Group Model Selection Using Marginal Correlations: The Good, the Bad and the Ugly

Waheed U. Bajwa
Dept. of Electrical and Computer Engineering
Rutgers, The State University of New Jersey
waheed.bajwa@rutgers.edu

Dustin G. Mixon
Dept. of Mathematics and Statistics
Air Force Institute of Technology
dustin.mixon@afit.edu

***Abstract*— Group model selection is the problem of determining a small subset of groups of predictors (e.g., the expression data of genes) that are responsible for majority of the variation in a response variable (e.g., the malignancy of a tumor). This paper focuses on group model selection in high-dimensional linear models, in which the number of predictors far exceeds the number of samples of the response variable. Existing works on high-dimensional group model selection either require the number of samples of the response variable to be significantly larger than the total number of predictors contributing to the response or impose restrictive statistical priors on the predictors and/or nonzero regression coefficients. This paper provides comprehensive understanding of a low-complexity approach to group model selection that avoids some of these limitations. The proposed approach, termed *Group Thresholding* (GroTh), is based on thresholding of marginal correlations of groups of predictors with the response variable and is reminiscent of existing thresholding-based approaches in the literature. The most important contribution of the paper in this regard is relating the performance of GroTh to a polynomial-time verifiable property of the predictors for the general case of arbitrary (random or deterministic) predictors and arbitrary nonzero regression coefficients.**

I. INTRODUCTION

A. Motivation and Background

One of the most fundamental of problems in statistical data analysis is to learn the relationship between the samples of a *dependent* or *response* variable (e.g., the malignancy of a tumor, the health of a network) and the samples of *independent* or *predictor* variables (e.g., the expression data of genes, the traffic data in the network). This problem was relatively easy in the data-starved world of yesteryears. We had n samples and p predictors, and our inability to observe too many variables meant that we lived in the "n greater than or equal to p" world. Times have changed now. The data-rich world of today has enabled us to simultaneously observe an unprecedented number of variables per sample. It is nearly impossible in many of these instances to collect as many, or more, samples as the number of predictors. Imagine, for example, collecting hundreds of thousands of thyroid tumors in a clinical setting. The "n smaller than p" world is no longer a theoretical construct in statistical data analysis. It has finally arrived; and it is here to stay.

This paper concerns statistical inference in the "n smaller than p" setting for the case when the response variable depends linearly on the predictors. Mathematically, a model of this form can be expressed as

$$y_i = \sum_{j=1}^{p} x_{i,j}\beta_j^0 + \varepsilon_i, \; i = 1, \dots, n. \tag{1}$$

Here, y_i denotes the i-th sample of the response variable, $x_{i,j}$ denotes the i-th sample of the j-th predictor, ε_i denotes the error in the model, and the parameters $\{\beta_j^0\}$ are called *regression coefficients*. This relationship between the samples of the response variable and those of the predictors can be expressed compactly in matrix-vector form as $\mathrm{y} = X\beta^0 + \varepsilon$. The matrix X in this form, termed the *design matrix*, is an $n \times p$ matrix whose j-th column comprises the n samples of the j-th predictor. In tumor classification, for example, an entry in the response variable y could correspond to the malignancy (expressed as a numerical number) of a tumor sample, while the corresponding row in X would correspond to the expression level of p genes in that tumor sample.

The linear model $\mathrm{y} = X\beta^0 + \varepsilon$, despite its mathematical simplicity, continues to make profound impacts in countless application areas [1]. Such models are used for various inferential purposes. In this paper, we focus on the problem of *model selection* in high-dimensional linear models, which involves determining a small subset of p predictors that are responsible for majority (or all) of the variation in the response variable y. High-dimensional model selection can be used to implicate a small number of genes in the development of cancerous tumors, identify a small number of genes that primarily affect prognosis of a disease, etc.

B. Group Model Selection and Our Contributions

There exist many applications in statistical model selection where the implication of a single predictor in the response variable implies presence of other related predictors in the true model. This happens, for instance, in the case of microarray data when the genes (predictors) share the same biological pathway [2]. In such situations, it is better to reformulate the problem of model selection in a "group" setting. Specifically, the response variable $\mathrm{y} = X\beta^0 + \varepsilon$ in high-dimensional linear models in group settings can be best explained by a small number of *groups* of predictors:

$$\mathrm{y} = \sum_{i=1}^{m} X_i\beta_i^0 + \varepsilon = \sum_{i\in\mathcal{K}} X_i\beta_i^0 + \varepsilon, \tag{2}$$

This work and the first author are supported in part by the National Science Foundation under grant CCF-1218942.

Algorithm 1 The Group Thresholding (GroTh) Algorithm for Group Model Selection

Input: An $n \times p$ design matrix X, response variable y, number of predictors per group r, and (group) model order k
Output: An estimate $\widehat{\mathcal{K}} \subset \{1, \dots, m\}$ of the true (group) model $\mathcal{K}$

$f \leftarrow \begin{bmatrix} X_1 & X_2 & \dots & X_m \end{bmatrix}^{\mathrm{T}} y$ {Compute marginal correlations}
$\left(\mathcal{I}, \{\|f_{(j)}\|_2\}\right) \leftarrow \text{SORT}\Big(\big(\{1, \dots, m\}, \{\|f_i\|_2 := \|X_i^{\mathrm{T}} y\|_2\}\big)\Big)$ {Sort groups of marginal correlations}
$\widehat{\mathcal{K}} \leftarrow \mathcal{I}[1:k]$ {Select model via group thresholding}

where X_i, an $n \times p_i$ submatrix of X, denotes the i-th group of predictors, β_i^0 denotes the group of p_i regression coefficients associated with the group of predictors X_i, and the set $\mathcal{K} := \{1 \leq i \leq m : \beta_i^0 \neq 0\}$ denotes the underlying true (group) model, corresponding to the $k := |\mathcal{K}| \ll m$ groups of predictors that explain y.

One of the main contributions of this paper is comprehensive understanding of a polynomial-time algorithm, which we term *Group Thresholding* (GroTh), that returns an estimate $\widehat{\mathcal{K}}$ of the true (group) model $\mathcal{K}$ for the general case of arbitrary (random or deterministic) design matrices and arbitrary nonzero regression coefficients. To this end, we make use of two computable geometric measures of *group coherence* of a design matrix—the *worst-case group coherence* μ_X^g and the *average group coherence* ν_X^g—to provide a *nonasymptotic* analysis of GroTh (Algorithm 1). We in particular establish that if X satisfies a verifiable *group coherence property* then, for all but a vanishingly small fraction of possible models $\mathcal{K}$, GroTh: (i) handles linear scaling of the total number of predictors contributing to the response, $\sum_{i \in \mathcal{K}} p_i = O(n)$,[1] and ($ii$) returns indices of the groups of predictors whose contributions to the response, $\{\|\beta_i^0\|_2\}_{i \in \mathcal{K}}$, are above a certain *self-noise floor* that is a function of both μ_X^g and $\|\beta^0\|_2$.

C. Relationship to Previous Work

The basic idea of using grouped predictors for inference in linear models has been explored by various researchers in recent years. Some notable works in this direction in the "$n \ll p$" setting include [3]–[9]. Despite these inspiring results, more work needs to be done for high-dimensional group model selection. This is because the results reported in [3]–[9] do not guarantee linear scaling of the total number of predictors contributing to the response for the case of arbitrary design matrices and nonzero regression coefficients.

The work in this paper is also related to another body of work in statistics and signal processing literature that studies the high-dimensional linear model $\mathrm{y} = X\beta^0 + \varepsilon$ for the restrictive case of X having a Kronecker structure: $X := A^{\mathrm{T}} \otimes I$ for some matrix A, where $\otimes$ denotes the Kronecker product. An incomplete list of works in this direction includes [10]–[16]. These restrictive works, however, also fail to guarantee linear scaling of the total number of predictors contributing to the response for the case of arbitrary nonzero regression coefficients.

