Group Regularized Estimation under Structural Hierarchy

Yiyuan She, Zhifeng Wang and He Jiang

Yiyuan She is Associate Professor, Department of Statistics, Florida State University, Tallahassee, FL 32306 (yshe@stat.fsu.edu). This work was supported in part by NSF grants DMS-1352259 and CCF-1617801. Zhifeng Wang is a Ph.D. student in the Department of Statistics from Florida State University (z.wang@stat.fsu.edu). Jiang He received Ph.D. degree in the Department of Statistics from Florida State University in 2015 (jiangsky2005@gmail.com). We would like to thank the editor, the associate editor and the anonymous referees for their careful comments and useful suggestions that significantly improved the quality of the paper.

Abstract

Variable selection for models including interactions between explanatory variables often needs to obey certain hierarchical constraints. Weak or strong structural hierarchy requires that the existence of an interaction term implies at least one or both associated main effects to be present in the model. Lately this problem has attracted a lot of attention, but existing computational algorithms converge slow even with a moderate number of predictors. Moreover, in contrast to the rich literature on ordinary variable selection, there is a lack of statistical theory to show reasonably low error rates of hierarchical variable selection. This work investigates a new class of estimators that make use of multiple group penalties to capture structural parsimony. We show that the proposed estimators enjoy sharp rate oracle inequalities, and give the minimax lower bounds in strong and weak hierarchical variable selection. A general-purpose algorithm is developed with guaranteed convergence and global optimality. Simulations and real data experiments demonstrate the efficiency and efficacy of the proposed approach.

1 Introduction

In statistical applications, it is often noticed that an additive model including main effects only is inadequate. Including some higher-order terms, such as interactions, in particular, are often of great help in prediction and modeling. Sometimes, interactions may be of independent interest; one example is the moderation analysis in behavioral sciences (Cohen et al., 2013). In this paper, we focus on the full quadratic model with all two-term interactions taken into account.

Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p] \in \mathbb{R}^{n\times p}$ be the (raw) predictor matrix and $\boldsymbol{y} \in \mathbb{R}^n$ be the response vector. We assume the following nonlinear additive regression model

$$\boldsymbol{y} = b_0^*\mathbf{1} + \sum_{1\le j\le p} b_j^*\boldsymbol{x}_j + \sum_{1\le j,k\le p} \phi_{jk}^*\boldsymbol{x}_j \odot \boldsymbol{x}_k + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$ and $\odot$ denotes the Hadamard product. The number of predictors is then $p+\binom{p}{2}$, posing a challenge in variable selection even when p is moderate. Moreover, in this scenario, statisticians are often interested in obtaining a model satisfying certain logical relations, such as the *structural hierarchy* discussed in Nelder (1977), McCullagh and Nelder (1989), and Hamada and Wu (1992). Hierarchy is a natural requirement in gene regulatory network studies (Davidson and Erwin, 2006), banded covariance matrix estimation (Bien et al., 2016) and lagged variable selection in time series. Hierarchical variable selection leads to reduced number of variables in measurement, referred to as *practical sparsity* (Bien et al., 2013). For instance, a model consisting of $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_1\boldsymbol{x}_2$ may be more parsimonious to practitioners than a model involving $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3\boldsymbol{x}_4$. In our setting, there are two types of hierarchy (Chipman, 1996; Bien et al., 2013): strong hierarchy (**SH**) and weak hierarchy (**WH**). Let ϕ_{jk} be the coefficient of $\boldsymbol{x}_j \odot \boldsymbol{x}_k$ and $\boldsymbol{x}_k \odot \boldsymbol{x}_j$. SH means that if an interaction term exists in the model, then both of its associated main effects must be present, i.e., $\phi_{jk} \neq 0 \rightarrow b_j \neq 0$ and $b_k \neq 0$, while WH requires that the inclusion of an interaction implies at least one of its associated main effects to be added into the model, i.e., $\phi_{jk} \neq 0 \rightarrow b_j \neq 0$ or $b_k \neq 0$. We will show in Section 3 that WH is relatively easy to realize

compared with SH. SH is invariant to linear transformations of predictors (Peixoto, 1990) and is the primary concern in this work.

It is a nontrivial task to maintain hierarchy in model selection using conventional approaches. LASSO (Tibshirani, 1996) may violate SH and WH as well. We refer to Nelder (1977), Peixoto (1987), Bickel et al. (2010), Wu et al. (2010), and Hao and Zhang (2014) for some well-developed multi-step procedures which, however, might be ad-hoc and greedy. This paper focuses on *regularization-based* approaches. The past works in this direction include SHIM (Choi et al., 2010), VANISH (Radchenko and James, 2010) and HL (Bien et al., 2013). SHIM reparametrizes ϕ_{jk} as $\rho_{jk} b_j b_k$ and enforces sparsity in both $\boldsymbol{b} = [b_j]$ and $\boldsymbol{\rho} = [\rho_{jk}]$. The formulation is motivating, and we could also use $\phi_{jk} = \rho_{jk}(b_j^2 + b_k^2)$ for WH. However, the corresponding optimization problem is nonconvex and the computational algorithm of SHIM is quite slow in large-scale problems. VANISH is one of the main motivations of our work and will be discussed in detail in Section 3. HL is a recent breakthrough in hierarchical variable selection. One of its key ideas is to enforce a magnitude constraint on the coefficients, $\|\boldsymbol{\phi}_j\|_1 \le |b_j|$, to make hierarchy naturally hold. Here, $\boldsymbol{\phi}_j$ is a vector of coefficients of the predictors $\boldsymbol{x}_j \odot \boldsymbol{x}_k$, $1 \le k \le p$. To handle the nonconvex constraint, Bien et al. (2013) rephrased it with the pesudo-positive and pesudo-negative parts b_j^+, b_j^- of b_j but dropped all zero-product constraints $b_j^+ b_j^- = 0, 1 \le j \le p$. The quality of such a convex relaxation seems to have no theoretical justification in the literature. In our experience, HL has excellent performance when the main effects are strong and p is not very large. But it can miss some interaction effects and become computationally prohibitive on large datasets. For example, when $p = 1000$, HL can take days to obtain a 20-point solution path.

In this work, we propose and study group regularized estimation under structural hierarchy (**GRESH**). In theory, we are able to establish non-asymptotic oracle inequalities to show the error rates of the proposed estimators are minimax optimal up to some logarithm factors. We come up with a new recipe to conquer the theoretical difficulties when analyzing overlapping regularization terms in pursuing structural parsimony. Moreover, we develop a computational algorithm which

guarantees the convergences of iterates and function values; it is not only efficient but also simple to implement.

The rest of the paper is organized as follows. Some notation and symbols are introduced in Section 2. Section 3 presents the general framework of GRESH. A fast computational algorithm with theoretical support is given in Section 4. Section 5 builds oracle inequalities for GRESH, and Section 6 shows the minimax optimal rates. In Section 7, simulation studies and real data analysis are conducted to show the prediction accuracy and computational efficiency of the proposed approach. All technical proofs, together with additional simulation studies and data analysis are given in the supplement.

2 Notation

We introduce some convenient notation and symbols to be used throughout the paper. First, for any matrix $\boldsymbol{A} = \left[\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p\right]^\top \in \mathbb{R}^{p\times n}$, define its $(2,1)$-norm, $(2,\infty)$-norm and ℓ_1-norm as $\|\boldsymbol{A}\|_{2,1} = \sum_{i=1}^p \|\boldsymbol{a}_i\|_2$, $\|\boldsymbol{A}\|_{2,\infty} = \max_{1\le i\le p}\|\boldsymbol{a}_i\|_2$ and $\|\boldsymbol{A}\|_1 = \|\operatorname{vec}(\boldsymbol{A})\|_1$, respectively, where vec is the standard vectorization operator. The spectral norm and Frobenius norm of $\boldsymbol{A}$ are denoted by $\|\boldsymbol{A}\|_2$ and $\|\boldsymbol{A}\|_F$, respectively. For any p-dimensional vector $\boldsymbol{a}$ that is divided into K groups with $\boldsymbol{a}_j$ representing the j-th subvector, its $(2,1)$-norm is defined by $\|\boldsymbol{a}\|_{2,1} = \sum_{j=1}^K \|\boldsymbol{a}_j\|_2$.

The following two operators diag and dg are introduced for notational simplicity. For a square matrix $\boldsymbol{A} := [a_{ij}]_{n\times n}$, $\operatorname{diag}(\boldsymbol{A}) := [a_{11}, \ldots, a_{nn}]^\top$, and for a vector $\boldsymbol{a} = [a_1, \ldots, a_n]^\top \in \mathbb{R}^n$, $\operatorname{diag}\{\boldsymbol{a}\}$ is defined as an $n \times n$ diagonal matrix with diagonal entries given by $a_1, \ldots, a_n$. Define $\operatorname{dg}(\boldsymbol{A}) := \operatorname{diag}\{\operatorname{diag}(\boldsymbol{A})\} = \operatorname{diag}\{[a_{11}, \ldots, a_{nn}]^\top\}$. We use $\boldsymbol{A}[\mathcal{I}, \mathcal{J}]$ to denote a submatrix of $\boldsymbol{A}$ with rows and columns indexed by $\mathcal{I}$ and $\mathcal{J}$, respectively.

