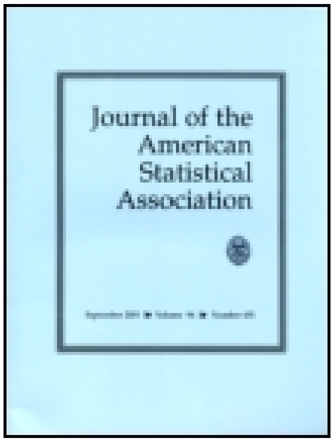


Journal of the American Statistical Association

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Homogeneity Pursuit

Zheng Tracy Ke, Jianqing Fan & Yichao Wu

Published online: 22 Apr 2015.

To cite this article: Zheng Tracy Ke, Jianqing Fan & Yichao Wu (2015) Homogeneity Pursuit, Journal of the American Statistical Association, 110:509, 175-194, DOI: 10.1080/01621459.2014.892882

To link to this article: http://dx.doi.org/10.1080/01621459.2014.892882

Homogeneity Pursuit

Zheng Tracy KE, Jianqing FAN, and Yichao WU

This article explores the homogeneity of coefficients in high-dimensional regression, which extends the sparsity concept and is more general and suitable for many applications. Homogeneity arises when regression coefficients corresponding to neighboring geographical regions or a similar cluster of covariates are expected to be approximately the same. Sparsity corresponds to a special case of homogeneity with a large cluster of known atom zero. In this article, we propose a new method called clustering algorithm in regression via data-driven segmentation (CARDS) to explore homogeneity. New mathematics are provided on the gain that can be achieved by exploring homogeneity. Statistical properties of two versions of CARDS are analyzed. In particular, the asymptotic normality of our proposed CARDS estimator is established, which reveals better estimation accuracy for homogeneous parameters than that without homogeneity exploration. When our methods are combined with sparsity exploration, further efficiency can be achieved beyond the exploration of sparsity alone. This provides additional insights into the power of exploring low-dimensional structures in high-dimensional regression: homogeneity and sparsity. Our results also shed lights on the properties of the fused Lasso. The newly developed method is further illustrated by simulation studies and applications to real data. Supplementary materials for this article are available online.

KEY WORDS: Clustering; Sparsity

1. INTRODUCTION

Driven by applications in genomics, image processing, etc., high dimensionality has become one of the major themes in statistics. See Bühlmann and van de Geer (2011) and references therein for an overview of recent developments in this area. To overcome the difficulty of fitting high-dimensional models, one usually assumes that the true parameters lie in a low-dimensional subspace. For example, many papers focus on *sparsity*, that is, only a small fraction of coefficients are nonzero (Tibshirani 1996; Chen, Donoho, and Saunders 1998). In this article, we consider a more general type of low-dimensional structure: *homogeneity*, that is, the regression coefficients share the same values in their unknown clusters. A motivating example is the gene network analysis, where it is assumed that genes cluster into groups which play similar roles in molecular processes (Kim and Xing 2009; Li and Li 2010). It can be modeled as a linear regression problem with groups of homogeneous coefficients. Similarly, in diagnostic lab tests, one often counts the number of positive results in a battery of medical tests, which implicitly assumes that their regression coefficients (impact) in the joint models are approximately the same. In spatial-temporal studies, it is not unreasonable to assume that the dynamics of neighboring geographical regions are similar, namely, their regression coefficients are clustered (Huang et al. 2010; Fan, Lv, and Qi 2011). In the same vein, financial returns of similar sectors of industry share similar loadings on risk factors.

Homogeneity is a more general assumption than sparsity, where the latter can be viewed as a special case of the former with a large group of 0-value coefficients. In addition, the atom 0 is known to data analysts. One advantage of assuming homogeneity rather than sparsity is that it enables us to possibly select more than n variables (n is the sample size). It is well known that the sparsity-based techniques, such as the Lasso, can select at most n variables. Moreover, identifying the homogeneous groups naturally provides a structure in the covariates, which can be helpful in scientific discoveries.

Regression under the homogeneity setting has been previously studied in the literature. Park, Hastie, and Tibshirani (2007) proposed a two-step method. Their method performs hierarchical clustering on the predictors, cuts the obtained dendrogram at an appropriate level, and treats the cluster averages as new predictors. The fused Lasso (Tibshirani et al. 2005; Friedman et al. 2007) can also be regarded as an effort of exploring homogeneity, with the assistance of neighborhoods defined according to either time or location. In this sense, our newly proposed methods are different since we do not know such a neighborhood a priori. The clustering of homogeneous coefficients is completely data-driven. For example, in the fused Lasso, where a complete ordering of the covariates is given, Tibshirani et al. (2005) used the L_1 penalty to penalize the pairwise differences of adjacent coordinates; in the case without a complete ordering, they suggest penalizing the pairs of "neighboring" nodes in the sense of a general distance measure. Bondell and Reich (2008) proposed the method OSCAR where a special octagonal shrinkage penalty is applied to each pair of coordinates to promote equal-value solutions. Shen and Huang (2010) developed an algorithm called grouping pursuit, where they used the truncated L_1 penalty to penalize the pairwise differences for all pairs of coordinates. In an extension, Zhu, Shen, and Pan (2013) considered simultaneous grouping pursuit and feature selection by including additional truncated L_1 penalties on the individual coefficients. Yang et al. (2012) explored simultaneous feature

Zheng Tracy Ke is graduate student (E-mail: *zke@princeton.edu*) and Jianqing Fan is Frederick L. Moore Professor (E-mail: *jqfan@princeton.edu*), Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544. Yichao Wu is Associate Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695. (E-mail: *wu@stat.ncsu.edu*) Ke and Fan's research works were partially supported by the National Institute of General Medical Sciences of the National Institutes of Health through grants R01-GM072611 and R01-GM100474 and National Science Foundation grant DMS-1206464. Wu's work was partially supported by NSF grants DMS-0905561 and DMS-1055210, and NIH grant NIH/NCI R01 CA-149569. The authors thank the Editor Professor Xuming He, the Associate Editor and two referees for their very helpful comments and suggestions.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jasa*.

© 2015 American Statistical Association
Journal of the American Statistical Association
March 2015, Vol. 110, No. 509, Theory and Methods
DOI: 10.1080/01621459.2014.892882

grouping and selection with the assistance of an undirected graph by penalizing the pairwise difference for each pair of coordinates that are connected by an edge in the graph. All the aforementioned methods either depend on a known ordering or graph of the covariates, which is sometimes not available, or use exhaustive pairwise penalties, which increase the computational complexity. Yang and He (2012) considered the homogeneity across coefficients of different percentile levels in quantile regression, and propose a Bayesian framework by using shrinkage priors to promote homogeneity. Although similar ideas may be applied to regression models, their settings are very different from ours, and there are no existing results on feature grouping for their method.

In this article, we propose a new method called clustering algorithm in regression via data-driven segmentation (CARDS) to explore homogeneity. The main idea of CARDS is to take advantage of a preliminary estimate without homogeneity structure and to shrink those coefficients that are estimated "closely," further toward each other to achieve homogeneity. In the basic version of CARDS, we first build an ordering of covariates from the preliminary estimate and run a penalized least squares afterward with fused penalties in the new ordering. The number of penalty terms is only $(p-1)$, compared to $p(p-1)/2$ in the exhaustive pairwise penalties. On the other hand, an advanced version of CARDS builds an "ordered segmentation" on the covariates, which can be viewed as a generalized ordering, and imposes "hybrid pairwise penalties," which can be viewed as a generalization of fused penalties. This version of CARDS tolerates possible misorderings in the preliminary estimate better and is thus more robust. Compared with other existing methods for homogeneity exploration, CARDS can successfully deal with the case of unordered covariates. At the same time, it avoids using exhaustive pairwise penalties and can be computationally more efficient than the grouping pursuit and OSCAR.

We study CARDS in details by providing some theoretical analysis. It reveals that the sum of squared errors of estimated coefficients is $O_p(K/n)$, where K is the number of true homogeneous groups. Therefore, the smaller the number of true groups is, the better precision it can achieve. In particular, when $K=p$, there is no homogeneity to explore and the result reduces to the case without grouping. Moreover, to exactly recover the true groups with high probability, the minimum signal strength (the gaps between different groups) is of the order $\max_k\{\sqrt{|A_k|\log(p)/n}\}$, where $|A_k|$'s are sizes of true groups. In addition, the asymptotic normality of our proposed CARDS estimator is established, which reveals better estimation accuracy than that without homogeneity exploration. Furthermore, our results can be combined with the sparsity-based results to provide additional insights into the power of exploring low-dimensional structure in high-dimensional regression: homogeneity and sparsity. As a byproduct, our analysis on the basic version of CARDS also establishes a framework for analyzing the fused type of penalties, which is new to our knowledge.

Throughout this article, we consider the following linear regression setting

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ is an $n \times p$ design matrix, $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is an $n \times 1$ vector of response, $\boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_p^0)^T$ denotes the true parameters of interest, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ contains independently and identically distributed noises with $E(\varepsilon_i) = 0$ and $E(\varepsilon_i^2) = \sigma^2$. We assume further that there is a partition of $\{1, 2, \ldots, p\}$ denoted as $\mathcal{A} = (A_0, A_1, \ldots, A_K)$ such that

$$\beta_j^0 = \beta_{A,k}^0 \quad \text{for all } j \in A_k, \tag{2}$$

where $\beta_{A,k}^0$ is the common value shared by all regression coefficients whose indices are in A_k. By default, $\beta_{A,0}^0 = 0$, so A_0 is the group of 0-value coefficients. This allows us to explore homogeneity and sparsity simultaneously. Write $\boldsymbol{\beta}_A^0 = (\beta_{A,1}^0, \ldots, \beta_{A,K}^0)^T$. Without loss of generality, we assume $\beta_{A,1}^0 < \beta_{A,2}^0 < \cdots < \beta_{A,K}^0$.

Our theory and methods are stated for the standard least-squares problem although they can be adapted to other more sophisticated models. For example, when forecasting housing appreciation in the United States (Fan, Lv, and Qi 2011), one builds the spatial-temporal model

$$Y_{it} = \mathbf{X}_{it}^T \boldsymbol{\beta}_i + \varepsilon_{it}, \tag{3}$$

in which i indicates a spatial location and t indicates time. It is expected that the $\boldsymbol{\beta}_i$'s are approximately the same for neighboring zip codes i and this type of homogeneity can be explored in a similar fashion. Similarly, when Y_{it} represents the return of a stock and $\mathbf{X}_{it} = \mathbf{X}_t$ stands for common market risk factors, one can assume a certain degree of homogeneity within each sector of industry; namely, the factor loading vector $\boldsymbol{\beta}_i$ is approximately the same for stocks belonging to the same sector of industry.

Throughout this article, $\mathbb{R}$ denotes the set of real numbers and $\mathbb{R}^p$ denotes the p-dimensional real Euclidean space. For any $a, b \in \mathbb{R}$, $a \vee b$ denotes the maximum between a and b. For any positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \gg b_n$ if $a_n/b_n \to \infty$ as $n \to \infty$. Given $1 \leq q < \infty$, for any vector $\mathbf{x}$, $\|\mathbf{x}\|_q = (\sum_j |x_j|^q)^{1/q}$ denotes the L_q-norm of $\mathbf{x}$ and $\|\mathbf{x}\|_\infty = \max_j\{|x_j|\}$. For any matrix $\mathbf{M}$, $\|\mathbf{M}\|_q = \max_{\mathbf{x}:\|\mathbf{x}\|_q=1} \|\mathbf{M}\mathbf{x}\|_q$ denotes the matrix L_q-norm of $\mathbf{M}$. In particular, $\|\mathbf{M}\|_\infty$ is the maximum absolute row sum of $\mathbf{M}$. We omit the subscript q when $q = 2$. $\|\mathbf{M}\|_{\max} = \max_{i,j}\{|M_{ij}|\}$ denotes the matrix max norm. When $\mathbf{M}$ is symmetric, $\lambda_{\max}(\mathbf{M})$ and $\lambda_{\min}(\mathbf{M})$ denote the maximum and minimum eigenvalues of $\mathbf{M}$, respectively.

The rest of the article is organized as follows. Section 2 describes CARDS, including the basic, advanced, and shrinkage versions. Section 3 studies theoretical properties of the basic version of CARDS, and Section 4 analyzes the advanced and shrinkage versions. Sections 5 and 6 present the results of simulation studies and real data analysis, respectively. Some concluding remarks are given in Section 7. Proofs can be found in the Appendix and online supplemental materials.

2. CARDS: A DATA-DRIVEN PAIRWISE SHRINKAGE PROCEDURE

2.1 Basic Version of CARDS

Without considering the homogeneity assumption (2), there are many methods available for fitting model (1). Let $\widetilde{\boldsymbol{\beta}}$ be such a preliminary estimator. A simple idea to generate homogeneity is as follows: first, rearrange the coefficients in $\widetilde{\boldsymbol{\beta}}$ in the ascending order; second, group together those adjacent indices whose coefficients in $\widetilde{\boldsymbol{\beta}}$ are close to each other; finally, force indices in each estimated group to share a common coefficient and refit model (1). A problem of this naive procedure is how to group the indices. Alternatively, we can run a penalized least squares to simultaneously extract the grouping structure and estimate coefficients. To shrink coefficients of adjacent indices (after reordering) toward homogeneity, we can add fused penalties, that is, $\{|\beta_{i+1} - \beta_i|, i = 1, \ldots, p-1\}$ are penalized. This leads to the following two-stage procedure:

- *Preordering*: Construct the rank statistics $\{\tau(j) : 1 \le j \le p\}$ such that $\tilde{\beta}_{\tau(j)}$ is the jth smallest value in $\{\tilde{\beta}_i, 1 \le i \le p\}$, that is,
$$\tilde{\beta}_{\tau(1)} \le \tilde{\beta}_{\tau(2)} \le \cdots \le \tilde{\beta}_{\tau(p)}. \tag{4}$$
- *Estimation*: Given a folded concave penalty function $p_\lambda(\cdot)$ (Fan and Li 2001) with a regularization parameter λ, the final estimate is given by
$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p-1} p_\lambda(|\beta_{\tau(j+1)} - \beta_{\tau(j)}|) \right\}. \tag{5}$$

We call this two-stage procedure basic CARDS (bCARDS).

At the first stage, bCARDS establishes a data-driven rank mapping $\tau(\cdot)$ based on the preliminary estimator $\widetilde{\boldsymbol{\beta}}$. At the second stage, only "adjacent" coefficient pairs in the order τ are penalized, resulting in only $(p-1)$ penalty terms in total. In addition, (5) does not require that $\beta_{\tau(j)} \le \beta_{\tau(j+1)}$. This allows coordinates in $\widehat{\boldsymbol{\beta}}$ to have a different order of increasing values from that in $\widetilde{\boldsymbol{\beta}}$.

With an appropriately large tuning parameter λ, $\widehat{\boldsymbol{\beta}}$ is a piecewise constant vector in the order τ and consequently its elements have homogeneous groups. In Section 3, we shall show that, if τ is consistent with the order of $\boldsymbol{\beta}^0$, that is,
$$\beta^0_{\tau(1)} \le \beta^0_{\tau(2)} \le \cdots \le \beta^0_{\tau(p)}, \tag{6}$$
then under some regularity conditions, $\widehat{\boldsymbol{\beta}}$ can consistently estimate the true coefficient groups of $\boldsymbol{\beta}^0$ with high probability.

When $p_\lambda(\cdot)$ is a folded-concave penalty function (e.g., SCAD, Fan and Li 2001, MCP, Zhang 2010), (5) is a nonconvex optimization problem. It is generally challenging to find the global minimizer. The local linear approximation (LLA) algorithm can be applied to find a local minimizer for any fixed initial solution; see Zou and Li (2008), Fan, Xue, and Zou (2012) and references therein for details. The coupling of the concave convex procedure (CCCP) can also be applied to produce a local minimizer; see Kim, Choi, and Oh (2008), Wang, Kim, and Li (2013) for a detailed explanation of CCCP.

2.2 Advanced Version of CARDS

To guarantee the success of bCARDS, (6) is an essential condition. It requires that whenever $\beta^0_i < \beta^0_j$, $\tau(i) < \tau(j)$ must hold. This imposes fairly strong conditions on the preliminary estimator $\widetilde{\boldsymbol{\beta}}$. For example, (6) can be violated if $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_\infty$ is larger than the minimum gap between groups. To relax such a restrictive requirement, we now introduce an advanced version of CARDS, where the main idea is to use less information from $\widetilde{\boldsymbol{\beta}}$ and to add more penalty terms in (5).

We first introduce the *ordered segmentation*, which can be viewed as a generalized ordering. Note that each rank mapping τ in bCARDS actually defines a partition of $\{1, \ldots, p\}$ into p disjoint sets $B_1, \ldots, B_p$ with $B_j = \{\tau(j)\}$ being a singleton. Similarly, we may divide $\{1, \ldots, p\}$ into $L (\le p)$ disjoint sets $B_1, \ldots, B_L$, where the B_l's are not necessarily singletons. We call such B_l's *segments*. The segments $B_1, \ldots, B_L$ are ordered, but the ordering of coordinates within each segment is not defined. This is similar to letter grades assigned to a course. A formal definition is as follows:

Definition 1. For an integer $1 \le L \le p$, the mapping $\Upsilon : \{1, \ldots, p\} \to \{1, \ldots, L\}$ is called an ordered segmentation if the sets $B_l \equiv \{1 \le j \le p : \Upsilon(j) = l\}$, $1 \le l \le L$, form a partition of $\{1, \ldots, p\}$.

When $L = p$, Υ is a one-to-one mapping and it defines a complete ordering.

Note that, in the basic version of CARDS, the preliminary estimator $\widetilde{\boldsymbol{\beta}}$ produces a complete rank mapping τ. Now in the advanced version of CARDS, instead of extracting a complete ordering, we only extract an ordered segmentation Υ from $\widetilde{\boldsymbol{\beta}}$. The analogue is similar to grading an exam: overall score rank (percentile rank) versus letter grade. Let $\delta > 0$ be a predetermined parameter. First, obtain the rank mapping τ as in (4) and find all indices $1 < i_2 < i_3 < \cdots < i_L$ such that the gaps
$$\tilde{\beta}_{\tau(j)} - \tilde{\beta}_{\tau(j-1)} > \delta, \qquad j = i_2, \ldots, i_L.$$
Then, construct the segments
$$B_l = \{\tau(i_l), \tau(i_l + 1), \ldots, \tau(i_{l+1} - 1)\}, \quad l = 1, \ldots, L, \tag{7}$$
where $i_1 = 1$ and $i_{L+1} = p + 1$. This process is indeed similar to the letter grades that we assign to a course. The intuition behind this construction is that when $\tilde{\beta}_{\tau(j+1)} \le \tilde{\beta}_{\tau(j)} + \delta$, that is, the estimated coefficients of two "adjacent coordinates" differ by only a small amount, we do not trust the ordering between them and group them into a same segment. Compared to the complete ordering τ, the ordered segments $\{B_1, \ldots, B_L\}$ use less information from $\widetilde{\boldsymbol{\beta}}$ and, hence, are less sensitive to the estimation error in $\widetilde{\boldsymbol{\beta}}$.

Given an ordered segmentation Υ, we wish to take advantage of the order of segments $B_1, \ldots, B_L$ and at the same time allow flexibility of order shuffling within each segment. Toward this goal, we introduce the *hybrid pairwise penalty*.

Definition 2. Given a penalty function $p_\lambda(\cdot)$ and tuning parameters λ_1 and λ_2, the hybrid pairwise penalty corresponding to an ordered segmentation Υ is
$$P_{\Upsilon,\lambda_1,\lambda_2}(\boldsymbol{\beta}) = \sum_{l=1}^{L-1} \sum_{i \in B_l, j \in B_{l+1}} p_{\lambda_1}(|\beta_i - \beta_j|) + \sum_{l=1}^{L} \sum_{i,j \in B_l} p_{\lambda_2}(|\beta_i - \beta_j|). \tag{8}$$

In (8), we call the first part *between-segment* penalty and the second part *within-segment* penalty. The within-segment penalty penalizes all pairs of indices in each segment, hence, it does not rely on any ordering within the segment. The between-segment penalty penalizes pairs of indices from two adjacent segments, and it can be viewed as a "generalized" fused penalty on segments.

