p)} = \max_{1\le i\le p} s_i$, $\varphi_{ij} = 2s_is_j/(s_i^2+s_j^2)$ and $\phi_{ij} = cor(\mathbf{x}_i, \mathbf{x}_j)$.

Theorem 2.2 entails the hierarchical grouping property for the proposed HEnet estimator. For instance, suppose $\mathbf{x}_i$ and $\mathbf{x}_j$ are highly correlated and, without loss of generality, we assume $\phi_{ij} \approx 1$ (if $\phi_{ij} \approx -1$ then consider $-\mathbf{x}_j$). Then, Theorem 2.2 indicates that $|\hat{\beta}_i - \hat{\beta}_j| \leq C\sqrt{1-\varphi_{ij}}$ where $C = \sqrt{2n}\|\mathbf{y}\|_2 s^{(p)}/(\lambda_1\delta)$. If $\mathbf{x}_i$ and $\mathbf{x}_j$ are also close (i.e., $\varphi_{ij} \approx 1$) in the hierarchical structure, it follows that $|\hat{\beta}_i - \hat{\beta}_j| \approx 0$. Thus, the upper bound in Theorem 2.2 provides the same quantitative description of the grouping effect as in Zou and Hastie (2005) and Bondell and Reich (2008). In the case when $\varphi_{ij} < 1$, the addition of the φ_{ij} term channels the information of the hierarchical structure into the grouping process.

We show next that the proposed HEnet estimator in (2.1) or (2.7) also has the property of model selection consistency under a sparsity assumption. Without loss of generality, we assume that the first q true parameters $\beta_j^0 \neq 0$ for $j \in \{1,\ldots,q\}$ and the rest $p-q$ true parameters $\beta_j^0 = 0$ for $j \in \{q+1,\ldots,p\}$. Write $\beta_{(1)}^0 = (\beta_1^0,\ldots,\beta_q^0)^T$ and $\beta_{(2)}^0 = (\beta_{q+1}^0,\ldots,\beta_p^0)^T = (0,\ldots,0)^T$. Also, let $\mathbf{X}_{(1)}$ and $\mathbf{X}_{(2)}$ be the first q and last $p-q$ columns of the design matrix $\mathbf{X}$, respectively, and $\Sigma_{ij} = \mathbf{X}_{(i)}^T\mathbf{X}_{(j)}/n$, for $i,j \in \{1,2\}$.

We define below a Hierarchical Elastic Irrepresentable Condition (HEIC):

HEIC. There exists a positive constant $\eta < 1$ such that

$$\left\|S_{(2)}\Sigma_{21}\tilde{\Sigma}_{11}^{-1}S_{(1)}^{-1}\{\mathrm{sgn}(\beta_{(1)}^0) + 2\delta\beta_{(1)}^0\}\right\|_\infty \leq 1-\eta, \tag{2.8}$$

where $S_{(1)} = \mathrm{diag}(s_1,\ldots,s_q)$, $S_{(2)} = \mathrm{diag}(s_{q+1},\ldots,s_p)$ and $\tilde{\Sigma}_{11} = \Sigma_{11} + \frac{\lambda_1\delta}{n}S_{(1)}^{-1}$.

Intuitively, the HEIC condition imposes a regularization constraint on the regression coefficients of $\mathbf{X}_2$ on $\mathbf{X}_1$ with hierarchical scores. The HEIC condition is an extension of the simple Elastic Irrepresentable Condition (EIC) proposed by Jia and Yu (2010):

EIC. There exists a positive constant $\eta < 1$ such that

$$\left\|\Sigma_{21}(\Sigma_{11} + \frac{\lambda_1\delta}{n}I)^{-1}(\mathrm{sgn}(\beta_{(1)}^0) + 2\delta\beta_{(1)}^0)\right\|_\infty \leq 1-\eta. \tag{2.9}$$

Specially, when all the scores $s_i \equiv 1$, we have $S_{(1)} = I_q$, $S_{(2)} = I_{p-q}$, and in this case the HEIC condition is exactly the EIC condition.

Jia and Yu (2010) proved model selection consistency of the conventional Enet estimator by assuming EIC. Similarly, under HEIC, we have the following theorem. A proof of the theorem is given in the appendix.

Theorem 2.3 (Model selection consistency). *Suppose* $\varepsilon \sim N(0,\sigma^2 I)$. *Then, under the HEIC condition, HEnet estimator* $\hat{\beta}$ *satisfies*

$$P(sgn(\hat{\beta}_{(1)}) = sgn(\beta_{(1)}^0), \hat{\beta}_{(2)} = 0) \to 1, \qquad as\ n \to \infty,$$

provided that the tuning parameters λ_1 *and* δ *are chosen such that*

(a) $\frac{1}{\beta_*}\left\{\sqrt{\frac{\sigma^2\log(q)}{nC_{min}}} + \|\tilde{\Sigma}_{11}^{-1}S_{(1)}^{-1}(\frac{\lambda_1\delta}{n}\beta_{(1)}^0 + \frac{\lambda_1}{2n}sgn(\beta_{(1)}^0))\|_\infty\right\} \to 0$, *where* C_{min} *is*

minimum eigenvalue of $\tilde{\Sigma}_{11}$, $\beta_* = \min\{|\beta^0_{(1)}|\}$, *and* $|\beta^0_{(1)}| = (|\beta^0_1|, \ldots, |\beta^0_q|)$
(b) $\sqrt{n \log(p-q)} = o(\lambda_1)$.

On the other hand, if the HEnet estimator $\hat{\beta}$ *is sign consistent (for some* (λ_1, δ)*), then we have that* (2.8) *holds with* $\eta = 0$, *i.e.,*

$$\left\| S_{(2)} \Sigma_{21} \tilde{\Sigma}_{11}^{-1} S_{(1)}^{-1} \{ sgn(\beta^0_{(1)}) + 2\delta \beta^0_{(1)} \} \right\|_\infty \leq 1. \tag{2.10}$$

Since, for any fixed α in (2.6), all the s_i's are bounded, the incorporation of the hierarchical scores in the Enet-type of penalty does not introduce additional complexity in the theoretical development compared to the original problem considered in Jia and Yu (2010), except that we need to include extra terms involving hierarchical scores.

3. Hierarchical covariate selection when a terminal node contains multiple predictors

When p is large, building a large hierarchical tree with each terminal node representing only one covariate typically causes overfitting problems. Many researchers have suggested to prune large hierarchical trees to prevent the overfitting problems yet still capture important structures; c.f., Breiman et al. (1984). The terminal nodes on a pruned tree often contain multiple covariates. In the example of the SLE dataset described in Section 1, all the 4779 covariates (genes) are divided into 28 modules (terminal nodes) with each module (node) having more than one gene. Indeed, the terminal nodes in most hierarchical clusters contain more than one member, whether it is by pruning, by its natural structure, or by other means.

The grouping by terminal nodes in a hierarchical tree imposes an additional layer of complication to our model selection problem. Let us call the group of covariates formed by a terminal node as *a terminal group*. We need to consider two levels of model selections: selection of the terminal groups and also selection of the covariates within each terminal group. We define an *important covariate* as any covariate with non-zero true coefficient and we also define an *important terminal group* as any terminal group with at least one important covariate. Our goal of model selection is to identify both the important terminal groups (terminal nodes) and also the important covariates (inside each important terminal node), while taking into account the given hierarchical tree structure.

In this section, we extend the developments in Section 2 to deal with this more general and also more challenging case. Under this context, the scoring scheme proposed in Section 2.2 is applied to the hierarchy of the terminal groups (terminal nodes), and thus every terminal group (terminal node) has an assigned score.

As pointed out by a reviewer, HEnet could be directly applied in this case by duplicating the weights for covariates within each group. This is similar to the situation where one uses lasso for a sparse group lasso problem (Simon et al , 2013), ignoring the grouping structure. Although we have not carefully investigated the effect of using HEnet in this case, we believe ignoring group structure would be inferior, as has been demonstrated in the group lasso literature.

3.1. Group Hierarchical Enet

Consider again the linear model (1.1). Suppose now the p covariates are clustered in a hierarchical tree of K terminal nodes. Let $G_k, k = 1, \ldots, K$ be the K non-overlapping subsets of the indices $(1, \ldots, p)$ of covariates corresponding to the K terminal nodes. Denote by $p_k = |G_k|$ the size of the k^{th} subset G_k, so we have $\sum_{k=1}^{K} p_k = p$. To simplify the notations and also without loss of generality, we assume the $p \times 1$ vector of regression coefficient $\beta = (\beta_1, \cdots, \beta_p)^T$ are arranged according to terminal groups, so that we can rewrite $(\beta_1, \cdots, \beta_p) = (\beta_{11}, \ldots, \beta_{1p_1}, \ldots, \beta_{k1}, \ldots, \beta_{kp_k})$, for $\beta_{kj}, j = 1, \ldots, p_k; k = 1, \ldots, K$. Correspondingly, we re-write $\mathbf{X} = (\mathbf{x}_{11}, \ldots, \mathbf{x}_{1p_1}, \ldots, \mathbf{x}_{k1}, \ldots, \mathbf{x}_{kp_k})$.

Following Wang et al (2009), we introduce a coefficient γ_k for the terminal group G_k by reparameterizing the β's as:

$$\beta_{kj} = \gamma_k \theta_{kj}, \ k = 1, \ldots, K, \ j = 1, \ldots, p_k. \tag{3.1}$$

Here, the parameter $\gamma_k \geq 0$ is a group level coefficient for the terminal group G_k, and the parameters $\theta_{kj}, j = 1, \ldots, p_k$ reflect different coefficients within G_k. When $\gamma_k > 0$, parameters $(\gamma_k^*, \theta_{k1}^*, \ldots, \theta_{kp_k}^*) \stackrel{\text{def}}{=} (c\gamma_k, \theta_{k1}/c, \ldots, \theta_{kp_k}/c)$ and parameters $(\gamma_k, \theta_{k1}, \ldots, \theta_{kp_k})$ are indistinguishable under model (1.1), for any non-zero constant $c \neq 0$. So we have a parameter identifiability problem. To resolve the problem, we impose the following constraint:

$$\sum_{j=1}^{p_k} \theta_{kj}^2 = 1 \qquad \text{for any } k \in \{k : \gamma_k > 0\}. \tag{3.2}$$

An explicit expression of γ_k and θ_{kj} can be derived based on (3.1) and (3.2). Specifically, when there is at least one nonzero β_{kj} in kth group,

$$\gamma_k = \sqrt{\sum_{j=1}^{p_k} \beta_{kj}^2} \quad \text{and} \quad \theta_{kj} = \frac{\beta_{kj}}{\gamma_k} = \frac{\beta_{kj}}{\sqrt{\sum_{j=1}^{p_k} \beta_{kj}^2}}; \tag{3.3}$$

Otherwise, in the case with $\beta_{kj} = 0$ for all $j = 1, \ldots, p_k$, we set $\gamma_k = 0$ and $\theta_{kj} = 0$ for $j = 1, \ldots, p_k$. Without loss of generality, we assume in our theoretical derivation that $\gamma_k > 0$, $k = 1, \ldots, r$, for the first r clusters and $\gamma_k = 0$, $k = r+1, \ldots, K$, for the remaining $K - r$ clusters. Furthermore, we let $q_k = |\{(k, j) : \beta_{kj} \neq 0\}|$ be the number of nonzero coefficients in the kth terminal group and write $q = \sum_{k=1}^{K} q_k$. In another words, q_k is the number of the important covariates in the kth terminal node and q is the total number of the important covariates in the entire model. We assume a sparsity condition that $q \ll n$.