For any arbitrary $\boldsymbol{b} \in \mathbb{R}^p$, $\boldsymbol{\Phi} \in \mathbb{R}^{p\times p}$, we define

$$\begin{aligned}
\mathcal{J}^{11}(\boldsymbol{b},\boldsymbol{\Phi}) &:= \{j \in [p] : b_j \neq 0,\ \boldsymbol{\phi}_j \neq \mathbf{0}\},\\
\mathcal{J}^{10}(\boldsymbol{b},\boldsymbol{\Phi}) &:= \{j \in [p] : b_j \neq 0,\ \boldsymbol{\phi}_j = \mathbf{0}\},\\
\mathcal{J}^{01}(\boldsymbol{b},\boldsymbol{\Phi}) &:= \{j \in [p] : b_j = 0,\ \boldsymbol{\phi}_j \neq \mathbf{0}\},\\
\mathcal{J}^{00}(\boldsymbol{b},\boldsymbol{\Phi}) &:= \{j \in [p] : b_j = 0,\ \boldsymbol{\phi}_j = \mathbf{0}\},\\
\mathcal{J}_e(\boldsymbol{\Phi}) &:= \{j \in [p^2] : (\mathrm{vec}\,(\boldsymbol{\Phi}))_j \neq 0\},\\
\mathcal{J}_G(\boldsymbol{b},\boldsymbol{\Phi}) &:= \{j \in [p] : b_j^2 + \|\boldsymbol{\phi}_j\|_2^2 \neq 0\},
\end{aligned} \tag{2}$$

where $[p] = \{1, \ldots, p\}$, and the coefficient vector $\boldsymbol{\phi}_j$ denotes the j-th column of $\boldsymbol{\Phi}$. With $|\cdot|$ standing for set cardinality, we define $J^{11}(\boldsymbol{b},\boldsymbol{\Phi}) := |\mathcal{J}^{11}(\boldsymbol{b},\boldsymbol{\Phi})|$, $J^{10}(\boldsymbol{b},\boldsymbol{\Phi}) := |\mathcal{J}^{10}(\boldsymbol{b},\boldsymbol{\Phi})|$, $J^{01}(\boldsymbol{b},\boldsymbol{\Phi}) := |\mathcal{J}^{01}(\boldsymbol{b},\boldsymbol{\Phi})|$, $J^{00}(\boldsymbol{b},\boldsymbol{\Phi}) := |\mathcal{J}^{00}(\boldsymbol{b},\boldsymbol{\Phi})|$, $J_e(\boldsymbol{\Phi}) := |\mathcal{J}_e(\boldsymbol{\Phi})|$ and $J_G(\boldsymbol{b},\boldsymbol{\Phi}) := |\mathcal{J}_G(\boldsymbol{b},\boldsymbol{\Phi})|$. Clearly, $J^{11} + J^{10} + J^{01} + J^{00} = p$, and $J^{11} + J^{10} + J^{01} = J_G$. In addition, under SH, $J_G(\boldsymbol{b},\boldsymbol{\Phi})$ equals the number of nonzero elements of $\boldsymbol{b}$. Given the true signal $(\boldsymbol{b}^*, \boldsymbol{\Phi}^*)$, the following abbreviated symbols are used: $J^{11*} = J^{11}(\boldsymbol{b}^*,\boldsymbol{\Phi}^*)$, $J^{10*} = J^{10}(\boldsymbol{b}^*,\boldsymbol{\Phi}^*)$, $J^{01*} = J^{01}(\boldsymbol{b}^*,\boldsymbol{\Phi}^*)$, $J_e^* = J_e(\boldsymbol{\Phi}^*)$, and $J_G^* = J_G(\boldsymbol{b}^*,\boldsymbol{\Phi}^*)$.

In the paper, we frequently use the concatenated coefficient matrix for convenience

$$\boldsymbol{\Omega} = [\boldsymbol{b}, \boldsymbol{\Phi}^\top]^\top, \tag{3}$$

and its j-th column is denoted by $\boldsymbol{\Omega}_j = [b_j, \boldsymbol{\phi}_j^\top]^\top$. Given $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_b$ and $\boldsymbol{\Omega}_\Phi$ stand for $(\boldsymbol{\Omega}[1,:])^\top$ and $\boldsymbol{\Omega}[2:(p+1),:]$, respectively.

Let $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p] \in \mathbb{R}^{n\times p}$ be the raw predictor matrix. Define

$$\bar{\boldsymbol{X}} = [\boldsymbol{x}_1 \odot \boldsymbol{x}_1, \ldots, \boldsymbol{x}_1 \odot \boldsymbol{x}_p, \ldots, \boldsymbol{x}_p \odot \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \odot \boldsymbol{x}_p] \in \mathbb{R}^{n\times p^2}, \tag{4}$$

$$\check{\boldsymbol{X}} = [\boldsymbol{x}_1, \boldsymbol{x}_1 \odot \boldsymbol{x}_1, \ldots, \boldsymbol{x}_1 \odot \boldsymbol{x}_p, \ldots, \boldsymbol{x}_p, \boldsymbol{x}_p \odot \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \odot \boldsymbol{x}_p] \in \mathbb{R}^{n\times(p^2+p)}. \tag{5}$$

Then $\bar{\boldsymbol{X}}$ consists of all interactions, and $\check{\boldsymbol{X}}$ includes all $p^2 + p$ predictors in the quadratic model. Given any subset $\mathcal{J} \subset [p]$, we abbreviate $\boldsymbol{X}[:,\mathcal{J}]$ as $\boldsymbol{X}_\mathcal{J}$. It is also easy to see that $\mathrm{diag}(\boldsymbol{X}\boldsymbol{\Phi}\boldsymbol{X}^\top) = \bar{\boldsymbol{X}}\,\mathrm{vec}\,(\boldsymbol{\Phi})$ and $\boldsymbol{X}\boldsymbol{b} + \mathrm{diag}(\boldsymbol{X}\boldsymbol{\Phi}\boldsymbol{X}^\top) = \check{\boldsymbol{X}}\,\mathrm{vec}\,(\boldsymbol{\Omega})$.

For any two real numbers a and b, $a \lesssim b$ means that $a \leq b$ holds up to a multiplicative numerical constant. For two equally sized matrices $\boldsymbol{A} = (a_{ij})$ and $\boldsymbol{B} = (b_{ij})$, $\boldsymbol{A} \geq \boldsymbol{B}$ means $a_{ij} \geq b_{ij}$, for $\forall\ i, j$.

3 Group regularized estimation under structural hierarchy

For simplicity, we assume for now that there exists no intercept term in the model. Then (1) can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b}^* + \mathrm{diag}(\boldsymbol{X}\boldsymbol{\Phi}^*\boldsymbol{X}^\top) + \boldsymbol{\varepsilon}, \tag{6}$$

where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ with $\sigma^2 > 0$, $\boldsymbol{y} = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$ is the response vector, and $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p] \in \mathbb{R}^{n\times p}$ is the design matrix consisting of main effects only.

We describe a general framework for hierarchical variable selection, referred to as group regularized estimation under structural hierarchy or **GRESH**. GRESH has two different types, depending on which objects to regularize. Denoting the squared error loss by ℓ, i.e.,

$$\ell(\boldsymbol{b}, \boldsymbol{\Phi}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} - \mathrm{diag}(\boldsymbol{X}\boldsymbol{\Phi}\boldsymbol{X}^\top)\|_2^2, \tag{7}$$

the first type is given by

$$\begin{aligned} \textbf{Type-A}: \min_{\boldsymbol{\Omega}=[\boldsymbol{b},\boldsymbol{\Phi}^\top]^\top \in \mathbb{R}^{(p+1)\times p}} & \ell(\boldsymbol{b}, \boldsymbol{\Phi}) + \lambda_1\|\boldsymbol{\Phi}\|_1 + \lambda_2 \sum_{j=1}^{p} \|[b_j, z(\boldsymbol{\phi}_j)]\|_q \\ \text{s.t. } & \boldsymbol{\Phi} = \boldsymbol{\Phi}^\top \text{ (for } \textbf{SH} \text{ only)}, \end{aligned} \tag{8}$$

where λ_1, λ_2 are regularization parameters, $1 < q \leq +\infty$ and $z(\boldsymbol{x})$ is a function satisfying the property that $z(\boldsymbol{x}) = \boldsymbol{0}$ implies $\boldsymbol{x} = \boldsymbol{0}$ for any vector $\boldsymbol{x} \in \mathbb{R}^p$. For instance, z can take the ℓ_r-norm function ($r > 0$)

$$z(\boldsymbol{x}) = \|\boldsymbol{x}\|_r, \tag{9}$$

or simply the identity function

$$z(\boldsymbol{x}) = \boldsymbol{x}^\top. \tag{10}$$

The first ℓ_1 penalty imposes elementwise sparsity on $\mathbf{\Phi}$ and the second group-ℓ_1 penalty enforces column sparsity in $\mathbf{\Omega}$. We argue that with the two penalties and the constraint, (8) can be used for strong hierarchical variable selection. Indeed, the sparsity of $\boldsymbol{b}$ comes from the second group-penalty alone, i.e., $b_j = 0$ implies $\|[b_j, z(\boldsymbol{\phi}_j)]\|_q = 0$ (with probability 1) or $z(\boldsymbol{\phi}_j) = 0$. By the property of the z function, $\boldsymbol{\phi}_j = \mathbf{0}$, and thus $\phi_{jk} = 0$. The symmetry condition indicates further that $\phi_{kj} = 0$. Hence $(\phi_{jk} + \phi_{kj})/2$, the coefficient for $\boldsymbol{x}_j \odot \boldsymbol{x}_k$, is zero. Consequently, whenever $b_j = 0$, $\boldsymbol{x}_j \odot \boldsymbol{x}_k$ will be removed from the model and SH is automatically obeyed. We can describe the reasoning as follows

$$\begin{aligned} b_j = 0 \Rightarrow \|[b_j, z(\boldsymbol{\phi}_j)]\|_q = 0 \Rightarrow z(\boldsymbol{\phi}_j) = 0 \Rightarrow \boldsymbol{\phi}_j = 0 \Rightarrow \phi_{jk} = 0 \\ \Rightarrow \phi_{kj} = 0 \Rightarrow \frac{\phi_{jk} + \phi_{kj}}{2} = 0. \end{aligned} \tag{11}$$

Without the symmetry constraint, we can only complete the argument in the first line of (11), and so SH dose not hold. But interestingly, WH is guaranteed, because from $b_j = b_k = 0$, we have $(\phi_{jk} + \phi_{kj})/2 = 0$. Therefore, WH gives a relatively simpler problem.

As pointed out by a reviewer, when the model contains an intercept, centering the response and the p raw predictors does not make it vanish due to the presence of nonlinear terms, and so $\mathbf{1}b_0 + \boldsymbol{X}\boldsymbol{b} + \text{diag}(\boldsymbol{X}\mathbf{\Phi}\boldsymbol{X}^\top)$ should be used to approximate $\boldsymbol{y}$. In the SH scenario, if at least one x-predictor is relevant, substituting $\boldsymbol{X}' = [\mathbf{1}, \boldsymbol{X}]$ for $\boldsymbol{X}$ in (7) suffices.