When $L = p$, each B_l is a singleton and (8) reduces to the fused penalty in (5). On the other hand, when $L = 1$, namely, no prior information about $\boldsymbol{\beta}$, there is only one segment $B_1 = \{1, \ldots, p\}$, and (8) reduces to the exhaustive pairwise penalty

$$P_\lambda^{\mathrm{TV}}(\boldsymbol{\beta}) = \sum_{1 \le i, j \le p} p_\lambda(|\beta_i - \beta_j|). \tag{9}$$

It is also called the total variation penalty (Harchaoui and Lévy-Leduc 2010), and the case with $p_\lambda(\cdot)$ being a truncated L_1 penalty is studied in Shen and Huang (2010). Thus, the penalty (8) is a generalization of both the fused penalty and the total variation penalty, which explains the name "hybrid."

The main motivation of introducing the hybrid pairwise penalty is to provide a set of intermediate versions between the fused penalty and the total variation penalty. When using pairwise penalties to promote homogeneity, we need to penalize "enough" pairs to guarantee that all true groups can be exactly recovered when the signal-to-noise ratio is sufficiently large. Given a consistent ordering, the fused penalty contains "just enough" pairs; but when the ordering is inconsistent, we have to penalize more pairs to achieve the aforementioned exact-group-recovery (see Section 2.3 for a numerical example). However, it may not be a good choice to include all pairs, that is, using the total variation penalty, as the large number of redundant pairs can result in statistical and computational inefficiency. The hybrid penalty is designed aiming to include "just enough" pairs that adapt to the available "partial" ordering information of an ordered segmentation.

Now, we discuss how the requirement (6) can be relaxed. If we let $B_j = \{\tau(j)\}$, then (6) can be written equivalently as $\max_{i \in B_j} \beta_i^0 \le \min_{i \in B_{j+1}} \beta_i^0$, for $1 \le j \le p - 1$. This definition can be generalized to the case B_j's are not singletons.

Definition 3. An ordered segmentation Υ preserves the order of $\boldsymbol{\beta}^0$ if $\max_{j \in B_l} \beta_j^0 \le \min_{j \in B_{l+1}} \beta_j^0$, for $l = 1, \ldots, L - 1$.

In the construction (7), even if (6) does not hold, it is still possible that the resulting Υ preserves the order of $\boldsymbol{\beta}^0$. Consider a toy example where $p = 4$, and $\beta_{\tau(1)}^0 = \beta_{\tau(2)}^0 = \beta_{\tau(4)}^0 < \beta_{\tau(3)}^0$ so that $\{\tau(1), \tau(2), \tau(4)\}$ and $\{\tau(3)\}$ are two true homogeneous groups in $\boldsymbol{\beta}^0$. Here τ ranks the coordinate $\tau(3)$ ahead of $\tau(4)$ based on the preliminary estimator $\widetilde{\boldsymbol{\beta}}$, but $\beta_{\tau(3)}^0 > \beta_{\tau(4)}^0$. So, τ fails to give a consistent ordering. However, as long as $\tilde{\beta}_{\tau(4)} \le \tilde{\beta}_{\tau(3)} + \delta$, $\tau(3)$ and $\tau(4)$ are grouped into the same segment in (7), say, $B_1 = \{\tau(1), \tau(2)\}$ and $B_2 = \{\tau(3), \tau(4)\}$. Then Υ still preserves the order of $\boldsymbol{\beta}^0$ according to Definition 3.

Now we formally introduce the advanced version of clustering algorithm in regression via data-driven segmentation. It consists of three steps, where the first two steps are very similar to the way that we assign letter grades based on scores of an exam.

- *Preliminary Ranking*: Given a preliminary estimate $\widetilde{\boldsymbol{\beta}}$, generate the rank mapping $\{\tau(j) : 1 \le j \le p\}$ such that $\tilde{\beta}_{\tau(1)} \le \tilde{\beta}_{\tau(2)} \le \cdots \le \tilde{\beta}_{\tau(p)}$.
- *Segmentation*: For a tuning parameter $\delta > 0$, construct an ordered segmentation Υ as described in (7).
- *Estimation*: For tuning parameters λ_1 and λ_2, compute the solution $\widehat{\boldsymbol{\beta}}$ that minimizes

$$Q_n(\boldsymbol{\beta}) = \frac{1}{2n} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + P_{\Upsilon, \lambda_1, \lambda_2}(\boldsymbol{\beta}). \tag{10}$$

We call this procedure advanced CARDS (aCARDS).

In Section 4, we shall show that if Υ preserves the order of $\boldsymbol{\beta}^0$, under certain conditions, $\widehat{\boldsymbol{\beta}}$ recovers the true homogeneous groups of $\boldsymbol{\beta}^0$ with high probability. Therefore, to guarantee the success of aCARDS, we need the existence of a $\delta > 0$ for the preliminary estimator $\widetilde{\boldsymbol{\beta}}$ such that the associated Υ preserves the order of $\boldsymbol{\beta}^0$. The above toy example shows that even when (6) fails, this condition can still hold. So aCARDS requires weaker conditions on $\widetilde{\boldsymbol{\beta}}$ than bCARDS. This is due to that the hybrid penalty contains penalty terms corresponding to more pairs of indices, hence, it is more robust to misordering in τ. In fact, bCARDS is a special case of aCARDS with $\delta = 0$.

2.3 Comparison of Two Versions of CARDS

In this section, we first use a numerical example to compare bCARDS and aCARDS. It reveals how the ordered segmentation and hybrid pairwise penalty (8) play a role in aCARDS. We then discuss how to choose between two versions of CARDS in real data analysis.

We generate a dataset with $p = 40$ predictors and $n = 100$ samples. The predictors are divided into two homogeneous groups, each of size 20. Let $\beta_j^0 = -0.2$ for j in Group 1 and $\beta_j^0 = 0.2$ for j in Group 2. $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are generated independently and identically from $N_p(\mathbf{0}, \mathbf{I})$, and $Y_i = \mathbf{X}_i^T \boldsymbol{\beta}^0 + \epsilon_i$ for $1 \le i \le n$, where $\epsilon_1, \ldots, \epsilon_n$ are independent noises with a standard normal distribution. In aCARDS, we take the ordinary least squares (OLS) estimator as the preliminary estimator. Figure 1 plots the sorted OLS coefficients for a realization. The estimated rank is not exactly consistent with the order of $\boldsymbol{\beta}^0$ since the predictors $\tau(17)$ and $\tau(18)$, which belong to Group 2, are mistakenly ranked ahead of some predictors in Group 1. If we use only the fused penalty, the terms that involve $\tau(17)$ and $\tau(18)$ are

$$p_\lambda\big(|\beta_{\tau(16)} - \beta_{\tau(17)}|\big) + p_\lambda\big(|\beta_{\tau(17)} - \beta_{\tau(18)}|\big) \\ + p_\lambda\big(|\beta_{\tau(18)} - \beta_{\tau(19)}|\big).$$

There are no penalty terms to shrink the coefficients of $\tau(17)$ and $\tau(18)$ toward being equal to the coefficients of other predictors in Group 2. Now, suppose that we extract an ordered segmentation from the OLS coefficients by taking $\delta = 0.3$; see Figure 1. Since it allows for arbitrary order reshuffling within the segment B_4, this ordered segmentation preserves the order of $\boldsymbol{\beta}^0$, that is, Definition 3 is satisfied. The hybrid pairwise penalty associated with this segmentation includes terms

$$p_{\lambda_1}\big(|\beta_{\tau(17)} - \beta_{\tau(23)}|\big) + p_{\lambda_1}\big(|\beta_{\tau(18)} - \beta_{\tau(23)}|\big)$$

between segments B_4 and B_5. So aCARDS will shrink the coefficients of $\tau(17)$ and $\tau(18)$ toward being equal to the coefficient of $\tau(23)$, a predictor in Group 2. Moreover, there are terms such as

$$p_{\lambda_1}\big(|\beta_{\tau(23)} - \beta_{\tau(24)}|\big) + p_{\lambda_1}\big(|\beta_{\tau(23)} - \beta_{\tau(25)}|\big) \\ + p_{\lambda_1}\big(|\beta_{\tau(23)} - \beta_{\tau(26)}|\big)$$

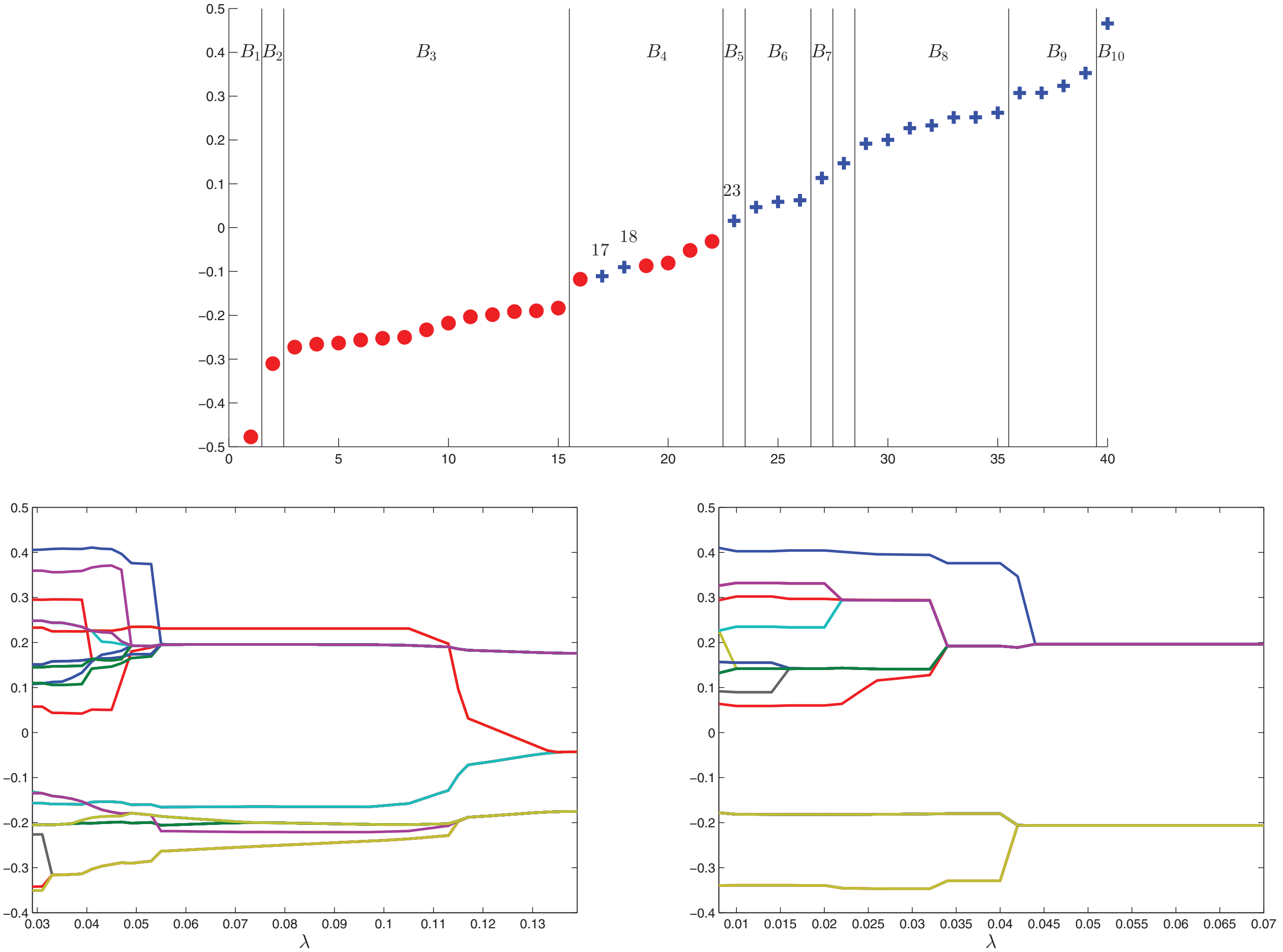


Figure 1. Illustration of the hybrid pairwise penalty and the aCARDS algorithm. Top panel: OLS coefficients and the associated ordered segmentation. Red dots and blue crosses represent predictors from Group 1 and Group 2, respectively. Bottom panel: Solution paths of bCARDS (left) and aCARDS (right) under misranking. The ranking and ordered segmentation are the same as in the top panel. For bCARDS, the horizontal axis represents the parameter λ. For aCARDS, the horizontal axis represents the between-segment parameter λ_1, where we fix the within-segment parameter $\lambda_2 = 0.02$. The vertical axis represents the estimated 40 regression coefficients, which are shrunk toward homogeneity (as the figures do not start from the smallest λ, we do not see initial 40 regression coefficients).

between segments B_5 and B_6. So aCARDS will also shrink the coefficient of $\tau(23)$ toward being equal to the coefficients of other predictors in Group 2. Eventually, aCARDS will shrink the coefficients of $\tau(17)$ and $\tau(18)$ toward being equal to the coefficients of many other predictors in Group 2. This example explains how the ordered segmentation and hybrid penalty help overcome issues caused by misranking in the preliminary estimator.

To better illustrate the effects of fused penalty and hybrid penalty under misranking, we fix the estimated rank and ordered segmentation from above, and compute the solution paths of both bCARDS and aCARDS. Note that the penalty terms in both (5) and (8) are now fixed (hence we do not need the parameter δ in aCARDS). For bCARDS, we let λ vary. For aCARDS, we set the within-segment parameter $\lambda_2 = 0.02$ and let the between-segment parameter λ_1 vary. Figure 1 displays the solution paths. We see that although bCARDS does not include the true grouping in the solution path owing to misranking, aCARDS still achieves the true grouping, which is a benefit of the hybrid penalty.

In practical data analysis, we need not differentiate between two versions of CARDS, but the tuning parameter selection process automatically tells us which version to use. This is because bCARDS is a special case of aCARDS with $\delta = 0$. We only need to include 0 in the candidates of the parameter δ and select δ in a data-driven manner (e.g., AIC, BIC, and GCV). We call the resulting method CARDS, which involves a data-driven selection between bCARDS and aCARDS.

2.4 CARDS Under Sparsity

In applications, we may need to explore homogeneity and sparsity simultaneously. Often the preliminary estimator $\widetilde{\boldsymbol{\beta}}$ takes into account the sparsity, namely it is obtained with a penalized least-squares method (Fan and Li 2001; Tibshirani et al. 2005) or sure independence screening (Fan and Lv 2008). Suppose $\widetilde{\boldsymbol{\beta}}$ has the sure screening property, that is, $S_0 \subset \widetilde{S}$ with high probability, where $\widetilde{S}$ and S_0 denote the support of $\widetilde{\boldsymbol\beta}$ and $\boldsymbol\beta^0$, respectively. We modify CARDS as follows: in the first two steps, using the nonzero elements of $\widetilde{\boldsymbol\beta}$, we can similarly construct a data-driven hybrid penalty only on coefficients of variables in $\widetilde{S}$; in the third step, we fix $\widehat{\boldsymbol\beta}_{\widetilde{S}^c} = \mathbf{0}$ and obtain $\widehat{\boldsymbol\beta}_{\widetilde{S}}$ by minimizing the following penalized least squares

$$Q_n^{\text{sparse}}(\boldsymbol\beta) = \frac{1}{2n}\|\boldsymbol y - \mathbf{X}_{\widetilde{S}}\boldsymbol\beta_{\widetilde{S}}\|^2 + P_{\Upsilon,\lambda_1,\lambda_2}(\boldsymbol\beta_{\widetilde{S}}) + \sum_{j\in\widetilde{S}} p_\lambda(|\beta_j|), \tag{11}$$

where $\mathbf{X}_{\widetilde{S}}$ is the submatrix of $\mathbf{X}$ restricted to columns in $\widetilde{S}$. In (11), the second term is the hybrid penalty to encourage homogeneity among coefficients of variables already selected in $\widetilde{\boldsymbol\beta}$, and the third term is the element-wise penalty to help further filter out falsely selected variables. We call this modified version shrinkage CARDS (sCARDS).

3. ANALYSIS OF THE BASIC CARDS

In this section, we analyze theoretical properties of bCARDS. Due to the limited space, we state the results here and only prove Theorems 1–3 in the Appendix, leaving the rest of the proofs to the online supplemental materials of this article.

3.1 Heuristics

We first provide some heuristics on why taking advantage of the homogeneity helps reduce the estimation error $\|\widehat{\boldsymbol\beta} - \boldsymbol\beta^0\|$. Consider an ideal case of orthogonal design $\mathbf{X}^T\mathbf{X} = n\mathbf{I}_p$ (necessarily $p \le n$). The ordinary least-squares estimator $\widehat{\boldsymbol\beta}^{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol y$ has the decomposition

$$\widehat{\beta}_j^{\text{ols}} = \beta_j^0 + z_j, \quad z_j \overset{\text{iid}}{\sim} N(0, n^{-1}), \quad j = 1,\ldots,p.$$

It is clear by the square-root law that $\|\widehat{\boldsymbol\beta}^{\text{ols}} - \boldsymbol\beta^0\| = O_P(\sqrt{p/n})$. Now, if there are K homogeneous groups in $\boldsymbol\beta^0$ and that we know the true groups, the original model (1) can be rewritten as

$$\boldsymbol y = \mathbf{X}_A\boldsymbol\beta_A^0 + \boldsymbol\varepsilon,$$

where $\boldsymbol\beta_A^0 = (\beta_{A,1}^0,\ldots,\beta_{A,K}^0)^T$ contains distinct values in $\boldsymbol\beta^0$, and $\mathbf{X}_A = (\mathbf{x}_{A,1},\ldots,\mathbf{x}_{A,K})$ is such that $\mathbf{x}_{A,k} = \sum_{j\in A_k}\mathbf{x}_j$. The corresponding ordinary least-squares estimator $\widehat{\boldsymbol\beta}_A^{\text{ols}} = (\mathbf{X}_A^T\mathbf{X}_A)^{-1}\mathbf{X}_A^T\boldsymbol y$ has the decomposition

$$\widehat{\beta}_{A,k}^{\text{ols}} = \beta_{A,k}^0 + \bar z_k, \quad \bar z_k\text{'s are independent,} \qquad \bar z_k \sim N(0, n^{-1}|A_k|^{-1}). \tag{12}$$

Here $\bar z_k = \frac{1}{|A_k|}\sum_{j\in A_k} z_j$ is the noise averaged over group k. The oracle estimator $\widehat{\boldsymbol\beta}^{\text{oracle}}$ is defined such that $\widehat{\beta}_j^{\text{oracle}} = \widehat{\beta}_{A,k}^{\text{ols}}$ for all $j \in A_k$. Then, by the square-root law,

$$\|\widehat{\boldsymbol\beta}^{\text{oracle}} - \boldsymbol\beta^0\|^2 = \sum_{k=1}^K |A_k||\widehat{\beta}_{A,k}^{\text{ols}} - \beta_{A,k}^0|^2 = O_p\Big(\sum_{k=1}^K |A_k|\cdot n^{-1}|A_k|^{-1}\Big) = O_p\big(K/n\big),$$

which immediately implies that $\|\widehat{\boldsymbol\beta}^{\text{oracle}} - \boldsymbol\beta^0\| = O_p(\sqrt{K/n})$.

The surprises of the results are two-fold. First, $\|\widehat{\boldsymbol\beta}^{\text{oracle}} - \boldsymbol\beta^0\|$ has the convergence rate $\sqrt{K/n}$ instead of $\sqrt{p/n}$. The point is that in (12) the noises are averaged, thanks to exploiting homogeneity, and consequently $\beta_{A,k}^0$ is estimated more accurately. The second surprise is that the rate has nothing to do with the sizes of true homogeneous groups. No matter whether we have K groups of equal size, or one dominating group with other $(K-1)$ small groups, the rate is always the same for the oracle estimator.