[1]Recall that $f(n) = O(g(n))$ if there exists positive C and n_0 such that for all $n > n_0$, $f(n) \leq Cg(n)$. Also, $f(n) = \Omega(g(n))$ if $g(n) = O(f(n))$, and $f(n) = \Theta(g(n))$ if $f(n) = O(g(n))$ and $g(n) = O(f(n))$.

Finally, note that the group model selection procedure studied in this paper is based on analyzing the *marginal correlations*, $X^{\mathrm{T}}\mathrm{y}$, of predictors with the response variable. Therefore, our work is algorithmically similar to the group thresholding approaches of [8], [13], [14]. The main appeal of such approaches is their low computational complexity of $O(np)$, which is much smaller than the typical computational complexity associated with other model selection procedures [17]. In addition to the scaling limitations of the total number of influential predictors discussed earlier, however, the works in [8], [13], [14] also incorrectly conclude that performance of thresholding-based approaches is inversely proportional to the dynamic range, $\frac{\max_{i \in \mathcal{K}} \|\beta_i^0\|_2}{\min_{i \in \mathcal{K}} \|\beta_i^0\|_2}$, of the nonzero groups of regression coefficients.

D. Mathematical Convention

The predictors and the response variable are assumed to be real valued throughout the paper, with the understanding that extensions to a complex-valued setting can be carried out in a straightforward manner. Uppercase letters are reserved for matrices, while lowercase letters are used for both vectors and scalars. Constants that do not depend upon the problem parameters (such as n, m, p, and k) are denoted by c_0, c_1, etc. The notation $[\![q]\!]$ for $q \in \mathbb{N}$ is a shorthand for the set $\{1, \dots, q\}$, while the notation $\stackrel{D}{=}$ signifies *equality in distribution*. The transpose operation is denoted by $(\cdot)^{\mathrm{T}}$ and the spectral norm of a matrix is denoted by $\|\cdot\|_2$. Finally, the $\ell_{p,q}$ norm of a vector $v^{\mathrm{T}} = \begin{bmatrix} v_1^{\mathrm{T}} & \dots & v_m^{\mathrm{T}} \end{bmatrix}$ with each $v_i \in \mathbb{R}^r$ is defined as $\|v\|_{p,q} := \left(\sum_{i=1}^m \|v_i\|_p^q\right)^{1/q}$ for $p, q \in (0, \infty]$, where $\|\cdot\|_p$ denotes the usual ℓ_p norm. Note that $\|v\|_{p,\infty} \equiv \max_i \|v_i\|_p$ and $\|v\|_{p,q} \equiv \|v\|_q$ for $r = 1$.

E. Organization

In Section II, we mathematically formulate the problem of group model selection, rigorously define the notions of worst-case group coherence, average group coherence and the group coherence property, and state and discuss the main result of the paper. In Section III, we prove the main result of the paper. Finally, we present some numerical results in Section IV and conclude in Section V.

II. Group Model Selection Using GroTh

A. Problem Formulation

The object of attention in this paper is the high-dimensional linear model $y = X\beta^0 + \varepsilon$ relating the response variable $y \in \mathbb{R}^n$ to the p ($\gg n$) predictors comprising the columns of the design matrix X. Since scalings of the columns of X can be absorbed into the regression vector β^0, we assume without loss of generality that the columns of X have unit ℓ_2 norms. There are three simplifying assumptions we make in this paper that will be relaxed in a sequel to this work. First, the modeling error is zero, $\varepsilon = 0$, and thus the response variable is exactly equal to a parsimonious linear combination of grouped predictors: $y = \sum_{i \in \mathcal{K}} X_i \beta_i^0$. Second, the groups of predictors $\{X_i\}_{i=1}^m$ are characterized by the same number of predictors per group: $X_i \in \mathbb{R}^{n \times r}$ with $r := \frac{p}{m} \leq n$. Third, the groups of predictors $\{X_i\}_{i=1}^m$ are orthonormalized: $X_i^{\mathrm{T}} X_i = I$.

The main goal of this paper is characterization of the performance of a group model selection procedure, termed GroTh, that returns an estimate $\widehat{\mathcal{K}}$ of the true model $\mathcal{K}$ by sorting the groups of marginal correlations $f_i := X_i^{\mathrm{T}} \mathrm{y}$ according to their ℓ_2-norms, $\|f_i\|_2$, in descending order and setting $\widehat{\mathcal{K}}$ to be indices of the first k sorted groups of marginal correlations (see Algorithm 1). Instead of focusing on the worst-case performance of GroTh, however, we seek to characterize its performance for an arbitrary (but fixed) set of nonzero (grouped) regression coefficients supported on *most* models. Specifically, we do not impose any statistical prior on the set of nonzero regression coefficients, while we assume that the true (group) model $\mathcal{K} := \{i \in [\![m]\!] : \beta_i^0 \neq 0\}$ is a uniformly random k-subset of $[\![m]\!]$. Finally, the metrics of goodness we use in this paper are the *false-discovery proportion* (FDP) and the *non-discovery proportion* (NDP), defined as

$$\mathsf{FDP}(\widehat{\mathcal{K}}) := \frac{|\widehat{\mathcal{K}} \setminus \mathcal{K}|}{|\widehat{\mathcal{K}}|} \quad \text{and} \quad \mathsf{NDP}(\widehat{\mathcal{K}}) := \frac{|\mathcal{K} \setminus \widehat{\mathcal{K}}|}{|\mathcal{K}|}, \qquad (3)$$

respectively. These two metrics have gained widespread usage in multiple hypotheses testing problems in recent years. In particular, the expectation of the FDP is the well-known *false-discovery rate* (FDR) [18], [19].

B. Main Result and Discussion

Heuristically, successful group model selection requires the groups of predictors contributing to the response variable to be sufficiently distinguishable from the ones outside the true model. In this paper, we capture the notion of distinguishability of predictors through two easily computable, global geometric measures of the design matrix, namely, the worst-case group coherence and the average group coherence. The worst-case group coherence of X is defined as

$$\mu_X^g := \max_{i,j \in [\![m]\!] : i \neq j} \|X_i^{\mathrm{T}} X_j\|_2, \qquad (4)$$

while the average group coherence of X is defined as

$$\nu_X^g := \frac{1}{m-1} \max_{i \in [\![m]\!]} \Big\| \sum_{j \in [\![m]\!] : j \neq i} X_i^{\mathrm{T}} X_j \Big\|_2. \qquad (5)$$

Note that μ_X^g is a trivial upper bound on ν_X^g. It is also worth pointing that the worst-case group coherence and its variants have existed in earlier literature [7], [20], but the average group coherence is defined for the first time in here.

The central thesis of this paper is that group model selection using GroTh can be successful if these two measures of group coherence of X are small enough. In particular, we address the question of how small should these two measures be in terms of the *group coherence property*.

Definition 1 (The Group Coherence Property)**.** The $n \times rm$ design matrix X is said to satisfy the group coherence property if the following two conditions hold for some positive constants c_μ and c_ν:

$$\mu_X^g \leq \frac{c_\mu}{\sqrt{\log m}} \quad \text{and} \qquad \text{(GroCP-1)}$$

$$\nu_X^g \leq c_\nu \mu_X^g \sqrt{\frac{r \log m}{n}}. \qquad \text{(GroCP-2)}$$

It is straightforward to observe from the above definition that the group coherence property is a global property of X that can be explicitly verified in polynomial time. Finally, we define $\beta_{(\ell)}^0$ to be the ℓ-th largest group of nonzero regression coefficients: $\|\beta_{(1)}^0\|_2 \geq \|\beta_{(2)}^0\|_2 \geq \cdots \geq \|\beta_{(k)}^0\|_2 > 0$. We are now ready to state the main result of this paper.