We focus on convex forms of GRESH in this work. But surely the ℓ_1 penalty and the group-ℓ_1 penalty in (8) can be replaced by their nonconvex alternatives; see, e.g., She (2012).

GRESH is related to some methods in the literature. HL makes a special case of (8) because one of its formulations corresponds to $q = \infty$ and $r = 1$, with a single regularization parameter being used. Another instance is given by $q = 2$ and $z(\boldsymbol{x}) = \boldsymbol{x}^\top$:

$$\min_{\mathbf{\Omega}=[\boldsymbol{b},\mathbf{\Phi}^\top]^\top} \ell(\boldsymbol{b}, \mathbf{\Phi}) + \lambda_1\|\mathbf{\Phi}\|_1 + \lambda_2\|\mathbf{\Omega}^\top\|_{2,1} \quad \text{s.t. } \mathbf{\Phi} = \mathbf{\Phi}^\top. \tag{12}$$

Bien et al. (2013) incorrectly described VANISH (Radchenko and James, 2010) in this form, without the symmetry condition. We will focus on (12) in the theoretical and computational studies of

Type-A GRESH.

As a matter of fact, Radchenko and James (2010) defined VANISH in a different way, which motivates another type of GRESH

$$\textbf{Type-B:} \min_{\boldsymbol{\Omega}=[\boldsymbol{b},\boldsymbol{\Phi}^\top]^\top} \ell(\boldsymbol{b},\boldsymbol{\Phi}) + \lambda_1 \sum_{1\le j,k\le p} \|\boldsymbol{\phi}_{jk}\boldsymbol{x}_j \odot \boldsymbol{x}_k\|_2 + \lambda_2 \sum_{j=1}^{p} \|[b_j\boldsymbol{x}_j, z(\boldsymbol{\phi}_{j1}\boldsymbol{x}_j \odot \boldsymbol{x}_1, \ldots, \boldsymbol{\phi}_{jp}\boldsymbol{x}_j \odot \boldsymbol{x}_p)]\|_q, \quad \text{s.t. } \boldsymbol{\Phi} = \boldsymbol{\Phi}^\top (\mathbf{SH} \text{ only}), \tag{13}$$

where λ_1, λ_2, q, and z are defined as in (8). Similarly, we can argue that (13) keeps hierarchy. When $q = 2$ and z takes the form of (10), the penalty part in (13) become

$$\lambda_1 \sum_{j,k} \|\boldsymbol{\phi}_{jk}\boldsymbol{x}_j \odot \boldsymbol{x}_k\|_2 + \lambda_2 \sum_j (\|b_j\boldsymbol{x}_j\|_2^2 + \sum_k \|\boldsymbol{\phi}_{jk}\boldsymbol{x}_j \odot \boldsymbol{x}_k\|_2^2)^{\frac{1}{2}}, \tag{14}$$

as considered by Radchenko and James (2010). VANISH constructs main effects and interactions from two small sets of orthonormal basis functions in a functional regression setting. We do not pose such a restriction on the design matrix, and p can be arbitrarily large.

The key difference between the two types of GRESH is that the penalties are imposed on the coefficients in (8), but on the terms in (13). A common practice before calling a shrinkage method is normalizing/standardizing all predictors, so that it is more reasonable to use a common regularization parameter in penalizing different coefficients. In this way, (8) builds a model on the normalized predictors and their interactions, while Type-B amounts to forming the overall design $\check{\boldsymbol{X}}$ first and then performing the standardization. They are not equivalent because in general, $(\boldsymbol{x}_j/\|\boldsymbol{x}_j\|_2) \odot (\boldsymbol{x}_k/\|\boldsymbol{x}_k\|_2) \neq (\boldsymbol{x}_j \odot \boldsymbol{x}_k)/\|\boldsymbol{x}_j \odot \boldsymbol{x}_k\|_2$. Then, which type of GRESH is preferable? An answer will be given in Section 5.

GRESH offers some general schemes for hierarchical variable selection. But it is no ordinary lasso or group lasso, since $\boldsymbol{\Phi}$ appears in both penalties as well as the symmetry constraint. The main goal of this paper is to tackle some computational and theoretical challenges arising from the overlapping regularization terms in high dimensions. In computation, we would like to develop

fast and scalable algorithms (cf. Section 4); in theory, how to treat the penalties and the constraint *jointly* to derive a sharp error bound for GRESH is intriguing and challenging (cf. Sections 5, 6).

4 Computation

It is perhaps natural to think of using the alternating direction method of multipliers (ADMM, cf. Boyd et al. (2011)) to deal with the computational challenges. ADMM recently gains its popularity among statisticians. In fact, Bien et al. (2016) designed an algorithm of HL based on ADMM, where one of the main ingredients is the augmented Lagrangian

$$\min_{\boldsymbol{b}^{\pm}\in\mathbb{R}^p,\boldsymbol{\Phi}\in\mathbb{R}^p} \ell(\boldsymbol{b}^+ - \boldsymbol{b}^-, \boldsymbol{\Phi}) + \lambda \mathbf{1}^\top(\boldsymbol{b}^+ + \boldsymbol{b}^-) + \lambda\|\boldsymbol{\Phi}\|_1 + \langle \boldsymbol{\Phi} - \boldsymbol{\Psi}, \boldsymbol{L}\rangle + \frac{\rho}{2}\|\boldsymbol{\Phi} - \boldsymbol{\Psi}\|_F^2,$$
$$\text{s.t. } \boldsymbol{\Psi} = \boldsymbol{\Psi}^\top, \|\boldsymbol{\phi}_j\|_1 \le b_j^+ + b_j^-, b_j^+ \ge 0, b_j^- \ge 0,\ 1 \le j \le p. \tag{15}$$

Here, $\boldsymbol{L}$ is a Lagrange multiplier matrix, and $\rho > 0$ is a given constant, sometimes referred as the penalty parameter. Although ADMM enjoys some nice convergence properties in theory, practically only when ρ is large enough can we obtain a solution with good statistical accuracy. But often the larger the value of ρ is, the slower the (primal) convergence is. For example, in the `R` package HierNet (version 1.6) for computing HL, $\rho = n$ is recommended, but for $p = 1000$, the algorithm may take several days to compute a single solution path. There are some empirical schemes on how to vary ρ during the iteration, but they are ad hoc and do not always behave well.

In this section, we consider a slightly more general optimization problem which includes both types of GRESH as particular instances

$$\min_{\boldsymbol{\Omega}=[\boldsymbol{b},\boldsymbol{\Phi}^\top]^\top} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} - \mathrm{diag}(\boldsymbol{Z}\boldsymbol{\Phi}\boldsymbol{Z}^\top)\|_2^2 + \|\boldsymbol{\lambda}_b \odot \boldsymbol{b}\|_1$$
$$+\|\boldsymbol{\Lambda}_\Phi \odot \boldsymbol{\Phi}\|_1 + \|\boldsymbol{\Lambda}_\Omega^\top \odot \boldsymbol{\Omega}^\top\|_{2,1} \quad \text{s.t.} \quad \boldsymbol{\Phi} = \boldsymbol{\Phi}^\top, \tag{16}$$

where $\boldsymbol{X}, \boldsymbol{Z} \in \mathbb{R}^{n\times p}$, and $\boldsymbol{\lambda}_b$, $\boldsymbol{\Lambda}_\Phi$ and $\boldsymbol{\Lambda}_\Omega$ are non-negative regularization vector and matrices. Let

$\bar{\boldsymbol{Z}} = [\boldsymbol{z}_1 \odot \boldsymbol{z}_1, \ldots, \boldsymbol{z}_1 \odot \boldsymbol{z}_p, \ldots, \boldsymbol{z}_p \odot \boldsymbol{z}_1, \ldots, \boldsymbol{z}_p \odot \boldsymbol{z}_p]$, and $\check{\boldsymbol{Z}} = [\boldsymbol{x}_1, \boldsymbol{z}_1 \odot \boldsymbol{z}_1, \ldots, \boldsymbol{z}_1 \odot \boldsymbol{z}_p, \ldots, \boldsymbol{x}_p, \boldsymbol{z}_p \odot \boldsymbol{z}_1, \ldots, \boldsymbol{z}_p \odot \boldsymbol{z}_p]$. We assume $\boldsymbol{\Lambda}_\Omega = \mathbf{1}\boldsymbol{\lambda}_\Omega^\top$ for some $\boldsymbol{\lambda}_\Omega \in \mathbb{R}^p$ in developing the algorithm. (But since our algorithm applies to a general $\bar{\boldsymbol{Z}} \in \mathbb{R}^{n \times p^2}$ that is not necessarily symmetric, this is without loss of generality.) In (16), the ℓ_1-type penalties are imposed on overlapping groups of variables. It is worth noting that the symmetry constraint considerably complicates the grouping structure. Without it, the variable groups can be shown to follow a tree structure, for which efficient algorithms can be developed on Jenatton et al. (2011) or Simon et al. (2013).

Our algorithm follows a different track than ADMM. The details are presented in Algorithm 1. Step 1 updates $\boldsymbol{\Xi}$ and results from a linearization-based surrogate function. Step 2 carries out a Dykstra-like splitting—see, e.g., Bauschke and Combettes (2008), by use of two proximity operators, $\vec{\boldsymbol{\Theta}}_S$ and Θ_S. Concretely, for any real number a, $\Theta_S(a;\lambda)$ is given by $\mathrm{sgn}(a)(|a|-\lambda)_+$ with $\mathrm{sgn}(\cdot)$ representing the sign function. For any vector $\boldsymbol{a}$, $\Theta_S(\boldsymbol{a};\lambda)$ is defined componentwise and the multivariate version $\vec{\boldsymbol{\Theta}}_S(\boldsymbol{a};\lambda)$ is given by $\boldsymbol{a}\Theta_S(\|\boldsymbol{a}\|_2;\lambda)/\|\boldsymbol{a}\|_2$ if $\boldsymbol{a} \neq \mathbf{0}$ and $\mathbf{0}$ otherwise.