3.2 Notations and Regularity Conditions

Let $\mathcal{M}_A$ be the subspace of $\mathbb{R}^p$ defined by

$$\mathcal{M}_A = \{\boldsymbol\beta \in \mathbb{R}^p : \beta_i = \beta_j, \text{ for any } i, j \in A_k, 1 \le k \le K\}.$$

For each $\boldsymbol\beta \in \mathcal{M}_A$, we can always write $\mathbf{X}\boldsymbol\beta = \mathbf{X}_A\boldsymbol\beta_A$, where $\mathbf{X}_A = (\mathbf{x}_{A,1},\ldots,\mathbf{x}_{A,K})$ is an $n\times K$ matrix such that its kth column $\mathbf{x}_{A,k} = \sum_{j\in A_k}\mathbf{x}_j$, and $\boldsymbol\beta_A$ is a K-dimensional vector with its kth component $\beta_{A,k}$ being the common coefficient in group A_k. Define the matrix $\mathbf{D} = \text{diag}(|A_1|^{1/2},\ldots,|A_K|^{1/2})$. We introduce the following conditions on the design matrix $\mathbf{X}$:

Condition 1. $\|\mathbf{x}_j\| = \sqrt{n}$, for $1 \le j \le p$. The eigenvalues of the $K\times K$ matrix $\frac{1}{n}\mathbf{D}^{-1}\mathbf{X}_A^T\mathbf{X}_A\mathbf{D}^{-1}$ are bounded below by $c_1 > 0$ and bounded above by $c_2 > 0$.

In the case of orthogonal designs, that is, $\mathbf{X}^T\mathbf{X} = n\mathbf{I}_p$, the matrix $\frac{1}{n}\mathbf{D}^{-1}\mathbf{X}_A^T\mathbf{X}_A\mathbf{D}^{-1}$ simplifies to $\mathbf{I}_K$, and $c_1 = c_2 = 1$.

Let $\rho(t) = \lambda^{-1}p_\lambda(t)$ and $\bar\rho(t) = \rho'(|t|)\text{sgn}(t)$. We assume that the penalty function $p_\lambda(\cdot)$ satisfies the following condition.

Condition 2. $p_\lambda(\cdot)$ is a symmetric function and it is nondecreasing and concave on $[0,\infty)$. $\rho'(t)$ exists and is continuous except for a finite number of t, and $\rho'(0+) = 1$. There exists a constant $a > 0$ such that $\rho(t)$ is a constant for all $|t| \ge a\lambda$.

We also assume that the noise vector $\boldsymbol\varepsilon = (\epsilon_1,\ldots,\epsilon_n)^T$ has sub-Gaussian tails.

Condition 3. For any vector $\mathbf{a} \in \mathbb{R}^n$ and $x > 0$, $P(|\mathbf{a}^T\boldsymbol\varepsilon| > \|\mathbf{a}\|x) \le 2e^{-c_3x^2}$, where c_3 is a positive constant.

Given the design matrix $\mathbf{X}$, let $\mathbf{X}_k$ be its submatrix formed by including only columns in A_k, for $1 \le k \le K$. For any vector $\mathbf{v} \in \mathbb{R}^q$, let $\text{DC}(\mathbf{v}) = \max_{1\le i\le q}|v_i - q^{-1}\sum_{j=1}^q v_j|$ be the "deviation from centrality." Define

$$\sigma_k = \lambda_{\max}\big(\tfrac{1}{n}\mathbf{X}_k^T\mathbf{X}_k\big) \quad\text{and}\quad \nu_k = \max_{\boldsymbol\mu\in\mathcal{M}_A:\|\boldsymbol\mu\|=1}\text{DC}\left(\tfrac{1}{n}\mathbf{X}_k^T\mathbf{X}\boldsymbol\mu\right), \tag{13}$$

where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue operator. In the case of orthogonal design, $\sigma_k = 1$ and $\nu_k = 0$. Let

$$b_n = \frac{1}{2}\min_{1\le k<l\le K}|\beta_{A,k}^0 - \beta_{A,l}^0|$$

be half of the minimum gap between groups in $\boldsymbol\beta^0$, and $\lambda = \lambda_n$ the tuning parameter in the penalty function.

3.3 Properties of bCARDS

When the true groups $A_1,\ldots,A_K$ are known, the oracle estimator is

$$\widehat{\boldsymbol\beta}^{\text{oracle}} = \arg\min_{\boldsymbol\beta\in\mathcal{M}_A}\left\{\frac{1}{2n}\|\boldsymbol y - \mathbf{X}\boldsymbol\beta\|^2\right\}.$$

Theorem 1. Suppose Conditions 1–3 hold, $K = o(n)$, and the preliminary estimator $\widetilde{\boldsymbol{\beta}}$ generates a rank mapping τ that is consistent with the order of $\boldsymbol{\beta}^0$, that is, (6) holds, with probability at least $1 - \epsilon_0$. If the half minimum gap between groups, b_n, satisfies that $b_n > a\lambda_n$, where a is the same as that in Condition 2, and

$$\lambda_n \gg \max_k \left\{ \sqrt{\sigma_k |A_k| \log(n \vee p)/n} + (1 + \nu_k |A_k|^{1/2}) \times \sqrt{K \log(n)/n} \right\}, \tag{14}$$

then with probability at least $1 - \epsilon_0 - n^{-1}K - (n \vee p)^{-1}$, the bCARDS objective function (5) has a strictly local minimizer $\widehat{\boldsymbol{\beta}}$ such that

- $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{\text{oracle}}$,
- $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| = O_p(\sqrt{K/n})$.

Theorem 1 shows that bCARDS includes the oracle estimator as a strictly local minimizer, with overwhelming probability. This strong oracle property is a stronger result than the oracle property in Fan and Li (2001).

The objective function (5) in bCARDS is nonconvex and may have multiple local minimizers. In practice, we apply the local linear approximation (LLA) algorithm (Zou and Li 2008) to solve it: start from an initial solution $\widehat{\boldsymbol{\beta}}^{(0)} = \widehat{\boldsymbol{\beta}}^{\text{initial}}$; at step m, update the solution by

$$\widehat{\boldsymbol{\beta}}^{(m)} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p-1} p'_\lambda\left(|\hat{\beta}^{(m-1)}_{\tau(j+1)} - \hat{\beta}^{(m-1)}_{\tau(j)}|\right) \cdot |\beta_{\tau(j+1)} - \beta_{\tau(j)}| \right\}.$$

Given $\widehat{\boldsymbol{\beta}}^{\text{initial}}$, this algorithm produces a unique sequence of estimators which converge to a certain local minimizer. Next, Theorem 2 shows that under certain conditions, the sequence of estimators produced by the LLA algorithm converge to the oracle estimator.

Theorem 2. Under conditions of Theorem 1, suppose $\rho'(\lambda_n) \geq a_0$ for some constant $a_0 > 0$, and that there exists an initial solution $\widehat{\boldsymbol{\beta}}^{\text{initial}}$ of (5) satisfying $\|\widehat{\boldsymbol{\beta}}^{\text{initial}} - \boldsymbol{\beta}^0\|_\infty \leq \lambda_n/2$. Then with probability at least $1 - \epsilon_0 - n^{-1}K - (n \vee p)^{-1}$, the LLA algorithm yields $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$ after one iteration, and it converges to $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$ after two iterations.

From Theorems 1 and 2, we conclude that bCARDS combined with the LLA algorithm yields the oracle estimator with overwhelming probability, provided that we have a good preliminary estimator $\widetilde{\boldsymbol{\beta}}$. Next, we discuss the choice of $\widetilde{\boldsymbol{\beta}}$.

Since we focus on dense problems in this section, the usual sparsity is not explicitly explored and the ordinary least squares estimator

$$\widehat{\boldsymbol{\beta}}^{\text{ols}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \right\}$$

can be used as the preliminary estimator. The following theorem shows that it induces a rank-consistent mapping with high probability.

Theorem 3. Suppose Condition 3 holds, $p = O(n^\alpha)$ and $\lambda_{\min}(\frac{1}{n}\mathbf{X}^T\mathbf{X}) \geq c_4$, where $0 < \alpha < 1$ and $c_4 > 0$ are constants. If $b_n > \sqrt{(2\alpha c_4/c_3)\log(n)/n}$, where b_n is the half of the minimum gap between groups in $\boldsymbol{\beta}^0$, then with probability at least $1 - O(n^{-\alpha})$, the rank mapping τ generated from $\widehat{\boldsymbol{\beta}}^{\text{ols}}$ is consistent with the order of $\boldsymbol{\beta}^0$.

When the rank mapping τ extracted from $\widetilde{\boldsymbol{\beta}}$ does not give a consistent order, that is, (6) does not hold, the penalty in (5) is no longer a "correct" penalty for promoting the true grouping structure. There is no hope that local minimizers of (5) exactly recover the true groups. However, if there is not too much misranking in τ, it is still possible to control $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|$. Given a rank mapping τ, define

$$K^*(\tau) = \sum_{j=1}^{p-1} 1\left\{\beta^0_{\tau(j)} \neq \beta^0_{\tau(j+1)}\right\}.$$

It is the number of coefficient "jumps" in $\boldsymbol{\beta}^0$ under the order given by τ. These "jumps" define subgroups $A'_1, A'_2, \ldots, A'_{K^*}$, each being a subset of one true group. Although different subgroups may share the same true coefficient, any two consecutive subgroups A'_k and A'_{k+1} have a gap in coefficient values. As a result, the above results apply to this *subgrouping structure*. The following theorem is a generalization and a direct application of the proof of Theorem 1 and its details are omitted.

Theorem 4. Suppose Conditions 1–3 hold, $K^*(\tau) = o(n)$, the half minimum gap $b_n > a\lambda_n$, and λ_n satisfies (14) with K replaced by $K^*(\tau)$. Then with probability at least $1 - \epsilon_0 - n^{-1}K - (n \vee p)^{-1}$, the bCARDS objective function (5) has a strictly local minimizer $\widehat{\boldsymbol{\beta}}$ such that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| = O_p(\sqrt{K^*(\tau)/n})$.

3.4 bCARDS With the L_1 Penalty

In the bCARDS formulation (5), $\rho(t)$ can also be the L_1 penalty function $\rho(t) = |t|$. It can be useful to get the initial solution $\widehat{\boldsymbol{\beta}}^{\text{initial}}$ for the LLA algorithm. However, $\rho(t) = |t|$ does not satisfy Condition 2. Hence, Theorem 1 does not apply and its associated properties requires additional studies.

We first relax the requirement that τ is consistent with the order of $\boldsymbol{\beta}^0$. Instead, we consider the case that "τ is consistent with groups in $\boldsymbol{\beta}^0$": there exists a permutation μ on $\{1, \ldots, K\}$ and $1 = i_1 < i_2 < \cdots < i_K < i_{K+1} = p + 1$ such that for $k = 1, \ldots, K$,

$$\beta^0_{\tau(i)} = \beta^0_{A,\mu(k)}, \quad i_k \leq i \leq i_{k+1} - 1. \tag{15}$$

When μ is the identical permutation, that is, $\mu(k) = k$, (15) is equivalent to (6) and τ is consistent with the order of $\boldsymbol{\beta}^0$. Under the condition (15), recovering the true groups is equivalent to locating coefficient jumps in $\boldsymbol{\beta}^0$ under the order given by τ, and these jumps can have positive or negative values.

To guarantee the exact recovery of jumps, we need a joint condition on $\mathbf{X}$ and $\boldsymbol{\beta}^0$, it is in the same spirit of the "irrepresentability" condition in Zhao and Yu (2006) but is specifically designed for the homogeneity setting. For notation simplicity, we change the indices of groups to let $\mu(k) = k$ for all k. Note that $\beta^0_1 < \beta^0_2 < \cdots < \beta^0_K$ does not hold with these new group indices.

For $k = 1, \ldots, K - 1$, write $d\beta^0_{A,k} = \beta^0_{A,k+1} - \beta^0_{A,k}$. Define the K-dimensional vector $\mathbf{d}_0$ by $d^0_1 = \text{sgn}(d\beta^0_{A,1})$, $d^0_K =$

$-\mathrm{sgn}(d\beta^0_{A,K-1})$ and

$$d^0_k = \mathrm{sgn}(d\beta^0_{A,k}) - \mathrm{sgn}(d\beta^0_{A,k-1}), \quad 2 \le k \le K-1.$$

Here $\mathbf{d}^0$ is the adjacent difference of the sign vector of jumps in $\boldsymbol{\beta}^0$. For example, suppose $K=4$ and the common coefficients in four groups satisfy $\beta^0_{A,2} - \beta^0_{A,1} > 0$, $\beta^0_{A,3} - \beta^0_{A,2} < 0$ and $\beta^0_{A,4} - \beta^0_{A,3} > 0$. Then $\mathbf{d}^0 = (1,-2,2,-1)$. Also, define the p-dimensional vector

$$\mathbf{b}^0 = \mathbf{X}^T\mathbf{X}_A(\mathbf{X}_A^T\mathbf{X}_A)^{-1}\mathbf{d}^0.$$

In the case of orthogonal design $\mathbf{X}^T\mathbf{X} = n\mathbf{I}_p$, $\mathbf{b}^0 \in \mathcal{M}_A$ and it has the form $b^0_j = d^0_k/|A_k|$ for $j \in A_k$. For each $j \in A_k$, let

$$A^1_{kj} = \{\tau(i) \in A_k : i \le j\}, \quad A^2_{kj} = \{\tau(i) \in A_k : i > j\}.$$

Namely, A^1_{kj} and A^2_{kj} contain indices in group k that are ranked above and below $\tau(j)$, respectively. Let $\theta_{kj} = |A^1_{kj}|/|A_k|$ be the proportion of indices in group k that are ranked above $\tau(j)$. Denote by $\overline{b}_{kj} = \frac{1}{|A^1_{kj}|}\sum_{\tau(i)\in A^1_{kj}} b^0_{\tau(i)}$ the average of elements in $\mathbf{b}^0$ over the indices in A^1_{kj}, and $\underline{b}_{kj} = \frac{1}{|A^2_{kj}|}\sum_{\tau(i)\in A^2_{kj}} b^0_{\tau(i)}$ the average over the indices in A^2_{kj}.

Condition 4. There exists a positive sequence $\{\omega_n\}$, which can go to 0, such that for $1 \le k \le K$, $j \in A_k$ and $j \ne j_{k+1} - 1$,

$$1 - \omega_n \ge \begin{cases} \left|\theta_{1j}\mathrm{sgn}(d\beta^0_{A,1}) + |A_1|^2\theta_{1j}(1-\theta_{1j})\left(\overline{b}_{1j} - \underline{b}_{1j}\right)\right|, \\ \left|(1-\theta_{kj})\mathrm{sgn}(d\beta^0_{A,k-1}) + \theta_{kj}\mathrm{sgn}(d\beta^0_{A,k}) \right. \\ \quad \left. + |A_k|^2\theta_{kj}(1-\theta_{kj})\left(\overline{b}_{kj} - \underline{b}_{kj}\right)\right|, \quad 2 \le k \le K-1, \\ \left|(1-\theta_{Kj})\mathrm{sgn}(d\beta^0_{A,K-1}) + |A_K|^2\theta_{Kj} \right. \\ \quad \left. (1-\theta_{Kj})\left(\overline{b}_{Kj} - \underline{b}_{Kj}\right)\right|. \end{cases} \tag{16}$$

In the case of orthogonal design $\mathbf{X}^T\mathbf{X} = n\mathbf{I}_p$, $\mathbf{b}^0 \in \mathcal{M}_A$ and $\overline{b}_{kj} - \underline{b}_{kj} = 0$ holds for all k and $j \in A_k$. Condition 4 reduces to

$$1 - \omega_n \ge \begin{cases} \left|\theta_{1j}\mathrm{sgn}(d\beta^0_{A,1})\right|, \\ \left|(1-\theta_{kj})\mathrm{sgn}(d\beta^0_{A,k-1}) + \theta_{kj}\mathrm{sgn}(d\beta^0_{A,k})\right|, \quad 2 \le k \le K-1, \\ \left|(1-\theta_{Kj})\mathrm{sgn}(d\beta^0_{A,K-1})\right|. \end{cases}$$

This is possible only when

$$\mathrm{sgn}(d\beta^0_{A,k-1}) \ne \mathrm{sgn}(d\beta^0_{A,k}), \quad 2 \le k \le K-1. \tag{17}$$

Noting that $1/|A_k| \le \theta_{kj} \le 1 - 1/|A_k|$, the associated ω_n can be chosen as $\min_k\{1/|A_k|\}$ when (17) holds.

Theorem 5. Suppose Conditions 1, 3, and 4 hold, $K = o(n)$, and the preliminary estimator $\widetilde{\boldsymbol{\beta}}$ generates an order τ that is consistent with groups in $\boldsymbol{\beta}^0$, that is, (15) holds, with probability at least $1 - \epsilon_0$. If the half minimum gap b_n and the tuning parameter λ_n satisfy

$$b_n \gg \sqrt{K\log(n)/n} + \lambda_n\left(\sum_{k=1}^{K}\frac{1}{|A_k|^2}\right)^{1/2}$$

and

$$\lambda_n \gg \omega_n^{-1}\max_k\left\{\sqrt{\sigma_k|A_k|\log(n\vee p)/n}\right\}, \tag{18}$$

then with probability at least $1 - \epsilon_0 - n^{-1}K - (n\vee p)^{-1}$, the bCARDS objective function (5) with $\rho(t) = |t|$ has a unique global minimizer $\widehat{\boldsymbol{\beta}}$ such that

- $\widehat{\boldsymbol{\beta}} \in \mathcal{M}_A$;
- $\mathrm{sgn}(\widehat{\beta}_{A,k+1} - \widehat{\beta}_{A,k}) = \mathrm{sgn}(\beta^0_{A,k+1} - \beta^0_{A,k})$, $\quad k = 1, \ldots, K-1$;
- $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| = O_p(\sqrt{K/n} + \gamma_n)$, where $\gamma_n = \lambda_n(\sum_{k=1}^K \frac{1}{|A_k|})^{1/2}$.

Compared to Theorem 1, there is an extra bias term in the estimation error $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|$, which is of order $\sqrt{K\log(n\vee p)/n}$. Moreover, to achieve the exact recovery, it requires Condition 4, which is restrictive. For example, in the case of orthogonal designs, it is required that all consecutive jumps (under the order given by τ) have oppositive signs.

4. ANALYSIS OF THE ADVANCED CARDS

In this section, we analyze aCARDS and its variant sCARDS. The proofs can be found in the online supplemental materials.

4.1 Properties of aCARDS

To guarantee the success of aCARDS, a key condition is that the ordered segmentation $\Upsilon = \{B_1, \ldots, B_L\}$ defined in (7) preserves the order of $\boldsymbol{\beta}^0$ in the sense of Definition 3. This allows the ranking of coefficients in $\widetilde{\boldsymbol{\beta}}$ to deviate from that in $\boldsymbol{\beta}^0$, but not too much: for some $\delta > 0$, whenever $\beta^0_i < \beta^0_j$, $\widetilde{\beta}_i \le \widetilde{\beta}_j + \delta$ must hold.

For given $A_1, \ldots, A_K$ and a segmentation $\Upsilon = \{B_1, \ldots, B_L\}$, define

$$\phi_k = |A_k| / \min\left\{|A_k|^3, \min_{l: B_l\cap A_k \ne \emptyset}\{|B_l|^2\}\right\}.$$

Here $1/|A_k|^2 \le \phi_k \le |A_k|$ for $1 \le k \le K$.

Theorem 6. Suppose Conditions 1–3 hold, $K = o(n)$, and the preliminary estimator $\widetilde{\boldsymbol{\beta}}$ and the tuning parameter δ_n together generate an ordered segmentation Υ that preserves the order of $\boldsymbol{\beta}^0$, with probability at least $1 - \epsilon_0$. If the half minimum gap b_n and the tuning parameters $(\lambda_{1n}, \lambda_{2n})$ in (10) satisfy that $b_n > a\max\{\lambda_{1n}, \lambda_{2n}\}$, where a is the same as that in Condition 2, and

$$\lambda_{1n} \gg \max_k\left\{\sqrt{\sigma_k\phi_k\log(n\vee p)/n} + (1 + \nu_k\phi_k^{1/2}) \times \sqrt{K\log(n)/n}\right\}, \tag{19}$$

and

$$\lambda_{2n} \gg \max_k\left\{|A_k|^{-1}\sqrt{\log(n\vee p)/n} + (1 + \nu_k|A_k|^{-1}) \times \sqrt{K\log(n)/n}\right\}, \tag{20}$$

then with probability at least $1 - \epsilon_0 - n^{-1}K - 2(n\vee p)^{-1}$, the aCARDS objective function (10) has a strictly local minimizer $\widehat{\boldsymbol{\beta}}$ such that

- $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{\text{oracle}}$,
- $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| = O_p(\sqrt{K/n})$.