Theorem 1 (Group Model Selection Using GroTh)**.** *Suppose the design matrix X satisfies the group coherence property with parameters c_μ and c_ν. Next, fix parameters $c_1 \geq 2$, $c_2 \in (0,1)$, and define parameter $c_3 := \frac{32\sqrt{2e}(2c_1-1)}{(1-c_2)(c_1-1)}$. Then, under the assumptions $c_1 rk \leq n$, $c_\mu < c_3^{-1}$, and $c_\nu \leq \sqrt{c_1} c_2 c_3$, we have with probability exceeding $1 - e^2 m^{-1}$ that*

$$\Big\{ i \in \mathcal{K} : \|\beta_i^0\|_2 \geq c_3 \mu_X^g \|\beta^0\|_2 \sqrt{\log m} \Big\} \subset \widehat{\mathcal{K}}, \qquad (6)$$

resulting in $\mathsf{FDP}(\widehat{\mathcal{K}}) \leq 1 - L/k$ and $\mathsf{NDP}(\widehat{\mathcal{K}}) \leq 1 - L/k$, where L is defined to be the largest integer for which the inequality $\|\beta_{(L)}^0\|_2 \geq c_3 \mu_X^g \|\beta^0\|_2 \sqrt{\log m}$ holds. Here, the probability is with respect to the uniform distribution of the true model $\mathcal{K}$ over all possible models.

A proof of this theorem is given in Section III. We now provide a brief discussion of the significance of this result. First, Theorem 1 indicates that a polynomial-time verifiable property, namely, the group coherence property, of the design matrix can be checked to ascertain whether GroTh, which has computational complexity of $O(np)$, is well suited for group model selection. Second, it states that if X satisfies the group coherence property then GroTh handles linear scaling of the total number of predictors contributing to the response, $rk = O(n)$, for all but a vanishingly small fraction $O(m^{-1})$ of models. This is in stark contrast to the earlier works [8], [13], [14] on thresholding-based approaches in high-dimensional linear models, which do not guarantee such linear scaling for the case of arbitrary nonzero regression coefficients. Note that while we do not provide in this paper explicit examples of design matrices satisfying the group coherence property, numerical results in Section IV show that the set of design matrices satisfying the group coherence property is not empty.

Finally, Theorem 1 offers a nice interpretation of the price one might have to pay in estimating the true model using only marginal correlations. Specifically, (6) in the theorem implies group thresholding of marginal correlations effectively gives rise to a *self-noise floor* of $O\left(\mu_X^g\|\beta^0\|_2\sqrt{\log m}\right)$. In words, the estimate $\widehat{\mathcal{K}}$ returned by GroTh is guaranteed to return the indices of all the groups of predictors whose contributions to the response variable (in the ℓ_2 sense) are above the self-noise floor of $O\left(\mu_X^g\|\beta^0\|_2\sqrt{\log m}\right)$ (cf. 6). This is again a significant improvement over the earlier works [8], [13], [14], which suggest that performance of thresholding-based approaches is inversely proportional to the dynamic range, $\frac{\max_{i\in\mathcal{K}}\|\beta_i^0\|_2}{\min_{i\in\mathcal{K}}\|\beta_i^0\|_2}$, of the nonzero groups of regression coefficients. In order to expand on this, we observe from (6) that

$$\|\beta_i^0\|_2 = \Omega\left(\mu_X^g\|\beta^0\|_2\sqrt{\log m}\right)$$
$$\iff \quad \frac{\|\beta_i^0\|_2^2}{\|\beta^0\|_2^2/k} = \Omega\left(k\left(\mu_X^g\right)^2\log m\right). \tag{7}$$

Theorem 1 and the left-hand side of (7) indicate that inclusion of the i-th group of predictors in the estimate $\widehat{\mathcal{K}}$ is in fact related to the ratio of the *energy contributed by the i-th group of predictors* to the *average energy contributed per group of nonzero predictors*: $\frac{\|\beta_i^0\|_2^2}{\|\beta^0\|_2^2/k}$. Further, this implies that an increase in the dynamic range that comes from a decrease in $\min_{i\in\mathcal{K}}\|\beta_i^0\|_2$ cannot affect the performance of GroTh too much since $\frac{\|\beta_i^0\|_2^2}{\|\beta^0\|_2^2/k}$ increases for most groups of predictors in this case. This is indeed confirmed by the numerical experiments reported in Section IV.

III. Proof of the Main Result

We begin by developing some notation to facilitate the forthcoming analysis. Notice that the p-dimensional vector of marginal correlations, $f = X^{\mathrm{T}}y$, can be written as m groups of r-dimensional marginal correlations: $f^{\mathrm{T}} = \begin{bmatrix} f_1^{\mathrm{T}} & \dots & f_m^{\mathrm{T}}\end{bmatrix}$ with the $r\times 1$ vector $f_i = X_i^{\mathrm{T}}y$. In the following, we use $X_{\mathcal{K}}$ (an $n\times rk$ submatrix of X), $\beta_{\mathcal{K}}^0$ (an $rk\times 1$ subvector of β^0), and $f_{\mathcal{K}} := X_{\mathcal{K}}^{\mathrm{T}}y = X_{\mathcal{K}}^{\mathrm{T}}X_{\mathcal{K}}\beta_{\mathcal{K}}^0$ (an $rk\times 1$ subvector of f) to denote the groups of predictors, groups of regression coefficients, and the marginal correlations corresponding to the true model $\mathcal{K}$, respectively. Similarly, we use $X_{\mathcal{K}^c}$ and $f_{\mathcal{K}^c} := X_{\mathcal{K}^c}^{\mathrm{T}}y = X_{\mathcal{K}^c}^{\mathrm{T}}X_{\mathcal{K}}\beta_{\mathcal{K}}^0$ to denote the groups of predictors and the marginal correlations corresponding to the complement set $\mathcal{K}^c := [\![m]\!]\setminus\mathcal{K}$, respectively.

A. Lemmata

Proof of Theorem 1 requires understanding the behaviors of the $rk\times 1$ group vector $\left(X_{\mathcal{K}}^{\mathrm{T}}X_{\mathcal{K}} - I\right)\beta_{\mathcal{K}}^0$ and the $r(m-k)\times 1$ group vector $X_{\mathcal{K}^c}^{\mathrm{T}}X_{\mathcal{K}}\beta_{\mathcal{K}}^0$. In this subsection, we state and prove two lemmas that help us toward this goal. We will then leverage these two lemmas to provide a proof of Theorem 1.

Before proceeding, recall that $\mathcal{K}$ is taken to be a uniformly random k-subset of $[\![m]\!]$, while the set of nonzero group regression coefficients $\{z_i\}_{i=1}^k := \{\beta_i^0 : i\in\mathcal{K}\}$ is considered to be deterministic (and fixed) but unknown. It therefore follows that the rk-dimensional group vector $\left(X_{\mathcal{K}}^{\mathrm{T}}X_{\mathcal{K}} - I\right)\beta_{\mathcal{K}}^0$ can be equivalently expressed as

$$\left(X_{\mathcal{K}}^{\mathrm{T}}X_{\mathcal{K}} - I\right)\beta_{\mathcal{K}}^0 \overset{D}{=} \left(X_{\Pi}^{\mathrm{T}}X_{\Pi} - I\right)z, \tag{8}$$

where $\bar{\Pi} := (\pi_1,\dots,\pi_m)$ is a random permutation of $[\![m]\!]$, $\Pi := (\pi_1,\dots,\pi_k)$ denotes the first k elements of $\bar{\Pi}$, $X_{\Pi} := \begin{bmatrix} X_{\pi_1} & \dots & X_{\pi_k}\end{bmatrix}$ is an $n\times rk$ submatrix of X, and $z^{\mathrm{T}} := \begin{bmatrix} z_1^{\mathrm{T}} & \dots & z_k^{\mathrm{T}}\end{bmatrix}$ is an $rk\times 1$ (group) vector of nonzero regression coefficients. Similarly, the $r(m-k)$-dimensional group vector $X_{\mathcal{K}^c}^{\mathrm{T}}X_{\mathcal{K}}\beta_{\mathcal{K}}^0$ can be expressed as

$$X_{\mathcal{K}^c}^{\mathrm{T}}X_{\mathcal{K}}\beta_{\mathcal{K}}^0 \overset{D}{=} X_{\Pi^c}^{\mathrm{T}}X_{\Pi}z \tag{9}$$

where $\Pi^c := (\pi_{k+1},\dots,\pi_m)$ denotes the last $m-k$ elements of $\bar{\Pi}$ and $X_{\Pi^c} := \begin{bmatrix} X_{\pi_{k+1}} & \dots & X_{\pi_m}\end{bmatrix}$ is an $n\times r(m-k)$ submatrix of X.