The GRESH algorithm is easy to implement and involves no complicated matrix operations such as matrix inversion. Moreover, it does not contain ad-hoc algorithmic parameters like ρ in ADMM, and need no line search. Theorem 1 provides a universal choice for τ and guarantees the global optimality of $\hat{\boldsymbol{\Omega}}$. In particular, strict iterate convergence, in addition to function-value convergence, can be established, which is considerably stronger than an "every accumulation point" type conclusion in many numerical studies. For clarity, we assume that the inner iteration runs till convergence, but this is unnecessary; see Remark 2 below.

Theorem 1. *Suppose* $\boldsymbol{\lambda}_b \geq \mathbf{0}$, $\boldsymbol{\Lambda}_\Phi \geq \mathbf{0}$, $\boldsymbol{\lambda}_\Omega > \mathbf{0}$. *For any* $\tau > \|\check{\boldsymbol{Z}}\|_2$ *and any starting point* $\boldsymbol{\Omega}^{(0)}$, *the sequence of iterates* $\{\boldsymbol{\Omega}^{(i)}\}$ *converges to a* globally *optimal solution of* (16).

Remark 1. The conclusion in the theorem holds for "$\ell_1+\ell_2$" type penalties as well (Zou and Hastie, 2005; Owen, 2007). For the associated proximity operators, see He et al. (2013). In hierarchical variable selection, adding an ℓ_2-type shrinkage is particularly helpful to compensate for model

Algorithm 1 The GRESH algorithm for solving the general problem (16)

Inputs:
Data:$\boldsymbol{X}$, $\boldsymbol{Z}$, $\boldsymbol{y}$. Regularization parameters: $\lambda_b, \boldsymbol{\Lambda}_\Phi, \boldsymbol{\lambda}_\Omega$.
Initialization:
$i \leftarrow 0$, τ large enough (say $\tau = \|\check{\boldsymbol{Z}}\|_2$);
$\lambda_b \leftarrow \lambda_b/\tau^2$, $\boldsymbol{\Lambda}_\Phi \leftarrow (\boldsymbol{\Lambda}_\Phi + \boldsymbol{\Lambda}_\Phi^\top)/(2\tau^2)$, $\boldsymbol{\lambda}_\Omega \leftarrow \boldsymbol{\lambda}_\Omega/\tau^2$.
repeat
1. $\boldsymbol{\Xi}_\Phi \leftarrow \boldsymbol{\Phi}^{(i)} + \boldsymbol{Z}^\top \mathrm{diag}\{\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}^{(i)} - \bar{\boldsymbol{Z}}\,\mathrm{vec}\,(\boldsymbol{\Phi}^{(i)})\}\boldsymbol{Z}/\tau^2$,
   $\boldsymbol{\Xi}_b \leftarrow \boldsymbol{b}^{(i)} + \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}^{(i)} - \bar{\boldsymbol{Z}}\,\mathrm{vec}\,(\boldsymbol{\Phi}^{(i)}))/\tau^2$, $\boldsymbol{\Xi} \leftarrow [\boldsymbol{\Xi}_b, \boldsymbol{\Xi}_\Phi^\top]^\top$;
2. $\boldsymbol{P} \leftarrow \boldsymbol{0}$, $\boldsymbol{Q} \leftarrow \boldsymbol{0}$;
   repeat
      (i) $\boldsymbol{\Omega}[:,k] \leftarrow \vec{\boldsymbol{\Theta}}_S(\boldsymbol{\Xi}[:,k] + \boldsymbol{P}[:,k]; \lambda_\Omega[k]), \forall k : 1 \le k \le p$;
      (ii) $\boldsymbol{P} \leftarrow \boldsymbol{P} + \boldsymbol{\Xi} - \boldsymbol{\Omega}$;
      (iii) $\boldsymbol{\Xi}[1,:] \leftarrow \Theta_S(\boldsymbol{\Omega}[1,:] + \boldsymbol{Q}[1,:]; \lambda_b^\top)$;
      (iv) $\boldsymbol{\Omega}[2{:}\mathrm{end},:] \leftarrow (\boldsymbol{\Omega}[2{:}\mathrm{end},:] + \boldsymbol{\Omega}[2{:}\mathrm{end},:]^\top)/2$;
      (v) $\boldsymbol{\Xi}[2{:}\mathrm{end},:] \leftarrow \Theta_S(\boldsymbol{\Omega}[2{:}\mathrm{end},:] + \boldsymbol{Q}[2{:}\mathrm{end},:]; \boldsymbol{\Lambda}_\Phi)$;
      (vi) $\boldsymbol{Q} \leftarrow \boldsymbol{Q} + \boldsymbol{\Omega} - \boldsymbol{\Xi}$;
   until convergence
3. $\boldsymbol{\Omega}^{(i+1)} \leftarrow \boldsymbol{\Xi}$;
4. $i \leftarrow i + 1$;
until convergence
Output $\hat{\boldsymbol{\Omega}}$

collinearity.

Remark 2. Neither the convergence of iterates nor the optimality guarantee requires the full convergence of the inner loop; see the Supplementary Material for more detail. Various stopping criteria can be employed, e.g., Schmidt et al. (2011). In our experience, running (i)–(vi) for a few steps (say 10) usually suffices.

Remark 3. Algorithm 1 and Theorem 1 can be extended beyond quadratic loss functions. When ℓ takes the binomial deviance in classification problems, the first step of Algorithm 1 becomes $\boldsymbol{\Xi}_b \leftarrow \boldsymbol{b}^{(i)} + \boldsymbol{X}^\top(\boldsymbol{y} - \pi(\boldsymbol{X}\boldsymbol{b}^{(i)} + \bar{\boldsymbol{Z}}\,\mathrm{vec}\,(\boldsymbol{\Phi}^{(i)})))/\tau^2$, $\boldsymbol{\Xi}_\Phi \leftarrow \boldsymbol{\Phi}^{(i)} + \boldsymbol{Z}^\top \mathrm{diag}\{\boldsymbol{y} - \pi(\boldsymbol{X}\boldsymbol{b}^{(i)} + \bar{\boldsymbol{Z}}\,\mathrm{vec}\,(\boldsymbol{\Phi}^{(i)}))\}\boldsymbol{Z}/\tau^2$, $\boldsymbol{\Xi} \leftarrow [\boldsymbol{\Xi}_b, \boldsymbol{\Xi}_\Phi^\top]^\top$, where $\pi(t) = 1/(1+\exp(-t))$ and extends componentwise to vectors. We can show that choosing $\tau > \|\check{\boldsymbol{Z}}\|_2/2$ guarantees the convergence of the algorithm.

Remark 4. We recommend applying Nesterov's first acceleration in implementations (Nesterov,

2007). In more detail, it uses a momentum update of $\boldsymbol{\Xi}$ in Step 1: If $i = 0$, $\boldsymbol{\Xi}_b \leftarrow \boldsymbol{b}^{(i)} + \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}^{(i)} - \bar{\boldsymbol{Z}}\,\mathrm{vec}\,(\boldsymbol{\Phi}^{(i)}))/\tau^2$, $\boldsymbol{\Xi}_\Phi \leftarrow \boldsymbol{\Phi}^{(i)} + \boldsymbol{Z}^\top\mathrm{diag}\{\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}^{(i)} - \bar{\boldsymbol{Z}}\,\mathrm{vec}\,(\boldsymbol{\Phi}^{(i)}))\}\boldsymbol{Z}/\tau^2$; if $i > 0$, $\boldsymbol{\Xi}_b \leftarrow (1-\omega_i)\boldsymbol{\Xi}_b + \omega_i(\boldsymbol{b}^{(i)} + \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}^{(i)} - \bar{\boldsymbol{Z}}\,\mathrm{vec}\,(\boldsymbol{\Phi}^{(i)}))/\tau^2)$, $\boldsymbol{\Xi}_\Phi \leftarrow (1-\omega_i)\boldsymbol{\Xi}_\Phi + \omega_i(\boldsymbol{\Phi}^{(i)} + \boldsymbol{Z}^\top\mathrm{diag}\{\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}^{(i)} - \bar{\boldsymbol{Z}}\,\mathrm{vec}\,(\boldsymbol{\Phi}^{(i)})\}\boldsymbol{Z}/\tau^2)$, where $\omega_i = (2i+3)/(i+3)$. Empirically, the number of iterations can be reduced by about 40% in comparison to the non-relaxed form.

5 Non-asymptotic analysis

In this section, given any $\boldsymbol{\Delta} \in \mathbb{R}^{(p+1)\times p}$ and $\mathcal{J}_G \subset [p]$, we use $\boldsymbol{\Delta}_{\mathcal{J}_G}$ to denote the submatrix $\boldsymbol{\Delta}[:, \mathcal{J}_G]$. Given any $\mathcal{J}_e \subset [p^2]$, $\|\,\mathrm{vec}\,(\boldsymbol{\Delta}_\Phi)_{\mathcal{J}_e}\|_1$ and $\|\,\mathrm{vec}\,(\boldsymbol{\Delta}_\Phi)_{\mathcal{J}_e}\|_2$ are abbreviated as $\|(\boldsymbol{\Delta}_\Phi)_{\mathcal{J}_e}\|_1$ and $\|(\boldsymbol{\Delta}_\Phi)_{\mathcal{J}_e}\|_2$, respectively, when there is no ambiguity.

In this multi-regularization setting, standard treatments of the stochastic term do not give sharp error rates. In particular, applying $\langle \boldsymbol{\varepsilon}, \bar{\boldsymbol{X}}\,\mathrm{vec}\,(\boldsymbol{\Delta}_\Phi)\rangle \le \|\bar{\boldsymbol{X}}^\top\boldsymbol{\varepsilon}\|_\infty\|\boldsymbol{\Delta}_\Phi\|_1$ and $\langle \boldsymbol{\varepsilon}, \check{\boldsymbol{X}}\,\mathrm{vec}\,(\boldsymbol{\Delta})\rangle \le \|\check{\boldsymbol{X}}^\top\boldsymbol{\varepsilon}\|_{2,\infty}\|\boldsymbol{\Delta}^\top\|_{2,1}$ as commonly used in the literature (Bickel et al., 2009; Lounici et al., 2011; Negahban et al., 2012; van de Geer, 2014) would yield a prediction error bound of the order $\sigma^2(J_e \log p + J_G p)$, which is, ironically, much worse than the error rate of LASSO or Group LASSO (G-LASSO). Our analysis relies on two interrelated inequalities derived from the statistical and computational properties of GRESH estimators. See the Supplementary Material for more technical detail.