Compared to Theorem 1, aCARDS not only imposes less restrictive conditions on $\widetilde{\boldsymbol{\beta}}$, but also requires a smaller minimum gap between true coefficients.

Next, we establish the asymptotic normality of the CARDS estimator. By Theorem 6, with overwhelming probability, aCARDS performs just like the oracle. In the oracle situation, for example, if $p = 5$ and there are three true groups $\{\beta_1, \beta_4\}$, $\{\beta_2\}$, and $\{\beta_3, \beta_5\}$, the accuracy of estimating $\boldsymbol{\beta}$ is the same as if we know the model:

$$Y = \beta_1(X_1 + X_4) + \beta_2 X_2 + \beta_3(X_3 + X_5) + \varepsilon.$$

Theorem 7. Given a positive integer q, let $\{\mathbf{B}_n\}$ be a sequence of matrices such that $\mathbf{B}_n \in \mathbb{R}^{q\times K}$, $\max_{\mathbf{v}\in\mathbb{R}^q:\|\mathbf{v}\|=1} \|\mathbf{X}_A^T(\mathbf{X}_A^T\mathbf{X}_A)^{-1}\mathbf{B}_n^T\mathbf{v}\|_4 = o(1)$, and $\mathbf{B}_n\mathbf{B}_n^T \to \mathbf{H}$, where $\mathbf{H}$ is a fixed $q \times q$ positive definite matrix. Suppose conditions of Theorem 6 hold and let $\widehat{\boldsymbol{\beta}}$ be the local minimizer of the aCARDS objective function (10) given in Theorem 6. Then

$$\mathbf{B}_n(\mathbf{X}_A^T\mathbf{X}_A)^{1/2}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_A^0) \xrightarrow{d} N(\mathbf{0}, \mathbf{H}),$$

where $\widehat{\boldsymbol{\beta}}_A$ is the K-dimensional vector of distinct values in $\widehat{\boldsymbol{\beta}}$.

Theorem 7 states the asymptotic normality of $\widehat{\boldsymbol{\beta}}_A$. Note that $\widehat{\boldsymbol{\beta}}$ duplicates elements in $\widehat{\boldsymbol{\beta}}_A$. We introduce the following corollary to compare the asymptotic covariance of $\widehat{\boldsymbol{\beta}}$ to that of $\widehat{\boldsymbol{\beta}}^{\text{ols}}$.

Corollary 1 Suppose conditions of Theorems 6 and 7 hold. Let $\widehat{\boldsymbol{\beta}}^{\text{ols}}$ and $\widehat{\boldsymbol{\beta}}$ be the ordinary least-squares estimator and CARDS estimator, respectively. Let $\mathbf{M}_n$ be the $p \times K$ matrix with $M_n(j,k) = (1/|A_k|^{1/2})1\{j \in A_k\}$. For any sequence of p-dimensional vectors $\{\mathbf{a}_n\}$,

- $v_{1n}^{-1/2}\mathbf{a}_n^T(\widehat{\boldsymbol{\beta}}^{\text{ols}} - \boldsymbol{\beta}^0) \xrightarrow{d} N(0,1)$ with $v_{1n} = \mathbf{a}_n^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}_n$;
- $v_{2n}^{-1/2}\mathbf{a}_n^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \xrightarrow{d} N(0,1)$ with $v_{2n} = \mathbf{a}_n^T\mathbf{M}_n(\mathbf{M}_n^T\mathbf{X}^T\mathbf{X}\mathbf{M}_n)^{-1}\mathbf{M}_n^T\mathbf{a}_n$.

Moreover, $v_{1n} \geq v_{2n}$.

4.2 Properties of sCARDS

In Section 2.4, we introduced sCARDS to explore both homogeneity and sparsity. In sCARDS, given a preliminary estimator $\widetilde{\boldsymbol{\beta}}$ and a parameter δ, we extract segments $B_1, \ldots, B_L$ such that $\cup_{l=1}^L B_l = \widetilde{S}$, where $\widetilde{S}$ is the support of $\widetilde{\boldsymbol{\beta}}$. Denote $B_0 = \{j : \widetilde{\beta}_j = 0\}$. In this case, we say $\Upsilon = \{B_0, B_1, \ldots, B_L\}$ preserves the order of $\boldsymbol{\beta}^0$ if

$$\max_{j\in B_0}|\beta_j^0| = 0 \quad \text{and} \quad \max_{j\in B_l}\beta_j^0 \leq \min_{j\in B_{l+1}}\beta_j^0, \qquad l = 1, \ldots, L-1. \tag{21}$$

This implies that $\widetilde{\boldsymbol{\beta}}$ has the sure screening property, and on those preliminarily selected variables, the data-driven segments preserve the order of true coefficients.

Suppose there is a group of zero coefficients in $\boldsymbol{\beta}^0$, namely, $\mathcal{A} = (A_0, A_1, \ldots, A_K)$. Let $\mathcal{M}_{\mathcal{A}}^*$ be the subspace of $\mathbb{R}^p$ defined by

$$\mathcal{M}_{\mathcal{A}}^* = \{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_i = 0, \text{ for any } i \in A_0;\ \beta_i = \beta_j, \\ \text{for any } i, j \in A_k, 1 \leq k \leq K\}.$$

The oracle estimator is

$$\widehat{\boldsymbol{\beta}}^{\text{oracle}} = \arg\min_{\beta\in\mathcal{M}_{\mathcal{A}}^*}\left\{\frac{1}{2n}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2\right\}.$$

Denote by S the support of $\boldsymbol{\beta}^0$, and write $s = |S|$ and $\tilde{s} = |\widetilde{S}|$. Define

$$b_n' = \frac{1}{2}\min\{|\beta_j^0| : \beta_j^0 \neq 0\}$$

to be the half minimum signal strength. For any matrix $\mathbf{M}$, $||\mathbf{M}||_{2,\infty} = \max_{||\mathbf{v}||=1}||\mathbf{M}\mathbf{v}||_\infty$.

Theorem 8. Suppose Conditions 1–3 hold, $s = o(n)$, $\log(p) = o(n)$, and the preliminary estimator $\widetilde{\boldsymbol{\beta}}$ and the tuning parameter δ_n together generate an ordered segmentation Υ that preserves the order of $\boldsymbol{\beta}^0$, that is, (21) holds, with probability at least $1 - \epsilon_0$. If b_n', b_n and the tuning parameters $(\lambda_{1n}, \lambda_{2n}, \lambda_n)$ satisfy that $b_n' = a\lambda_n$, $b_n > a\max\{\lambda_{1n}, \lambda_{2n}\}$ and

$$\lambda_n \gg \sqrt{\log(n\vee\tilde{s})/n} + \|\mathbf{X}_{\widetilde{S}}^T\mathbf{X}_S\|_{2,\infty}\sqrt{K\log(n)/n},$$

$$\lambda_{1n} \gg \max_k\left\{\sqrt{\sigma_k\phi_k\log(n\vee s)/n} + (1+\nu_k\phi_k^{1/2})\sqrt{K\log(n)/n}\right\},$$

$$\lambda_{2n} \gg \max_k\left\{|A_k|^{-1}\sqrt{\log(n\vee s)/n} + (1+\nu_k|A_k|^{-1}) \times\sqrt{K\log(n)/n}\right\}.$$

Then with probability at least $1 - \epsilon_0 - n^{-1}K - (n\vee s)^{-1} - (n\vee\tilde{s})^{-1}$, the sCARDS objective function (11) has a strictly local minimizer $\widehat{\boldsymbol{\beta}}$ such that

- $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{\text{oracle}}$,
- $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| = O_p(\sqrt{K/n})$.

The preliminary estimator $\widetilde{\boldsymbol{\beta}}$ can be chosen, for example, as the SCAD estimator

$$\widehat{\boldsymbol{\beta}}^{\text{scad}} = \arg\min\left\{\frac{1}{2n}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^p p_{\lambda'}(|\beta_j|)\right\}, \tag{22}$$

where $p_{\lambda'}(\cdot)$ is the SCAD penalty function (Fan and Li 2001). The following theorem is a direct result of Theorem 2 in Fan and Lv (2011), and the proof is omitted.

Theorem 9. Under Conditions 1 and 3, if $s = o(n)$, $\lambda_n' \gg n^{-1/2}[\log(n)]^2$ and $b_n' \gg n^{-1/2}\max\{\sqrt{\log p}, \|\frac{1}{n}\mathbf{X}_{S^c}^T\mathbf{X}_S\|_\infty\sqrt{\log n}\}$, then with probability at least $1 - o(1)$, there exists a strictly local minimizer $\widehat{\boldsymbol{\beta}}^{\text{scad}}$ and $\delta_n = O(\log(n)/n)$ which together generate a segmentation preserving the order of $\boldsymbol{\beta}^0$.

5. SIMULATION STUDIES

We conduct numerical experiments to study the performance of two versions of CARDS, bCARDS, and aCARDS, and their variant sCARDS. The goal is to investigate the performance of CARDS under different situations. Experiments 1–4 are based on the linear regression model $Y_i = \mathbf{X}_i^T\boldsymbol{\beta}^0 + \epsilon_i$, with Experiments 1–3 exploring the homogeneity only and Experiment 4 exploring the homogeneity and sparsity simultaneously. Experiment 5 is based on the spatial-temporal model $Y_{it} = \mathbf{X}_t^T\boldsymbol{\beta}_i^0 + \epsilon_{it}$.

In all experiments, $\{\mathbf{X}_i : 1 \leq i \leq n\}$ or $\{\mathbf{X}_t : 1 \leq t \leq T\}$ are generated independently and identically from the multivariate standard Gaussian distributions, and $\{\epsilon_i : 1 \leq i \leq n\}$ or $\{\epsilon_{it} : 1 \leq i \leq p, 1 \leq t \leq T\}$ are IID samples of $N(0,1)$. All results are based on 100 repetitions.

Experiment 1: Consider the linear regression model with $p = 60$ and $n = 100$. Predictors are divided into four groups, each of size 15. The true regression coefficients shared within each group are $-2r$, $-r$, r, and $2r$, respectively. Different values of $r > 0$ lead to various signal-to-noise ratios. Here we let r take values in $\{1, 0.8, 0.5\}$, corresponding to high, moderate, and low signal-to-noise ratio, respectively.

We compare the performance of six different methods: Oracle, ordinary least squares (OLS), bCARDS, aCARDS, total variation (TV), fused Lasso (fLasso). Oracle is the least-squares estimator knowing the true groups. aCARDS and bCARDS are described in Section 2; here we let the penalty function $p_\lambda(\cdot)$ be the SCAD penalty with $a = 3.7$, and take the OLS estimator as the preliminary estimator. TV uses the exhaustive pairwise penalty (9), where $p_\lambda(\cdot)$ is also the SCAD penalty with $a = 3.7$. The fused Lasso is based on an order generated from ranking the OLS coefficients. In this sense, the fused Lasso is essentially bCARDS with the Lasso penalty $p_\lambda(t) = \lambda|t|$. Tuning parameters of all these methods are selected via Bayesian information criteria (BIC).

Prediction performance of different methods is evaluated in terms of the average model error over an independent test set of size 10,000. The model error is the prediction error subtracted by the variance of ϵ_{it}, and it better reflects the performance of statistical methods. In addition, to measure how close the estimated grouping structure approaches the true one, we introduce the normalized mutual information (NMI), which is a common measure for similarity between clusterings (Fred and Jain 2003). Suppose $\mathbb{C} = \{C_1, C_2, \ldots\}$ and $\mathbb{D} = \{D_1, D_2, \ldots,\}$ are two sets of disjoint clusters of $\{1, \ldots, p\}$, define

$$\mathrm{NMI}(\mathbb{C}, \mathbb{D}) = \frac{I(\mathbb{C};\mathbb{D})}{[H(\mathbb{C}) + H(\mathbb{D})]/2},$$

where $I(\mathbb{C};\mathbb{D}) = \sum_{k,j}(|C_k \cap D_j|/p)\log(p|C_k \cap D_j|/|C_k||D_j|)$ is the mutual information between $\mathbb{C}$ and $\mathbb{D}$, and $H(\mathbb{C}) = -\sum_k(|C_k|/p)\log(|C_k|/p)$ is the entropy of $\mathbb{C}$. $\mathrm{NMI}(\mathbb{C}, \mathbb{D})$ takes values on $[0, 1]$, and larger NMI implies the two groupings are closer. In particular, $\mathrm{NMI} = 1$ means that the two groupings are exactly the same.

Figure 2 shows boxplots of the average model error and NMI for six different methods. We observe that except for the case of weak signals ($r = 0.5$), two versions of CARDS outperform other methods since they lead to smaller average model error and larger NMI. bCARDS is performing especially well in achieving low model errors, even in the case $r = 0.5$. aCARDS has a better performance in terms of NMI, which indicates that it is better in recovering the true grouping structure. The possible reason that aCARDS does not perform as well as bCARDS in model errors is that aCARDS has more tuning parameters and selection of these tuning parameters in simulations may be nonoptimal.

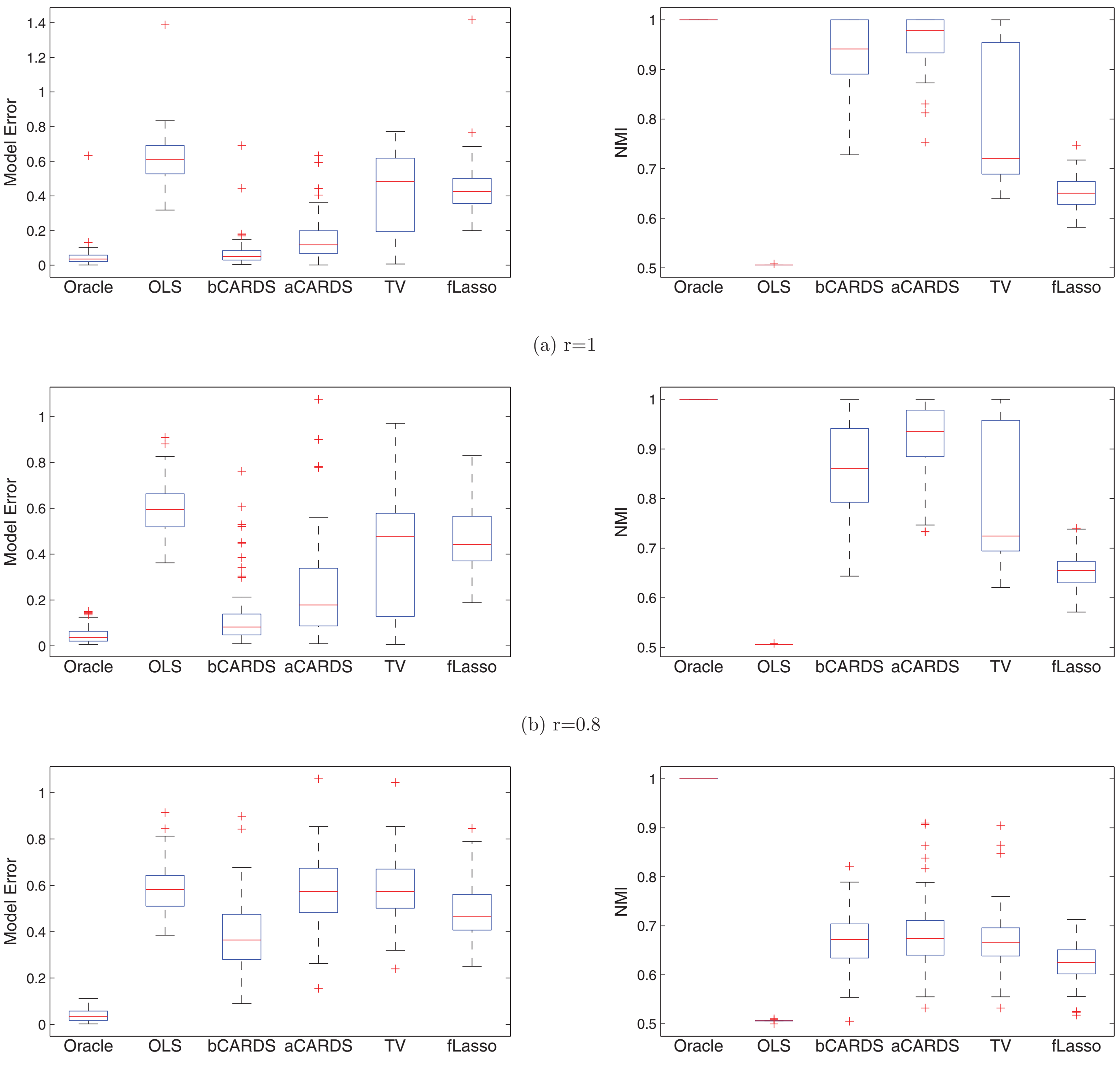
Experiment 2: The setting of this experiment is the same as in Experiment 1, except that the homogeneous groups have nonequal sizes. In Experiment 2a, the predictors are divided into four groups of size 1, 15, 15, and 29. The four distinct regression coefficients are $-4r$, $-r$, r, and $2r$, respectively. Here, the first group is a singleton. In Experiment 2b, there is one dominating group of size 50 with a common regression coefficient $-2r$. The other 10 predictors have the regression coefficients 0, $\frac{2}{9}$, $\frac{4}{9}$, $\frac{6}{9}$, $\ldots$, 2, respectively. In both subexperiments, we take $r = 1$ and 0.7 to represent two levels of signal-to-noise ratio. Besides the six methods compared in Experiment 1, we also implement a data-driven selection between bCARDS and aCARDS, as described in Section 2.3. In detail, we select the parameter δ via BIC (for each value of δ, the other parameters are also selected via BIC). The resulting method is called CARDS.

Figure 3 displays results for Experiment 2a. It suggests that both bCARDS and aCARDS outperform other methods, so does CARDS. Figure 4 displays results for Experiment 2b. We see that the total variation and fused Lasso cannot improve the OLS estimator in terms of the average model error. bCARDS also performs unsatisfactorily, possibly due to misranking in the preliminary estimate. But aCARDS performs much better than OLS and other methods. After a data-driven selection between bCARDS and aCARDS, the resulting method CARDS also outperforms other methods.

Experiment 3: We use this experiment to investigate how the performance of two versions of CARDS is affected by misranking in the preliminary estimate. The setting is the same as Experiment 1 and we fix $r = 1$ (so $n = 100$, $p = 60$, and there are four equal-size groups with true regression coefficients equal to -2, -1, 1, and 2, respectively). For each dataset $(\mathbf{X}, \boldsymbol{y})$, we generate 11 different preliminary ranks as follows: for each σ in $\{1, 1.2, 1.4, \ldots, 3\}$, we generate $\mathbf{z} \sim N(\mathbf{X}\boldsymbol{\beta}^0, \sigma^2\mathbf{I}_n)$ independently of $\boldsymbol{y}$ conditional on $\mathbf{X}$, and then use the OLS estimator associated with $\mathbf{z}$ to get a preliminary rank. A larger value of σ tends to yield a "worse" preliminary rank. We use K^* defined in Section 3.3 to quantify the level of misranking. Recall that K^* is the total number of jumps in true regression coefficients under the preliminary rank, and $K^* > 3$ means there exists misordering in the preliminary rank. We generated 100 datasets and 11 preliminary ranks for each dataset so that the results are based on 1100 repetitions. We compare the performance of bCARDS and aCARDS with that of the total variation (TV), where TV does not use any information from the preliminary rank. Figure 5 contains boxplots of the average model error as K^* changes. First, we see that two versions of CARDS are quite robust to the increase of K^* and always outperform the total variation. Second, the model error of bCARDS increases as K^* increases, which provides empirical evidence for the claim in Theorem 4 about the effect of misranking to bCARDS. Third, when K^* is large, aCARDS has a better performance than bCARDS, because the hybrid pairwise penalty can tolerate a higher level of misranking than the fused penalty.