Lemma 1. *Fix $c_1\geq 2$ and $\epsilon\in(0,1)$. Next, assume $k\leq \min\{\epsilon^2(\nu_X^g)^{-2}+1, c_1^{-1}m\}$ and let $\Pi = (\pi_1,\dots,\pi_k)$ denote the first k elements of a random permutation of $[\![m]\!]$. Then for any fixed $rk\times 1$ group vector $z^{\mathrm{T}} := \begin{bmatrix} z_1^{\mathrm{T}} & \dots & z_k^{\mathrm{T}}\end{bmatrix}$*

$$\Pr\left(\left\|\left(X_{\Pi}^{\mathrm{T}}X_{\Pi} - I\right)z\right\|_{2,\infty} \geq \epsilon\|z\|_2\right) \leq e^2 k\exp\Big(-c_4\big(\epsilon - \nu_X^g\sqrt{k-1}\big)^2(\mu_X^g)^{-2}\Big), \tag{10}$$

where $c_4 := \frac{(c_1-1)^2}{1024e(2c_1-1)^2}$ is an absolute constant.

Proof. The proof of this lemma relies heavily on Banach-space-valued Azuma's inequality stated in the Appendix. To begin, note that

$$\|\left(X_{\Pi}^{\mathrm{T}}X_{\Pi} - I\right)z\|_{2,\infty} \equiv \max_{i\in[\![k]\!]}\Big\|\sum_{\substack{j=1\\ j\neq i}}^k X_{\pi_i}^{\mathrm{T}}X_{\pi_j}z_j\Big\|_2. \tag{11}$$

We next fix an $i\in[\![k]\!]$ and define the event $\mathcal{A}_i' := \{\pi_i = i'\}$ for $i'\in[\![k]\!]$. Then conditioned on $\mathcal{A}_i'$, we have

$$\Pr\Big(\Big\|\sum_{\substack{j=1\\ j\neq i}}^k X_{\pi_i}^{\mathrm{T}}X_{\pi_j}z_j\Big\|_2 \geq \epsilon\|z\|_2\Big|\mathcal{A}_i'\Big) = \Pr\Big(\Big\|\sum_{\substack{j=1\\ j\neq i}}^k X_{i'}^{\mathrm{T}}X_{\pi_j}z_j\Big\|_2 \geq \epsilon\|z\|_2\Big|\mathcal{A}_i'\Big). \tag{12}$$

In order to make use of the concentration inequality in Proposition 1 in the Appendix for upper bounding (12), we construct an $\mathbb{R}^r$-valued Doob martingale on $\sum_{j\neq i}X_{i'}^{\mathrm{T}}X_{\pi_j}z_j$. We first define $\Pi^{-i} := (\pi_1,\dots,\pi_{i-1},\pi_{i+1},\dots,\pi_k)$ and then define the Doob martingale $(M_0,M_1,\dots,M_{k-1})$ as follows:

$$M_0 := \sum_{\substack{j=1\\ j\neq i}}^k X_{i'}^{\mathrm{T}}\mathbb{E}\big[X_{\pi_j}\big|\mathcal{A}_i'\big]z_j, \quad \text{and}$$

$$M_\ell = \sum_{\substack{j=1\\ j\neq i}}^k X_{i'}^{\mathrm{T}}\mathbb{E}\big[X_{\pi_j}\big|\pi_{1\to\ell}^{-i},\mathcal{A}_i'\big]z_j,\ \ \ell = 1,\dots,k-1,$$

where $\pi_{1\to\ell}^{-i}$ denotes the first ℓ elements of Π^{-i}. The next step involves showing that the constructed martingale has bounded ℓ_2 differences. In order for this, we use π_ℓ^{-i} to denote the ℓ-th element of Π^{-i} and define

$$M_\ell(u) := \sum_{\substack{j=1\\ j\neq i}}^{k} X_{i'}^{\mathrm{T}}\mathbb{E}\big[X_{\pi_j}\big|\pi_{1\to\ell-1}^{-i}, \pi_\ell^{-i}=u, \mathcal{A}_i'\big]z_j \tag{13}$$

for $u \in [\![m]\!]$ and $\ell = 1,\dots,k-1$. It can then be established using techniques very similar to the ones used in the *method of bounded differences* for scalar-valued martingales that [21], [22]

$$\|M_\ell - M_{\ell-1}\|_2 \le \sup_{u,v}\|M_\ell(u) - M_\ell(v)\|_2. \tag{14}$$

In order to upper bound $\|M_\ell(u)-M_\ell(v)\|_2$, we first define an $n\times r$ random matrix

$$\begin{aligned}\widetilde{X}_{\ell,j}^{u,v} :=& \mathbb{E}\big[X_{\pi_j}\big|\pi_{1\to\ell-1}^{-i}, \pi_\ell^{-i}=u, \mathcal{A}_i'\big]\\ &- \mathbb{E}\big[X_{\pi_j}\big|\pi_{1\to\ell-1}^{-i}, \pi_\ell^{-i}=v, \mathcal{A}_i'\big].\end{aligned} \tag{15}$$

Next, we notice that for every $j > \ell+1, j\neq i$, the random variable π_j conditioned on $\{\pi_{1\to\ell-1}^{-i}, \pi_\ell^{-i}=u, \mathcal{A}_i'\}$ has a uniform distribution over $[\![m]\!]\setminus\{\pi_{1\to\ell-1}^{-i}, u, i'\}$, while π_j conditioned on $\{\pi_{1\to\ell-1}^{-i}, \pi_\ell^{-i}=v, \mathcal{A}_i'\}$ has a uniform distribution over $[\![m]\!]\setminus\{\pi_{1\to\ell-1}^{-i}, v, i'\}$. Therefore, we get

$$\widetilde{X}_{\ell,j}^{u,v} = \frac{1}{m-\ell-1}\left(X_u - X_v\right), \quad j>\ell+1, j\neq i. \tag{16}$$

In order to evaluate $\widetilde{X}_{\ell,j}^{u,v}$ for $j\le \ell+1, j\neq i$, we consider three cases for the index i. In the first case of $i\le\ell$, it can be seen that $\widetilde{X}_{\ell,j}^{u,v}=0$ for every $j\le\ell$ and $\widetilde{X}_{\ell,j}^{u,v}=X_u-X_v$ for $j=\ell+1$. In the second case of $i=\ell+1$, it can similarly be seen that $\widetilde{X}_{\ell,j}^{u,v}=0$ for every $j<\ell$ and $j=\ell+1$, while $\widetilde{X}_{\ell,j}^{u,v}=X_u-X_v$ for $j=\ell$. In the final case of $i>\ell+1$, it can be argued that $\widetilde{X}_{\ell,j}^{u,v}=0$ for every $j<\ell$, $\widetilde{X}_{\ell,j}^{u,v}=X_u-X_v$ for $j=\ell$, and $\widetilde{X}_{\ell,j}^{u,v}=\frac{1}{m-\ell-1}\left(X_u-X_v\right)$ for $j=\ell+1$. Consequently, regardless of the initial choice of i, we have

$$\begin{aligned}&\|M_\ell(u)-M_\ell(v)\|_2\\ &\quad\equiv \Big\|\sum_{j\neq i} X_{i'}^{\mathrm{T}}\widetilde{X}_{\ell,j}^{u,v}z_j\Big\|_2 \overset{(a)}{\le} \sum_{\substack{j\ge\ell\\ j\neq i}}\|X_{i'}^{\mathrm{T}}\widetilde{X}_{\ell,j}^{u,v}\|_2\|z_j\|_2\\ &\quad\overset{(b)}{\le} 2\mu_X^g\Big(\|z_\ell\|_2+\|z_{\ell+1}\|_2+\sum_{\substack{j>\ell+1\\ j\neq i}}\frac{\|z_j\|_2}{m-\ell-1}\Big),\end{aligned} \tag{17}$$

where (a) is due to the triangle inequality and the submultiplicative nature of the induced norm, while (b) primarily follows since $\|X_{i'}^{\mathrm{T}}X_u - X_{i'}^{\mathrm{T}}X_v\|_2 \le 2\mu_X^g$. We now have from (14) and (17) that $\|M_\ell - M_{\ell-1}\|_2\le a_\ell$ with

$$a_\ell := 2\mu_X^g\Big(\|z_\ell\|_2+\|z_{\ell+1}\|_2+\sum_{\substack{j>\ell+1\\ j\neq i}}\frac{\|z_j\|_2}{m-\ell-1}\Big). \tag{18}$$