Similar to the HEnet discussed in Section 2, we propose to consider a penalized likelihood estimator

$$\begin{aligned}(\hat{\gamma}, \hat{\theta}) = \text{argmin}_{\gamma,\theta} \|\mathbf{y} - \sum_{k=1}^{K} \sum_{j=1}^{p_k} \gamma_k \theta_{kj} \mathbf{x}_{kj}\|_2^2 \\ + \lambda_1 \left\{ \sum_{k=1}^{K} \frac{\gamma_k + \delta \gamma_k^2}{s_k} \right\} + \lambda_2 \sum_{k=1}^{K} \frac{\gamma_k \|\theta_k\|_1}{s_k},\end{aligned} \tag{3.4}$$

subject to the constraint (3.2). Here, λ_1, λ_2 and δ are the tuning parameters, and $\theta_k = (\theta_{k1}, \ldots, \theta_{kp_k})^T$. In (3.4), the first penalty term $\{\sum_{k=1}^K \frac{\gamma_k + \delta\gamma_k^2}{s_k}\}$ is an Enet-type of penalty but on the terminal-group-level (terminal-node-level) parameter γ_k, which encourages a sparse group selection. The second penalty term $\sum_{k=1}^K \frac{\gamma_k \|\theta_k\|_1}{s_k}$ is a LASSO-type penalty on coefficients θ_k, which encourages a selection of important covariates within each terminal group (terminal node).

We refer the estimator obtained from (3.4) as the *Group Hierarchical Enet (GHEnet)* estimator. This GHEnet estimator is an extension of HEnet estimator defined in (2.7). Specifically, if there is only one predictor within each group, i.e. $p_k = 1$ for all $k = 1, \ldots, K$, we go back to the simplified case discussed in Section 2. Note that, in this special case, the reparameterization (3.1) becomes

$$\beta_{k1} = \gamma_k \text{sign}(\beta_{k1}) \text{ with } \gamma_k = |\beta_{k1}| \text{ and } \theta_{k1} = \text{sign}(\beta_{k1}) \text{ for } k = 1, \ldots, K.$$

Plugging the above into (3.4), it leads us back to (2.7).

3.2. Theoretical results

3.2.1. Hierarchical grouping property

Denote by $\tilde{\mathbf{x}}_k = \sum_{j=1}^{p_k} \hat{\theta}_{kj} \mathbf{x}_{kj}$, for $k = 1, \ldots, K$, where $\hat{\theta}_{kl}$ are the elements of $\hat{\theta}$. The definition (3.4) and the constraint (3.2) imply that $\|\tilde{\mathbf{x}}_k\|_2^2 = n$ for $\{k : \hat{\gamma}_i > 0\}$. We may view $\tilde{\mathbf{x}}_k$ as an "overall predictor vector" that represents terminal group G_k. If we define a correlation of the terminal groups (terminal nodes) G_k and $G_{k'}$ as $\phi_{kk'} = \text{cor}(\tilde{\mathbf{x}}_k, \tilde{\mathbf{x}}_{k'}) = \sum_{j=1}^{p_k} \sum_{j'=1}^{p_{k'}} \hat{\theta}_{kj} \hat{\theta}_{k'j'} \text{cor}(\mathbf{x}_{kj}, \mathbf{x}_{k'j'})$, a weighted average of the correlation coefficients of all paired individual predictors between nodes G_k and $G_{k'}$. By treating $\tilde{\mathbf{x}}_k$'s as $\mathbf{x}_i$'s in (2.7), we provide below in Theorem 3.1 a hierarchical grouping property for the terminal groups (terminal nodes). A proof of Theorem 3.1 is given in the appendix.

Theorem 3.1. *(i) Let $(\hat{\theta}, \hat{\gamma})$ be the estimator in (3.4). Suppose $\hat{\gamma}_k \hat{\gamma}_{k'} > 0$, then*

$$\begin{aligned} |\hat{\gamma}_k - \hat{\gamma}_{k'}| &\leq \frac{\|\mathbf{y}\|_2}{\lambda_1 \delta} \|s_k \tilde{\mathbf{x}}_k - s_{k'} \tilde{\mathbf{x}}_{k'}\|_2 + \frac{\lambda_2}{2\lambda_1 \delta} \left| \|\hat{\theta}_k\|_1 - \|\hat{\theta}_{k'}\|_1 \right| \\ &\leq \frac{\|\mathbf{y}\|_2}{\lambda_1 \delta} \|s_k \tilde{\mathbf{x}}_k - s_{k'} \tilde{\mathbf{x}}_{k'}\|_2 + \frac{\lambda_2 \max_{1 \leq k \leq K} \sqrt{p_k}}{2\lambda_1 \delta}. \end{aligned}$$

Furthermore, if $\lambda_2 \max_{1 \leq k \leq K} \sqrt{p_k} = o(n)$, we have

$$|\hat{\gamma}_k - \hat{\gamma}_{k'}| \leq \frac{\sqrt{n} \|\mathbf{y}\|_2}{\lambda_1 \delta} \left(\frac{\|s_k \tilde{\mathbf{x}}_k - s_{k'} \tilde{\mathbf{x}}_{k'}\|_2}{\sqrt{n}} + o_p(1) \right).$$

(ii) Suppose further that the predictors are orthogonal within each group such that $\mathbf{x}_{kj}^T \mathbf{x}_{kj'} = 0$ for $j \neq j'$. Then, we can bound the term $\|s_k \tilde{\mathbf{x}}_k - s_{k'} \tilde{\mathbf{x}}_{k'}\|_2 / \sqrt{n}$ by $s^{(K)} \sqrt{2(1 - \varphi_{kk'} \phi_{kk'})}$ and

$$|\hat{\gamma}_k - \hat{\gamma}_{k'}| \leq \frac{\sqrt{n} \|\mathbf{y}\|_2 s^{(K)}}{\lambda_1 \delta} \left(\sqrt{2(1 - \varphi_{kk'} \phi_{kk'})} + o_p(1) \right).$$

Here, $s^{(K)} = \max_{1\le k\le K} s_k$ *and* $\varphi_{kk'} = 2s_k s_{k'}/(s_k^2 + s_{k'}^2)$*, for any* $1 \le k, k' \le K$*.*

Comparing with Theorem 2.2, the additional term in the inequalities of Theorem 3.1 comes from the derivative of the Lasso penalty on individual level coefficients. This additional term for the individual level coefficients is dominated by the original term for grouping of terminal nodes when $\lambda_2 \max_{1\le k\le K} \sqrt{p_k} = o(n)$. Thus, asymptotically, we have a similar statement of the hierarchical grouping property as the the simplified case discussed in Section 2. In particular, if two terminal groups, say k and k', are highly correlated (i.e., $\phi_{kk'} \approx 1$) and they also are close in the hierarchical tree (i.e., $\varphi_{kk'} \approx 1$), then we have $|\hat{\gamma}_k - \hat{\gamma}_{k'}| \approx 0$. Using the terminology in Definition 2.2, the terminal nodes G_k and $G_{k'}$ are *grouped*.

In the simple case with only one covariate for each terminal node, Theorem 3.1 reduces back to Theorem 2.2. So it is a generalization of Theorem 2.2.

Theorem 3.1 only concern about $\hat{\gamma}_k$ and $\hat{\gamma}_{k'}$. We can also compare $\hat{\beta}_k$ with $\hat{\beta}_{k'}$, where $\hat{\beta}_a$ ($a = k$ or k') is the p_a-vector for the coefficients in the terminal group a. A proof of Theorem 3.2 is given in the appendix.

Theorem 3.2. *Let* $\hat{\beta}_k^{(1)}$ *be a* $r \times 1$ *subvector of* $\hat{\beta}_k$ *and* $\hat{\beta}_{k'}^{(1)}$ *be a* $r \times 1$ *subvectors of* $\hat{\beta}_{k'}$*, both of which contain* r *nonzero coefficients, with* $r \le \min\{q_k, q_{k'}\}$*. Let* $\mathbf{X}_k^{(1)}$ *and* $\mathbf{X}_{k'}^{(1)}$ *be the corresponding submatrices of* $\mathbf{X}_k$ *and* $\mathbf{X}_{k'}$ *containing the columns associated with the nonzero coefficients, where* $\mathbf{X}_k$ *and* $\mathbf{X}_{k'}$ *are the submatrices of* $\mathbf{X}$ *containing the predictors in the terminal groups* k *and* k' *respectively. Suppose also* $\operatorname{sgn}(\hat{\beta}_k^{(1)}) = \operatorname{sgn}(\hat{\beta}_{k'}^{(1)})$ *component-wise (otherwise change signs of the predictors). We have*

$$\|\hat{\beta}_k^{(1)} - \hat{\beta}_{k'}^{(1)}\|_2 \le \frac{\|\mathbf{y}\|_2}{\lambda_1 \delta} \|s_k \mathbf{X}_k^{(1)} - s_{k'} \mathbf{X}_{k'}^{(1)}\|_F,$$

where $\|.\|_F$ *for a matrix denotes its Frobenius norm.*

3.2.2. Model selection consistency

To show that the GHEnet approach also has model selection consistency, we express (3.4) in terms of β. By plugging (3.3) into (3.4), we have:

$$P_n(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_1 \sum_{k=1}^{K} \frac{\|\beta_k\|_2 + \delta\|\beta_k\|_2^2}{s_k} + \lambda_2 \sum_{k=1}^{K} \frac{\|\beta_k\|_1}{s_k} \tag{3.5}$$

where $\beta_k = (\beta_{k1}, \ldots, \beta_{kp_k})^T$. Let us denote by the true regression coefficients $\beta_{kj}^0 = \gamma_k^0 \theta_{kj}^0$ for $k = 1, \ldots, K, j = 1, \ldots, p_k$. Also, the true parameter values $\gamma_k^0 > 0$ for $k = 1, \ldots, r$ and $\gamma_k^0 = 0$ for $k = r+1, \ldots, K$. We note that the proposal is reminiscent of sparse group lasso, with the important difference being that we try to take into account hierarchical structure information, besides that has an extra bridge penalty based on elastic net approach.