First, let's consider the **Type-A** problem (12), with λ_1, λ_2 redefined:

$$\min_{\boldsymbol{\Omega}=[\boldsymbol{b},\boldsymbol{\Phi}^\top]^\top} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} - \bar{\boldsymbol{X}}\,\mathrm{vec}\,(\boldsymbol{\Phi})\|_2^2 + \lambda_1\|\check{\boldsymbol{X}}\|_2\|\boldsymbol{\Phi}\|_1 + \lambda_2\|\check{\boldsymbol{X}}\|_2\|\boldsymbol{\Omega}^\top\|_{2,1} \text{ s.t. } \boldsymbol{\Phi} = \boldsymbol{\Phi}^\top. \tag{17}$$

Let $\hat{\boldsymbol{\Omega}} = [\hat{\boldsymbol{b}}, \hat{\boldsymbol{\Phi}}^\top]^\top$ be any global minimizer of (17). We are interested in its prediction accuracy measured by $M(\hat{\boldsymbol{b}} - \boldsymbol{b}^*, \hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}^*)$, where

$$M(\boldsymbol{b}, \boldsymbol{\Phi}) = \|\boldsymbol{X}\boldsymbol{b} + \bar{\boldsymbol{X}}\,\mathrm{vec}\,(\boldsymbol{\Phi})\|_2^2. \tag{18}$$

The predictive learning perspective is always legitimate in evaluating the quality of the estimator regardless of the signal-to-noise ratio. To guarantee small predictor errors when using a convex method, the design matrix must satisfy certain incoherence conditions, one of the most popular being the restricted eigenvalue (RE) assumption (Bickel et al., 2009; Lounici et al., 2011). In the following, we give an extension of RE in the hierarchy setting, with the restricted cone defined with both ℓ_1 and group-ℓ_1 penalties. A less intuitive but technically much less demanding condition is used in the proof.

Assumption $\mathcal{A}(\mathcal{J}_e, \mathcal{J}_G, \vartheta, \kappa)$. Given $\mathcal{J}_e \subset [p^2]$, $\mathcal{J}_G \subset [p]$, $\kappa \geq 0$ and a constant $\vartheta \geq 0$, for any $\boldsymbol{\Delta} = [\boldsymbol{\Delta}_b, \boldsymbol{\Delta}_\Phi^\top]^\top \in \mathbb{R}^{(p+1)\times p}$ satisfying $\boldsymbol{\Delta}_\Phi = \boldsymbol{\Delta}_\Phi^\top$ and $\|(\boldsymbol{\Delta}_\Phi)_{\mathcal{J}_e^c}\|_1 + \|(\boldsymbol{\Delta}_{\mathcal{J}_G^c})^\top\|_{2,1} \leq (1+\vartheta)(\|(\boldsymbol{\Delta}_\Phi)_{\mathcal{J}_e}\|_1 + \|(\boldsymbol{\Delta}_{\mathcal{J}_G})^\top\|_{2,1})$, the following inequality holds

$$\kappa \|\check{\boldsymbol{X}}\|_2^2 (\|(\boldsymbol{\Delta}_\Phi)_{\mathcal{J}_e}\|_2^2 + \|\boldsymbol{\Delta}_{\mathcal{J}_G}\|_F^2) \leq \|\boldsymbol{X}\boldsymbol{\Delta}_b + \bar{\boldsymbol{X}} \operatorname{vec}(\boldsymbol{\Delta}_\Phi)\|_2^2. \tag{19}$$

The rate choices of the regularization parameters play a major role in prediction. We choose λ_1 and λ_2 in (17) according to

$$\lambda_1 = A_1 \sigma \sqrt{\log(ep)}, \quad \lambda_2 = A_2 \sigma \sqrt{\log(ep)}, \tag{20}$$

where A_1, A_2 are large constants. (20) is quite different from the typical choice in group-ℓ_1 penalization; see Remark 2 for more detail.

The following theorem states a non-asymptotic oracle inequality as well as a model cardinality bound for GRESH estimators. For convenience, we use abbreviated symbols $\hat{J}_G = J_G(\hat{\boldsymbol{b}}, \hat{\boldsymbol{\Phi}})$ and $\hat{J}_e = J_e(\hat{\boldsymbol{\Phi}})$ for the estimate, and $J_G = J_G(\boldsymbol{b}, \boldsymbol{\Phi})$ and $J_e = J_e(\boldsymbol{\Phi})$ for the reference signal.

Theorem 2. *Assume* $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. *Let* $\hat{\boldsymbol{\Omega}} = [\hat{\boldsymbol{b}}, \hat{\boldsymbol{\Phi}}^\top]^\top$ *be a global minimizer of* (17). *Then under* (20)*, for any sufficiently large constants* A_1, A_2*, the following oracle inequality holds for any* $(\boldsymbol{b}, \boldsymbol{\Phi}) \in \mathbb{R}^p \times \mathbb{R}^{p\times p}$

$$\mathbb{E}[M(\hat{\boldsymbol{b}} - \boldsymbol{b}^*, \hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}^*)] \lesssim M(\boldsymbol{b} - \boldsymbol{b}^*, \boldsymbol{\Phi} - \boldsymbol{\Phi}^*) + (1 \vee \frac{1}{\kappa}) \sigma^2 (J_e + J_G) \log p + \sigma^2, \tag{21}$$

provided that $(\boldsymbol{X}, \bar{\boldsymbol{X}}, \check{\boldsymbol{X}})$ *satisfies* $\mathcal{A}(\mathcal{J}_e, \mathcal{J}_G, \vartheta, \kappa)$ *for some* $\kappa > 0$ *and some constant* $\vartheta \geq 0$. *Furthermore, under the same regularity condition, the overall sparsity of the obtained model is controlled by*

$$\mathbb{E}[\hat{J}_e] + \mathbb{E}[\hat{J}_G] \lesssim \{M(\boldsymbol{b}^* - \boldsymbol{b}, \boldsymbol{\Phi}^* - \boldsymbol{\Phi}) + \sigma^2\}/\{\sigma^2 \log(ep)\} + J_e + J_G. \tag{22}$$

Remark 1. Letting $\boldsymbol{b} = \boldsymbol{b}^*$ and $\boldsymbol{\Phi} = \boldsymbol{\Phi}^*$ in (21), we obtain an error bound no larger than $\sigma^2(J_e^* + J_G^*) \log p$ (omitting constant factors). This indicates that GRESH not only guarantees SH, but can give an error rate as low as that of LASSO. The existence of the bias term $M(\boldsymbol{b} - \boldsymbol{b}^*, \boldsymbol{\Phi} - \boldsymbol{\Phi}^*)$ makes our results applicable to approximately sparse signals, which is of practical significance. The theorem does not require the spectral norms of the design matrices $\boldsymbol{X}$, $\bar{\boldsymbol{X}}$ and $\check{\boldsymbol{X}}$ to be bounded above by $O(\sqrt{n})$ as assumed in, for example, Zhang and Huang (2008) and Bickel et al. (2009). In addition, the true signal $\boldsymbol{\Omega}^*$ and the reference signal $\boldsymbol{\Omega}$ in the theorem need not obey SH.

Remark 2. It is widely acknowledged that the penalty parameter for a grouped ℓ_1 penalty should be adjusted by the group size (Yuan and Lin, 2006). In fact, λ_2 would be of order $\sigma\sqrt{p + \log p}$ from Lounici et al. (2011) and Wei and Huang (2010), in light of the fact that there are p groups of size $(p + 1)$ in $\|\boldsymbol{\Omega}\|_{2,1}$. Perhaps surprisingly, this parameter choice becomes suboptimal in hierarchical variable selection. In fact, due to the presence of multiple penalties, we show in the proof that (20) suffices to suppress the noise, which in turn leads to a reduced error rate. Such a novel finding is owing to the careful treatment of the stochastic term, which is generally applicable to overlap group lasso (Jenatton et al., 2011). The conclusion that λ_1 and λ_2 are of essentially the same rate also facilitates parameter tuning, since one just needs to search along a one-dimensional grid.

Remark 3. Theorem 2 can be extended to sub-Gaussian $\mathrm{vec}(\boldsymbol{\varepsilon})$ with mean 0 and its ψ_2-norm bounded by σ, which covers more noise distributions. High probability form results of the prediction error can be obtained as well: (21) and (22), without the expectation and the additive σ^2 term, hold with probability at least $1 - Cp^{-c\min\{A_1^2, A_2^2\}}$ for some universal constants C and c, and so $\hat{J}_e + \hat{J}_G \lesssim J_e^* + J_G^*$ with high probability. Moreover, in the Supplementary Material, we show how

to adapt our proof to deliver a coordinatewise error bound which can be used for recovering the sparsity pattern of the true signal.

Remark 4. For the WH version of (17) (without the symmetry condition), following the lines of the proof of Theorem 2, we can show its error rate is of the order $\sigma^2\{J_e^w(\mathbf{\Phi}) + J_G^w(\mathbf{\Omega})\}\log p$, where $J_G^w(\mathbf{\Omega}) = J_G(\mathbf{\Omega}')$, $J_e^w(\mathbf{\Phi}) = J_e(\mathbf{\Phi}')$, with $\mathbf{\Omega}' = [\boldsymbol{b}, \mathbf{\Phi}'^\top]^\top$ and $\phi'_{kj} = \phi_{kj} + \phi_{jk}$ for $k \geq j$ and 0 otherwise. The associated regularity condition uses $\mathcal{J}_e^w$, $\mathcal{J}_G^w$, in place of $\mathcal{J}_e$, $\mathcal{J}_G$, respectively, and does not require $\mathbf{\Delta}_\Phi$ to be symmetric. Details are not reported in the paper.