The next step needed to upper bound (12) involves providing an upper bound on $\|M_0\|_2$. To this end, note that

$$\begin{aligned}\|M_0\|_2 &\overset{(c)}{=} \Big\|\sum_{j\neq i}X_{i'}^{\mathrm{T}}\Big(\frac{1}{m-1}\sum_{\substack{q=1\\ q\neq i'}}^{m}X_q\Big)z_j\Big\|_2\\ &\le \Big\|\frac{1}{m-1}\sum_{\substack{q=1\\ q\neq i'}}^{m}X_{i'}^{\mathrm{T}}X_q\Big\|_2\Big\|\sum_{j\neq i}z_j\Big\|_2\\ &\overset{(d)}{\le} \nu_X^g\sum_{j\neq i}\|z_j\|_2 \le \nu_X^g\sqrt{k-1}\|z\|_2,\end{aligned} \tag{19}$$

where (c) follows since π_j conditioned on $\mathcal{A}_{i'}$ has a uniform distribution over $[\![m]\!]\setminus\{i'\}$ and (d) is a consequence of the definition of average group coherence. Finally, we note from [23, Lemma B.1] that $\rho_{\mathcal{B}}(\tau)$ defined in Proposition 1 satisfies $\rho_{\mathcal{B}}(\tau)\le\tau^2/2$ for $(\mathcal{B},\|\cdot\|)\equiv(L_2(\mathbb{R}^r),\|\cdot\|_2)$. Consequently, under the assumption that $k\le\epsilon^2(\nu_X^g)^{-2}+1$, it can be seen from our construction of the Doob martingale that

$$\begin{aligned}&\Pr\Big(\Big\|\sum_{\substack{j=1\\ j\neq i}}^{k}X_{i'}^{\mathrm{T}}X_{\pi_j}z_j\Big\|_2\ge\epsilon\|z\|_2\Big|\mathcal{A}_i'\Big)\\ &\quad\le \Pr\Big(\big\|M_{k-1}-M_0\big\|_2\ge\big(\epsilon-\nu_X^g\sqrt{k-1}\big)\|z\|_2\Big|\mathcal{A}_i'\Big)\\ &\quad\overset{(e)}{\le} e^2\exp\left(-\frac{c_0\big(\epsilon-\nu_X^g\sqrt{k-1}\big)^2\|z\|_2^2}{\sum_{\ell=1}^{k-1}a_\ell^2}\right),\end{aligned} \tag{20}$$

where (e) follows from Banach-space-valued Azuma's inequality stated in the Appendix. Further, it can be established using (18) through tedious algebraic manipulations that

$$\begin{aligned}\sum_{\ell=1}^{k-1}a_\ell^2 &\le \left(16+\frac{4k^2}{(m-k)^2}+\frac{16k}{m-k}\right)(\mu_X^g)^2\|z\|_2^2\\ &\overset{(f)}{\le} 4(2+(c_1-1)^{-1})^2(\mu_X^g)^2\|z\|_2^2,\end{aligned} \tag{21}$$

where (f) follows from the condition $k\le m/c_1$. Combining all these facts together, we finally obtain from (20) and (21) the following concentration inequality:

$$\begin{aligned}&\Pr\Big(\big\|\big(X_\Pi^{\mathrm{T}}X_\Pi-I\big)z\big\|_{2,\infty}\ge\epsilon\|z\|_2\Big)\\ &\quad\overset{(g)}{\le} k\Pr\Big(\Big\|\sum_{\substack{j=1\\ j\neq i}}^{k}X_{\pi_i}^{\mathrm{T}}X_{\pi_j}z_j\Big\|_2\ge\epsilon\|z\|_2\Big)\\ &\quad= k\sum_{i'=1}^{m}\Pr\Big(\Big\|\sum_{\substack{j=1\\ j\neq i}}^{k}X_{i'}^{\mathrm{T}}X_{\pi_j}z_j\Big\|_2\ge\epsilon\|z\|_2\Big|\mathcal{A}_i'\Big)\Pr(\mathcal{A}_i')\\ &\quad\overset{(h)}{\le} e^2k\exp\Big(-c_2\big(\epsilon-\nu_X^g\sqrt{k-1}\big)^2(\mu_X^g)^{-2}\Big),\end{aligned} \tag{22}$$

where $c_4 := c_0/4(2+(c_1-1)^{-1})^2$, (g) follows from the union bound and the fact that π_i's are identically distributed, while (h) follows since π_i has a uniform distribution over the set $[\![m]\!]$. ■

Lemma 2. *Fix* $c_1 \geq 2$ *and* $\epsilon \in (0,1)$. *Next, assume* $k \leq \min\{\epsilon^2(\nu_X^g)^{-2}, c_1^{-1}m\}$, *and let* $\Pi = (\pi_1, \dots, \pi_k)$ *and* $\Pi^c = (\pi_{k+1}, \dots, \pi_m)$ *denote the first* k *elements and the last* $(m-k)$ *elements of a random permutation of* $[\![m]\!]$, *respectively. Then for any fixed* $rk \times 1$ *group vector* $z^{\mathrm{T}} := \begin{bmatrix} z_1^{\mathrm{T}} & \dots & z_k^{\mathrm{T}} \end{bmatrix}$

$$\Pr\Big(\big\|X_{\Pi^c}^{\mathrm{T}} X_\Pi z\big\|_{2,\infty} \geq \epsilon \|z\|_2\Big) \leq e^2(m-k)\exp\Big(-c_5\big(\epsilon - \nu_X^g\sqrt{k}\big)^2(\mu_X^g)^{-2}\Big), \quad (23)$$

where $c_5 := \frac{(c_1-1)^2}{1024ec_1^2}$ *is an absolute constant.*

Proof. The proof of this lemma is similar to that of Lemma 1 and also relies on Proposition 1 in the Appendix. To begin, we use π_i^c to denote the i-th element of Π^c and note

$$\|X_{\Pi^c}^{\mathrm{T}} X_\Pi z\|_{2,\infty} \equiv \max_{i \in [\![m-k]\!]} \big\|\sum_{j=1}^k X_{\pi_i^c}^{\mathrm{T}} X_{\pi_j} z_j\big\|_2. \quad (24)$$

We next fix an $i \in [\![m-k]\!]$ and define $\mathcal{A}_i' := \{\pi_i^c = i'\}$ for $i' \in [\![m-k]\!]$. Then conditioned on $\mathcal{A}_i'$, we again have the following simple equality:

$$\Pr\Big(\big\|\sum_{j=1}^k X_{\pi_i^c}^{\mathrm{T}} X_{\pi_j} z_j\big\|_2 \geq \epsilon\|z\|_2 \Big| \mathcal{A}_i'\Big) = \Pr\Big(\big\|\sum_{j=1}^k X_{i'}^{\mathrm{T}} X_{\pi_j} z_j\big\|_2 \geq \epsilon\|z\|_2 \Big| \mathcal{A}_i'\Big). \quad (25)$$