Define $A = \{(k, j) : \beta_{kj}^0 \ne 0\}$ the index set of all important covariates and define $B_1 = \{(k, j) : \gamma_k^0 > 0, \beta_{kj}^0 = 0\}$ the index set of non-important

covariates within important terminal groups, and $B_2 = \{(k,j) : \gamma_k^0 = 0\}$ the index set of non-important terminal groups. Set $B = B_1 \cup B_2$, which is the index set of all non-important covariates. Furthermore, corresponding to the set A, we denote by a $|A| \times |A|$ diagonal score matrix $S_A = \text{diag}\{s_k \,|\, (k,j) \in A\}$, a $n \times |A|$ design matrix $\mathbf{X}_A = (\mathbf{x}_{kj}, (k,j) \in A)$, and the vector of nonzero coefficients $\beta_A = (\beta_{kj}, (k,j) \in A)^T$. Similarly, corresponding to the set B, we have $S_B = \text{diag}\{s_k \,|\, (k,j) \in B\}$ and $\mathbf{X}_B = (\mathbf{x}_{kj}, (k,j) \in B)$. Finally, we denote $\Sigma_{AA} = \mathbf{X}_A^T \mathbf{X}_A / n$, and $\Sigma_{BA} = \mathbf{X}_B^T \mathbf{X}_A / n$. Similar notations are defined when B is replaced by B_1 or B_2.

We propose below a Generalized Hierarchical Elastic Irrepresentable Condition (GHEIC), an extension of the HEIC discussed in Section 2.3:

GHEIC. There exists a positive constant η, η' with $\eta' < \eta$ such that

$$\|\frac{\lambda_1}{\lambda_2} S_B \Sigma_{BA} \Sigma^{-1} S_A^{-1}\|_\infty < \eta' \tag{3.6}$$

$$\left\| S_B \Sigma_{BA} \Sigma^{-1} S_A^{-1} \left(\frac{2\lambda_1 \delta}{\lambda_2} \beta_A^0 + \text{sgn}(\beta_A^0) \right) \right\|_\infty < (1 - \eta), \tag{3.7}$$

where $\Sigma = \Sigma_{AA} + \frac{\lambda_1 \delta}{n} S_A^{-1}$.

The second condition (3.7) in GHEIC mimics the HEIC condition by imposing a regularization constraint on the hierarchical-score-weighted regression coefficient of $\mathbf{X}_B$ on $\mathbf{X}_A$. The first condition (3.6) in GHEIC is introduced due to the L_2-norm term in the penalty function (3.5). Also, in the GHEIC, we only require Σ to be invertible instead of Σ_{AA}. We further note that (3.6) can be trivially satisfied if we choose λ_2 large enough, although λ_2 should also satisfy condition (a) in Theorem 3.3 below. A simple sufficient condition for both (3.6) and (3.7) is

$$\|S_B \Sigma_{BA} \Sigma^{-1} S_A^{-1}\|_\infty < \left(\frac{\lambda_1}{\lambda_2} + \frac{2\lambda_1 \delta}{\lambda_2} \|\beta_A^0\|_\infty + 1 \right)^{-1}.$$

Theorem 3.3 below establishes the covariate selection consistency of the GHEnet estimator. A proof of the theorem is given in the appendix.

Theorem 3.3 (**Model selection consistency**). *Suppose* $\varepsilon \sim N(0, \sigma^2 I)$. *Assume GHEIC (3.6) and (3.7) hold. Then, the GHEnet estimator* $\hat{\beta}$ *satisfies:*

$$P(sgn(\hat{\beta}_A) = sgn(\beta_A^0), \hat{\beta}_B = 0) \to 1, \qquad as\ n \to \infty,$$

provided that the tuning parameters λ_1, λ_2 *and* δ *are chosen such that*

(a) $\frac{1}{\beta_*} \left\{ \sqrt{\frac{\sigma^2 \log(q)}{n C_{min}}} + (\lambda_1/n) \|\Sigma^{-1} S_A^{-1}\|_\infty + \|\Sigma^{-1} S_A^{-1} (\frac{\lambda_1 \delta}{n} \beta_A^0 + \frac{\lambda_2}{2n} sgn(\beta_A^0))\|_\infty \right\}$ $\to 0$, *where* C_{min} *is minimum eigenvalue of* Σ *and* $\beta_* = \min\{|\beta_A^0|\}$;

(b) $\sqrt{n \log(p-q)} = o(\lambda_2)$.

4. Computational algorithm

In this section, we focus on the computational issues and modify two existing computing algorithms in the literature to obtain the HEnet and GHEnet estimators proposed in (2.7) and (3.4). Algorithm A below modifies the conventional

Enet algorithm by Zou and Hastie (2005) for the Enet estimator to obtain the HEnet estimator. Algorithm B follows Wang et al (2009) and provides an algorithm for the GHEnet estimator. Since the hierarchical scores proposed in (2.6) are bounded and positive, the algorithms can be justified following simple algebra and the existing literature (i.e., Zou and Hastie (2005) and Wang et al (2009)); thus the proofs are omitted here.

We have the following two algorithms for a given data set $(\mathbf{y}, \mathbf{X})$ and known hierarchical scores $(s_1, \ldots, s_p)$.

Algorithm A (HEnet estimation)

Step 1: Define an artificial data set $(\mathbf{y}^*, \mathbf{X}^*)$ by

$$\mathbf{X}^* = \begin{pmatrix} \mathbf{X}S \\ \sqrt{\lambda_1 \delta} S^{1/2} \end{pmatrix}, \quad \mathbf{y}^* = \begin{pmatrix} \mathbf{y} \\ 0 \end{pmatrix},$$

where $S = \mathrm{diag}(s_1, \ldots, s_p)$.
Step 2: Solve the lasso problem for all λ_1,

$$\hat{\beta}^* = \mathrm{argmax}_\beta \|\mathbf{y}^* - \mathbf{X}^* \beta\|_2^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j|.$$

Step 3: Output $\hat{\beta}_j = \hat{\beta}_j^* s_j, j = 1, \ldots, p$.

Algorithm B (GHEnet estimation)

Step 1: Obtain an initial value $\gamma_k^{(0)}$ for each γ_k; for example, $\gamma_k^{(0)} = 1$. Also, set $m = 1$.
Step 2: At the mth iteration,
(a) Let $\tilde{\mathbf{x}}_{kj} = \gamma_k^{(m-1)} \mathbf{x}_{kj}$ and estimate $\theta_{kj}^{(m)}$ by

$$\theta_{kj}^{(m)} = \mathrm{argmin}_\theta \|\mathbf{y} - \sum_{k=1}^{K} \sum_{j=1}^{p_k} \theta_{kj} \tilde{\mathbf{x}}_{kj}\|_2^2 + \lambda_2 \sum_{k=1}^{K} \sum_{j=1}^{p_k} \frac{\gamma_k^{(m-1)} |\theta_{kj}|}{s_k}$$

subject to $\sum_{l=1}^{p_k} \theta_{kj}^2 = 1$ for $\{k : \gamma_k^{(m-1)} > 0\}$, and set $\theta_{kj}^{(m)} = 0$ for $\{k : \gamma_k^{(m-1)} = 0\}$;
(b) Let $\tilde{\mathbf{x}}_k = \sum_{j=1}^{p_k} \theta_{kj}^{(m)} \mathbf{x}_{kj}$ and estimate $\gamma_k^{(m)}$ by

$$\gamma_k^{(m)} = \mathrm{argmin}_\gamma \|\mathbf{y} - \sum_{k=1}^{K} \gamma_k \tilde{\mathbf{x}}_k\|_2^2 + \lambda_1 \{ \sum_{k=1}^{K} \frac{\gamma_k (1 + \|\theta_k^{(m)}\|_1 \lambda_2 / \lambda_1)}{s_k} + \delta \sum_{k=1}^{K} \frac{\gamma_k^2}{s_k} \}.$$

Step 3: Set $m = m + 1$, and repeat Steps 2 and 3 until the algorithm converges.

Note that setting $\theta_{kj}^{(m)} = 0$ for $\{k : \gamma_k^{(m-1)} = 0\}$ in Step 2(a) of algorithm B will not cause any problem to the minimizations because the objective function is independent of θ_{kj} for $\{k : \gamma_k^{(m-1)} = 0\}$. The two minimizations in Step 2 (a)

and (b) of Algorithm B can be solved using a constrained quadratic programming algorithm. In addition, the values of the tuning parameters $\boldsymbol{\lambda} = (\lambda_1, \delta)$ in Algorithm A or $(\lambda_1, \delta, \lambda_2)$ in Algorithm B are critical. The results are certainly sensitive to the choice of these values. In practice, these can be chosen via the AIC criterion $\text{AIC}(\boldsymbol{\lambda}) = \log(\|\mathbf{y} - \mathbf{X}\hat{\beta}(\boldsymbol{\lambda})\|_2^2/n) + 2\|\hat{\beta}(\boldsymbol{\lambda})\|_0/n$, or the BIC criterion $\text{BIC}(\boldsymbol{\lambda}) = \log(\|\mathbf{y} - \mathbf{X}\hat{\beta}(\boldsymbol{\lambda})\|_2^2/n) + \log(n)\|\hat{\beta}(\boldsymbol{\lambda})\|_0/n$. Here, $\|\hat{\beta}\|_0$ is the number of non-zero $\hat{\beta}$'s. In the our numerical studies in the next section, the tuning parameters are chosen using the BIC criterion, which appears to perform satisfactorily in our numerical results.

In the above algorithms, we assume that the hierarchical scores s_i are provided. In our numerical examples, the hierarchical trees are given and we use (2.6) to calculate s_i with $\tau = 3$ plus a given α. As discussed in Section 2.2, the parameter α can be viewed as a tuning control of the "degree" on how much the hierarchical structure is allowed to impact our estimation. It may be selected based on empirical studies. An illustrative example is provided in the following numerical studies section.

5. Numerical studies

5.1. Simulation studies

Two simulation examples are carried out to evaluate the performance of the proposed methods. In Example 1, we consider a small scale dataset to test the HEnet estimator proposed in Section 2, in which each terminal node contains only one covariate. In Example 2, we investigate the performance of GHEnet estimator proposed in Section 3 by considering a relatively large dataset. The second example mimics the motivating Systemic Lupus Erythematosus (SLE) data example, in which each terminal node contains multiple covariates. In both examples, several simulation settings are considered to cover different scenarios and α values. The number of repetitions in each simulation setting is 1000.

5.1.1. Example 1

The data in this example consist of $n = 20$ samples with $p = 6$ covariates. The true parameter $\beta = (0, 0, 0, 3, 3, 3)$ and $\sigma = 1$. Thus the true model is

$$\mathbf{y} = \mathbf{x}_4\beta_4 + \mathbf{x}_5\beta_5 + \mathbf{x}_6\beta_6 + \varepsilon.$$

We consider two scenarios with different hierarchical and covariance structure:

Scenario 1. The covariates are assumed to have a balanced hierarchical structure as in Figure 2 (a). Each row vector of the design matrix $\mathbf{X}$ is generated from the standard normal distribution with covariance matrix $\text{Cov}(\mathbf{x}_i, \mathbf{x}_j) = 0.9^{1\{i \neq j\}}$ for $(i, j) \in \{1, 2, 3\}$ and $(i, j) \in \{4, 5, 6\}$. For any other pair of (i, j), $\text{Cov}(\mathbf{x}_i, \mathbf{x}_j) = 0$.