Similarly, we can derive an oracle inequality for GRESH estimators of Type-B. Let $\boldsymbol{X}^s$ be the column-scaled $\boldsymbol{X}$ such that the ℓ_2-norm of each of its columns equals 1. $\bar{\boldsymbol{X}}^s$ and $\check{\boldsymbol{X}}^s$ are similarly defined. The corresponding coefficients, denoted by $\boldsymbol{b}^s$ and $\mathbf{\Phi}^s$, satisfy $\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^s\boldsymbol{b}^s$, $\bar{\boldsymbol{X}}\,\mathrm{vec}\,(\mathbf{\Phi}) = \bar{\boldsymbol{X}}^s\,\mathrm{vec}\,(\mathbf{\Phi}^s)$. Let $(\hat{\boldsymbol{b}}^s, \hat{\mathbf{\Phi}}^s)$ be a global minimizer of the scaled **Type-B** problem: $\min_{\mathbf{\Omega}^s=[\boldsymbol{b}^s,\mathbf{\Phi}^{s\top}]^\top} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}^s\boldsymbol{b}^s - \bar{\boldsymbol{X}}^s\,\mathrm{vec}\,(\mathbf{\Phi}^s)\|_2^2 + \lambda_1\|\check{\boldsymbol{X}}^s\|_2\|\mathbf{\Phi}^s\|_1 + \lambda_2\|\check{\boldsymbol{X}}^s\|_2\|(\mathbf{\Omega}^s)^\top\|_{2,1}$ s.t. $\mathbf{\Phi}^s = (\mathbf{\Phi}^s)^\top$. As aforementioned, the problem can not be reduced to (17) because $\bar{\boldsymbol{X}}^s[:, jk]$ does not equal $\boldsymbol{X}^s[:, j] \odot \boldsymbol{X}^s[:, k]$ in general.

Assumption $\mathcal{A}'(\mathcal{J}_e, \mathcal{J}_G, \vartheta, \kappa')$. Given $\mathcal{J}_e \subset [p^2]$, $\mathcal{J}_G \subset [p]$, and positive constants κ' and ϑ, for any $\mathbf{\Delta} = [\mathbf{\Delta}_b, \mathbf{\Delta}_\Phi^\top]^\top$ satisfying $\mathbf{\Delta}_\Phi = \mathbf{\Delta}_\Phi^\top$ and $\|(\mathbf{\Delta}_\Phi)_{\mathcal{J}_e^c}\|_1 + \|(\mathbf{\Delta}_{\mathcal{J}_G^c})^\top\|_{2,1} \leq (1+\vartheta)(\|(\mathbf{\Delta}_\Phi)_{\mathcal{J}_e}\|_1 + \|(\mathbf{\Delta}_{\mathcal{J}_G})^\top\|_{2,1})$, the following inequality holds

$$\kappa'\|\check{\boldsymbol{X}}^s\|_2^2(\|(\mathbf{\Delta}_\Phi)_{\mathcal{J}_e}\|_2^2 + \|\mathbf{\Delta}_{\mathcal{J}_G}\|_F^2) \leq \|\boldsymbol{X}^s\mathbf{\Delta}_b + \bar{\boldsymbol{X}}^s\,\mathrm{vec}\,(\mathbf{\Delta}_\Phi)\|_2^2.$$

Theorem 2' *Under the same conditions as in Theorem 2 and with* $\mathcal{A}'(\mathcal{J}_e, \mathcal{J}_G, \vartheta, \delta'_{\mathcal{J}_e,\mathcal{J}_G})$ *in place of* $\mathcal{A}(\mathcal{J}_e, \mathcal{J}_G, \vartheta, \kappa)$, (21) *and* (22) *hold.*

The error bounds of the two types of GRESH are of the same order, but their regularity conditions place different requirements on the design. We performed extensive simulation studies to

compare $\mathcal{A}$ and $\mathcal{A}'$, and found that for the same ϑ, $\kappa < \kappa'$ usually holds, which suggests the penalization on the basis of terms seems more appropriate than that on the coefficients. Therefore, we recommend Type-B regularization for hierarchical variable selection.

6 Minimax lower bound and error rate comparison

In this section, we show that in a minimax sense, the error rate we obtained in Theorem 2 is minimax optimal up to some logarithm factors. Consider two signal classes having hierarchy and joint sparsity:

$$\mathbf{SH}(J_G, J_e) = \{\mathbf{\Omega} = [\boldsymbol{b}, \mathbf{\Phi}^\top]^\top : \mathbf{\Omega} \text{ obeys SH}, \mathbf{\Phi} = \mathbf{\Phi}^\top, J_G(\mathbf{\Omega}) \le J_G, J_e(\mathbf{\Phi}) \le J_e\}, \tag{23}$$

$$\mathbf{WH}(J_G, J_e) = \{\mathbf{\Omega} = [\boldsymbol{b}, \mathbf{\Phi}^\top]^\top : \mathbf{\Omega} \text{ obeys WH}, J_G^w(\mathbf{\Omega}) \le J_G, J_e^w(\mathbf{\Phi}) \le J_e\}, \tag{24}$$

where $1 \le J_G \le p$, $1 \le J_e \le pJ_G$. Recall the definitions of J_G^w and J_e^w in Remark 4 following Theorem 2. Let $\ell(\cdot)$ be a nondecreasing loss function with $\ell(0) = 0$, $\ell \not\equiv 0$. Under some regularity assumptions, we study the minimax lower bounds for strong and weak hierarchical variable selection.

Assumption $\mathcal{B}^{\mathcal{S}}(J_G, J_e)$. For any $\mathbf{\Omega} = [\boldsymbol{b}, \mathbf{\Phi}^\top]^\top \in \mathbb{R}^{(p+1)\times p}$ satisfying that $\mathbf{\Phi}$ is symmetric, $J_e(\mathbf{\Phi}) \le J_e$ and $J_G(\mathbf{\Omega}) \le J_G$, $\underline{\kappa}\|\mathbf{\Omega}\|_F^2 \le \|\check{\boldsymbol{X}}\,\mathrm{vec}\,(\mathbf{\Omega})\|_2^2 \le \overline{\kappa}\|\mathbf{\Omega}\|_F^2$ holds, where $\underline{\kappa}/\overline{\kappa}$ is a positive constant.

Assumption $\mathcal{B}^{\mathcal{W}}(J_G, J_e)$. For any $\mathbf{\Omega} = [\boldsymbol{b}, \mathbf{\Phi}^\top]^\top \in \mathbb{R}^{(p+1)\times p}$ satisfying that $J_e^w(\mathbf{\Phi}) \le J_e$ and $J_G^w(\mathbf{\Omega}) \le J_G$, $\underline{\kappa}\|\mathbf{\Omega}\|_F^2 \le \|\check{\boldsymbol{X}}\,\mathrm{vec}\,(\mathbf{\Omega})\|_2^2 \le \overline{\kappa}\|\mathbf{\Omega}\|_F^2$ holds, where $\underline{\kappa}/\overline{\kappa}$ is a positive constant.

Theorem 3. *(i) Strong hierarchy. Assume* $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b}^* + \bar{\boldsymbol{X}}\,vec\,(\mathbf{\Phi}^*) + \boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, $J_G \ge 2$, $p \ge 2$, $J_e \ge 1$, $n \ge 1$, $J_G \le p/2$, $J_e \le J_G^2/2$, *and* $\mathcal{B}^{\mathcal{S}}(2J_G, 2J_e)$ *is satisfied. Then there exist positive*

constants C, c (depending on $\ell(\cdot)$ only) such that

$$\inf_{\hat{\mathbf{\Omega}}} \sup_{\mathbf{\Omega}^* \in SH(J_G, J_e)} \mathbb{E}[\ell(M(\hat{\boldsymbol{b}} - \boldsymbol{b}^*, \hat{\mathbf{\Phi}} - \mathbf{\Phi}^*)/(CP_o(J_e, J_G)))] \geq c > 0, \tag{25}$$

where $\hat{\mathbf{\Omega}}$ denotes any estimator, and

$$P_o(J_e, J_G) = \sigma^2\{J_e \log(eJ_G^2/J_e) + J_G \log(ep/J_G)\}. \tag{26}$$

(ii) Weak hierarchy. Let $J_G \geq 1$, $J_e \geq 1$, $n \geq 1$, $p \geq 2$, $J_G \leq p/2$, $J_e \leq J_G p/2$. Under the same model assumption and $\mathcal{B}^{\mathcal{W}}(2J_G, 2J_e)$, (25) *holds if $SH(J_G, J_e)$ is replaced by $WH(J_G, J_e)$ and P_o is replaced by*

$$P_o'(J_e, J_G) = \sigma^2\{J_e \log(eJ_G p/J_e) + J_G \log(ep/J_G)\}. \tag{27}$$

We give some examples of ℓ to illustrate the conclusion. For SH, using the indicator function $\ell(u) = 1_{u \geq 1}$, we know that for any estimator $(\hat{\boldsymbol{b}}, \hat{\mathbf{\Phi}})$,

$$M(\hat{\boldsymbol{b}} - \boldsymbol{b}^*, \hat{\mathbf{\Phi}} - \mathbf{\Phi}^*) \gtrsim \sigma^2 \left(J_e \log(eJ_G^2/J_e) + J_G \log(ep/J_G)\right)$$

occurs with positive probability, under some mild conditions. For $\ell(u) = u$, Theorem 3 shows that the risk $\mathbb{E}[M(\hat{\boldsymbol{b}} - \boldsymbol{b}^*, \hat{\mathbf{\Phi}} - \mathbf{\Phi}^*)]$ is bounded from below by $P_o(J_e, J_G)$ up to some multiplicative constant. Because $J_G \leq J_e \leq J_G p$ and $J_G \leq p$, it is easy to see that the minimax rates are no larger than the error rate obtained in Theorem 2.