In order to upper bound (25) using Proposition 1, we now construct an $\mathbb{R}^r$-valued Doob martingale $(M_0, M_1, \dots, M_k)$ on $\sum_j X_{i'}^{\mathrm{T}} X_{\pi_j} z_j$ as follows:

$$M_0 := \sum_{j=1}^k X_{i'}^{\mathrm{T}} \mathbb{E}\big[X_{\pi_j}\big|\mathcal{A}_i'\big] z_j, \quad \text{and}$$

$$M_\ell = \sum_{j=1}^k X_{i'}^{\mathrm{T}} \mathbb{E}\big[X_{\pi_j}\big|\pi_{1\to\ell}, \mathcal{A}_i'\big] z_j, \ \ell = 1, \dots, k,$$

where $\pi_{1\to\ell}$ denotes the first ℓ elements of Π. The next step in the proof involves showing $\|M_\ell - M_{\ell-1}\|_2$ is bounded for all $\ell \in [\![k]\!]$. To do this, we define

$$M_\ell(u) = \sum_{j=1}^k X_{i'}^{\mathrm{T}} \mathbb{E}\big[X_{\pi_j}\big|\pi_{1\to\ell-1}, \pi_\ell = u, \mathcal{A}_i'\big] z_j \quad (26)$$

for $u \in [\![k]\!]$ and once again resort to the argument in Lemma 1 that $\|M_\ell - M_{\ell-1}\|_2 \leq \sup_{u,v}\|M_\ell(u) - M_\ell(v)\|_2$. Further, we define an $n \times r$ random matrix

$$\widetilde{X}_{\ell,j}^{u,v} := \mathbb{E}\big[X_{\pi_j}\big|\pi_{1\to\ell-1}, \pi_\ell = u, \mathcal{A}_i'\big] - \mathbb{E}\big[X_{\pi_j}\big|\pi_{1\to\ell-1}, \pi_\ell = v, \mathcal{A}_i'\big] \quad (27)$$

and notice that $\widetilde{X}_{\ell,j}^{u,v} = 0$ for $j < \ell$, $\widetilde{X}_{\ell,j}^{u,v} = X_u - X_v$ for $j = \ell$, and $\widetilde{X}_{\ell,j}^{u,v} = \frac{1}{m-\ell-1}(X_u - X_v)$ for $j > \ell$. It then follows from this discussion that

$$\|M_\ell(u) - M_\ell(v)\|_2 \leq \sum_{j=1}^k \|X_{i'}^{\mathrm{T}} \widetilde{X}_{\ell,j}^{u,v}\|_2 \|z_j\|_2 \overset{(a)}{\leq} 2\mu_X^g\Big(\|z_\ell\|_2 + \frac{\sum_{j>\ell}\|z_j\|_2}{m-\ell-1}\Big), \quad (28)$$

where (a) is primarily due to $\|X_{i'}^{\mathrm{T}} X_u - X_{i'}^{\mathrm{T}} X_v\|_2 \leq 2\mu_X^g$. We have now established that $\|M_\ell - M_{\ell-1}\|_2 \leq a_\ell$ with

$$a_\ell := 2\mu_X^g\Big(\|z_\ell\|_2 + \frac{\sum_{j>\ell}\|z_j\|_2}{m-\ell-1}\Big), \ \ell \in [\![k]\!]. \quad (29)$$

The final bound we need in order to utilize Proposition 1 is that on $\|M_0\|_2$. Similar to (19) in Lemma 1, however, it is straightforward to show that $\|M_0\|_2 \leq \nu_X^g\sqrt{k}\|z\|_2$.

It now follows from our construction of the Doob martingale, Proposition 1 in the Appendix, [23, Lemma B.1] and the assumption $k \leq \epsilon^2(\nu_X^g)^{-2}$ that

$$\begin{aligned}\Pr\Big(\big\|\sum_{j=1}^k X_{i'}^{\mathrm{T}} X_{\pi_j} z_j\big\|_2 \geq \epsilon\|z\|_2\Big|\mathcal{A}_i'\Big) &\leq \Pr\Big(\big\|M_k - M_0\big\|_2 \geq \big(\epsilon - \nu_X^g\sqrt{k}\big)\|z\|_2\Big|\mathcal{A}_i'\Big) \\ &\leq e^2\exp\left(-\frac{c_0\big(\epsilon - \nu_X^g\sqrt{k}\big)^2\|z\|_2^2}{\sum_{\ell=1}^k a_\ell^2}\right). \quad (30)\end{aligned}$$

In addition, it can be shown using (29) and the assumption $k \leq m/c_1$ that $\sum_{\ell=1}^k a_\ell^2 \leq 4(1 + (c_1-1)^{-1})^2(\mu_X^g)^2\|z\|_2^2$. Combining all these facts together, we obtain the claimed result as follows:

$$\begin{aligned}&\Pr\Big(\big\|X_{\Pi^c}^{\mathrm{T}} X_\Pi z\big\|_{2,\infty} \geq \epsilon\|z\|_2\Big) \\ &\overset{(b)}{\leq} (m-k)\Pr\Big(\big\|\sum_{j=1}^k X_{\pi_i^c}^{\mathrm{T}} X_{\pi_j} z_j\big\|_2 \geq \epsilon\|z\|_2\Big) \\ &= (m-k)\sum_{i'=1}^m \Pr\Big(\big\|\sum_{j=1}^k X_{i'}^{\mathrm{T}} X_{\pi_j} z_j\big\|_2 \geq \epsilon\|z\|_2\Big|\mathcal{A}_i'\Big)\Pr(\mathcal{A}_i') \\ &\overset{(c)}{\leq} e^2(m-k)\exp\Big(-c_3\big(\epsilon - \nu_X^g\sqrt{k}\big)^2(\mu_X^g)^{-2}\Big), \quad (31)\end{aligned}$$

where $c_5 := c_0/4(1 + (c_1-1)^{-1})^2$, (b) follows from the union bound and the fact that π_i^c's are identically distributed, while (c) follows since π_i^c has a uniform distribution over the set $[\![m]\!]$. ■

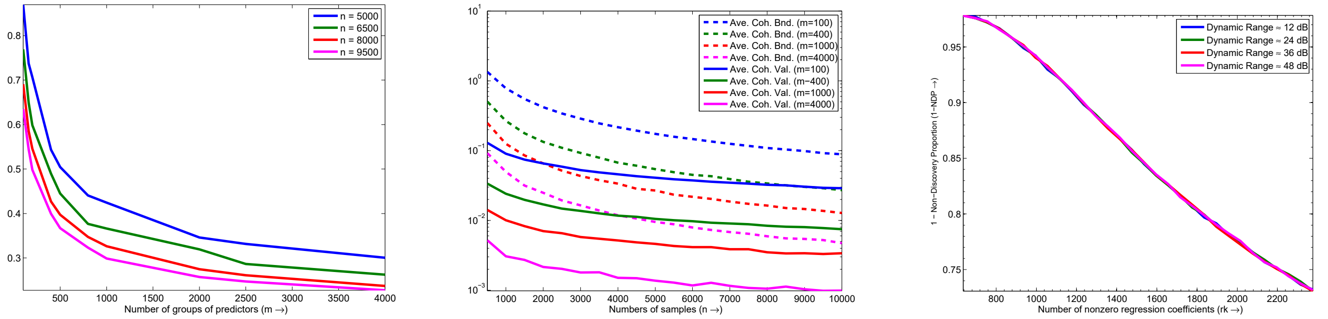
B. Proof of Theorem 1

Define $\widetilde{\mathcal{K}} := \Big\{i \in \mathcal{K} : \|\beta_i^0\|_2 \geq c_3\mu_X^g\|\beta^0\|_2\sqrt{\log m}\Big\}$. In order to prove this theorem, we need to understand the behavior of the marginal correlations corresponding to the restricted model $\widetilde{\mathcal{K}}$ and the marginal correlations corresponding to the complement set $\mathcal{K}^c$. To this end, recall the definition of L from the statement of the theorem and note that

$$\begin{aligned}\min_{i\in\widetilde{\mathcal{K}}}\|f_i\|_2 &= \min_{i\in\widetilde{\mathcal{K}}}\|\beta_i^0 + (X_i^{\mathrm{T}} X_{\mathcal{K}}\beta_{\mathcal{K}}^0 - \beta_i^0)\|_2 \\ &\geq \min_{i\in\widetilde{\mathcal{K}}}\|\beta_i^0\|_2 - \max_{i\in\widetilde{\mathcal{K}}}\|(X_i^{\mathrm{T}} X_{\mathcal{K}}\beta_{\mathcal{K}}^0 - \beta_i^0)\|_2 \\ &= \|\beta_{(L)}^0\|_2 - \|(X_{\mathcal{K}}^{\mathrm{T}} X_{\mathcal{K}} - I)\beta_{\mathcal{K}}^0\|_{2,\infty}. \quad (32)\end{aligned}$$