Scenario 2. The covariates are assumed to have an unbalanced hierarchical structure as Figure 2 (b). Each row vector of the design matrix $\mathbf{X}$ is generated

from the standard normal distribution with covariance matrix $\text{Cov}(\mathbf{x}_i, \mathbf{x}_j) = 0.9^{1\{i \neq j\}}$.

In our proposed approach, we have a tuning parameter α to control how much impact the hierarchy tree will have on the model selection results. When $\alpha = 0$, $s_i \equiv 1$. In this case, the tree has no impact and the HEnet reduces to the conventional Enet. Note that, the scale weight defined in (2.4) $\varphi_{jk} = \frac{2s_j s_k}{s_j^2 + s_k^2}$ is a decreasing function of α. In addition, $\varphi_{jk} = 1$ for $\alpha = 0$ and $\lim_{\alpha \to \infty} \varphi_{jk} = 0$ for any pair of $(\mathbf{x}_j, \mathbf{x}_k)$ with different ancestors. We plot φ_{jk} against α for all pairs (j, k) in Figure 3. For each plot, the upper shade corresponds to the 75% or higher quantiles of all φ_{jk} and the lower shade is the 50% or lower quantiles of all φ_{jk}, at each fixed α value and over the range of $0 \leq \alpha \leq 40$. We use the plots for assist us choosing the tuning α value empirically. For scenario 1, we first pick $\alpha = 15$ (marked with a vertical red line) which represents the largest range of φ_{jk} meanwhile keeping the smallest φ_{jk} away from 0. Then we pick $\alpha = 8$ (marked with the second vertical red line) to test the impact of different ranges of φ_{jk} on the model selection. Similarly, we choose $\alpha \in \{5, 10\}$ for Scenario 2. We also include $\alpha = 0.1$ case in both Scenario to illustrate the idea that almost no hierarchy is contributed to the model selection result for small α.

We benchmark our proposed HEnet estimators, with several α values, against three existing estimators: Enet, Lasso and OSCAR. The numerical results are summarized in Table 1. The notation P_{ij}^S represents the frequency on which $\mathbf{x}_i$ and $\mathbf{x}_j$ being selected together, and the notation P_{ij}^D represents the frequency on which $\mathbf{x}_i$ and $\mathbf{x}_j$ being dropped from the model together. We can see that for Scenario 1, $P_{45}^S, P_{46}^S, P_{56}^S$ are very close under Enet, Lasso, OSCAR and HEnet with $\alpha = 0.1$ because pairwise correlations are the same within important covariates and (almost) no hierarchical structure is considered. When α increases to 8 and 15, the hierarchy impacts on the model selection result. Specifically, in Figure 2 (a) and among the $\binom{3}{2} = 3$ pairs of the three important covariates $\{\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$, the pair $\{\mathbf{x}_5, \mathbf{x}_6\}$ is closer than the other two. Such structure leads to higher value of P_{56}^S than P_{45}^S and P_{46}^S, i.e. $\{\mathbf{x}_5, \mathbf{x}_6\}$ is more likely selected together than other two pairs of important covariates. Similar results are found in $P_{12}^D, P_{13}^D, P_{23}^D$ with $\alpha \in \{8, 15\}$ for the pairs of non-important covariates. In particular, $\{\mathbf{x}_1, \mathbf{x}_2\}$ is more likely dropped together than the other two pairs, because they are closer in the hierarchical structure.

For Scenario 2, we also have the similar results. The $P_{45}^S, P_{46}^S, P_{56}^S$ are very close under Enet, Lasso, OSCAR and HEnet with $\alpha = 0.1$. The pair $\{\mathbf{x}_5, \mathbf{x}_6\}$ are closer than other two pairs of important covariates on the hierarchical tree from Figure 2 (b). Thus P_{56}^S is higher than P_{45}^S and P_{46}^S with $\alpha = 5$ or 10. For the non-important covariates, $\{\mathbf{x}_2, \mathbf{x}_3\}$ is closer than other two pairs thus is more likely dropped together. Overall, the hierarchical grouping property of our proposed estimator is well illustrated by the above results.

We use two measures, sensitivity and specificity, to evaluate the covariate selection performance. Sensitivity is defined as the proportion of the selected covariates that are the important covariates. Specificity is defined as the proportion of the excluded covariates that are unimportant covariates. From Table 1,

TABLE 1
Frequency of group selection, selection specificity and sensitivity under the settings of Example 1

Scenario 1								
	Important Pairs			Non-important Pairs				
Method	P^S_{45}/P^D_{45}	P^S_{46}/P^D_{46}	P^S_{56}/P^D_{56}	P^S_{12}/P^D_{12}	P^S_{13}/P^D_{13}	P^S_{23}/P^D_{23}	Spec.	Sens.
Enet	.741/.006	.737/.014	.724/.009	.071/.719	.067/.717	.061/.709	.823	.866
HEnet($\alpha = .1$)	.739/.007	.735/.014	.726/.008	.071/.721	.067/.718	.061/.712	.824	.867
HEnet($\alpha = 8$)	.672/.015	.685/.011	.793/.002	.051/.829	.050/.815	.048/.809	.882	.849
HEnet($\alpha = 15$)	.654/.007	.667/.011	.828/.001	.047/.843	.046/.827	.045/.820	.892	.860
Lasso	.745/.006	.746/.013	.733/.009	.078/.680	.075/.677	.066/.670	.800	.870
OSCAR	.744/.006	.744/.013	.730/.008	.076/.681	.074/.679	.066/.672	.801	.870
Scenario 2								
	Important Pairs			Non-important Pairs				
Method	P^S_{45}/P^D_{45}	P^S_{46}/P^D_{46}	P^S_{56}/P^D_{56}	P^S_{12}/P^D_{12}	P^S_{13}/P^D_{13}	P^S_{23}/P^D_{23}	Spec.	Sens.
Enet	.595/.039	.615/.040	.613/.034	.124/.398	.118/.396	.133/.392	.640	.784
HEnet($\alpha = .1$)	.612/.034	.629/.038	.626/.033	.096/.410	.097/.409	.134/.398	.665	.794
HEnet($\alpha = 5$)	.625/.015	.650/.020	.718/.012	.007/.713	.008/.708	.057/.725	.872	.837
HEnet($\alpha = 10$)	.661/.015	.688/.021	.730/.011	.005/.740	.005/.748	.073/.768	.890	.826
Lasso	.591/.039	.612/.042	.609/.039	.122/.401	.117/.399	.131/.394	.642	.782
OSCAR	.568/.047	.592/.048	.582/.051	.125/.409	.119/.397	.131/.393	.640	.766

there is no significant advantage of HEnet in Scenario 1 in terms of sensitivity and specificity. HEnet with $\alpha = 5, 10$ outperform others in Scenario 2.

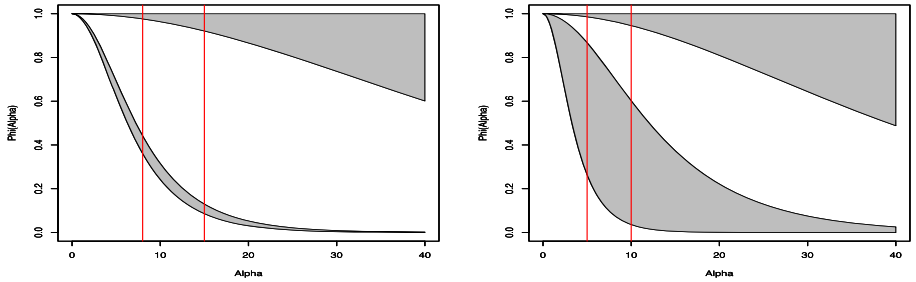

FIG 3. *Plot of the scale weight φ as a function of α. The left panel is for Scenario 1 and right panel is for Scenario 2.*

5.1.2. Example 2

The second example consists of $n = 50$ samples with $p = 4000$ covariates, which is the similar size as the SLE dataset. The covariates are distributed among 20 terminal groups $(G_1, \ldots, G_{20})$ with 200 covariates in each group. The true parameters are $\beta_{1,j} = \beta_{2,j} = \beta_{3,j} = 2$, for $j = 1, \ldots, 10$ and all other $\beta_{k,j} = 0$. Also, $\sigma = 0.25$. Thus the true model is

$$\mathbf{y} = \sum_{i=1}^{10} \mathbf{x}_{1,i}\beta_{1,i} + \sum_{i=1}^{10} \mathbf{x}_{2,i}\beta_{2,i} + \sum_{i=1}^{10} \mathbf{x}_{3,i}\beta_{3,i} + \varepsilon.$$

Each row vector of the design matrix $\mathbf{X}$ is generated from the standard normal distribution with covariance matrix $\mathrm{Cov}(\mathbf{x}_{ki}, \mathbf{x}_{k'j}) = 0.7^{1\{i\neq j\}} \cdot 0.9^{1\{k\neq k'\}}$, $k, k' = 1, \ldots, 20$, $i, j = 1, \ldots, 200$. We assume that a hierarchical structure on top of the 20 terminal groups is as Figure 4.

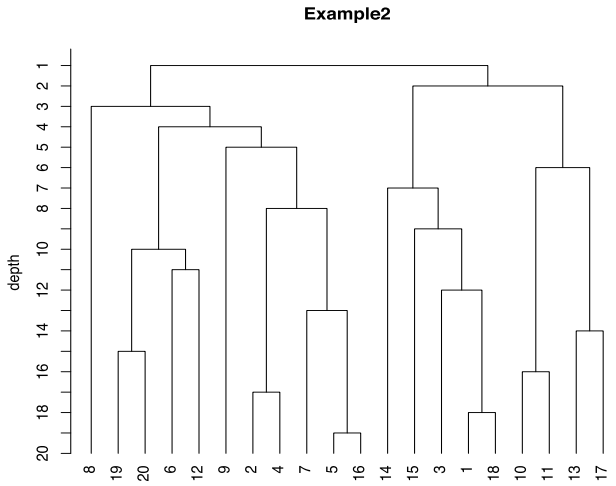

FIG 4. *Hierarchical structure for Example 2.*

We still denote by P_{ij}^S and P_{ij}^D the frequencies on which terminal groups G_i and G_j are selected or dropped together, respectively. Here, we define a terminal group being selected if and only if at least one of the covariates in the group is selected. To demonstrate the hierarchical grouping property of our proposed GHEnet estimator, we calculate P_{ij}^S and P_{ij}^D for the important group pairs $\{i, j\} = \{1, 2\}, \{1, 3\}, \{2, 3\}$ and the pairs among the non-important groups $\{4, \ldots, 20\}$. We benchmark our proposed estimator with $\alpha \in \{0, 0.1, 0.2, 0.3, 0.4, 1, 20\}$ against a Group Lasso and Sparse-Group Lasso (SGL) approach where the covariates are pre-grouped by the same terminal groups (as a set of parallel groups) with the hierarchical information ignored. The choice of α is based on a similar empirical study as Example 1 (the details are omitted to avoid repetitions). We use GEnet to indicate GHEnet with $\alpha = 0$, i.e., no hierarchical structure is considered. Note that GEnet is also similar to SGL with added advantage in dealing with strongly correlated covariates attributed to the ridge penalty.

