A SPLIT-AND-CONQUER APPROACH FOR ANALYSIS OF EXTRAORDINARILY LARGE DATA

Xueying Chen and Min-ge Xie

Department of Statistics, Rutgers University, Piscataway, New Jersey, USA

Summary. If there are extraordinarily large data, too large to fit into a single computer or too expensive to perform a computationally intensive data analysis, what should we do? To deal with this problem, we propose in this paper a *split-and-conquer* approach and illustrate it using several computationally intensive penalized regression methods, along with a theoretical support. Consider a regression setting of generalized linear models with n observations and p covariates, in which n is extraordinarily large and p is either bounded or goes to ∞ at a certain rate of n. We propose to randomly split the data of size n into K subsets of size $O(n/K)$. For each subset of data, we perform a penalized regression analysis and the results from each of the K subsets are then combined to obtain an overall result. We show that under mild conditions the combined overall result still retains desired properties of many commonly used penalized estimators, such as the model selection consistency and asymptotic normality. When K is well controlled, we also show that the combined result is asymptotically equivalent to the result of analyzing the entire data all at once (assuming that there is a super computer that could carry out such an analysis). In addition, the split-and-conquer approach involves a random splitting and a systemic combining. We demonstrate that the approach has an inherent advantage of being more resistant to false model selections caused by spurious correlations. Similar to what reported in the literature, we can establish an upper bound for the expected number of falsely selected variables and a lower bound for the expected number for truly selected variables. Furthermore, when a computational intensive algorithm is used in the sense that its computing expense is at the order of $O(n^a p^b)$, $a > 1$ and $b \geq 0$, we show that the split-and-conquer approach can substantially reduce computing time and computer memory requirement. The proposed methodology is illustrated numerically using both simulation and real data examples.

Key Words Big data, Combining Results from Independent Analyses, Distributed Computing, Generalized linear models, Large sample theory, Penalized regression

1. Introduction

Consider a generalized linear model:

$$E(y_i) = g(\boldsymbol{x}_i^T \boldsymbol{\beta}), i = 1, \ldots, n$$

where y_i is a response variable and $\boldsymbol{x}_i$ is a $p \times 1$ vector of explanatory variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, and g is a link function. Both the sample size n and the number of parameters p can be potentially very large. We assume that, given $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, the conditional distribution of $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ follows the canonical exponential distribution:

$$f(\boldsymbol{y}; \boldsymbol{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} f_0(y_i; \theta_i) = \prod_{i=1}^{n} \left\{ c(y_i) \exp\left[\frac{y_i\theta_i - b(\theta_i)}{\phi} \right] \right\}, \tag{1}$$

where $\theta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}, i = 1, \ldots, n$ and ϕ is a nuisance dispersion parameter. The log-likelihood function $\log f(\boldsymbol{y}; \boldsymbol{X}, \boldsymbol{\beta})$ is then given by

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X}) = [\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} - \mathbf{1}^T\boldsymbol{b}(\boldsymbol{X}\boldsymbol{\beta})]/n, \tag{2}$$

where $\boldsymbol{b}(\boldsymbol{\theta}) = (b(\theta_1), \ldots, b(\theta_n))^T$ for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$ and the function $b(\cdot)$ is a smooth function with the second derivative. In the case when p is large (or grows with n) and $\boldsymbol{\beta}$ is sparse (i.e., many elements of $\boldsymbol{\beta}$ are zero), a penalized likelihood estimator is often used, which is defined as, in a general form,

$$\hat{\boldsymbol{\beta}}^{(a)} = \text{argmax}_{\boldsymbol{\beta}} \left\{ \ell(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X})/n - \rho(\boldsymbol{\beta}; \lambda_a) \right\}. \tag{3}$$

Here, ρ is the penalty function with tuning parameter λ_a. Also, to distinguish the estimator obtained from a split-and-conquer approach to be proposed, we use in (3) the superscript $^{(a)}$ to indicate the estimator is obtained by analyzing the *entire* data $(\boldsymbol{y}, \boldsymbol{X})$ all at once. Depending on the choice of penalty function $\rho(\boldsymbol{\beta}; \lambda_a)$, we have LASSO estimator (Tibshirani, 1996; Chen et al., 2001), LARS algorithm (Efron et al., 2004), SCAD estimator (Fan and Li, 2001) and MCP estimators (Zhang, 2010), among others.