In addition, we trivially have

$$\max_{i\in\mathcal{K}^c}\|f_i\|_2 = \max_{i\in\mathcal{K}^c}\|X_i^{\mathrm{T}} X_{\mathcal{K}}\beta_{\mathcal{K}}^0\|_2 = \|X_{\mathcal{K}^c}^{\mathrm{T}} X_{\mathcal{K}}\beta_{\mathcal{K}}^0\|_{2,\infty}. \quad (33)$$



(a) Plots of $\mu_X^g\sqrt{\log m}$ (b) Plots of ν_X^g (solid) and $\mu_X^g\sqrt{r\log m/n}$ (dashed) (c) Plots of $1-\mathsf{NDP}$ for GroTh

Fig. 1. Numerical experiments validating main result of the paper. Together, (a) and (b) illustrate that the set of design matrices satisfying the group coherence property is not empty. Further, (c) illustrates that the performance of GroTh is not exactly a function of the dynamic range $\frac{\max_{i\in\mathcal{K}}\|\beta_i^0\|_2}{\min_{i\in\mathcal{K}}\|\beta_i^0\|_2}$.

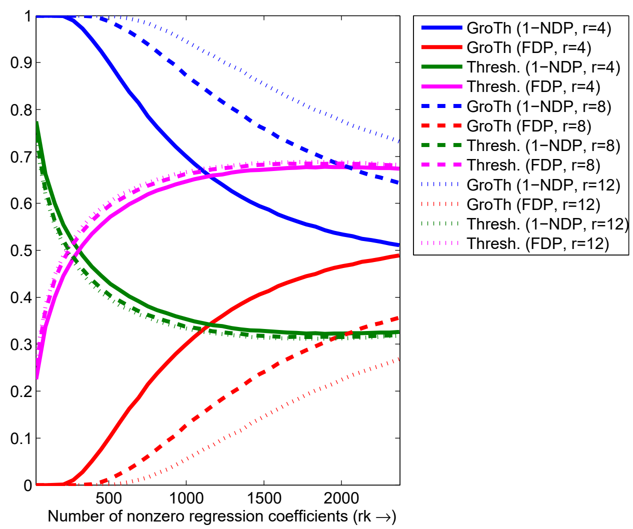


Fig. 2. Comparison between the performances of GroTh and thresholding of individual marginal correlations that ignores the grouping of predictors.

It is easy to argue using (32) and (33) that

$$\|\beta^0_{(L)}\|_2 > \|(X_{\mathcal{K}}^{\mathrm{T}}X_{\mathcal{K}}-I)\beta^0_{\mathcal{K}}\|_{2,\infty} + \|X_{\mathcal{K}^c}^{\mathrm{T}}X_{\mathcal{K}}\beta^0_{\mathcal{K}}\|_{2,\infty} \tag{34}$$

is a sufficient condition for the proof of the theorem. To see this, note that (34) implies $\min_{i\in\widetilde{\mathcal{K}}}\|f_i\|_2 > \max_{i\in\mathcal{K}^c}\|f_i\|_2$. This in turn means $\widetilde{\mathcal{K}}\subset\widehat{\mathcal{K}}$, since $L\le k$, resulting in $\mathsf{FDP}(\widehat{\mathcal{K}})\le 1-L/k$ and $\mathsf{NDP}(\widehat{\mathcal{K}})\le 1-L/k$.

The next step in the proof is therefore establishing that the sufficient condition (34) holds in our case. It is easy to show using Lemmas 1–2 and the union bound that

$$\|(X_{\mathcal{K}}^{\mathrm{T}}X_{\mathcal{K}}-I)\beta^0_{\mathcal{K}}\|_{2,\infty} + \|X_{\mathcal{K}^c}^{\mathrm{T}}X_{\mathcal{K}}\beta^0_{\mathcal{K}}\|_{2,\infty} \ge \epsilon\|\beta^0\|_2 \tag{35}$$

with probability $\delta\le e^2 m\exp\Big(-c_4\big(\epsilon-\nu_X^g\sqrt{k}\big)^2(\mu_X^g)^{-2}\Big)$ as long as $k\le\min\{\epsilon^2(\nu_X^g)^{-2}, c_1^{-1}m\}$ for $c_1\ge 2$ and $\epsilon\in(0,1)$. We now fix $\epsilon = c_3\mu_X^g\sqrt{\log m}$ and claim that (35) holds with probability $\delta\le e^2m^{-1}$ under the assumptions of the theorem. Notice that validity of this claim implies the sufficient condition (34) holds with probability $1-\delta\ge 1-e^2m^{-1}$ as long as $\|\beta^0_{(L)}\|_2\ge c_3\mu_X^g\|\beta^0\|_2\sqrt{\log m}$.

In order to complete the proof, we therefore need only establish the claim that (35) holds for $\epsilon = c_3\mu_X^g\sqrt{\log m}$ with probability $\delta\le e^2m^{-1}$. In this regard, note: (i) $\epsilon<1$ because of (GroCP-1) with $c_\mu < c_3^{-1}$ and (ii) $\sqrt{k}\nu_X^g\le c_2\epsilon$ because of $c_1rk\le n$ and (GroCP-2) with $c_\nu\le\sqrt{c_1}c_2c_3$. It then follows that (35) holds for $\epsilon = c_3\mu_X^g\sqrt{\log m}$ with probability $\delta\le e^2m^{1-c_4(1-c_2)^2c_3^2}$. The proof now trivially follows by noting that $c_4(1-c_2)^2c_3^2 = 2$ for the chosen value of c_3. ■

IV. Numerical Results

In this section, we report the outcomes of some numerical experiments that validate Theorem 1. The $n\times p$ matrix X in all these experiments is created as follows. First, we generate m of $n\times r$ matrices $\widetilde{X}_i$ whose entries are drawn independently from a standard normal distribution. Next, we use the Gram–Schmidt process to orthonormalize $\widetilde{X}_i$'s and stack the resulting orthonormal X_i's into an $n\times p$ design matrix X.

The first set of experiments reported in Fig. 1(a) and Fig. 1(b) confirms that the set of design matrices satisfying the group coherence property is not empty. Specifically, Fig. 1(a) plots $\mu_X^g\sqrt{\log m}$ as a function of m for $p=20000$ and four different values of n. It can be seen from this figure that $\mu_X^g\sqrt{\log m}=O(1)$, which verifies (GroCP-1). Further, Fig. 1(b) plots both ν_X^g (solid lines) and $\mu_X^g\sqrt{r\log m/n}$ (dashed lines) as a function of n for $p=20000$ and four different values of m. It can be seen from this figure that $\nu_X^g = O\big(\mu_X^g\sqrt{r\log m/n}\big)$, which verifies (GroCP-2).