A comparison of some popular methods follows, where we can see the benefits of hierarchical variable selection. LASSO, in our context, solves $\min_{\boldsymbol{b}, \mathbf{\Phi}; \mathbf{\Phi} = \mathbf{\Phi}^\top} \ell(\boldsymbol{b}, \mathbf{\Phi}) + \lambda(\|\boldsymbol{b}\|_1 + \|\mathbf{\Phi}\|_1)$. From Bickel et al. (2009), the estimator has a prediction error of the order $\sigma^2(J_e + J^{10} + J^{11}) \log p$. G-LASSO, with the optimization problem defined by $\min_{\boldsymbol{b}, \mathbf{\Phi}; \mathbf{\Phi} = \mathbf{\Phi}^\top} \ell(\boldsymbol{b}, \mathbf{\Phi}) + \lambda\|\mathbf{\Omega}^\top\|_{2,1}$, automatically maintains SH and has an error rate of $\sigma^2 J_G p$ (Lounici et al., 2011). In general, there is no clear winner between the two. Let's turn to a particularly interesting case where $J_e^* \ll J_G^* p$, i.e., the

existence of a main effect in the model does not indicate that all its associated interactions must be relevant. In this scenario, LASSO always outperforms G-LASSO, although it does not possess the SH property. GRESH achieves the same low error rate and guarantees hierarchy, because under SH, $J^{11} + J^{10} = J_G$.

The error rate proved in (21) does not always beat that of G-LASSO, because only large values of A are considered in Theorem 2. Yet, even in the worst case when $J_e^* \asymp J_G^* p$, GRESH is only a logarithmic factor worse. In practical data analysis, there will be no performance loss, because when $\lambda_1 = 0$, GRESH degenerates to G-LASSO.

7 Experiments

7.1 Simulations

In this part, we perform some simulation studies to compare the performance of HL and GRESH (of Type B, cf. (14)) in terms of prediction accuracy, selection consistency, and computational efficiency. We use a Toeplitz design to generate all main predictors, with the correlation between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ given by $0.5^{|i-j|}$. The true coefficients $\boldsymbol{b}^*$ and $\boldsymbol{\Phi}^*$ (symmetric) are generated according to the following three setups.

Example 1. $n = 40$, $p = 100$ or 200 (and so $p+\binom{p}{2} = 5050$ or 20100). $\boldsymbol{b}^* = [3, 1.5, 0, 0, 2, 2, 0, \cdots, 0]^\top$, $\boldsymbol{\Phi}^* = \mathbf{0}$, $\sigma^2 = 1$. No interactions are relevant to the response variable. SH is satisfied.

Example 2. $n = 150$, $p = 50$ or 100 (and so $p+\binom{p}{2} = 1275$ or 5050). $\boldsymbol{b}^* = [3, 3, 3, 3, 3, 3, 3, 3, 3, 0, \cdots, 0]^\top$, $\boldsymbol{\Phi}^*[1,2] = \boldsymbol{\Phi}^*[1,3] = \boldsymbol{\Phi}^*[4,5] = \boldsymbol{\Phi}^*[4,6] = \boldsymbol{\Phi}^*[7,8] = \boldsymbol{\Phi}^*[7,9] = 3$, $\sigma^2 = 1$. The model involves both main and interaction effects and obeys SH.

Example 3. $n = 100$, $p = 50$ or 100 (and so $p+\binom{p}{2} = 1275$ or 5050). $\boldsymbol{b}^* = [1, 1, 1, 1, 1, 1, 1, 0, \cdots, 0]^\top$, $\boldsymbol{\Phi}^*[i,j] = 5$, $1 \le i, j \le 5$, $i \ne j$, $\boldsymbol{\Phi}^*[4,5] = \boldsymbol{\Phi}^*[4,6] = \boldsymbol{\Phi}^*[4,7] = 5$, $\sigma^2 = 1$. The true model does

not have very strong main effects but satisfies SH.

All regularization parameters are tuned on a (separate) large validation dataset containing 10K observations. There is no need to perform a full two-dimensional grid search to find the optimal parameters in GRESH. Rather, motivated by Theorem 2, we set $\lambda_2 = c\lambda_1$, and chose $c = 0.5$ according to experience. Because of the convex nature of the problem, pathwise computation with warm starts is used. After variable selection, a ridge regression model is always refitted to be used for prediction. The official `R` package for HL is HierNet, implemented in `C`. We set `strong=TRUE` and post-calibrate HL by a restricted ridge refitting, which substantially enhances its accuracy. To make a fair comparison between HL and GRESH, we use the same error tolerance (1e-5) and the same number of grid values (20). All other algorithmic parameters in HierNet are set to their default values. Given each setup, we repeat the experiment for 50 times and evaluate the performance of each algorithm according to the measures defined below. The test error (Err) is the mean squared error between the true mean of $\boldsymbol{y}$ and its estimate; for robustness and stability, we report the median test error from all runs. The joint detection (JD) rate is the fraction of $|\{(i,j) : \Omega^*_{ij} \neq 0\}| \subseteq |\{(i,j) : \hat{\Omega}_{ij} \neq 0\}|$ among all experiments. The missing (M) rate and the false alarm (FA) rate are the mean of $|\{(i,j) : \Omega^*_{ij} \neq 0, \hat{\Omega}_{ij} = 0\}|/|\{(i,j) : \Omega^*_{ij} \neq 0\}|$ and the mean $|\{(i,j) : \Omega^*_{ij} = 0, \hat{\Omega}_{ij} \neq 0\}|/|\{(i,j) : \Omega^*_{ij} = 0\}|$, respectively. The path computational cost is the average running time of an algorithm in seconds. All the experiments were run on a PC with 3.2GHz CPU, 32GB memory and 64-bit Windows 8.1. Table 2 and Table 3 summarize the statistical and computational results.

From Table 2, GRESH and HL behaved equally well in Example 1, the model of which contains main effects only, but GRESH is faster. In Example 2 and Example 3, the two methods show more differences; see their test errors and joint identification rates, for example. We also noticed that GRESH often gave a more parsimonious model. When the main effects are weak as in Example 3, HL may miss some genuine interaction effects. Overall, GRESH showed comparable or better test

errors. In fact, this is observed even when SH is not satisfied (results not shown in the table). We suspect that the performance differences between HL and GRESH largely result from the fact that HL compares $|b_j|$ with $\|\boldsymbol{\phi}_j\|_1$, the ℓ_1-norm of the overall $\boldsymbol{\phi}_j$, to realize SH, while (14) groups $b_j\boldsymbol{x}_j$, $\phi_{j1}\boldsymbol{x}_j \odot \boldsymbol{x}_1$, ..., and $\phi_{jp}\boldsymbol{x}_j \odot \boldsymbol{x}_p$, on the term basis, to select main effects.

The computational times in Table 3 show the scalability of each algorithm as p varies. When $p = 1000$, there are 500500 variables in total, and so HL became computationally prohibitive, also evidenced by Lim and Hastie (2015). GRESH offered impressive computation gains in the experiment.

7.2 Comparison with ADMM

This part shows that directly applying ADMM does not give a scalable algorithm for solving the optimization problem (16) which has a large number of groups with large group size. The detailed algorithm design is given in the Supplementary Material. We set $\rho = 1$ in ADMM and compared it to Algorithm 1. The results are reported in Table 4 and Table 5. The statistical performances of the two algorithms are close. This is reasonable because they solve the same optimization problem. However, ADMM is much slower. In the experiments, ADMM became infeasible when $p = 200$ or larger.

7.3 Real data example

We performed hierarchical variable selection on the California housing data (Pace and Barry, 1997). The dataset consists of 9 summary characteristics for 20640 neighborhoods in California. The response variable is the median house value in each neighborhood. Following Hastie et al. (2009), we obtained eight household-related predictor variables: median income, housing median age, average number of rooms and bedrooms per household, population, average occupancy (population/households), latitude, and longitude, denoted by `MedInc`, `Age`, `AvgRms`, `AvgBdrms`, `Popu`,

AvgOccu, Lat, and Long, respectively. Similar to Ravikumar et al. (2007) and Radchenko and James (2010), 50 nuisance features generated as standard Gaussian random variables were added, to make the problem more challenging. The full quadratic model on this enlarged dataset contains 3422 unknowns.

To prevent from getting over-optimistic error estimates, we used a hierarchical cross-validation procedure where an outer 10-fold cross-validation (CV) is for performance evaluation and the inner 10-fold CVs are for parameter tuning. We managed to run both HL and GRESH for hierarchical variable selection, with the estimates post-calibrated by a local ridge fitting as described in Section 7.1. It took us approximately one and half days to complete the CV experiment for HL, and about 1.6 hours for GRESH. The median and mean test errors of the models obtained by HL are 530.8 and 553.5, respectively, and the average number of selected variables is 31.2. GRESH gave 516.9 and 521.1 for the median and mean test errors, , respectively, and selected 17.1 variables on average, about half of the model size of HL.

To help the reader get an intuition of the selection frequencies of all predictors, we display heat maps in Figure 1. The two heat maps in the top panel include all variables, and the bottom panel only shows the heat maps restricted to the original covariates and their interactions. According to the figure, both methods successfully removed most of the artificially added noisy features. On average, only 9.3 nuisance covariates exist in the models obtained by HL, and 5.4 in the GRESH models. The heat maps of GRESH are however neater. The HL selection results are less parsimonious and are perhaps more difficult to interpret. The nonlinear terms in GRESH include the interaction between MedInc and Age, in addition to the quadratic effects of MedInc, Age, and AvgBdrms. Popu and all its associated interaction terms never got selected by GRESH. The insignificance of Popu can be confirmed by more elaborate analysis based on gradient boosting (Hastie et al., 2009).

References

Bauschke, H. H. and Combettes, P. L. (2008). "A Dykstra-like Algorithm for Two Monotone Operators". *Pacific Journal of Optimization*, 4(3):383–391.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). "Simultaneous Analysis of Lasso and Dantzig Selector". *The Annals of Statistics*, pages 1705–1732.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2010). "Hierarchical Selection of Variables in Sparse High-Dimensional Regression". In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, pages 56–69. Institute of Mathematical Statistics.

Bien, J., Bunea, F., and Xiao, L. (2016). "Convex Banding of the Covariance Matrix". *Journal of the American Statistical Association*, 111(514):834–845.