In this paper, we consider a big data situation in which the data size is extraordinarily large, too large to fit into a single computer or be analyzed with available computing resources. We propose a split-and-conquer approach to solve the problem and illustrate it using the aforementioned penalized regression methods. Specifically, we split the entire dataset into K subsets of smaller sample sizes. Each subset is then analyzed separately (assuming the analysis can be performed on the smaller subsets). A set of K results are obtained. Subsequently, the K results are combined to obtain a final result. Our task is to investigate whether the combined overall result can be the same or as good as the result that is obtained from analyzing the entire dataset all at once and, if conditions are needed, what they are. As a general methodology and for easy practical implementation (especially without additional programming or analytical efforts in model fitting), we assume in our development that the same method (including software) is used to analyze each subset data as well as the entire data all together (if we would have enough computing power to do so).

This split-and conquer approach is general and can be applied to a broad range of examples. To facilitate our discussion, we focus our developments on a general penalized regression setting considered in the review article of Fan and Lv (2011), which covers almost all commonly used penalty functions in the current penalized regression practice such as LASSO, SCAD, MCP and others.

Under the setting, Fan and Lv (2011) show that the penalized estimators under the generalized linear models (3) have good asymptotic properties, such as model selection consistency and asymptotic normality, etc. We investigate in this paper specifically whether the combined result from the proposed split-and-conquer method using the corresponding penalized regression still retains these desired properties and, if so, under which conditions. We assume that each subset contains enough data to provide a meaningful inference (e.g., consistency, etc.) for the unknown model parameters.

The idea behind the proposed split-and-conquer approach is straightforward. Its essence can be illustrated using the simple example of the regular Gaussian linear regression where we have finite p and non-sparse $\boldsymbol{\beta}$. Assume that $\boldsymbol{X}^T\boldsymbol{X}$ is invertible, the ordinary least squares estimator using entire data all at once is

$$\hat{\boldsymbol{\beta}}^{(a)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y},$$

When we split the dataset into K pieces and assume $\boldsymbol{X}_k^T\boldsymbol{X}_k$ is invertible, the ordinary least squares estimator obtained from the k^{th} subset is $\hat{\boldsymbol{\beta}}_k = (\boldsymbol{X}_k^T\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^T\boldsymbol{y}_k$, for $k = 1,\ldots,K$. Here, $\boldsymbol{X}_k$ is the design matrix and $\boldsymbol{y}_k$ is the response vector for the data in the k^{th} subset. These K least square estimators can be combined to form a new estimator

$$\hat{\boldsymbol{\beta}}^{(c)} = (\sum_{k=1}^{K} S_k^{-1})^{-1}\sum_{k=1}^{K} S_k^{-1}\hat{\boldsymbol{\beta}}_k = (\sum_{k=1}^{K}\boldsymbol{X}_k^T\boldsymbol{X}_k)^{-1}\sum_{k=1}^{K}\boldsymbol{X}_k^T\boldsymbol{y}_k = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y},$$

where $S_k \overset{\mathrm{d}}{=} (\boldsymbol{X}_k^T\boldsymbol{X}_k)^{-1} \propto \mathrm{var}(\hat{\boldsymbol{\beta}}_k)$. This $\hat{\boldsymbol{\beta}}^{(c)}$ is exactly identical to $\hat{\boldsymbol{\beta}}^{(a)}$, thus we can get the same estimator using the split-and-conquer approach! In order to obtain the ordinary least squares estimator in each subset, we have assumed that $\boldsymbol{X}_k^T\boldsymbol{X}_k$ is invertible. This assumption is slightly stronger than the original design matrix assumption that $\boldsymbol{X}^T\boldsymbol{X}$ is invertible. This example suggests that we can obtain the same estimator using the split-and-conquer approach and pay a small price of requiring a slightly stronger assumption on the design matrix.

We investigate in this paper whether we have any similar results to support the split-and-conquer approach for the penalized estimators (3) and under the more general generalized linear models (1). Specifically, we prove that, under some mild conditions and with a suitable choice of K, our combined estimator using the split-and-conquer approach is asymptotically equivalent to the penalized estimator obtained from analyzing the entire data all at once. Here, the number of splitting K should be relatively large so that each subset is small enough and can be analyzed using computing resources available to us. But it can not be too large and we require each subset contains enough data to provide a meaningful estimator for the unknown regression parameter $\boldsymbol{\beta}$. The combined estimator is model selection consistent as long as the penalized estimators from the imposed penalty function are model selection consistent. When asymptotic normality is attainable, the combined estimator does not lose any efficiency through the split-and-conquer process, in the sense that it has the same asymptotic variance as the penalized estimator using the entire data all at once.