From Table 2, P_{12}^S, P_{13}^S and P_{23}^S are not materially different under GEnet, as well as, Group Lasso and SGL because between-group correlations are the same within important groups and no hierarchical structure is considered. When α increases, the fact that $\{G_1, G_3\}$ is closer than the other two pairs ($\{G_1, G_2\}$ and $\{G_2, G_3\}$) on Figure 4 leads to higher value of P_{13}^S than those of P_{12}^S and P_{23}^S. This result indicates that $\{G_1, G_3\}$ is more likely selected together than other two pairs of important groups. Similar results can be found in non-important group pairs P_{ij}^S and P_{ij}^D with $\{i, j\} \in \{4, \ldots, 20\}$ (due to space limitation, the

numerical results are not reported in Table 2 but is available by request to the first author). These results demonstrate the proposed GHEnet estimator has indeed the hierarchical grouping property.

We also report in Table 2 the sensitivity and specificity at both terminal-group and individual levels to assess the model selection performance. Both GHEnet and Group Lasso perform well with the terminal group specificity reaching 100%. GHEnet outperforms Group Lasso in covariate selection specificity, since Group Lasso tends to select a larger model and can not perform selections within given groups. GHEnet performs slightly better than SGL in terms of group and covariate specificity. The terminal group sensitivity of GHEnet is better than Group Lasso and SGL for small α values. But for relatively large α, the terminal group sensitivity of GHEnet is relatively small as expected, since the hierarchical tree structure is used in the selection and it impacts the selection process. The covariate selection sensitivity is also relatively low for GHEnet because GHEnet performs model selection at covariate level and the performance deteriorates when the correlation between those covariates is high. Such behavior is consistent with the HEnet example above. On the other hand, Group Lasso and SGL either selects all the covariates or a large number of covariates within selected groups resulting in high covariate sensitivity. In addition and as we can anticipate, incorporating group structure is generally bad for variable sensitivity, and incorporating hierarchical information is generally bad for the overall group sensitivity. GEnet without taking into account the hierarchical structure works well in this example in terms of variable selection due to that group 2 is far away from groups 1 and 3 and thus with large α group 2 tends to be not selected together with the other two groups.

Example 2 used groups with equal sizes for illustration. Based on our observation from a number of cases (including the real data example), the performance under unbalanced group sizes appears to be similar (as long as the sample sizes are not extremely unbalanced). For the interest of space, an example of unbalanced group sizes is omitted in the simulation study.

In summary, the above simulation studies in both examples have demonstrated the hierarchical grouping property and model selection consistency of our proposed estimators.

5.2. Analysis of SLE dataset in PBMC study

In the blood genomic studies reported in Chaussabel et al. (2008), PBMC samples are obtained from $n = 47$ individuals with Systemic Lupus Erythematosus (SLE) condition. Transcriptional profiles were generated with Affymetrix U133A and U133B GeneChips (> 44000 probe sets). The gene intensity signal is assessed and normalized using Microarray Suite, Version 5.0 for each probe set. Then logarithmic transformation is performed on the gene intensity level. Among these 44000+ transcripts, 4779 of them considered "present" are selected as the input of the module-construction algorithm. Total 28 modules (terminal groups) are formed. Within each module, the transcripts are coordinately expressed, i.e. highly correlated and usually have similar functions.

TABLE 2
Frequency of covariates being grouped, selection specificity and sensitivity under the settings of Example 2

Method	Important Pairs			Specificity		Sensitivity	
	P^S_{12}/P^D_{12}	P^S_{13}/P^D_{13}	P^S_{23}/P^D_{23}	Grp	Var	Grp	Var
GEnet	.910/.000	.914/.000	.917/.003	1.000	.992	.956	.562
GHEnet($\alpha = .1$)	.886/.000	.949/.000	.889/.002	1.000	.992	.953	.560
GHEnet($\alpha = .2$)	.796/.002	.960/.001	.801/.000	1.000	.992	.926	.541
GHEnet($\alpha = .3$)	.465/.001	.988/.000	.468/.004	1.000	.991	.819	.446
GHEnet($\alpha = .4$)	.212/.000	.996/.000	.212/.002	1.000	.990	.736	.373
GHEnet($\alpha = 1$)	.000/.001	.998/.000	.000/.001	1.000	.990	.666	.309
GHEnet($\alpha = 20$)	.000/.000	1.000/.000	.000/.000	1.000	.990	.666	.309
Group Lasso	.788/.008	.797/.010	.792/.009	1.000	.807	.892	.892
SGL	.843/.026	.862/.023	.860/.030	0.939	.934	.914	.873

Note: GEnet corresponds to $\alpha = 0$ in (3.4), i.e. no hierarchical information is considered.

A hierarchical clustering algorithm is applied with "complete" linkage on the correlation structure among these 28 modules, and the resulting hierarchy tree structure is shown in Figure 1. On average, each terminal node represents about 170 transcripts. The goal is to identify the modules and the transcripts within these modules that are potentially related to the individual's disease index: SLE disease-activity index (SLEDAI).

Our approach is to use regression analysis under the framework of Section 3. We treat the SLEDAI as the response covariate and all the transcripts as the predictors. Also, the interaction between different modules can be captured by hierarchical clustering. We deploy our proposed GHEnet procedure to perform the gene selection at both module and individual level. Since unlike the simulation studies we don't know which modules or transcripts are truly important in this dataset, we are not able to report P^S_{ij}, P^D_{ij}, sensitivity and specificity for our model selection. To examine the impact of hierarchical scores, we perform a sensitivity analysis by trying different α values.

Summarized in Table 3 Part I are the selected modules and number of genes for $\alpha \in \{0, 0.01, 0.1, 5, 10\}$. Table 3 Part II lists the functionality of these modules. The choice of α is based on an empirical study similar to that produces Figure 3 of Example 1. From Table 3 Part I, we can see how the identified modules evolve as α increases. The module identification results are the same for $\alpha = 0$ and $\alpha = 0.01$. As α increases, the hierarchical structure starts impacting the module selection results as expected: modules #7, #20, #12 are dropped sequentially. The detailed sequential information can help better design confirmatory experiments for medical researchers.

We would like to comment that the example is used to demonstrate how we can take into account the hierarchical structure in a real data analysis and how a hierarchical structure can impact the analysis outcomes. The actual impact and implications depend on whether the hierarchical structure is informative with respect to the underlying true covariate structure. Without any input from experts with domain knowledge, it is difficult to interpret the biological meanings and implications of the analytic results.

TABLE 3
Part I: Sensitivity Analysis: identified Modules under different $\alpha's$

	Modules	# of Genes
$\alpha = 0$	#7, #10, #12, #20	37
$\alpha = 0.01$	#7, #10, #12, #20	37
$\alpha = 0.1$	#10, #12, #20	37
$\alpha = 5$	#10, #12	32
$\alpha = 10$	#10	24

Part II: Functionality of selected modules

Module	Functionality
#7	MHC/Ribosomal proteins. Almost exclusively formed by genes encoding MHC class I molecules (HLA-A,B,C,G,E)+Beta 2-microglobulin (B2M) or ribosomal proteins (RPLs,RPSs).
#10	Neutrophils. This set includes genes encoding innate molecules that are found in neutrophil granules (lactotransferrin: LTF, defensin: DEAF1, bacterial permeability increasing protein: BPI, cathelicidin antimicrobial protein: CAMP.).
#12	Ribosomal proteins. Includes genes encoding ribosomal proteins (RPLs, RPSs), Eukaryotic Translation Elongation Factor-family members (EEFs), and Nucleolar proteins (NPM1, NOAL2, NAP1L1).
#20	Interferon-inducible. This set includes interferon-inducible genes: antiviral molecules (OAS1/2/3/L, GBP1, G1P2, EIF2AK2/PKR, MX1, PML), chemokines (CXCL10/IP-10), signaling molecules (STAT1, STAt2, IRF7, ISGF3G).

6. Concluding remarks

In this paper, we develop a new methodology to incorporate hierarchical structures among the covariates in model selection problems. We construct hierarchical scores from a known hierarchical tree and use them as weights on the penalty function in Elastic net approach. The resulting estimator is proved to have a desirable hierarchical grouping property, and at the same time provides consistent model selection. Although we present our idea through Elastic net penalty, the construction of hierarchical score is independent of the choice of penalty functions. We believe we can combine the hierarchical scores with other types of penalty function which encourages grouping, for example the OSCAR and other procedures.

There are a few questions that are worth further investigation regarding the construction of hierarchical trees, the scoring weights and the development. First, we have found a specific way of constructing hierarchical scores s_i that possess desired properties. It is conceivable that other weights with the same desired properties may exist. But the s_i constructed in Section 2.2 is the only set that we can find so far that is easy to construct and also intuitive. Second, the internal nodes are assumed to have distinct depths throughout the development. Sometimes some of these depths can be tied with equal values. In this case, our scoring method may not directly apply depending on the situations. For example, consider a full binary tree in which the ancestors of all terminal nodes have the same depths, and we would have assigned the same scores to all

terminal nodes by the present method. The equal weights do not reflect that some pairs of terminal nodes are closer than others, which may be viewed undesirable depending on the applications. To overcome the problem, one may alternate and break the ties, i.e., equal depth values, to reflect the hierarchy of closeness before applying our method. Of course, it is also of interest to investigate whether we can find appropriate scoring method without alternating these tie values. Third, we treat α as a tuning parameter because the proposed methodology is not intended to improve the model selection performance but rather to provide a practical way to incorporate the hierarchical information. The parameter α control the impact of hierarchical information on the model selection outcome. User of this approach is recommended to try different level of α to produce sequential information. Such information can help better design, for example in confirmatory experiments for medical researchers. On the other hand, in some situations, it may be desirable to have a data-driven way of choosing α which we leave as future works. Fourth, we assumed the hierarchical tree is given and fixed, while in practice it is typically constructed from data and thus random. How to take into account the uncertainty in the tree structure is an important but challenging problem. Finally, in the theoretical results in this paper, we only focus on consistency of model selection. In some applications, the identification of important covariates is perhaps more important than the estimation of the values of the coefficients. Furthermore, one could also study the convergence rate (or even minimax rate) of the estimators, for which sparse eigenvalue conditions may be needed.

Appendix

A.1. Proof of Lemma 2.1

Proof of Lemma 2.1. (i) When $s_j = s_k$, we have $\varphi_{jk} = 1$ and thus $|\hat{\beta}_j - \hat{\beta}_k| \to 0$ as $\phi_{jk} \to 1$ by (2.3) and (2.4). Therefore, if a pair of predictors have the same ancestors, they are likely grouped together if they are also highly correlated. Thus, Condition C1 ensures P1.