Experiment 4: This experiment explores the homogeneity and sparsity simultaneously. Consider the linear regression model with $p = 100$ and $n = 150$. Among the 100 predictors, 60 are important ones and their coefficients are the same as those in Experiment 1. Besides, there are 40 unimportant predictors whose coefficients are all equal to 0.

We implemented sCARDS and compared its performance with different oracle estimators, Oracle, Oracle0, and OracleG, as well as ordinary least squares (OLS), SCAD, shrinkage total variation (sTV), and fused Lasso (fLasso). The three oracles are defined with different levels of prior information: the Oracle knows both the important predictors and the true groups, the Oracle0 only knows which are important predictors; and the OracleG only knows the true groups (it treats all unimportant predictors as one group with unknown coefficients). sCARDS



(a) r=1

(b) r=0.8

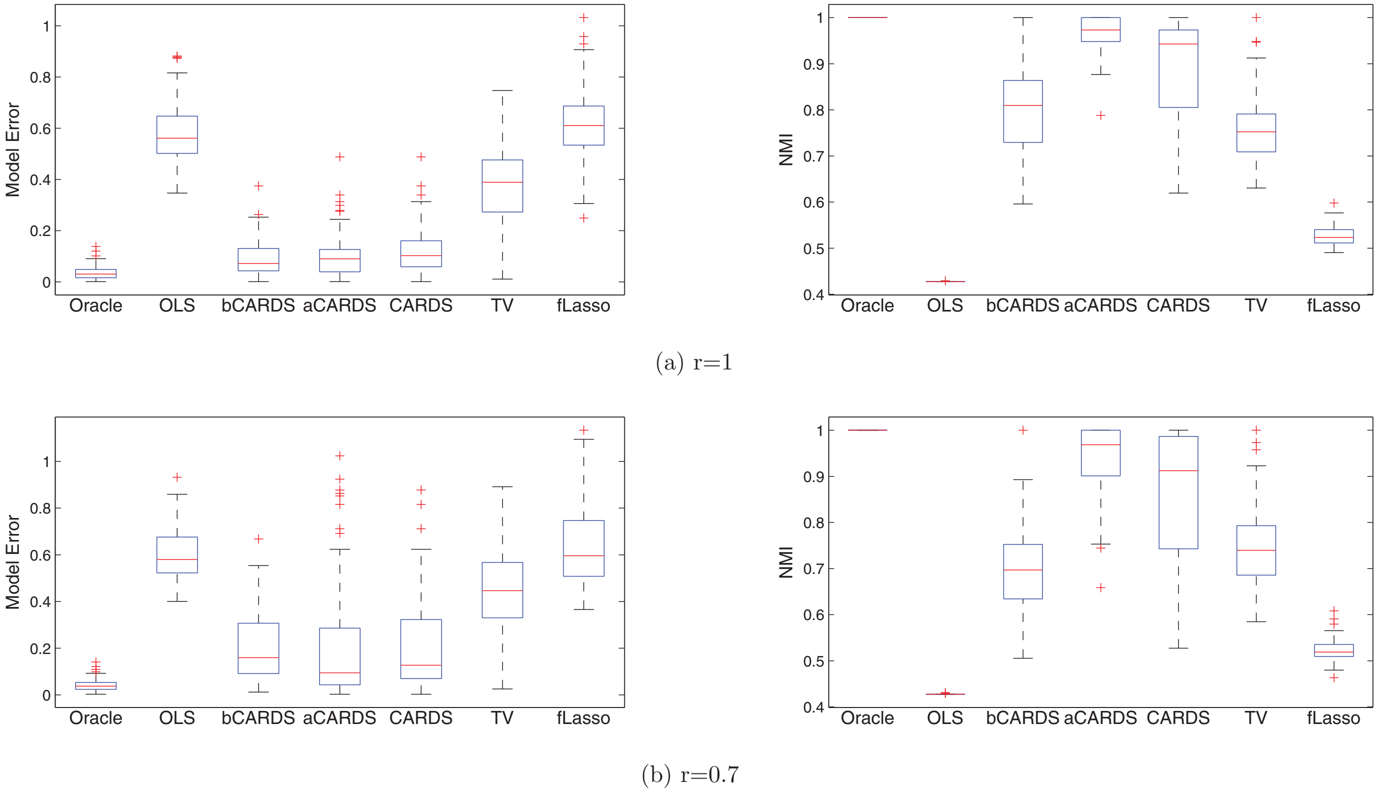
(c) r=0.5

Figure 2. The average model error and normalized mutual information in Experiment 1, where $p = 60$, $n = 100$, and there are four equal-size coefficient groups.

is as described in Section 2; while implementing it, we take the SCAD estimator as the preliminary estimator. The shrinkage total variation is an extension of TV by adding both the elementwise SCAD penalty and exhaustive pairwise SCAD penalty. The fused Lasso used here has both the elementwise L_1 penalty and fused pairwise L_1 penalty.

Figure 6 displays the boxplots of average model errors for $r = 1$ and $r = 0.7$. First, by comparing model errors of the three oracles, we see a significant advantage of taking into account both homogeneity and sparsity over pure sparsity. Moreover, the results of Oracle0 and OracleG show that exploring the group structure is more important than sparsity. Second, sCARDS achieves a smaller model error than OLS, SCAD, and fused Lasso. The performance of sCARDS and sTV in terms of model error are comparable. However, they are different in feature selection. Figure 7 contains frequency histograms of the number of falsely selected features for two methods. In about 17% of the repetitions, sTV fails to shrink coefficients of the 40 unimportant predictors to 0.

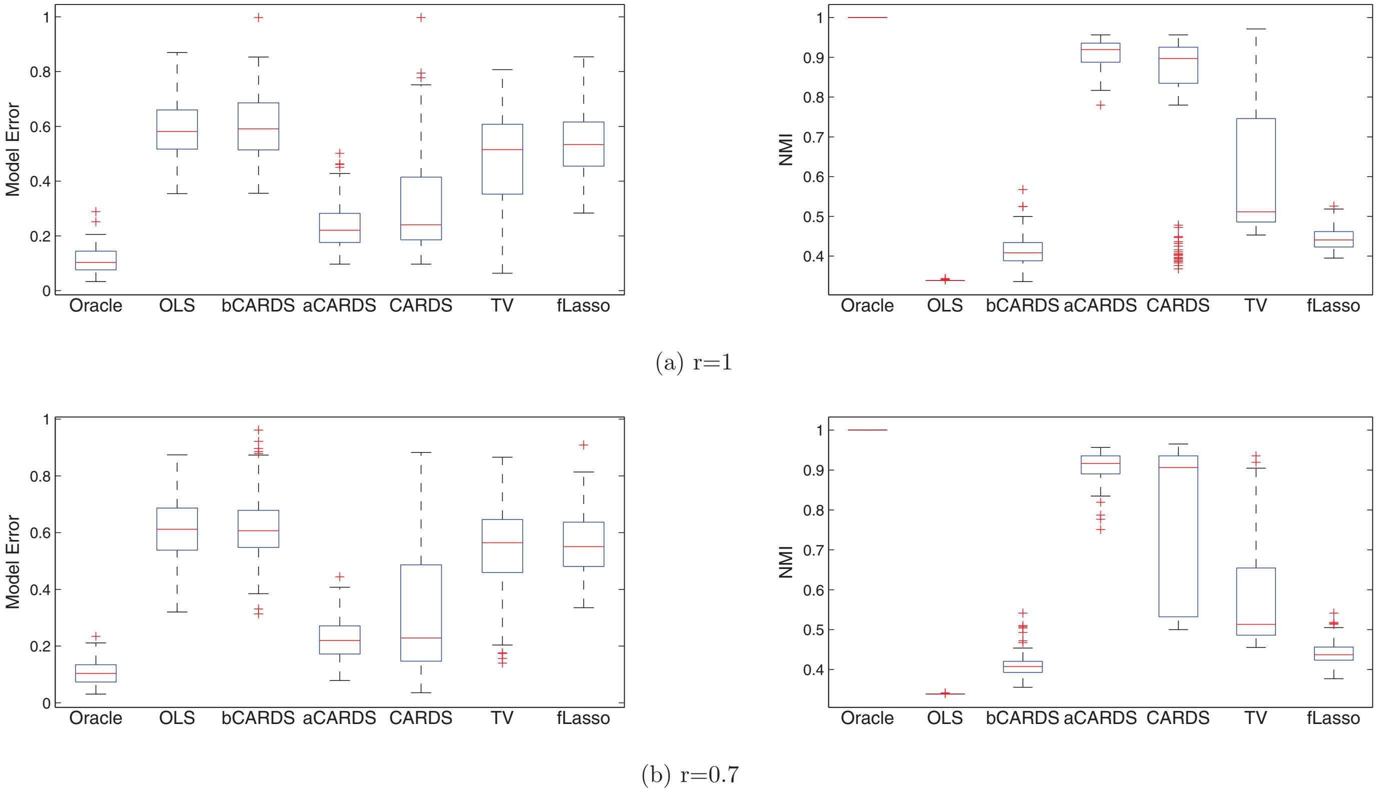
Experiment 5: We consider a special case of the spatial-temporal model (3), $Y_{it} = \mathbf{X}_t^T \boldsymbol{\beta}_i + \epsilon_{it}$, $i = 1, \ldots, p$, that is, the predictors are the same for all spatial locations. There are $p = 100$ different locations and $k = 5$ common predictors (so each $\boldsymbol{\beta}_i$ has a dimension 5). We assume only the spatial



(a) r=1

(b) r=0.7

Figure 3. The average model error and normalized mutual information in Experiment 2a, where $p = 60, n = 100$, and there are four coefficient groups of size 1, 15, 15, and 29.



(a) r=1

(b) r=0.7

Figure 4. The average model error and normalized mutual information in Experiment 2b, where $p = 60, n = 100$, and there is one dominating group of size 50.

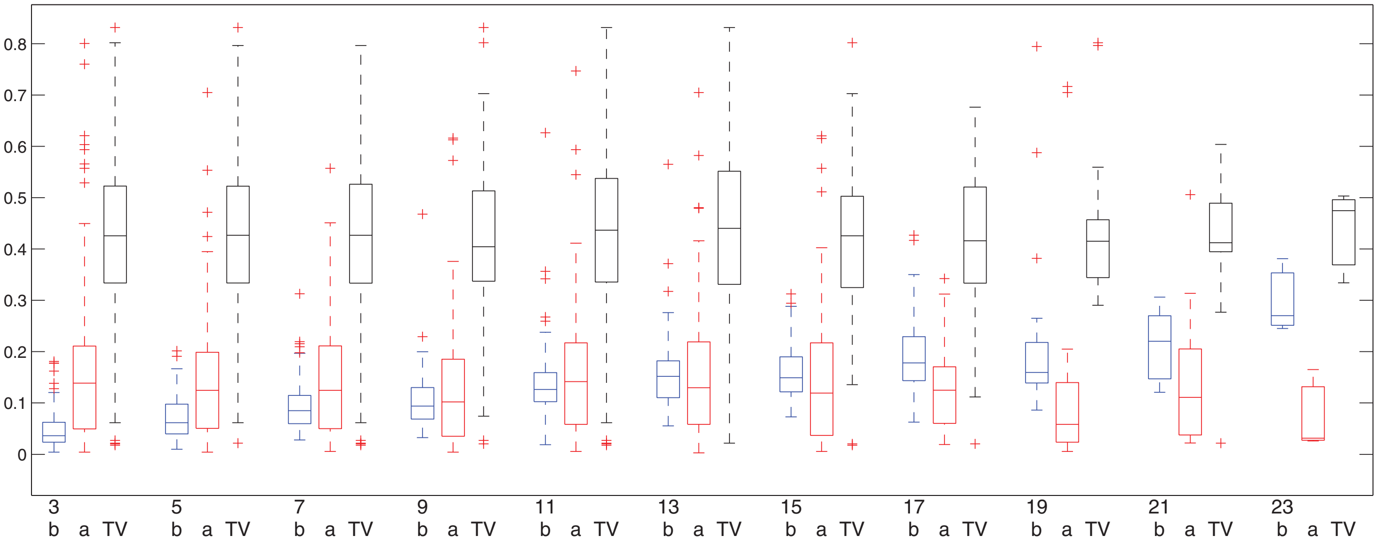


Figure 5. The average model error in Experiment 3. The horizontal axis represents K^*, and "b," "a," and "TV" are short for bCARDS, aCARDS, and total variation.

homogeneity in the regression coefficients, that is, for each predictor j ($j = 1, \ldots, 5$), the coefficients, $\{\beta_{i,j}, 1 \le i \le 100\}$, are divided into four groups of equal size 25, where coefficients in the same group share a similar value. In simulations, we let

$$\beta_{i,j} = \omega_j + (-2)I_{1\le i\le 25} + (-1)I_{26\le i\le 50} + I_{51\le i\le 75} + 2I_{76\le i\le 100}, \quad 1 \le j \le 5,$$

where $\{\omega_j = 0.1 \times (j-1), 1 \le j \le 5\}$ are location-independent constants. In this experiment, instead of varying the signal-to-noise ratio directly, we equivalently change T, the total number of time points.

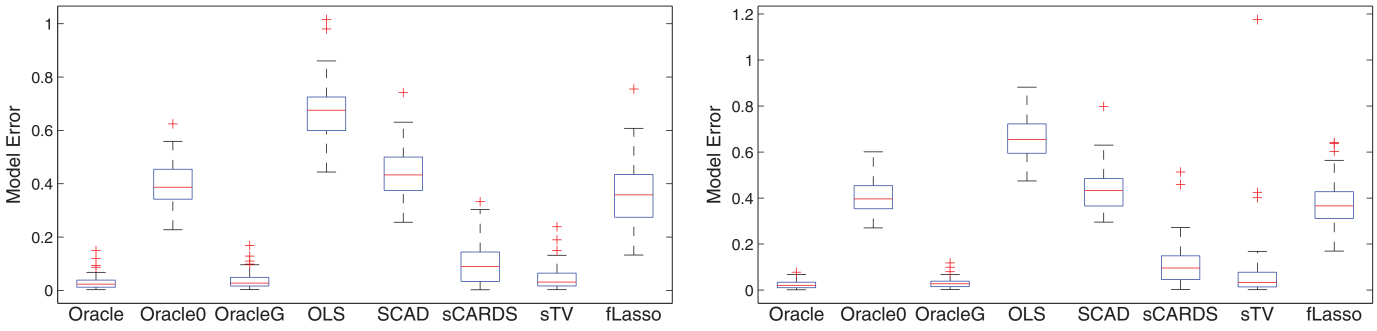
We extend aCARDS to this model by adding the hybrid pairwise penalty on coefficients of the same predictor at different locations, and still call the method aCARDS. The total variation (TV) and fused Lasso (fLasso) can be extended to this model in a similar way. The Oracle is the maximum likelihood estimator which knows the true groups for each predictor. We aim to compare the performance of Oracle, OLS, aCARDS, total variation, and fused Lasso.

Figure 8 displays the results. We see that aCARDS achieves significantly lower model errors than OLS, due to exploring homogeneity. Moreover, it has a better performance than the total variation and fused Lasso, particularly when $T = 50, 80$. aCARDS also estimates well the true grouping structure; when $T = 50, 80$, the normalized mutual information is larger than 0.95 in most repetitions.

6. REAL DATA ANALYSIS

6.1 S&P500 Returns

In this study, we fit a homogeneous Fama-French model for stock returns: $Y_{it} = \alpha_i + \mathbf{X}_t^T \boldsymbol{\beta}_i^0 + \epsilon_{it}$, where $\mathbf{X}_t$ contains three Fama-French factors at time t, Y_{it} is the excess return of stocks and ϵ_{it} are idiosyncratic noises. We collected daily returns of 410 stocks, which were always included in the components of the S&P500 index in the period December 1, 2010 to December 1, 2011 ($T = 254$). We applied CARDS as in Experiment 5, except that the intercepts α_j's were also penalized for sparsity. The sparsity of α_j's is supported by the capital asset pricing model (CAPM) and its extension, multifactor pricing model, in financial econometric theories. The tuning parameters were chosen via generalized cross-validation (GCV). Table 1 shows the number of fitted coefficient groups on three factors and the number of nonzero intercepts. We then used daily returns of the same stocks in the period December 1, 2011 to July 1, 2012 ($T = 146$) to evaluate the prediction error. Let



(a) r=1

(b) r=0.7

Figure 6. The average model error and normalized mutual information in Experiment 4, where $p = 100$, $n = 100$, and there are 60 important predictors divided into four equal-size groups.

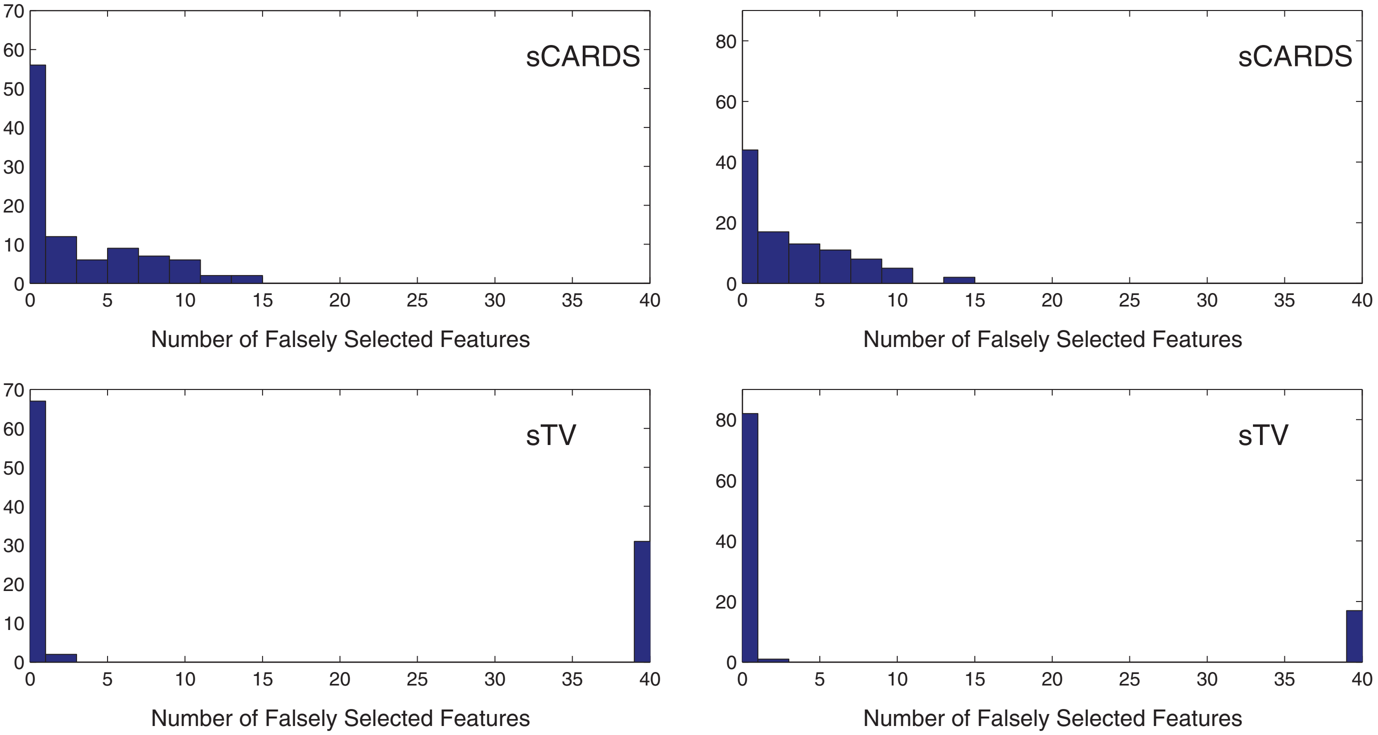


Figure 7. Feature selection of sCARDS and shrinkage total variation (sTV) in Experiment 4 (left: $r = 1$; right: $r = 0.7$), where there are 40 unimportant predictors.