Bien, J., Taylor, J., and Tibshirani, R. (2013). "A Lasso for Hierarchical Interactions". *The Annals of Statistics*, 41(3):1111–1141.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers". *Foundations and Trends® in Machine Learning*, 3(1):1–122.

Chipman, H. (1996). "Bayesian Variable Selection with Related Predictors". *Canadian Journal of Statistics*, 24(1):17–36.

Choi, N. H., Li, W., and Zhu, J. (2010). "Variable Selection With the Strong Heredity Constraint and its Oracle Property". *Journal of the American Statistical Association*, 105(489):354–364.

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2013). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Routledge.

Davidson, E. H. and Erwin, D. H. (2006). "Gene Regulatory Networks and the Evolution of Animal Body Plans". *Science*, 311(5762):796–800.

Hamada, M. and Wu, C. J. (1992). "Analysis of Designed Experiments With Complex Aliasing". *Journal of Quality Technology*, 24(3):130–137.

Hao, N. and Zhang, H. H. (2014). "Interaction Screening for Ultra-High Dimensional Data". *Journal of the American Statistical Association*, 109(1):1285–1301.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer-Verlag, New York, 2nd edition.

He, Y., She, Y., and Wu, D. (2013). "Stationary-Sparse Causality Network Learning". *Journal of Machine Learning Research*, 14(1):3073–3104.

Jenatton, R., Mairal, J., Obozinski, G., and Bach, F. (2011). "Proximal Methods for Hierarchical Sparse Coding". *Journal of Machine Learning Research*, 12:2297–2334.

Lim, M. and Hastie, T. (2015). "Learning Interactions via Hierarchical Group-Lasso Regularization". *Journal of Computational and Graphical Statistics*, 24(3):627–654.

Lounici, K., Pontil, M., Van De Geer, S., and Tsybakov, A. B. (2011). "Oracle Inequalities and Optimal Inference Under Group Sparsity". *The Annals of Statistics*, 39(4):2164–2204.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. London England Chapman and Hall.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). "A Unified Framework for High-Dimensional Analysis of *M*-Estimators With Decomposable Regularizers". *Statistical Science*, 27(4):538–557.

Nelder, J. A. (1977). "A Reformulation of Linear Models". *Journal of the Royal Statistical Society. Series A (General)*, pages 48–77.

Nesterov, Y. (2007). "Gradient Methods for Minimizing Composite Objective Function". Technical report, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE).

Owen, A. B. (2007). "A Robust Hybrid of Lasso and Ridge Regression". *Contemporary Mathematics*, 443:59–72.

Pace, R. K. and Barry, R. (1997). "Sparse Spatial Autoregressions". *Statistics & Probability Letters*, 33(3):291–297.

Peixoto, J. L. (1987). "Hierarchical Variable Selection in Polynomial Regression Models". *The American Statistician*, 41(4):311–313.

Peixoto, J. L. (1990). "A Property of Well-Formulated Polynomial Regression Models". *The American Statistician*, 44(1):26–30.

Radchenko, P. and James, G. M. (2010). "Variable Selection Using Adaptive Nonlinear Interaction Structures in High Dimensions". *Journal of the American Statistical Association*, 105(492):1541–1553.

Ravikumar, P. D., Liu, H., Lafferty, J. D., and Wasserman, L. A. (2007). "SpAM: Sparse Additive Models". In *In Advances in Neural Information Processing Systems 20*, pages 1201–1208. MIT Press.

Schmidt, M., Roux, N. L., and Bach, F. R. (2011). "Convergence Rates of Inexact Proximal-Gradient Methods for Convex Optimization". In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 24*, pages 1458–1466.

She, Y. (2012). "An Iterative Algorithm for Fitting Nonconvex Penalized Generalized Linear Models With Grouped Predictors". *Computational Statistics & Data Analysis*, 56(10):2976–2990.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). "A Sparse-Group Lasso". *Journal of Computational and Graphical Statistics*, 22(2):231–245.

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

van de Geer, S. (2014). "Weakly Decomposable Regularization Penalties and Structured Sparsity". *Scandinavian Journal of Statistics*, 41(1):72–86.

Wei, F. and Huang, J. (2010). "Consistent Group Selection in High-Dimensional Linear Regression". *Bernoulli*, 16(4):1369–1384.

Wu, J., Devlin, B., Ringquist, S., Trucco, M., and Roeder, K. (2010). "Screen and Clean: A Tool for Identifying Interactions in Genome-Wide Association Studies". *Genetic epidemiology*, 34(3):275–285.

Yuan, M. and Lin, Y. (2006). "Model Selection and Estimation in Regression With Grouped Variables". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, C.-H. and Huang, J. (2008). "The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression". *The Annals of Statistics*, pages 1567–1594.

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

Table 1: Error rate comparison between LASSO, G-LASSO, and GRESH, where σ^2 and other constant factors are omitted.

LASSO	$(J_e + J^{11} + J^{10}) \log p$
G-LASSO	$J_G p$
GRESH	$(J_G + J_e) \log p$
Minimax:	$J_G \log(ep/J_G) + J_e \log(eJ_G^2/J_e)$

Table 2: Statistical performance of HL and GRESH, measured in test error, joint detection rate, missing rate, and false alarm rate on simulation data. All numbers are multiplied by 100.

	Ex 1 (p = 100)				**Ex 2** (p = 50)				**Ex 3** (p = 50)			
	Err	JD	M	FA	Err	JD	M	FA	Err	JD	M	FA
HL	13.7	100	0.00	0.00	14.2	95	0.24	0.31	68.0	90	0.65	2.71
GRESH	13.6	100	0.00	0.00	11.6	100	0.00	0.12	26.2	100	0.00	0.23
	Ex 1 (p = 200)				**Ex 2** (p = 100)				**Ex 3** (p = 100)			
	Err	JD	M	FA	Err	JD	M	FA	Err	JD	M	FA
HL	13.8	100	0.00	0.00	17.3	90	0.48	0.11	92.2	15	24.03	1.44
GRESH	13.5	100	0.00	0.00	14.3	100	0.00	0.05	26.9	100	0.00	0.08

Table 3: Path computation costs of GRESH and HL when $p = 200$ and 1000. The computational times are in seconds unless otherwise specified.

	$p = 200$			$p = 1000$		
	Ex1	Ex2	Ex3	Ex1	Ex2	Ex3
HL	574	3057	2.9 hours	7.6 hours	—	—
GRESH	110	158	128	1066	2.5 hours	3194

Table 4: Statistical performance of GRESH and ADMM, measured in test error, joint detection rate, missing rate, and false alarm rate.

	Ex 1 ($p = 50$)				**Ex 2** ($p = 50$)				**Ex 3** ($p = 50$)			
	Err	JD	M	FA	Err	JD	M	FA	Err	JD	M	FA
GRESH	11.6	100	0.00	0.02	11.6	100	0.00	0.12	26.2	100	0.00	0.23
ADMM	11.7	100	0.00	0.02	11.9	100	0.00	0.12	26.3	100	0.00	0.22
	Ex 1 ($p = 100$)				**Ex 2** ($p = 100$)				**Ex 3** ($p = 100$)			
	Err	JD	M	FA	Err	JD	M	FA	Err	JD	M	FA
GRESH	13.6	100	0.00	0.00	14.3	100	0.00	0.05	26.9	100	0.00	0.08
ADMM	16.8	100	0.00	0.01	14.5	100	0.00	0.06	27.1	100	0.00	0.09

Table 5: Path computation costs of GRESH and ADMM.

	$p = 50$			$p = 100$		
	Ex1	Ex2	Ex3	Ex1	Ex2	Ex3
GRESH	8	11	9	27	36	30
ADMM	66	70	60	1027	1571	1439

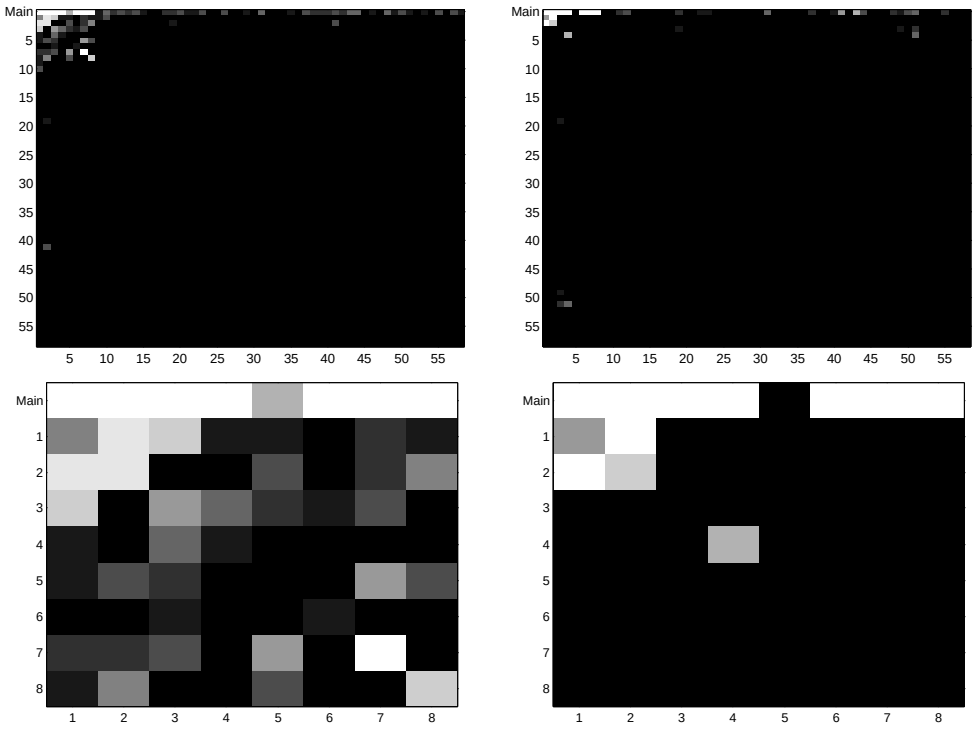

Figure 1: Top panel: heat maps of HL (left) and GRESH (right) on California housing data. Bottom panel: heat maps of HL (left) and GRESH (right) restricted to the original 8 variables and their interactions.