(ii) When we have $\min(s_i/s_j, s_j/s_i) > \min(s_i/s_k, s_k/s_i)$, we have $1 \geq \varphi_{ij} > \varphi_{ik} \geq 0$ by (2.4). It follows that $1 - \varphi_{ij}\phi_{ij} < 1 - \varphi_{ik}\phi_{ik}$ if $\phi_{ij} = \phi_{ik} > 0$. Thus, in this case and by (2.3), $\mathbf{x}_i$ and $\mathbf{x}_j$ are more likely grouped than $\mathbf{x}_i$ and $\mathbf{x}_k$ (in the sense that $|\hat{\beta}_i - \hat{\beta}_j|$ has a smaller upper bound than $|\hat{\beta}_i - \hat{\beta}_k|$'s as in (2.3)). The conclusion in (ii) follows.

A.2. Proof of Theorem 2.1

Proof of Theorem 2.1. Let's first prove $\alpha = 1$ case, where

$$s_i = \sum_{l=1}^{p-1} \mathbf{v}_i(l)\tau^{-l}, \quad \text{for} \quad i = 1, \ldots, p.$$

By definition of the binary vector $\mathbf{v}$, if x_j and x_k have the same ancestors, we have $\mathbf{v}_j = \mathbf{v}_k$. It follows immediately $s_j = s_k$. Conversely, when $\tau \geq 3$, there is a unique one-to-one correspondence between the score s_i and its binary vector representation $\mathbf{v}_i$. Thus, if $s_j = s_k$, we have $\mathbf{v}_j = \mathbf{v}_k$. It follows that the predictors x_j and x_k have the same ancestors. We have proved that Condition C1 holds for the set of scores s_i.

Now, without loss of generality, let us assume that the set of the ancestors of two predictors, say x_i and x_k, is a subset of the ancestor set of predictors x_i and x_j. There exist two integers L_1 and L_2 with $1 \leq L_2 < L_1 \leq p-1$ such that

$$\mathbf{v}_i(l) = \mathbf{v}_j(l) = \mathbf{v}_k(l),\ l = 1, \ldots, L_2 - 1, \quad \mathbf{v}_i(L_2) = \mathbf{v}_j(L_2) = \mathbf{v}_k(L_2) = 1,$$

$$\mathbf{v}_i(l) = \mathbf{v}_j(l),\ l = L_2 + 1, \ldots, L_1 - 1, \quad \mathbf{v}_i(L_1) = \mathbf{v}_j(L_1) = 1,$$

and

$$\text{the set } \{l | L_2 < l \leq L_1 \text{ and } \mathbf{v}_i(l) \neq \mathbf{v}_k(l)\} \text{ is not empty.}$$

Since $\mathbf{v}_i(l) = \mathbf{v}_j(l)$ for $l = 1, \ldots, L_1$, we have following inequality

$$\min\left(\frac{s_i}{s_j}, \frac{s_j}{s_i}\right) \geq \frac{\sum_{l=1}^{L_1} \mathbf{v}_i(l)\tau^{-l}}{\sum_{l=1}^{L_1} \mathbf{v}_i(l)\tau^{-l} + \sum_{l=L_1+1}^{p-1} \tau^{-l}} > \frac{\sum_{l=1}^{L_1} \mathbf{v}_i(l)\tau^{-l}}{\sum_{l=1}^{L_1} \mathbf{v}_i(l)\tau^{-l} + \frac{1}{\tau-1}\tau^{-L_1}}.$$

Also, denote by $L_* = \inf\{L_2 < l \leq L_1 | \mathbf{v}_i(l) \neq \mathbf{v}_k(l)\}$. The binary vectors $\mathbf{v}_i(l) = \mathbf{v}_k(l)$ for $l = 1, \ldots, L_* - 1$. We can show that, when $\tau \geq 3$, we have

$$\begin{aligned}
&\min\left(\frac{s_i}{s_k}, \frac{s_k}{s_i}\right) \\
&\quad \leq \frac{\sum_{l=1}^{L_*-1} \mathbf{v}_i(l)\tau^{-l} + \sum_{l=L_*+1}^{p-1} \tau^{-l}}{\sum_{l=1}^{L_*-1} \mathbf{v}_i(l)\tau^{-l} + \tau^{-L_*}} \leq \frac{\sum_{l=1}^{L_1-1} \mathbf{v}_i(l)\tau^{-l} + \sum_{l=L_1+1}^{p-1} \tau^{-l}}{\sum_{l=1}^{L_1-1} \mathbf{v}_i(l)\tau^{-l} + \tau^{-L_1}} \\
&\quad < \frac{\sum_{l=1}^{L_1-1} \mathbf{v}_i(l)\tau^{-l} + \frac{1}{\tau-1}\tau^{-L_1}}{\sum_{l=1}^{L_1-1} \mathbf{v}_i(l)\tau^{-l} + \tau^{-L_1}} \leq \frac{\sum_{l=1}^{L_1} \mathbf{v}_i(l)\tau^{-l}}{\sum_{l=1}^{L_1} \mathbf{v}_i(l)\tau^{-l} + \frac{1}{\tau-1}\tau^{-L_1}}.
\end{aligned}$$

Thus, we have

$$\min\left(\frac{s_i}{s_j}, \frac{s_j}{s_i}\right) > \min\left(\frac{s_i}{s_k}, \frac{s_k}{s_i}\right).$$

Condition 2 holds for the set of s_i when $\tau \geq 3$.

Finally, for any $\alpha > 0$, the new score s_i is just a simple power transformation of the aforementioned s_i. Because the power function is monotonic and the score function is positive, Conditions C1 and C2 still hold for any new score s_i with $\alpha > 0$.

A.3. Proof of Theorem 2.3

Proof of Theorem 2.3. For a constant $0 < d < 1$, we define a set $E_d^1 = \{\beta | \|\beta_{(1)} - \beta_{(1)}^0\|_\infty \leq (1-d)\beta_*, \beta_{(2)} = 0\}$, we first prove that there exist a $\hat{\beta} \in E_d^1$ that is a

solution of

$$\nabla \left\{ \|\mathbf{y} - \sum_{j=1}^{p} \beta_j \mathbf{x}_j\|_2^2 + \lambda_1 \sum_{j=1}^{p} \frac{|\beta_j| + \delta \beta_j^2}{s_j} \right\} = 0.$$

The above equation can be expressed as the following equations:

$$-2\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\beta}) + \lambda_1 \frac{z_j}{s_j} + 2\lambda_1 \delta \frac{\hat{\beta}_j}{s_j} = 0, \tag{A.1}$$

for any $j = 1, \ldots, p$. Here, $z_j = \text{sgn}(\hat{\beta}_j)$ if $\hat{\beta}_j \neq 0$ and $|z_j| \leq 1$ if $\hat{\beta}_j = 0$. By substituting $\mathbf{y}$ with $\mathbf{X}\beta^0 + \varepsilon$ and limiting to $j = 1, \ldots, q$, (A.1) becomes

$$-2\mathbf{X}_{(1)}^T(\mathbf{X}_{(1)}(\beta_{(1)}^0 - \hat{\beta}_{(1)}) + \varepsilon) + \lambda_1 S_{(1)}^{-1} \text{sgn}(\beta_{(1)}^0) + 2\lambda_1 \delta S_{(1)}^{-1} \hat{\beta}_{(1)} = 0. \tag{A.2}$$

We can further rewrite equation (A.2) as

$$\beta_{(1)}^0 - \hat{\beta}_{(1)} = \tilde{\Sigma}_{11}^{-1} \left(-\frac{\mathbf{X}_{(1)}^T \varepsilon}{n} + \frac{\lambda_1 \delta}{n} S_{(1)}^{-1} \beta_{(1)}^0 + \frac{\lambda_1}{2n} S_{(1)}^{-1} \text{sgn}(\beta_{(1)}^0) \right), \tag{A.3}$$

which is bounded by $(1-d)\beta_*$, if the following inequality holds

$$\|\tilde{\Sigma}_{11}^{-1} \mathbf{X}_{(1)}^T \varepsilon / n\|_\infty + \|\tilde{\Sigma}_{11}^{-1} \left(\frac{\lambda_1 \delta}{n} S_{(1)}^{-1} \beta_{(1)}^0 + \frac{\lambda_1}{2n} S_{(1)}^{-1} \text{sgn}(\beta_{(1)}^0) \right) \|_\infty \leq (1-d)\beta_*. \tag{A.4}$$

Let $U_i = e_i^T \tilde{\Sigma}_{11}^{-1} \mathbf{X}_{(1)}^T \varepsilon / n$, where e_i is the vector with 1 in the ith position and zeroes elsewhere. Then U_i is a normal random covariate with mean 0 and variance

$$\text{Var}(U_i) = \sigma^2 e_i^T \tilde{\Sigma}_{11}^{-1} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \tilde{\Sigma}_{11}^{-1} e_i / n^2 \leq \sigma^2 e_i^T \tilde{\Sigma}_{11}^{-1} e_i / n \leq \frac{\sigma^2}{n C_{\min}}.$$

By standard Gaussian maximum inequality, we have

$$E(\|\tilde{\Sigma}_{11}^{-1} \mathbf{X}_{(1)}^T \varepsilon / n\|_\infty) \leq 3\sqrt{\frac{\sigma^2 \log(q)}{n C_{\min}}}.$$

So by Chebyshev inequality and Condition (a), we have

$$\begin{aligned}
&P\left(\|\tilde{\Sigma}_{11}^{-1} \mathbf{X}_{(1)}^T \varepsilon / n\|_\infty + \|\Sigma_{11}^{-1} \left(\frac{\lambda_1 \delta}{n} S_{(1)}^{-1} \beta_{(1)}^0 + \frac{\lambda_1}{2n} S_{(1)}^{-1} \text{sgn}(\beta_{(1)}^0) \right) \|_\infty > (1-d)\beta_* \right) \\
&\leq \frac{1}{(1-d)\beta_*} E(\|\tilde{\Sigma}_{11}^{-1} \mathbf{X}_{(1)}^T \varepsilon / n\|_\infty + \|\Sigma_{11}^{-1} \left(\frac{\lambda_1 \delta}{n} S_{(1)}^{-1} \beta_{(1)}^0 + \frac{\lambda_1}{2n} S_{(1)}^{-1} \text{sgn}(\beta_{(1)}^0) \right) \|_\infty) \\
&\leq \frac{1}{(1-d)\beta_*} \left\{ 3\sqrt{\frac{\sigma^2 \log(q)}{n C_{\min}}} + \|\Sigma_{11}^{-1} \left(\frac{\lambda_1 \delta}{n} S_{(1)}^{-1} \beta_{(1)}^0 + \frac{\lambda_1}{2n} S_{(1)}^{-1} \text{sgn}(\beta_{(1)}^0) \right) \|_\infty \right\} \\
&\to 0.
\end{aligned}$$

Thus, the inequality (A.4) holds in probability.