$\widehat{y}_{it}$ and y_{it} be the fitted and observed excess returns of stock i at time $t = 1, \ldots, 146$, respectively. Define the discounted cumulative sum of squared estimation errors (cRSS) at time t as $\text{cRSS}_t = \sum_{s=1}^{t} \rho^{\lfloor s/10 \rfloor} \sum_i (\widehat{y}_{it} - y_{it})^2$, where we take $\rho = 0.95$. Figure 9 shows the percentage improvement in cRSS_t of the CARDS estimator over the OLS estimator. We see that CARDS achieves a smaller discounted cRSS compared to OLS at most time points, especially in the "very-close" and "far-away" future. The North American Industry Classification System (NAICS) classifies these 410 companies into 18 different industry sectors. Figure 10(a) shows the OLS coefficients on the factor "book-to-market ratio." We can see that stocks belonging to Sector 3 "Utilities" (29 stocks in total) have very close OLS coefficients, and 17 stocks in this sector were clustered into one group in CARDS estimator. Figure 10(b) shows the percentage improvement in cRSS_t only for stocks in this sector, where the improvement is more significant.

6.2 Polyadenylation Signals

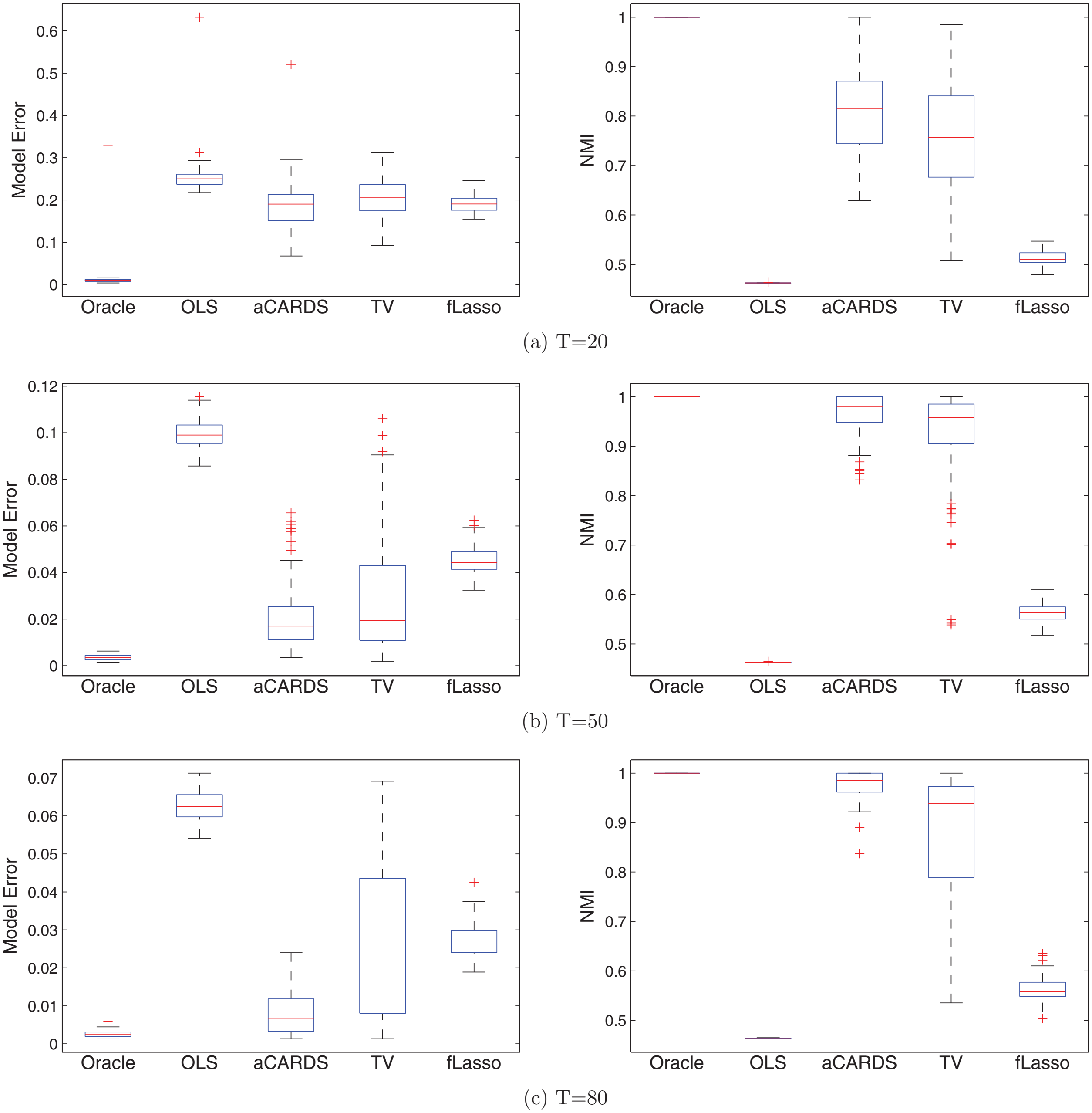
CARDS can be easily extended to more general settings such as generalized linear models (McCullagh and Nelder 1989) although we have focused on the linear regression model so far. In this section, we apply CARDS to a logistic regression example. This study tried to predict polyadenylation signals (PASes) in human DNA and mRNA sequences by analyzing features around them. The dataset was first used in Legendre and Gautheret (2003) and later analyzed by Liu et al. (2003), and it is available at *http://datam.i2r.a-star.edu.sg/datasets/krbd/SequenceData/Polya.html*. There is one training dataset and five testing datasets. To avoid any platform bias, we use the training dataset only. It has 4418 observations each with 170 predictors and a binary response. The binary response indicates whether a terminal sequence is classified as a "strong" or "weak" polyA site, and the predictors are features from the upstream (USE) and downstream (DSE) sequence elements. We randomly select 2000 observations to perform model estimation and use the rest to evaluate performance. Our numerical analysis consists of the following steps. Step 1 is to apply the L_1-penalized logistic regression to these 2000 observations with all 170 predictors and use AIC to select an appropriate regularization parameter. In Step 2, we use the logistic regression coefficients obtained in Step 1 as our preliminary estimate and apply sCARDS accordingly. Average prediction error (and standard error in parentheses) over 40 random splitting are reported in Table 2. We also

Table 1. Number of groups in fitting the S&P500 data

Fama-French factors	No. of coef. groups
"Market return"	41
"Market capitalization"	32
"Book-to-market ratio"	56
Intercept	60

Table 2. Results of the PASes data

	sCARDS	SCAD	sTV	fLasso
Prediction error	0.2449 (.0015)	0.2458 (.0014)	0.2757 (.0026)	0.2445 (0.0010)
No. of nonzero coef. groups	5.5000	21.6250	5.7500	5.5000
No. of selected features	73.2750	21.6250	40.3500	86.8750



(a) T=20

(b) T=50

(c) T=80

Figure 8. The average model error and normalized mutual information in Experiment 5, where the model is a spatial-temporal regression model with $p = 100$ locations and $k = 5$ common predictors.

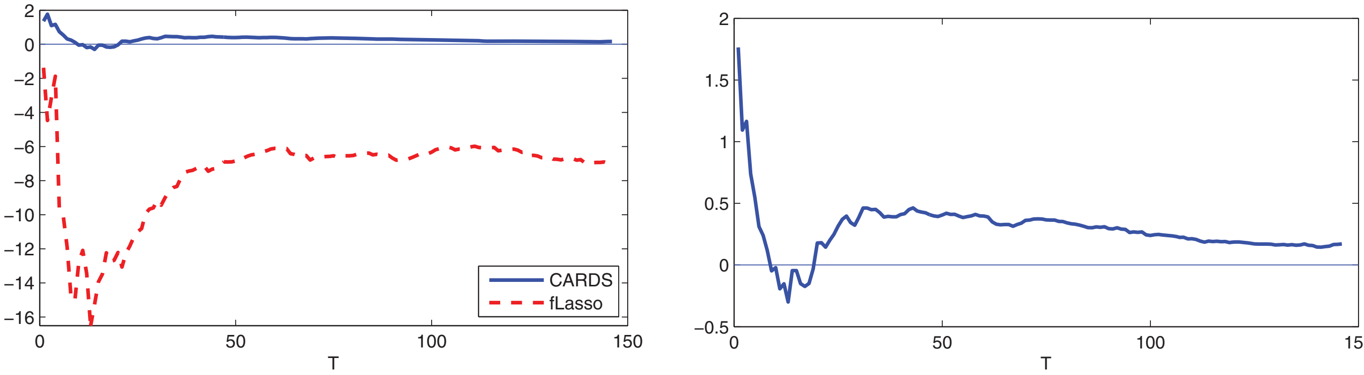


Figure 9. Comparison of the cumulative sum of squared prediction errors of the S&P500 data from December 1, 2011 to July 1, 2012. The vertical axis is the percent improvement on the prediction error relative to OLS, defined by $100(\mathrm{cRSS}_t^{\mathrm{ols}} - \mathrm{cRSS})/\mathrm{cRSS}_t^{\mathrm{ols}}$. The right panel is a zoom-in of the results for CARDS.

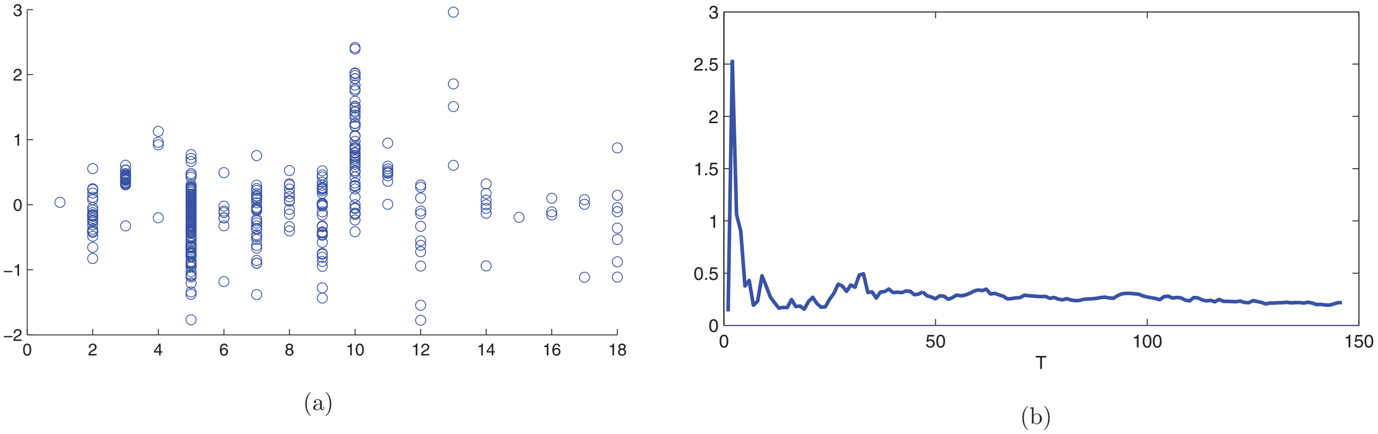


Figure 10. (a) OLS coefficients on the "book-to-market ratio" factor. The x-axis represents different sectors. (b) percent improvement of the prediction error relative to OLS for stocks in the sector "Utilities" (Sector 2).

report the average number of nonzero coefficient groups and the average number of selected features. It shows that sCARDS lead to a smaller prediction error when compared with the shrinkage total variation (sTV). In addition, the sCARDS has fewer groups of nonzero coefficients but more selected features, which implies that we can include more predictors while fixing the degrees of freedom at a small value. Note that in this example, the fused Lasso (fLasso) has a similar performance as the sCARDS. In Section 5, we remarked that the fused Lasso is essentially bCARDS with the Lasso penalty $p_\lambda(t) = \lambda|t|$.

7. CONCLUSION

In this article, we explored homogeneity of coefficients in high-dimensional regression. We proposed a new method called clustering algorithm in regression via data-driven segmentation (CARDS) to estimate regression coefficients and to detect homogeneous groups. The implementation of CARDS does not need any geographical information (neighborhoods, distance, graphs, etc.) a priori, which distinguishes it from other methods in similar settings and makes it more general for applications. A modification of CARDS, sCARDS, can be used to explore homogeneity and sparsity simultaneously. Our theoretical results show that better estimation accuracy can be achieved by exploring homogeneity. In particular, when the number of homogeneous groups is small, the power of exploring homogeneity and sparsity simultaneously is much larger than that of exploring sparsity only, which is also confirmed in our simulation studies.

Methodologically, CARDS has two main innovations. First, it takes advantage of a preliminary estimate by extracting from which either an estimated ranking or an estimated ordered segmentation. Second, it introduces the so-called "hybrid pairwise penalty" to adapt to available partial ordering information. The hybrid pairwise penalty not only is robust to misordering, but also avoids statistical and computational inefficiency due to penalizing too many pairs. These ideas about handling homogeneity can be applied to much broader situations than linear regression, if we combine the hybrid pairwise penalty with appropriate loss functions. For example, CARDS can be extended to generalized linear models (GLM) when homogeneity appears.

To promote homogeneity, CARDS takes advantage of a preliminary estimate. Such idea can be generalized. Instead of extracting a complete raking or an ordered segmentation, we may also apply clustering methods to coefficients of the preliminary estimate, such as k-mean algorithm or hierarchical clustering algorithm, to help construct data-driven penalties and further promote homogeneity.

This article only considers the case where predictors in each homogeneous group have equal coefficients. In a more general situation, coefficients of predictors in the same group are close but not exactly equal. The idea of data-driven pairwise penalties still applies, but instead of using the class of folded concave penalty functions, we may need to use penalty functions which are smooth at the origin, for example, the L_2 penalty function. Another possible approach is to use posterior-type estimators combined with, say, a Gaussian prior on the coefficients. These are beyond the scope of this article and we leave them as future work.

PROOFS

A.1 Proof of Theorem 1

Introduce the mapping $T : \mathcal{M}_A \to \mathbb{R}^K$, where $T(\boldsymbol{\beta})$ is the K-dimensional vector whose kth coordinate equals to the common value of β_j for $j \in A_k$. Note that T is a bijection and T^{-1} is well-defined for any $\boldsymbol{\mu} \in \mathbb{R}^K$. Also, introduce the mapping $T^* : \mathbb{R}^p \to \mathbb{R}^K$, where $T^*(\boldsymbol{\beta})_k = \frac{1}{|A_k|}\sum_{j\in A_k}\beta_j$. We see that $T^* = T$ on $\mathcal{M}_A$, and $T^{-1} \circ T^*$ is the orthogonal projection from $\mathbb{R}^p$ to $\mathcal{M}_A$. Denote $\boldsymbol{\mu}^0 = T(\boldsymbol{\beta}^0)$ and $\widehat{\boldsymbol{\mu}}^{\text{oracle}} = T(\widehat{\boldsymbol{\beta}}^{\text{oracle}})$.

Denote $L_n(\boldsymbol{\beta}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ and $P_n(\boldsymbol{\beta}) = \lambda_n \sum_{j=1}^{p-1} \rho(\beta_{\tau(j+1)} - \beta_{\tau(j)})$, so that we can write $Q_n(\boldsymbol{\beta}) = L_n(\boldsymbol{\beta}) + P_n(\boldsymbol{\beta})$. For any $\boldsymbol{\mu} \in \mathbb{R}^K$, let

$$L_n^A(\boldsymbol{\mu}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}_A\boldsymbol{\mu}\|^2, \qquad P_n^A(\boldsymbol{\mu}) = \lambda_n \sum_{k=1}^{K-1} \rho(\mu_{k+1} - \mu_k),$$

and define $Q_n^A(\boldsymbol{\mu}) = L_n^A(\boldsymbol{\mu}) + P_n^A(\boldsymbol{\mu})$. Note that when τ is consistent with the order of $\boldsymbol{\beta}^0$, there exist $1 = j_1 < j_2 < \cdots < j_K < j_{K+1} = p + 1$ such that $A_k = \{\tau(j_k), \tau(j_k + 1), \ldots, \tau(j_{k+1} - 1)\}$ for $1 \le k \le K$. Then $Q_n(\boldsymbol{\beta}) = Q_n^A(T(\boldsymbol{\beta}))$ and $Q_n^A(\boldsymbol{\mu}) = Q_n(T^{-1}(\boldsymbol{\mu}))$ for any $\boldsymbol{\beta} \in \mathcal{M}_A$ and $\boldsymbol{\mu} \in \mathbb{R}^K$.

In the first part of the proof, we show $\|\widehat{\boldsymbol{\beta}}^{\text{oracle}} - \boldsymbol{\beta}^0\| = O_p(\sqrt{K/n})$. By definition and direct calculations,

$$\|\widehat{\boldsymbol{\beta}}^{\text{oracle}} - \boldsymbol{\beta}^0\| = \|\mathbf{D}(\widehat{\boldsymbol{\mu}}^{\text{oracle}} - \boldsymbol{\mu}^0)\|, \quad \widehat{\boldsymbol{\mu}}^{\text{oracle}} - \boldsymbol{\mu}^0 = (\mathbf{X}_A^T\mathbf{X}_A)^{-1}\mathbf{X}_A^T\boldsymbol{\varepsilon}.$$

Therefore, we can write

$$\|\widehat{\boldsymbol{\beta}}^{\text{oracle}} - \boldsymbol{\beta}^0\| = \|(\mathbf{D}^{-1}\mathbf{X}_A^T\mathbf{X}_A\mathbf{D}^{-1})^{-1}\mathbf{D}^{-1}\mathbf{X}_A^T\boldsymbol{\varepsilon}\|. \tag{A.1}$$

From Condition 3.1, $\|(\mathbf{D}^{-1}\mathbf{X}_A^T\mathbf{X}_A\mathbf{D}^{-1})^{-1}\| \le (c_1 n)^{-1}$ and $\text{tr}(\mathbf{D}^{-1}\mathbf{X}_A^T\mathbf{X}_A\mathbf{D}^{-1}) \le c_2 nK$. By the Markov inequality, for any $\delta > 0$,

$$P\left(\|\mathbf{D}^{-1}\mathbf{X}_A^T\boldsymbol{\varepsilon}\| > \sqrt{\frac{c_2 nK}{\delta}}\right) \le \frac{E\|\mathbf{D}^{-1}\mathbf{X}_A^T\boldsymbol{\varepsilon}\|^2}{c_2 nK/\delta} = \frac{\text{tr}(\mathbf{D}^{-1}\mathbf{X}_A^T\mathbf{X}_A\mathbf{D}^{-1})}{c_2 nK/\delta} \le \delta. \tag{A.2}$$

Combining the above, we have shown that with probability at least $1-\delta$, $\|\widehat{\boldsymbol{\beta}}^{\text{oracle}} - \boldsymbol{\beta}^0\| \le C\delta^{-1/2}\sqrt{K/n}$. This proves $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| = O_p(\sqrt{K/n})$.