The second set of experiments reported in Fig. 1(c) confirms that the performance of GroTh is not exactly a function of the dynamic range. In these experiments, corresponding to $p=15000$, $n=3000$ and $r=12$, all but one group of nonzero regression coefficients $\{\beta_i^0\}_{i\in\mathcal{K}}$ are normalized to have unit ℓ_2 norms, while one randomly selected group of nonzero regression coefficients is normalized to yield specified dynamic range. Fig. 1(c) plots $1-\mathsf{NDP}$ (averaged over 500 random realizations of the true model $\mathcal{K}$) for GroTh under this setup as a function of rk for four different values of dynamic range. It can be seen from this figure that the performance of GroTh indeed does not change with the dynamic range, because of the reasons outlined earlier in Section II.

The final set of experiments reported in Fig. 2 illustrates that GroTh performs better than thresholding of the individual marginal correlations that ignores the grouping of predictors. In these experiments, corresponding to $p = 15000$ and $n = 3000$, all groups of nonzero regression coefficients $\{\beta_i^0\}_{i\in\mathcal{K}}$ have unit ℓ_2 norms, but individual nonzero regression coefficients do not necessarily have same magnitudes. Fig. 2 plots FDP and $1 - \mathsf{NDP}$ (averaged over 500 random realizations of the true model $\mathcal{K}$) for both GroTh and (individual) thresholding under this setup as a function of rk for three different values of r. It can be seen from this figure that thresholding of individual marginal correlations performs almost identically for different r. Performance of GroTh, on the other hand, improves with an increase in r.

V. Conclusions

In this paper, we have provided a comprehensive understanding of Group Thresholding (GroTh) for high-dimensional group model selection. In particular, we have established that the performance of GroTh can be characterized in terms of a global geometric property of the design matrix that is explicitly verifiable in polynomial time. Results reported in this paper have also enhanced our understanding of thresholding-based approaches in high-dimensional linear models that rely on marginal correlations between the predictors and the response variable. In the future, we plan on extending this work by deriving fundamental bounds on worst-case and average group coherences, providing explicit examples of design matrices that satisfy the group coherence property, understanding the effects of modeling error, and relaxing the assumption of orthonormal groups of predictors.

Appendix
Banach-Space-Valued Azuma's Inequality

In this appendix, we state a Banach-space-valued concentration inequality from [24] that is central to this paper.

Proposition 1 (Banach-Space-Valued Azuma's Inequality)**.** *Fix $s > 0$ and assume that a Banach space $(\mathcal{B}, \|\cdot\|)$ satisfies*

$$\rho_{\mathcal{B}}(\tau) := \sup_{\substack{u,v\in\mathcal{B}\\ \|u\|=\|v\|=1}} \left\{\frac{\|u+\tau v\| + \|u-\tau v\|}{2} - 1\right\} \le s\tau^2$$

for all $\tau > 0$. Let $\{M_k\}_{k=0}^{\infty}$ be a $\mathcal{B}$-valued martingale satisfying the pointwise bound $\|M_k - M_{k-1}\| \le a_k$ for all $k \in \mathbb{N}$, where $\{a_k\}_{k=1}^{\infty}$ is a sequence of positive numbers. Then for every $\delta > 0$ and $k \in \mathbb{N}$, we have

$$\Pr\left(\|M_k - M_0\| \ge \delta\right) \le e^{\max\{s,2\}} \exp\left(-\frac{c_0\delta^2}{\sum_{\ell=1}^{k} a_\ell^2}\right),$$

where $c_0 := \frac{e^{-1}}{256}$ is an absolute constant.

Remark 1. Theorem 1.5 in [24] does not explicitly specify c_0 and also states the constant in front of $\exp(\cdot)$ to be e^{s+2}. Proposition 1 stated in its current form, however, can be obtained from the proof of Theorem 1.5 in [24].

References

[1] A. C. Rencher and G. B. Schaalje, *Linear Models in Statistics*, 2nd ed. Hoboken, NJ: John Wiley & Sons, 2008.

[2] M. R. Segal, K. D. Dahlquist, and B. R. Conklin, "Regression approaches for microarray data analysis," *J. Comput. Biol.*, vol. 10, no. 6, pp. 961–980, Jul. 2004.

[3] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc. Ser. B*, vol. 68, no. 1, pp. 49–67, 2006.

[4] F. Bach, "Consistency of the group lasso and multiple kernel learning," *J. Machine Learning Res.*, vol. 9, no. 6, pp. 1179–1225, Jun. 2008.

[5] Y. Nardi and A. Rinaldo, "On the asymptotic properties of the group lasso estimator for linear models," *Electron. J. Stat.*, vol. 2, pp. 605–633, 2008.

[6] J. Huang and T. Zhang, "The benefit of group sparsity," *Ann. Statist.*, vol. 38, no. 4, pp. 1978–2004, Aug. 2010.

[7] Y. C. Eldar, P. Kuppinger, and H. Bölcksei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Trans. Signal Processing*, vol. 58, no. 6, pp. 3042–3054, Jun. 2010.

[8] Z. Ben-Haim and Y. C. Eldar, "Near-oracle performance of greedy block-sparse estimation techniques from noisy measurements," *IEEE J. Select. Topics Signal Processing*, vol. 5, no. 5, pp. 1032–1047, Sep. 2011.

[9] E. Elhamifar and R. Vidal, "Block-sparse recovery via convex optimization," *IEEE Trans. Signal Processing*, vol. 60, no. 8, pp. 4094–4107, Aug. 2012.

[10] S. Cotter, B. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Processing*, vol. 53, no. 7, pp. 2477–2488, Jul. 2005.

[11] J. Tropp, A. Gilbert, and M. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Processing*, vol. 86, no. 3, pp. 572–588, Apr. 2006.

[12] J. Tropp, "Algorithms for simultaneous sparse approximation. Part II: Convex relaxation," *Signal Processing*, vol. 86, no. 3, pp. 589–602, Apr. 2006.

[13] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst, "Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms," *J. Fourier Anal. Appl.*, vol. 14, no. 5-6, pp. 655–687, Dec. 2008.

[14] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Trans. Inform. Theory*, vol. 56, no. 1, pp. 505–519, Jan. 2010.

[15] G. Obozinski, M. Wainwright, and M. Jordan, "Support union recovery in high-dimensional multivariate regression," *Ann. Statist.*, vol. 39, no. 1, pp. 1–47, Jan. 2011.

[16] M. Davies and Y. Eldar, "Rank awareness in joint sparse recovery," *IEEE Trans. Inform. Theory*, vol. 58, no. 2, pp. 1135–1146, Feb. 2012.

[17] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space [with comments, rejoinder]," *J. Roy. Statist. Soc. Ser. B*, vol. 70, no. 5, pp. 849–911, Nov. 2008.

[18] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Statist. Soc. Ser. B*, vol. 57, no. 1, pp. 289–300, 1995.

[19] F. Abramovich, Y. Benjamini, D. L. Donoho, and I. M. Johnstone, "Adapting to unknown sparsity by controlling the false discovery rate," *Ann. Statist.*, vol. 34, no. 2, pp. 584–653, 2006.

[20] W. U. Bajwa, R. Calderbank, and M. F. Duarte, "On the conditioning of random block subdictionaries," Department of Computer Science, Duke University, Technical Report TR-2010-06, Jun. 2010. [Online]. Available: http://www.rci.rutgers.edu/~wub1/pubs/TR2010_block_subdict.pdf

[21] C. McDiarmid, "On the method of bounded differences," in *Surveys in Combinatorics*, J. Siemons, Ed. Cambridge University Press, 1989, pp. 148–188.

[22] R. Motwani and P. Raghavan, *Randomized Algorithms*. New York, NY: Cambridge University Press, 1995.

[23] M. Donahue, C. Darken, L. Gurvits, and E. Sontag, "Rates of convex approximation in non-Hilbert spaces," in *Constructive Approximation*. New York, NY: Springer, Jun. 1997, vol. 13, no. 2, pp. 187–220.

[24] A. Naor, "On the Banach-space-valued Azuma inequality and small set isoperimetry of Alon–Roichman graphs," *Combinatorics, Probability and Computing*, vol. 21, no. 04, pp. 623–634, Jul. 2012.