Now, consider $j = q+1, \ldots, p$. The equation (A.1) becomes

$$-2S_{(2)}\mathbf{X}_{(2)}^T(\mathbf{y} - \mathbf{X}_{(1)}\hat{\beta}_{(1)}) + \lambda_1 \mathbf{u} = 0, \tag{A.5}$$

where $\mathbf{u} = (u_j)_{j=q+1,\ldots,p}$ with $|u_j| \leq 1$. We can also further rewrite equation (A.5) as

$$\|2S_{(2)}\mathbf{X}_{(2)}^T\mathbf{X}_{(1)}(\beta_{(1)}^0 - \hat{\beta}_{(1)}) + 2S_{(2)}\mathbf{X}_{(2)}^T\varepsilon\|_\infty \leq \lambda_1. \tag{A.6}$$

By plugging (A.3) into (A.6), we have

$$\|2S_{(2)}\mathbf{X}_{(2)}^T\left(I - \frac{\mathbf{X}_{(1)}\tilde{\Sigma}_{11}^{-1}\mathbf{X}_{(1)}^T}{n}\right)\varepsilon + \lambda_1 S_{(2)}\Sigma_{21}\tilde{\Sigma}_{11}^{-1}S_{(1)}^{-1}(\operatorname{sgn}(\beta_{(1)}^0) + 2\delta\beta_{(1)}^0)\|_\infty \leq \lambda_1. \tag{A.7}$$

We first prove sufficiency. By HEIC condition, the second term in the above expression can be bounded by

$$\|\lambda_1 S_{(2)}\Sigma_{21}\tilde{\Sigma}_{11}^{-1}S_{(1)}^{-1}(\operatorname{sgn}(\beta_{(1)}^0) + 2\delta\beta_{(1)}^0)\|_\infty \leq (1-\eta)\lambda_1.$$

Thus, (A.5) is implied by

$$\|\mathbf{X}_{(2)}^T\left(I - \frac{\mathbf{X}_{(1)}\tilde{\Sigma}_{11}^{-1}\mathbf{X}_{(1)}^T}{n}\right)\varepsilon\|_\infty/\sqrt{n} \leq \frac{\eta\lambda_1}{2\sqrt{n}\max_{j=q+1,\ldots,p} s_j}. \tag{A.8}$$

By condition (b), the right hand side of (A.8) is $O(\sqrt{n}\beta_*)$. Following the classical standard Gaussian tail bound and Bonferroni's inequality, we have

$$\begin{aligned} & P\left(\|\mathbf{X}_{(2)}^T\left(I - \frac{\mathbf{X}_{(1)}\tilde{\Sigma}_{11}^{-1}\mathbf{X}_{(1)}^T}{n}\right)\varepsilon\|_\infty > C\lambda_1\right) \\ \leq & \sum_{j=q+1,\ldots,p} P\left(\left|\mathbf{X}_j^T\left(I - \frac{\mathbf{X}_{(1)}\tilde{\Sigma}_{11}^{-1}\mathbf{X}_{(1)}^T}{n}\right)\varepsilon\right| > C\lambda_1\right) \\ \leq & (p-q)\exp\{-C'\lambda_1^2/n\} \to 0. \end{aligned}$$

Thus, inequality (A.8) holds in probability and, therefore, the $\hat{\beta} \in E_d^1$ is a solution to (A.1) for any j. Finally, the objective function in (2.7) is globally convex function in β, thus the estimator $\hat{\beta} \in E_d^1$ is actually the global minimizer of (2.7).

Now we prove necessity by contradiction. If is easy to see that if $\hat{\beta}$ is a sign consistent solution, then (A.3) and (A.6) hold, which again leads to (A.7). If (2.10) fails, then we assume, without loss of generality, that the first element of $\lambda_1 S_{(2)}\Sigma_{21}\tilde{\Sigma}_{11}^{-1}S_{(1)}^{-1}(\operatorname{sgn}(\beta_{(1)}^0) + 2\delta\beta_{(1)}^0)$ is no smaller than λ_1. Since

$$\mathbf{X}_{(2)}^T\left(I - \frac{\mathbf{X}_{(1)}\tilde{\Sigma}_{11}^{-1}\mathbf{X}_{(1)}^T}{n}\right)\frac{\varepsilon}{\sigma}$$

has a Gaussian distribution centered at 0, there is non-vanishing probability that the first element is positive. Therefore, inequality (A.6) does not hold with positive probability. This contradicts with the sign consistency assumption.

This completes the proof.

A.4. Proof of Theorem 3.1

Proof of Theorem 3.1. By taking derivatives on (3.4) with respect to γ_k and $\gamma_{k'}$ respectively and setting them to 0, we have

$$0 = -2s_k\tilde{\mathbf{x}}_k^T(\mathbf{y} - \sum_r \hat{\gamma}_r\tilde{\mathbf{x}}_r) + \lambda_2\|\hat{\theta}_k\|_1 + \lambda_1(1 + 2\delta\hat{\gamma}_k),$$

and

$$0 = -2s_{k'}\tilde{\mathbf{x}}_{k'}^T(\mathbf{y} - \sum_r \hat{\gamma}_r\tilde{\mathbf{x}}_r) + \lambda_2\|\hat{\theta}_{k'}\|_1 + \lambda_1(1 + 2\delta\hat{\gamma}_{k'}).$$

It follows that

$$|\hat{\gamma}_k - \hat{\gamma}_{k'}| \le \frac{1}{\lambda_1\delta}\|\mathbf{y} - \sum_r \hat{\gamma}_r\tilde{\mathbf{x}}_r\|_2 \cdot \|s_k\tilde{\mathbf{x}}_k - s_{k'}\tilde{\mathbf{x}}_{k'}\|_2 + \frac{\lambda_2}{2\lambda_1\delta}\Big|\|\hat{\theta}_k\|_1 - \|\hat{\theta}_{k'}\|_1\Big|. \tag{A.9}$$

Thus, the first inequality in Theorem 3.1 holds. The second inequality follows from $\|\hat{\theta}_k\|_1 \le \sqrt{p_k}\|\hat{\theta}_k\|_2 = \sqrt{p_k}$. The third inequality follows from simple algebra and the fact that $\|\mathbf{y}\|_2 = O(\sqrt{n})$.

A.5. Proof of Theorem 3.2

Proof of Theorem 3.2. Taking derivatives on (3.5) with respect to β_k and $\beta_{k'}$ gives

$$-2s_k\mathbf{X}_k^T(\mathbf{y} - \mathbf{X}\hat{\beta}) + \lambda_1\frac{\hat{\beta}_k}{\|\hat{\beta}_k\|_2} + 2\lambda_1\delta\hat{\beta}_k + \lambda_2\mathrm{sgn}(\hat{\beta}_k) = 0,$$

and

$$-2s_{k'}\mathbf{X}_{k'}^T(\mathbf{y} - \mathbf{X}\hat{\beta}) + \lambda_1\frac{\hat{\beta}_{k'}}{\|\hat{\beta}_{k'}\|_2} + 2\lambda_1\delta\hat{\beta}_{k'} + \lambda_2\mathrm{sgn}(\hat{\beta}_{k'}) = 0.$$

where $\mathrm{sgn}(\hat{\beta}_k) = (\mathrm{sgn}(\hat{\beta}_{k1}), \ldots, \mathrm{sgn}(\hat{\beta}_{kq_k}))^T$. Here, $(\hat{\beta}_{k1}, \ldots, \hat{\beta}_{kq_k})$ is the estimates of $(\beta_{k1}, \ldots, \beta_{kq_k})$, the q_k nonzero elements of β_k.

Consider the two sets of r equations corresponding the subvectors (assuming $r \le \min\{q_k, q_{k'}\}$). It follows from the above two equations that

$$\|(\frac{\lambda_1}{\|\beta_k^{(1)}\|_2} + 2\lambda_1\delta)\beta_k^{(1)} - (\frac{\lambda_1}{\|\beta_{k'}^{(1)}\|_2} + 2\lambda_1\delta)\beta_{k'}^{(1)}\|_2 \le 2\|s_k\mathbf{X}_k^{(1)} - s_{k'}\mathbf{X}_{k'}^{(1)}\|_F\|\mathbf{y}\|_2.$$

Define the mapping $f(\beta) = (\frac{\lambda_1}{\|\beta\|_2} + 2\lambda_1\delta)\beta$ for any vector β. We have $\frac{\partial}{\partial\beta} f(\beta) = \lambda_1 \frac{\|\beta\|_2^2 I - \beta\beta^T}{\|\beta\|_2^3} + 2\lambda_1\delta I$. We also note here that $\|\beta\|_2^2 I - \beta\beta^T$ is nonnegative-definite. Thus, the left hand side of the above displayed inequality is

$$\|f(\beta_k^{(1)}) - f(\beta_{k'}^{(1)})\|_2 = \|\frac{\partial f(\beta^*)}{\partial\beta}(\beta_k^{(1)} - \beta_{k'}^{(1)})\|_2 \geq 2\lambda_1\delta\|\beta_k^{(1)} - \beta_{k'}^{(1)}\|,$$

where in the equality above we used the mean value theorem and β^* lies between $\beta_k^{(1)}$ and $\beta_{k'}^{(1)}$. Therefore,

$$\|\beta_k^{(1)} - \beta_{k'}^{(1)}\|_2 \leq \frac{\|\mathbf{y}\|_2}{\lambda_1\delta}\|s_i\mathbf{X}_k^{(1)} - s_j\mathbf{X}_{k'}^{(1)}\|_F.$$

A.6. Proof of Theorem 3.3

Proof of Theorem 3.3. Similar to Lemma 2 of Obozinski et al (2011), we only need to prove that there exists a triple $(\hat\beta, \hat Z, \hat W)$ with $\hat\beta = (\hat\beta_1^T, \ldots, \hat\beta_K^T)^T$, $\hat Z = (\hat Z_1^T, \ldots, \hat Z_K^T)^T$, $\hat W = (\hat W_1^T, \ldots, \hat W_K^T)^T$, that satisfies the conditions

$$\begin{aligned} 0 &= -2s_k\mathbf{x}_{kj}^T(\mathbf{y} - \mathbf{X}\hat\beta) + \lambda_1\hat Z_{kj} + 2\lambda_1\delta\hat\beta_{kj} + \lambda_2\hat W_{kj} && \text{(A.10)}\\ \operatorname{sgn}(\hat\beta_A) &= \operatorname{sgn}(\beta_A^0) && \text{(A.11)}\\ \hat\beta_B &= 0, && \text{(A.12)} \end{aligned}$$

where

$$\begin{aligned} \hat Z_k &= \hat\beta_k/\|\hat\beta_k\|_2 \text{ if } \|\hat\beta_k\|_2 > 0 && \text{(A.13)}\\ \|\hat Z_k\| &< 1 \text{ if } \|\hat\beta_k\|_2 = 0 && \text{(A.14)}\\ \hat W_{kj} &= \operatorname{sgn}(\hat\beta_{kj}) \text{ if } \hat\beta_{kj} \neq 0 && \text{(A.15)}\\ |\hat W_{kj}| &< 1 \text{ if } \hat\beta_{kj} = 0. && \text{(A.16)} \end{aligned}$$

Next we construct the triple as follows. First, let $\hat\beta$ be the minimizer of (3.5) subject to the constraint $\beta_B = 0$.