Furthermore, we show a result that will be frequently used in later proofs:

$$\|\widehat{\boldsymbol{\beta}}^{\text{oracle}} - \boldsymbol{\beta}^0\| \le C\sqrt{K\log(n)/n}, \quad \text{with probability } \ge 1 - n^{-1}K. \tag{A.3}$$

Write $\mathbf{D}^{-1}\mathbf{X}_A^T\boldsymbol{\varepsilon} = (\mathbf{v}_1^T\boldsymbol{\varepsilon}, \dots, \mathbf{v}_k^T\boldsymbol{\varepsilon})^T$, where $\mathbf{v}_k = \mathbf{X}_A\mathbf{D}^{-1}\mathbf{e}_k$ and $\mathbf{e}_k$ is the unit vector with 1 on the kth coordinate and 0 elsewhere. Observing that $\|\mathbf{v}_k\|^2$ is the kth diagonal of the matrix $\mathbf{D}^{-1}\mathbf{X}_A^T\mathbf{X}_A\mathbf{D}^{-1}$, we have $\|\mathbf{v}_k\| \le \sqrt{c_2 n}$. It follows from Condition 3.3 and the union bound that

$$P\left(\|\mathbf{D}^{-1}\mathbf{X}_A^T\boldsymbol{\varepsilon}\|_\infty > \sqrt{c_2 c_3^{-1} n\log(2n)}\right) \le \sum_{k=1}^K P\left(\|\mathbf{v}_k^T\boldsymbol{\varepsilon}\| > \|\mathbf{v}_k\|\sqrt{c_3^{-1}\log(2n)}\right) \le n^{-1}K.$$

Since $\|\mathbf{D}^{-1}\mathbf{X}_A^T\boldsymbol{\varepsilon}\| \le K^{1/2}\|\mathbf{D}^{-1}\mathbf{X}_A^T\boldsymbol{\varepsilon}\|_\infty$,

$$\|\mathbf{D}^{-1}\mathbf{X}_A^T\boldsymbol{\varepsilon}\| \le \sqrt{c_2 c_3^{-1} Kn\log(2n)}, \quad \text{with probability } \ge 1 - n^{-1}K. \tag{A.4}$$

Then (A.3) follows by combining (A.1) and (A.4).

In the second part of the proof, we show that $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$ is a strictly local minimizer of $Q_n(\boldsymbol{\beta})$ with probability at least $1 - \epsilon_0 - n^{-1}K - (n \vee p)^{-1}$. By assumption, there is an event E_1 such that $P(E_1^c) \le \epsilon_0$ and over the event E_1, τ is consistent with the order of $\boldsymbol{\beta}^0$. Consider the neighborhood of $\boldsymbol{\beta}^0$:

$$\mathcal{B} = \left\{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^0\| < 2C\sqrt{K\log(n)/n}\right\}.$$

By (A.3), there is an event E_2 such that $P(E_2^c) \le n^{-1}K$ and over the event E_2, $\|\widehat{\boldsymbol{\beta}}^{\text{oracle}} - \boldsymbol{\beta}^0\| \le C\sqrt{K\log(n)/n}$. Hence, $\widehat{\boldsymbol{\beta}}^{\text{oracle}} \in \mathcal{B}$ over the event E_2.

For any $\boldsymbol{\beta} \in \mathcal{B}$, write $\boldsymbol{\beta}^*$ as its orthogonal projection to $\mathcal{M}_A$. We aim to show

(a) Over the event $E_1 \cap E_2$,

$$Q_n(\boldsymbol{\beta}^*) \ge Q_n(\widehat{\boldsymbol{\beta}}^{\text{oracle}}), \quad \text{for any } \boldsymbol{\beta} \in \mathcal{B}, \tag{A.5}$$

and the inequality is strict whenever $\boldsymbol{\beta}^* \ne \widehat{\boldsymbol{\beta}}^{\text{oracle}}$.

(b) There is an event E_3 such that $P(E_3^c) \le (n \vee p)^{-1}$. Over the event $E_1 \cap E_2 \cap E_3$, there exists $\mathcal{B}_n$ ($\mathcal{B}_n \subset \mathcal{B}$), a neighborhood of $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$, such that

$$Q_n(\boldsymbol{\beta}) \ge Q_n(\boldsymbol{\beta}^*), \quad \text{for any } \boldsymbol{\beta} \in \mathcal{B}_n, \tag{A.6}$$

and the inequality is strict whenever $\boldsymbol{\beta} \ne \boldsymbol{\beta}^*$.

Combining (a) and (b), $Q_n(\boldsymbol{\beta}) \ge Q_n(\widehat{\boldsymbol{\beta}}^{\text{oracle}})$ for any $\boldsymbol{\beta} \in \mathcal{B}_n$, a neighborhood of $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$, and the inequality is strict whenever $\boldsymbol{\beta} \ne \widehat{\boldsymbol{\beta}}^{\text{oracle}}$. This proves that $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$ is a strictly local minimizer of Q_n, over the event $E_1 \cap E_2 \cap E_3$.

It remains to show (a) and (b). Consider (a) first. We claim that

$$P_n^A(T^*(\boldsymbol{\beta})) = 0 \quad \text{for any } \boldsymbol{\beta} \in \mathcal{B}. \tag{A.7}$$

To see this, for a given $\boldsymbol{\beta} \in \mathcal{B}$, write $\boldsymbol{\mu} = T^*(\boldsymbol{\beta})$. It suffices to check $|\mu_{k+1} - \mu_k| > a\lambda_n$ for $k = 1, \dots, K-1$. Note that $|\mu_{k+1} - \mu_k| \ge \min_{i \in A_k, j \in A_{k+1}} |\beta_i - \beta_j| \ge \min_{i,j} |\beta_i^0 - \beta_j^0| - 2\|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|_\infty \ge 2b_n - 2C\sqrt{K\log(n)/n}$. Since $b_n > a\lambda_n \gg \sqrt{K\log(n)/n}$, it is easy to see that $|\mu_{k+1} - \mu_k| > a\lambda_n$.

Using (A.7), we see that $Q_n^A(T^*(\boldsymbol{\beta})) = L_n^A(T^*(\boldsymbol{\beta}))$, for all $\boldsymbol{\beta} \in \mathcal{B}$. Since $Q_n^A = Q_n \circ T^{-1}$ and $T^{-1} \circ T^*$ is the orthogonal projection from $\mathbb{R}^p$ to $\mathcal{M}_A$, for any $\boldsymbol{\beta} \in \mathcal{B}$,

$$Q_n(\boldsymbol{\beta}^*) = Q_n(T^{-1} \circ T^*(\boldsymbol{\beta})) = Q_n^A(T^*(\boldsymbol{\beta})) = L_n^A(T^*(\boldsymbol{\beta})). \tag{A.8}$$

In particular, noting that $\widehat{\boldsymbol{\beta}}^{\text{oracle}} \in \mathcal{B}$ and its orthogonal projection to $\mathcal{M}_A$ is itself, the above further implies

$$Q_n(\widehat{\boldsymbol{\beta}}^{\text{oracle}}) = L_n^A(\widehat{\boldsymbol{\mu}}^{\text{oracle}}). \tag{A.9}$$

By definition and the fact that $\frac{\partial^2 L_n^A(\boldsymbol{\mu})}{\partial\boldsymbol{\mu}\partial\boldsymbol{\mu}^T} = \frac{1}{n}\mathbf{X}_A^T\mathbf{X}_A$ is positive definite, $\widehat{\boldsymbol{\mu}}^{\text{oracle}}$ is the unique global minimizer of $L_n^A(\boldsymbol{\mu})$. As a result,

$$L_n^A(T^*(\boldsymbol{\beta})) \ge L_n^A(\widehat{\boldsymbol{\mu}}^{\text{oracle}}), \tag{A.10}$$

and the inequality is strict whenever $T^*(\boldsymbol{\beta}) \ne \widehat{\boldsymbol{\mu}}^{\text{oracle}}$, that is, $\boldsymbol{\beta}^* \ne T^{-1}(\widehat{\boldsymbol{\mu}}^{\text{oracle}}) = \widehat{\boldsymbol{\beta}}^{\text{oracle}}$. Combining (A.8)–(A.10) gives (a).

Second, consider (b). For a positive sequence $\{t_n\}$ to be determined, let

$$\mathcal{B}_n = \mathcal{B} \cap \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{\text{oracle}}\| \le t_n\}.$$

Since $\boldsymbol{\beta}^*$ is the orthogonal projection of $\boldsymbol{\beta}$ to $\mathcal{M}_A$, $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \le \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|$ for any $\boldsymbol{\beta}' \in \mathcal{M}_A$. In particular, $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \le \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{\text{oracle}}\|$. As a result, to show (A.6), it suffices to show

$$Q_n(\boldsymbol{\beta}) \ge Q_n(\boldsymbol{\beta}^*), \quad \text{for any } \boldsymbol{\beta} \text{ such that } \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \le t_n, \tag{A.11}$$

and the inequality is strict whenever $\boldsymbol{\beta} \ne \boldsymbol{\beta}^*$.

To show (A.11), write $\boldsymbol{\mu} = T^*(\boldsymbol{\beta})$ so that $\boldsymbol{\beta}^* = T^{-1}(\boldsymbol{\mu})$. By Taylor expansion,

$$\begin{aligned} Q_n(\boldsymbol{\beta}) - Q_n(\boldsymbol{\beta}^*) &= -\frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^m)^T\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + \sum_{j=1}^p \frac{\partial P_n(\boldsymbol{\beta}^m)}{\partial\beta_{\tau(j)}}(\beta_{\tau(j)} - \beta^*_{\tau(j)}) \\ &\equiv I_1 + I_2, \end{aligned}$$

where $\boldsymbol{\beta}^m$ is in the line between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$. Consider I_2 first. Direct calculations yield

$$\frac{\partial P_n(\boldsymbol{\beta})}{\partial\beta_{\tau(j)}} = \begin{cases} -\lambda_n\bar{\rho}(\beta_{\tau(2)} - \beta_{\tau(1)}), & j = 1 \\ \lambda_n\bar{\rho}(\beta_{\tau(j)} - \beta_{\tau(j-1)}) - \lambda_n\bar{\rho}(\beta_{\tau(j+1)} - \beta_{\tau(j)}), & 2 \le j \le p-1 \\ \lambda_n\bar{\rho}(\beta_{\tau(p)} - \beta_{\tau(p-1)}), & j = p, \end{cases}$$

where $\bar{\rho}(t) = \rho'(|t|)\text{sgn}(t)$ and $\rho(t) = \lambda^{-1}p_\lambda(t)$. Plugging it into I_2 and rearranging the sum, we obtain

$$I_2 = \lambda_n \sum_{j=1}^{p-1} \bar{\rho}(\beta^m_{\tau(j+1)} - \beta^m_{\tau(j)})\big[(\beta_{\tau(j+1)} - \beta_{\tau(j)}) - (\beta^*_{\tau(j+1)} - \beta^*_{\tau(j)})\big]. \tag{A.12}$$

When $\tau(j)$ and $\tau(j+1)$ belong to the same group, $\beta^*_{\tau(j)} = \beta^*_{\tau(j+1)}$, and hence the sign of $(\beta^m_{\tau(j+1)} - \beta^m_{\tau(j)})$ is the same as the sign of $(\beta_{\tau(j+1)} -$

$\beta_{\tau(j)})$ if neither of them is 0. In addition, recall that $A_k = \{\tau(j_k), \tau(j_k+1), \dots, \tau(j_{k+1}-1)\}$ for all $1 \le k \le K$, for some indices $1 = j_1 < j_2 < \cdots < j_K < j_{K+1} = p+1$. Combining the above, we can rewrite

$$\begin{aligned} I_2 = {}& \lambda_n \sum_{k=1}^{K} \sum_{j=j_k}^{j_{k+1}-2} \rho'(|\beta^m_{\tau(j+1)} - \beta^m_{\tau(j)}|)|\beta_{\tau(j+1)} - \beta_{\tau(j)}| \\ & + \lambda_n \sum_{k=2}^{K} \bar{\rho}(|\beta^m_{\tau(j_k)} - \beta^m_{\tau(j_k-1)}|)\big[(\beta_{\tau(j_k)} - \beta_{\tau(j_k-1)}) \\ & - (\beta^*_{\tau(j_k)} - \beta^*_{\tau(j_k-1)})\big]. \end{aligned}$$

First, since $\boldsymbol{\beta}^0 \in \mathcal{M}_A$ and $\boldsymbol{\beta}^*$ is the orthogonal projection of $\boldsymbol{\beta}$ to $\mathcal{M}_A$, $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^0\| \le \|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|$. Hence, $\boldsymbol{\beta} \in \mathcal{B}$ implies $\boldsymbol{\beta}^*, \boldsymbol{\beta}^m \in \mathcal{B}$. By repeating the proof of (A.7), we can show $\bar{\rho}(|\beta^m_{\tau(j_k)} - \beta^m_{\tau(j_k-1)}|) = 0$ for $2 \le k \le K$. So the second term in I_2 disappears. Second, in the first term of I_2, since $|\beta^m_{\tau(j+1)} - \beta^m_{\tau(j)}| \le 2\|\boldsymbol{\beta}^m - \boldsymbol{\beta}^*\|_\infty \le 2\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\infty \le 2t_n$, it follows by concavity that $\rho'(|\beta^m_{\tau(j+1)} - \beta^m_{\tau(j)}|) \ge \rho'(2t_n)$. Together, we have

$$I_2 \ge \lambda_n \sum_{k=1}^{K} \sum_{j=j_k}^{j_{k+1}-2} \rho'(2t_n)|\beta_{\tau(j+1)} - \beta_{\tau(j)}|. \tag{A.13}$$

Next, we simplify I_1. Let $\mathbf{z} = \mathbf{z}(\boldsymbol{\beta}^m) = \mathbf{X}^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}^m)$ and write $I_1 = -\frac{1}{n}\mathbf{z}^T(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$. For any fixed k and l such that $\tau(l) \in A_k$ and $l \ne j_{k+1} - 1$, let $A^1_{kl} = \{\tau(j) \in A_k : j \le l\}$ and $A^2_{kl} = \{\tau(j) \in A_k : j > l\}$. Regarding that $\beta^*_{\tau(i)} = \frac{1}{|A_k|}\sum_{j=j_k}^{j_{k+1}-1} \beta_{\tau(j)}$ for $i \in A_k$, we can reexpress I_1 as

$$\begin{aligned} I_1 &= -\frac{1}{2}\sum_{k=1}^{K}\sum_{i=j_k}^{j_{k+1}-1} \frac{1}{n} z_{\tau(i)}\big[\beta_{\tau(i)} - \beta^*_{\tau(i)}\big] - \frac{1}{2}\sum_{k=1}^{K}\sum_{j=j_k}^{j_{k+1}-1} \frac{1}{n} z_{\tau(j)} \\ &\quad \times \big[\beta_{\tau(j)} - \beta^*_{\tau(j)}\big] \\ &= -\sum_{k=1}^{K} \frac{1}{2n|A_k|} \sum_{i,j=j_k}^{j_{k+1}-1} z_{\tau(i)}\big[\beta_{\tau(i)} - \beta_{\tau(j)}\big] - \sum_{k=1}^{K} \frac{1}{2n|A_k|} \\ &\quad \times \sum_{i,j=j_k}^{j_{k+1}-1} z_{\tau(j)}\big[\beta_{\tau(j)} - \beta_{\tau(i)}\big] \\ &= -\sum_{k=1}^{K} \frac{1}{2n|A_k|} \sum_{i,j=j_k}^{j_{k+1}-1} \big[z_{\tau(j)} - z_{\tau(i)}\big]\big[\beta_{\tau(j)} - \beta_{\tau(i)}\big] \\ &= -\sum_{k=1}^{K} \frac{1}{n|A_k|} \sum_{j_k \le i < j = j_{k+1}-1} \big[z_{\tau(j)} - z_{\tau(i)}\big] \sum_{i \le l < j} \big[\beta_{\tau(l+1)} - \beta_{\tau(l)}\big] \\ &= -\sum_{k=1}^{K} \frac{1}{n|A_k|} \sum_{l=j_k}^{j_{k+1}-2} \big[\beta_{\tau(l+1)} - \beta_{\tau(l)}\big] \\ &\quad \times \left[|A^1_{kl}| \sum_{j \in A^2_{kl}} z_{\tau(j)} - |A^2_{kl}| \sum_{i \in A^1_{kl}} z_{\tau(i)}\right] \\ &\equiv \sum_{k=1}^{K}\sum_{l=j_k}^{j_{k+1}-2} w_{\tau(l)}(\mathbf{z})\big[\beta_{\tau(l+1)} - \beta_{\tau(l)}\big], \end{aligned} \tag{A.14}$$

where for any vector $\mathbf{v} \in \mathbb{R}^p$,

$$w_{\tau(l)}(\mathbf{v}) = n^{-1}\left[\frac{|A^2_{kl}|}{|A_k|} \sum_{j \in A^1_{kl}} v_{\tau(j)} - \frac{|A^1_{kl}|}{|A_k|} \sum_{j \in A^2_{kl}} v_{\tau(j)}\right].$$

We aim to bound $|w_{\tau(l)}(\mathbf{z})|$. Let $\boldsymbol{\eta} = \mathbf{X}^T\mathbf{X}(\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)$, $\boldsymbol{\eta}^m = \mathbf{X}^T\mathbf{X}(\boldsymbol{\beta}^m - \boldsymbol{\beta}^*)$ and write $\mathbf{z} = \mathbf{X}^T\boldsymbol{\varepsilon} - \boldsymbol{\eta} - \boldsymbol{\eta}^m$. First, $w_{\tau(l)}(\mathbf{v})$ is a linear function of $\mathbf{v}$. Second, since $\boldsymbol{\beta}^m$ lies between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$, we have $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^m\| \le \|\boldsymbol{\beta}^* - \boldsymbol{\beta}\| \le t_n$. It follows that $\|\boldsymbol{\eta}^m\| \le \lambda_{\max}(\mathbf{X}^T\mathbf{X})t_n$. Moreover, $|w_{\tau(l)}(\mathbf{v})| \le (|A_k|/n)\|\mathbf{v}\|_\infty \le (p/n)\|\mathbf{v}\|$ for all $\mathbf{v}$. Combining the above yields

$$\begin{aligned} |w_{\tau(l)}(\mathbf{z})| &\le |w_{\tau(l)}(\mathbf{X}^T\boldsymbol{\varepsilon})| + |w_{\tau(l)}(\boldsymbol{\eta})| + \sup_{\mathbf{v}:\|\mathbf{v}\| \le \lambda_{\max}(\mathbf{X}^T\mathbf{X})t_n} |w_{\tau(l)}(\mathbf{v})| \\ &\le |w_{\tau(l)}(\mathbf{X}^T\boldsymbol{\varepsilon})| + |w_{\tau(l)}(\boldsymbol{\eta})| + (p/n)\lambda_{\max}(\mathbf{X}^T\mathbf{X}) \cdot t_n. \end{aligned} \tag{A.15}$$

First, we bound the term $w_{\tau(l)}(\mathbf{X}^T\boldsymbol{\varepsilon})$. Let E_3 be the event that

$$\max_{\tau(l) \in A_k} |w_{\tau(l)}(\mathbf{X}^T\boldsymbol{\varepsilon})| \le n^{-1/2}\sqrt{\sigma_k|A_k|\log(2(n \vee p))/c_3}, \quad k = 1, \dots, K, \tag{A.16}$$

where we recall σ_k is the maximum eigenvalue of $n^{-1}\mathbf{X}^T\mathbf{X}$ restricted to the (A_k, A_k)-block. Given $\tau(l)$, we can express $w_{\tau(l)}(\mathbf{X}^T\boldsymbol{\varepsilon})$ as

$$w_{\tau(l)}(\mathbf{X}^T\boldsymbol{\varepsilon}) = \mathbf{a}^T_{\tau(l)}\boldsymbol{\varepsilon}, \quad \text{where } \mathbf{a}_{\tau(l)} = n^{-1}\left(\frac{|A^2_{kl}|}{|A_k|}\mathbf{X}_{A^1_{kl}}\mathbf{1}_{A^1_{kl}} - \frac{|A^1_{kl}|}{|A_k|}\mathbf{X}_{A^2_{kl}}\mathbf{1}_{A^2_{kl}}\right).$$

Write $L_1 = |A^1_{kl}|$ and $L_2 = |A^2_{kl}|$, so that $|A_k| = L_1 + L_2$. It is observed that $\|\mathbf{X}_{A^1_{kl}}\mathbf{1}_{A^1_{kl}}\|^2 \le n\sigma_k\|\mathbf{1}_{A^1_{kl}}\|^2 \le n\sigma_k L_1$. Using the fact that $(a+b)^2 \le 2(a^2+b^2)$ for any real values a, b, we have $\|\mathbf{a}_{\tau(l)}\|^2 \le 2n^{-1}\sigma_k(L_2^2L_1/|A_k|^2 + L_1^2L_2/|A_k|^2) = 2\sigma_kL_1L_2/(n|A_k|) \le \sigma_k|A_k|/(2n)$. Applying Condition 3.3 and the probability union bound,

$$\begin{aligned} P(E_3^c) &\le \sum_{k=1}^{K}\sum_{\tau(l) \in A_k} P\left(|w_{\tau(l)}(\mathbf{X}^T\boldsymbol{\varepsilon})| > n^{-1/2}\sqrt{\sigma_k|A_k|\log(2(n \vee p))/c_3}\right) \\ &\le \sum_{1 \le j \le p} P(|\mathbf{a}_j^T\boldsymbol{\varepsilon}| > \|\mathbf{a}_j\|\sqrt{2\log(2(n \vee p))/c_3}) < (n \vee p)^{-1}. \end{aligned} \tag{A.17}$$

Second, we bound the term $w_{\tau(l)}(\boldsymbol{\eta})$. Observing that for any vector $\mathbf{v}$, $w_{\tau(l)}(\mathbf{v}) = w_{\tau(l)}(\mathbf{v} - \bar{v}_k\mathbf{1})$, where $\bar{v}_k$ is the mean of $\{v_j, j \in A_k\}$, we have

$$\begin{aligned} |w_{\tau(l)}(\mathbf{v})|^2 &\le 2\left(\frac{|A^2_{kl}|^2|A^1_{kl}|}{n^2|A_k|^2} + \frac{|A^1_{kl}|^2|A^2_{kl}|}{n^2|A_k|^2}\right)\big(\max_{j \in A_k}|v_j - \bar{v}_k|\big)^2 \\ &\le \frac{|A_k|}{2n^2}\big(\max_{j \in A_k}|v_j - \bar{v}_k|\big)^2. \end{aligned}$$

Since $\boldsymbol{\eta} = \mathbf{X}^T\mathbf{X}(\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)$ and $\boldsymbol{\beta}^* - \boldsymbol{\beta}^0 \in \mathcal{M}_A$, we have $\max_{j \in A_k}|\eta_j - \bar{\eta}_k| \le n\nu_k\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^0\|$, where ν_k is defined in (13). As a result,

$$\max_{\tau(l) \in A_k} |w_{\tau(l)}(\boldsymbol{\eta})| \le \frac{\nu_k}{\sqrt{2}}|A_k|^{1/2} \cdot \|\boldsymbol{\beta}^* - \boldsymbol{\beta}^0\| \le C\nu_k\sqrt{K|A_k|\log(n)/n}, \tag{A.18}$$

where the last inequality is because we consider $\boldsymbol{\beta} \in \mathcal{B}$ in (A.6), and $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^0\| \le \|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|$ (noticing that $\boldsymbol{\beta}^*$ is the orthogonal projection of $\boldsymbol{\beta}$ onto $\mathcal{M}_A$).