By the first order optimality condition for $\hat\beta_A$, we have for $(k, j) \in A$,

$$-2s_k\mathbf{x}_{kj}^T(\varepsilon - \mathbf{X}_A(\hat\beta_A - \beta_A^0)) + \lambda_1\hat Z_{kj} + 2\lambda_1\delta\hat\beta_{kj} + \lambda_2\hat W_{kj} = 0,$$

where $\hat Z_k$ and $\hat W_{kj}$ satisfy (A.13)-(A.16). and it follows that

$$\begin{aligned} & 2n(S_A\Sigma_{AA} + \frac{\lambda_1\delta}{n}I)(\hat\beta_A - \beta_A^0)\\ = \;& 2S_A\mathbf{X}_A^T\varepsilon - \lambda_1\tilde Z - 2\lambda_1\delta\beta_A^0 - \lambda_2\hat W_A, \end{aligned}$$

where $\tilde Z = (\tilde Z_1^T, \ldots, \tilde Z_r^T)^T$ and $\tilde Z_k$ is the subvector of $\hat Z_k$ associated with nonzero $\hat\beta_{kj}$. Equivalently, we have

$$\hat\beta_A - \beta_A^0 = (S_A\Sigma_{AA} + \frac{\lambda_1\delta}{n}I)^{-1}\,.$$

$$\left(\frac{S_A\mathbf{X}_A^T\varepsilon}{n}-\frac{\lambda_1}{2n}\tilde{Z}-\frac{\lambda_1\delta}{n}\beta_A^0-\frac{\lambda_2}{2n}\hat{W}_A\right). \tag{A.17}$$

Thus we can get (A.11) if

$$\|(S_A\Sigma_{AA}+\frac{\lambda_1\delta}{n}I)^{-1}\left(\frac{S_A\mathbf{X}_A^T\varepsilon}{n}-\frac{\lambda_1}{2n}\tilde{Z}-\frac{\lambda_1\delta}{n}\beta_A^0-\frac{\lambda_2}{2n}\mathrm{sgn}(\beta_A^0)\right)\|_\infty<\beta_*. \tag{A.18}$$

We have

$$\|(S_A\Sigma_{AA}+\frac{\lambda_1\delta}{n}I)^{-1}\frac{S_A\mathbf{X}_A^T\varepsilon}{n}\|_\infty=O_p\left(\sqrt{\frac{\sigma^2\log q}{nC_{\min}}}\right)=o(\beta_*),$$

and using that $\|\hat{Z}_k\|_\infty\le 1$, we have

$$\|(S_A\Sigma_{AA}+\frac{\lambda_1\delta}{n}I)^{-1}\frac{\lambda_1}{2n}\tilde{Z}\|_\infty\le\|(S_A\Sigma_{AA}+\frac{\lambda_1\delta}{n}I)^{-1}\frac{\lambda_1}{2n}\|_\infty=o(\beta_*)$$

by assumption (a). Together with other expressions in assumption (a), (A.18) is proved.

What is left is to show that $\hat{Z}_k$ and $\hat{W}_{kj}$ can be constructed for $(k,j)\in B_1$ and $q_k=0$, respectively, that satisfies (A.10).

When $q_k>0$, $\hat{Z}_k$ is already defined above as $\hat{Z}_k=\hat{\beta}_k/\|\hat{\beta}_k\|_2$ and thus $\hat{Z}_{kj}=0$ for $(k,j)\in B_1$. Thus to demonstrate (A.10) for $(k,j)\in B_1$, we only need to show that

$$\lambda_2|\hat{W}_{kj}|:=|2s_k\mathbf{x}_{kj}^T(\varepsilon-\mathbf{X}(\hat{\beta}-\beta^0))|<\lambda_2.$$

Plugging (A.17) into the above and denoting $\Sigma=\Sigma_{AA}+\frac{\lambda_1\delta}{n}S_A^{-1}$, we need to show that

$$\begin{aligned}&\Big|2s_k\mathbf{x}_{kj}^T(I-\mathbf{X}_A\Sigma^{-1}\mathbf{X}_A^T/n)\varepsilon\\&\quad+2s_k\mathbf{x}_{kj}^T\mathbf{X}_A\Sigma^{-1}S_A^{-1}\left(\frac{\lambda_1}{2n}\tilde{Z}+\frac{\lambda_1\delta}{n}\beta_A^0+\frac{\lambda_2}{2n}\mathrm{sgn}(\beta_A^0)\right)\Big|<\lambda_2.\end{aligned} \tag{A.19}$$

Similar to the proof of (A.8), we have

$$\max_{(k,j)\in B_1}|\mathbf{x}_{kj}^T(I-\mathbf{X}_A\Sigma^{-1}\mathbf{X}_A^T/n)\varepsilon|=O_p(\sqrt{n\log(p-q)})=o(\lambda_2).$$

Furthermore,

$$\max_{(k,j)\in B_1}|s_k\mathbf{x}_{kj}^T\mathbf{X}_A\Sigma^{-1}S_A^{-1}\frac{\lambda_1}{n}\tilde{Z}|\le\lambda_1\|S_{B_1}\Sigma_{B_1A}\Sigma^{-1}S_A^{-1}\|_\infty<\eta'\lambda_2,$$

and

$$\max_{(k,j)\in B}\left|s_k\mathbf{x}_{kj}^T\mathbf{X}_A\Sigma^{-1}S_A^{-1}\left(\frac{2\lambda_1\delta}{n}\beta_A^0+\frac{\lambda_2}{n}\mathrm{sgn}(\beta_A^0)\right)\right|<(1-\eta)\lambda_2.$$

Combining the above three displayed equations and assumption (b) shows (A.19).

Finally, when $q_k = 0$ $(k > r)$, we need to show that

$$2\mathbf{X}_k^T(I - \mathbf{X}_A\Sigma^{-1}\mathbf{X}_A^T/n)\varepsilon + 2\mathbf{X}_k^T\mathbf{X}_A\Sigma^{-1}S_A^{-1}\left(\frac{\lambda_1}{2n}\tilde{Z} + \frac{\lambda_1\delta}{n}\beta_A^0 + \frac{\lambda_2}{2n}\mathrm{sgn}(\beta_A^0)\right) \tag{A.20}$$

can be represented as $\lambda_1\hat{Z}_k + \lambda_2\hat{W}_k$ with $\|\hat{Z}_k\| < 1$ and $\|\hat{W}_k\|_\infty < 1$. Same as in the proof of (A.19), absolute values of the components of (A.20) are smaller than λ_2 and thus we can take $\hat{Z}_k = 0$ and define $\hat{W}_k$ appropriately.

References

Beissbarth, T. and Speed, T. P. (2004), "GOstat: find statistically overrepresented Gene Ontologies within a group of genes," *Bioinformatics*, **20**, 1464–1465.

Bien. J., Taylor. J. and Tibshirani R., "A LASSO for hierarchical interactions" *The Annals of Statistics*, **41**, 1111–1141. MR3113805

Bondell, H. and Reich, B. (2008), "Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR" *Biometrics*. **64**, 115–123. MR2422825

Breiman, L., Friedman, J., Olshen, R. and Stone, C.(1984), "Classification and Regression Trees," *Wadsworth International Group*. MR0726392

Chaussabel, D., Quinn, C., Shen, J., Patel, P., Glaser, C., Baldwin, N., Stichweh, D., Blankenship, D., Li, L., Munagala, I. et al. (2008), "A modular analysis framework for blood genomic studies: application to systemic lupus erythematosus," *Immunity*, **29**, 150–164.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R.(2004), "Least angle regression," *Annals of Statistics*, **32**, 407–499. MR2060166

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

Fan, J. and Lv, J. (2008), "Sure independence screening for ultra-high dimensional feature space," *Journal of the Royal Statistical Society B*, **70**, 849–911. MR2530322

Fan, J. and Lv, J. (2011), "Non-concave penalized likelihood with NP-Dimensionality," *IEEE Transactions on Information Theory*, **57**, 5467–5484. MR2849368

Frank, I. E. and Friedman, J. H. (1993), "A statistical view of some chemometrics regression tools," *Technometrics*, **35**, 109–148.

Jia, J. and Yu, B. (2010), "On model selection consistency of the elastic net when $p \gg n$," *Statistica Sinica*, **20**, 595–611. MR2682632

Jenatton, R., Audibert, J. and Bach, F. (2011), "Structured variable selection with sparsity-inducing norms," *Journal of Machine Learning Research*, **12**, 2777–2824. MR2854347

Lv, J. and Fan, Y. (2009), "A unified approach to model selection and sparse recovery using regularized least squares," *The Annals of Statistics*, **37**, 3498–3528. MR2549567

Huang, J., Breheny, P., Ma, S. and Zhang, C.-H. (2010). "The Mnet method for variable selection," *Technical report # 402*, Department of Statistics and Actuarial Science, Univeristy of Iowa.

Huang, J., Ma, S., Li, H. and Zhang, C. (2011), "The sparse Laplacian shrinkage estimator for high-dimensional regression," *Annals of Statistics*, **2011**, 2021–2046. MR2893860

Huang, J., Ma, S., Xie, H. and Zhang, C. (2009), "A group bridge approach for variable selection," *Biometrika*, **96**, 339–355. MR2507147

Huang, J., Zhang, T. and Metaxas, D. (2011), "Learning with structured sparsity," *Journal of Machine Learning Research*, **12**, 3371–3412. MR2877603

Nei, M. (1973), "Analysis of gene diversity in subdivided populations," *PNAS*, **70**, 3321–3323.

Obozinski, G., Wainwright, M. and Jordan, M. (2011), "Support union recovery in high-dimensional multivariate regression," *Annals of Statistics*, **39**, 1–47. MR2797839

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2013), "A sparse-group lasso," *Journal of Computational and Graphical Statistics*, **22**, 231–245. MR3173712

Tibshirani, R.(1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B,* **58**, 267–288. MR1379242

Wang, S., Nan, B., Zhou, N. and Zhu, J. (2009), "Hierachically penalized Cox regression with grouped variables," *Biometrika*, **96**, 307–322. MR2507145

Yuan, M. and Lin, Y. (2006), "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B,* **68**, 49–67. MR2212574

Yuan, M., Joseph, V. and Zou. H.(2009), "Structured variable selection and estimation," *The Annals of Applied Statistics,* **3**, 1738–1757. MR2752156

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B,* **67**, 301–320. MR2137327

Zhang, C. (2010), "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics,* **38** 894–942. MR2604701

Zhao, P., Rocha, G. and Yu, B. (2009), "The composite absolute penalties family for grouped and hierarchical variable selection," *The Annals of Statistics,* **37** 3468–3497. MR2549566