Combining (A.14)–(A.18), we find that over the event $E_1 \cap E_2 \cap E_3$,

$$\begin{aligned} |I_1| &\le \sum_{k=1}^{K}\sum_{l=j_k}^{j_{k+1}-2}\left[C\left(\sqrt{\frac{\sigma_k|A_k|\log(n \vee p)}{n}} + \nu_k\sqrt{\frac{K|A_k|\log(n)}{n}}\right) + \frac{p\lambda_{\max}(\mathbf{X}^T\mathbf{X})}{n}t_n\right]|\beta_{\tau(l+1)} - \beta_{\tau(l)}| \\ &\le \sum_{k=1}^{K}\sum_{l=j_k}^{j_{k+1}-2}\left(\frac{\lambda_n}{2} + \frac{p\lambda_{\max}(\mathbf{X}^T\mathbf{X})}{n}t_n\right)|\beta_{\tau(l+1)} - \beta_{\tau(l)}|, \end{aligned} \tag{A.19}$$

where we have used condition (14) on λ_n.

From (A.13) and (A.19), over the event $E_1 \cap E_2 \cap E_3$,

$$\begin{aligned} &\inf_{\boldsymbol{\beta} \in \mathcal{B}:\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \le t_n} \big[Q_n(\boldsymbol{\beta}) - Q_n(\boldsymbol{\beta}^*)\big] \\ &\quad \ge \sum_{k=1}^{K}\sum_{l=j_k}^{j_{k+1}-2}\left[\frac{\lambda_n}{2} - g_n(t_n)\right]|\beta_{\tau(l+1)} - \beta_{\tau(l)}|, \end{aligned}$$

where $g_n(t_n) = n^{-1}p\lambda_{\max}(\mathbf{X}^T\mathbf{X})t_n - \lambda_n[1 - \rho'(2t_n)]$. Since $\rho'(0+) = 1$, $g_n(0+) = 0$. So we can always choose t_n sufficiently small to make sure $|g_n(t_n)| < \lambda_n/2$; consequently, the right-hand side is nonnegative, and strictly positive when $\sum_{k=1}^K \sum_{l=j_k}^{j_{k+1}-2} |\beta_{\tau(l+1)} - \beta_{\tau(l)}| > 0$, that is, $\boldsymbol{\beta} \neq \boldsymbol{\beta}^*$. This proves (b). □

A.2 Proof of Theorem 2

First, we show that the LLA algorithm yields $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$ after one iteration. Let E_1 be the event that the ranking τ is consistent with the order of $\boldsymbol{\beta}^0$, E_2 the event that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| \leq C\sqrt{K\log(n)/n}$, and E_3 the event that (A.16) holds. We have shown that $P(E_1 \cap E_2 \cap E_3) \geq 1 - \epsilon_0 - n^{-1}K - (n \vee p)^{-1}$. It suffices to show that over the event $E_1 \cap E_2 \cap E_3$, the LLA algorithm gives $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$ after the first iteration.

Let $w_j = \rho'(|\widehat{\beta}^{\text{initial}}_{\tau(j+1)} - \widehat{\beta}^{\text{initial}}_{\tau(j)}|)$. At the first iteration, the algorithm minimizes

$$Q_n^{\text{initial}}(\boldsymbol{\beta}) \equiv \frac{1}{2n}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_n \sum_{j=1}^{p-1} w_j|\beta_{\tau(j+1)} - \beta_{\tau(j)}|.$$

This is a convex function, hence it suffices to show that $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$ is a strictly local minimizer of Q_n^{initial}. Using the same notations as in the proof of Theorem 1, for any $\boldsymbol{\beta} \in \mathbb{R}^p$, write $\boldsymbol{\beta}^* = T^{-1} \circ T^*(\boldsymbol{\beta})$ as its orthogonal projection to $\mathcal{M}_A$. Let $\mathcal{B} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^0\| \leq C\sqrt{K\log(n)/n}\}$, and for a sequence $\{t_n\}$ to be determined, consider the neighborhood of $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$ defined by $\mathcal{B}_n = \{\boldsymbol{\beta} \in \mathcal{B} : \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{\text{oracle}}\| \leq t_n\}$. It suffices to show

$$Q_n^{\text{initial}}(\boldsymbol{\beta}) \geq Q_n^{\text{initial}}(\boldsymbol{\beta}^*) \geq Q_n^{\text{initial}}(\widehat{\boldsymbol{\beta}}^{\text{oracle}}), \qquad \text{for any } \boldsymbol{\beta} \in \mathcal{B}_n, \tag{A.20}$$

and the first inequality is strict whenever $\boldsymbol{\beta} \neq \boldsymbol{\beta}^*$, and the second inequality is also strict whenever $\boldsymbol{\beta} \neq \widehat{\boldsymbol{\beta}}^{\text{oracle}}$.

We first show the second inequality in (A.20). For $\tau(j)$ and $\tau(j+1)$ in different groups, $|\beta^0_{\tau(j+1)} - \beta^0_{\tau(j)}| > 2b_n$. In addition, $\|\widehat{\boldsymbol{\beta}}^{\text{initial}} - \boldsymbol{\beta}^0\|_\infty \leq \lambda_n/2$. Hence, $|\widehat{\beta}^{\text{initial}}_{\tau(j+1)} - \widehat{\beta}^{\text{initial}}_{\tau(j)}| \geq 2b_n - \lambda_n > a\lambda_n$, and it follows that $w_j = 0$. On the other hand, for $\tau(j)$ and $\tau(j+1)$ in the same group, $\beta_{\tau(j+1)} - \beta_{\tau(j)} = 0$ whenever $\boldsymbol{\beta} \in \mathcal{M}_A$. Consequently,

$$Q_n^{\text{initial}}(\boldsymbol{\beta}) = \frac{1}{2n}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = L_n(\boldsymbol{\beta}), \qquad \text{for } \boldsymbol{\beta} \in \mathcal{M}_A.$$

It is easy to see that $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$ is the unique global minimizer of L_n constrained on $\mathcal{M}_A$. So the second inequality in (A.20) holds.

Next, consider the first inequality in (A.20). We apply Taylor expansion to $Q_n^{\text{initial}}(\beta) - Q_n^{\text{initial}}(\boldsymbol{\beta}^*)$, and rearrange the sums as in (A.12). Then, for some $\boldsymbol{\beta}^m$ that lies in the line between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$,

$$\begin{aligned}
&Q_n^{\text{initial}}(\boldsymbol{\beta}) - Q_n^{\text{initial}}(\boldsymbol{\beta}^*)\\
&\quad = \lambda_n \sum_{j=1}^{p-1} w_j \cdot \operatorname{sgn}(\beta^m_{\tau(j+1)} - \beta^m_{\tau(j)})\big[(\beta_{\tau(j+1)} - \beta_{\tau(j)}) - (\beta^*_{\tau(j+1)} - \beta^*_{\tau(j)})\big]\\
&\qquad - \frac{1}{n}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}^m)^T\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \equiv J_1 + J_2.
\end{aligned}$$

We first simplify J_1. Note that $w_j = 0$ when $\tau(j)$ and $\tau(j+1)$ are in different groups. When $\tau(j)$ and $\tau(j+1)$ are in the same A_k, first, $\beta^*_{\tau(j+1)} = \beta^*_{\tau(j)}$, and $[\beta^m_{\tau(j+1)} - \beta^m_{\tau(j)}]$ has the same sign as $[\beta_{\tau(j+1)} - \beta_{\tau(j)}]$; second, $|\widehat{\beta}^{\text{initial}}_{\tau(j+1)} - \widehat{\beta}^{\text{initial}}_{\tau(j)}| \leq 2\|\widehat{\boldsymbol{\beta}}^{\text{initial}} - \boldsymbol{\beta}^0\|_\infty \leq \lambda_n$, and hence $w_j \geq \rho'(\lambda_n) \geq a_0$. Combining the above yields

$$J_1 = \lambda_n \sum_{k=1}^{K} \sum_{j=j_k}^{j_{k+1}-2} w_j|\beta_{\tau(j+1)} - \beta_{\tau(j)}| \geq a_0\lambda_n \sum_{k=1}^{K} \sum_{j=j_k}^{j_{k+1}-2} |\beta_{\tau(j+1)} - \beta_{\tau(j)}|. \tag{A.21}$$

Next, we simplify J_2. Denote $\mathbf{z} = \mathbf{X}^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}^m)$. Similarly to (A.14)–(A.19), we find that

$$J_2 = -\sum_{k=1}^{K} \sum_{l=j_k}^{j_{k+1}-2} w_{\tau(l)}(\mathbf{z})\big[\beta_{\tau(l+1)} - \beta_{\tau(l)}\big],$$

where over the event E_3, for any $j_k \leq l \leq j_{k+1} - 2$,

$$\begin{aligned}
|w_{\tau(l)}(\mathbf{z})| \leq{}& C\left(\sqrt{\frac{\sigma_k|A_k|\log(n \vee p)}{n}} + \nu_k\sqrt{\frac{K|A_k|\log(n)}{n}}\right)\\
&+ \frac{p\lambda_{\max}(\mathbf{X}^T\mathbf{X})}{n}t_n.
\end{aligned}$$

From the condition on λ_n, the sum of the first two terms is upper bounded by $a_0\lambda_n/3$ for large n. We choose $t_n = a_0 n\lambda_n/(3p\lambda_{\max}(\mathbf{X}^T\mathbf{X}))$. It follows that

$$|J_2| \leq \sum_{k=1}^{K} \sum_{l=j_k}^{j_{k+1}-2} \frac{2a_0\lambda_n}{3}|\beta_{\tau(l+1)} - \beta_{\tau(l)}|. \tag{A.22}$$

Combining (A.21) and (A.22), over the event $E_1 \cap E_2 \cap E_3$,

$$Q_n^{\text{initial}}(\boldsymbol{\beta}) - Q_n^{\text{initial}}(\boldsymbol{\beta}^*) \geq \frac{a_0\lambda_n}{3} \sum_{k=1}^{K} \sum_{l=j_k}^{j_{k+1}-2} |\beta_{\tau(l+1)} - \beta_{\tau(l)}| \geq 0.$$

This proves the first inequality in (A.20).

Second, we show that over the event $E_1 \cap E_2 \cap E_3$, at the second iteration, the LLA algorithm still yields $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$ and, therefore, it converges to $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$. We have shown that after the first iteration, the algorithm outputs $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$. It then treats $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$ as the initial solution for the second iteration. So it suffices to check

$$\|\widehat{\boldsymbol{\beta}}^{\text{oracle}} - \boldsymbol{\beta}^0\|_\infty \leq \lambda_n/2.$$

This is true because over the event E_1, $\|\widehat{\boldsymbol{\beta}}^{\text{oracle}} - \boldsymbol{\beta}^0\| \leq C\sqrt{K\log(n)/n} \ll \lambda_n$. □

A.3 Proof of Theorem 3

It suffices to show that, with probability at least $1 - O(n^{-\alpha})$, $\beta_i^0 < \beta_j^0$ implies $\widehat{\beta}_i^{\text{ols}} \leq \widehat{\beta}_j^{\text{ols}}$ for any $1 \leq i, j \leq p$. When $\beta_i^0 < \beta_j^0$, necessarily $\beta_j^0 - \beta_i^0 \geq 2b_n$. Moreover, $\widehat{\beta}_j^{\text{ols}} - \widehat{\beta}_i^{\text{ols}} \geq (\beta_j^0 - \beta_i^0) - 2\|\widehat{\boldsymbol{\beta}}^{\text{ols}} - \boldsymbol{\beta}^0\|_\infty$. So it suffices to show that $\|\widehat{\boldsymbol{\beta}}^{\text{ols}} - \boldsymbol{\beta}^0\|_\infty \leq b_n$ with probability at least $1 - O(n^{-\alpha})$.

From direct calculations, $\boldsymbol{\beta}^{\text{ols}} = \boldsymbol{\beta}^0 + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}$. Write $\widehat{\beta}_j^{\text{ols}} - \beta_j^0 = \mathbf{a}_j^T\boldsymbol{\varepsilon}$, where $\mathbf{a}_j = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{e}_j$ for $j = 1, \ldots, p$. Then $\|\mathbf{a}_j\|^2 = \mathbf{e}_j^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{e}_j \leq c_4 n^{-1}$. By Condition 3 and applying the union bound,

$$\begin{aligned}
&P\left(\|\widehat{\boldsymbol{\beta}}^{\text{ols}} - \boldsymbol{\beta}^0\|_\infty > b_n\right)\\
&\quad \leq P\left(\|\widehat{\boldsymbol{\beta}}^{\text{ols}} - \boldsymbol{\beta}^0\|_\infty > \sqrt{(2\alpha c_4/c_3)\log(n)/n}\right)\\
&\quad \leq \sum_{j=1}^{p} P\left(|\mathbf{a}_j^T\boldsymbol{\varepsilon}| > \|\mathbf{a}\|\sqrt{2\alpha\log(n)/c_3}\right) \leq 2pn^{-2\alpha}.
\end{aligned}$$

Since $p = O(n^\alpha)$, $2pn^{-2\alpha} = O(n^{-\alpha})$. This completes the proof. □

SUPPLEMENTARY MATERIALS

The supplementary materials contain technical proofs for Theorems 5, 6, 7, 8, and Corollary 1.

[Received May 2013. Revised October 2013.]

REFERENCES

Bondell, H. D., and Reich, B. J. (2008), "Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors With OSCAR," *Biometrics*, 64, 115–123. [175]

Bühlmann, P., and van de Geer, S. (2011), *Statistics for High-Dimensional Data*, Berlin: Springer. [175]

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1998), "Atomic Decomposition by Basis Pursuit," *SIAM Journal on Scientific Computing*, 20, 33–61. [175]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of American Statistical Association*, 96, 1348–1360. [177,179,181,183]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultra-High Dimensional Feature Space," *Journal of Royal Statistical Society,* Series B, 70, 849–911. [179]

——— (2011), "Nonconcave Penalized Likelihood With NP-Dimensionality," *IEEE Transactions on Information Theory*, 57, 5467–5484. [183]

Fan, J., Lv, J., and Qi, L. (2011), "Sparse High-Dimensional Models in Economics," *Annual Review of Economics*, 3, 291–317. [175,176]

Fan, J., Xue, L., and Zou, H. (2012), "Strong Oracle Optimality of Folded Concave Penalized Estimation," unpublished manuscript, available at *http://arxiv.org/abs/1210.5992*. [177]

Fred, A., and Jain, A. K. (2003), "Robust Data Clustering," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3, 128–136. [184]

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007), "Pathwise Coordinate Optimization," *The Annals of Applied Statistics*, 1, 302–332. [175]

Harchaoui, Z., and Lévy-Leduc, C. (2010), "Multiple Change-Point Estimation With a Total Variation Penalty," *Journal of the American Statistical Association*, 105, 1480–1493. [178]

Huang, H.-C., Hsu, N.-J., Theobald, D. M., and Breidt, F. J. (2010), "Spatial Lasso With Applications to GIS Model Selection," *Journal of Computational and Graphical Statistics*, 19, 963–983. [175]

Kim, S., and Xing, E. P. (2009), "Statistical Estimation of Correlated Genome Associations to a Quantitative Trait Network," *PLos Genetics*, 5, e1000587. [175]

Kim, Y., Choi, H., and Oh, H.-S. (2008), "Smoothly Clipped Absolute Deviation on High Dimensions," *Journal of the American Statistical Association*, 103, 1665–1673. [177]

Legendre, M., and Gautheret, D. (2003), "Sequence Determinants in Human Polyadenylation Site Selection," *BMC Genomics*, 4, 7–15. [188]

Li, C., and Li, H. (2010), "Variable Selection and Regression Analysis for Graph-Structured Covariates With an Application to Genomics," *The Annals of Applied Statistics*, 4, 1498–1516. [175]

Liu, H., Han, H., Li, J., and Wong, L. (2003), "An In-Silico Method for Prediction of Polyadenylation Signals in Human Sequences," *Genome Informatics Series*, 14, 84–93. [188]

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman & Hall. [188]

Park, M. Y., Hastie, T., and Tibshirani, R. (2007), "Averaged Gene Expressions for Regression," *Biostatistics*, 8, 212–227. [175]

Shen, X., and Huang, H.-C. (2010), "Grouping Pursuit Through a Regularization Solution Surface," *Journal of the American Statistical Association*, 105, 727–739. [175,178]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society,* Series B, 58, 267–288. [175]

Tibshirani, S., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and Smoothness via the Fused Lasso," *Journal of the Royal Statistical Society,* Series B, 67, 91–108. [175,179]

Wang, L., Kim, Y., and Li, R. (2013), "Calibrating Nonconvex Penalized Regression in Ultra-High Dimension," *The Annals of Statistics*, 5, 2505–2536. [177]

Yang, S., Yuan, L., Lai, Y.-C., Shen, X., Wonka, P., and Ye, J. (2012), "Feature Grouping and Selection Over an Undirected Graph," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 922–930. [175]

Yang, Y., and He, X. (2012), "Bayesian Empirical Likelihood for Quantile Regression," *The Annals of Statistics*, 40, 1102–1131. [176]

Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [177]

Zhao, P., and Yu, B. (2006), "On Model Selection of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563. [181]

Zhu, Y., Shen, X., and Pan, W. (2013), "Simultaneous Grouping Pursuit and Feature Selection over an Undirected Graph," *Journal of the American Statistical Association*, 108, 713–725. [175]

Zou, H., and Li, R. (2008), "One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models" (with discussion), *The Annals of Statistics*, 36, 1509–1566. [177,181]