Although the combined estimator may not be exactly the same as the one using the entire data all at once, it is asymptotically equivalent to, thus as efficient as, the penalized estimator obtained by analyzing the entire data all at once. The price that we need to pay is to require to have a slight stronger assumption on the design matrix, a little larger coefficient signals and/or a slower growth rate of p. This is similar to but more involved than the regular Gaussian linear regression example described in the previous paragraph.

In the literature, improvements over the regular penalized estimators in model selection are reported through a majority voting and averaging operation when results from a finite number of random split subsets are combined; see, e.g., Meinshausen and Buhlmann (2010). In the proposed split-and-conquer approach, we can establish an upper bound for the expected number of falsely selected variables and a lower bound for the expected number of truly selected variables, which are consistent with those reported in the literature. For examples, Fan et al. (2010) propose refitted cross-validation to attenuate false correlations among the random errors and explanatory variables (i.e., spurious variables); Meinshausen and Buhlmann (2010) introduce a stability selection and an exact error control bound through a combination of subsampling and model selection algorithms; and Shah and Samworth (2012) propose a variant of stability selection with improved error control property. Similarly, the split-and-conquer approach provides, as a byproduct, a resistance to selection errors caused by spurious correlations and keeps a large amount of variables that are in the true model at the same time. This control on the selected variables are not typically available for conventional penalized estimators by analyzing the entire data.

Furthermore, when a computational intensive algorithm is used with computing expenses at the order of $O(n^a p^b)$, $a > 1$ and $b \geq 0$, the split-and-conquer approach has a very attractive feature for practical applications — it can substantially reduce computing time and memory requirement. For instance, consider the example of linear regression with L_1 norm penalty function, where the LARS (Efron et al., 2004) algorithm has been considered by many (e.g., Yuan and Lin (2006); Zou and Hastie (2005)) as a fast and efficient algorithm to solve the LASSO problem. Efron et al. (2004) reports that the LARS algorithm requires the order of $O(n^a)$ with $a > 1$ computations when $p \geq n$; see more details in Section 3.4 later. The computing time can be costly when both n and p are extraordinarily large. In this case, we show, both mathematically and numerically, that the proposed split-and-conquer approach with LARS can save up to $(1 - 1/K^{(a-1)})\%$ computing time, where K is the number of splitting. This result holds under a general setting and is discussed in Section 3.4. Furthermore, we provide in Section 4 several numerical examples across a number of different models and penalized methods, and demonstrate that the proposed split and conquer approach can save substantial computing time while producing comparable estimators.

The split and conquer approach is intuitive, and a similar practice can also be found in the computer sciences community under the name of parallel and distributed computing (see, e.g., Andrews (2000)). Most of the the research in computer sciences focuses on the aspects such as accessing to a shared memory, exchanging information between processors, or identifying parallel components within an algorithm, etc.; See, e.g. Ahmed et al. (2012). More recently, there are also several papers that provide discussions on the performance of combined results. For examples, Mackey et al. (2011) propose a divide-and-conquer method for matrix factorization, which partitions a large-scale matrix into submatrix, then factors each submatrix and combines submatrix estimates. Zhang et al. (2013) provide a divide-and-combine method for kernel ridge regression, which divides a dataset into several subsets, then obtains a kernel ridge regression estimate from each subset and uses an arithmetic average to combine the local estimates. See also Zhang et al. (2012), Agarwal and Duchi (2012); Ahmed et al. (2012); Duchi et al. (2012). In this paper, we use a weighted combining method and study the statistical performance and computing issues of the proposed method. Different from the others, we also focus and provide a direct comparison between the combined results from our method and the corresponding results using the entire dataset all at once. We also focus on statistical issues such as convergence, efficient estimation, and provided discussions on computing time when computationally intensive approaches are involved.

The rest of this article is organized as follows. Section 2 proposes a split-and-conquer approach and a combined estimator under the generalized linear models. Section 3 studies theoretical properties of the combined estimator and also investigate issues related to error bound controls and computing time. Section 4 demonstrates the utilities of the proposed methodology using both simulation studies and an example of real data application in cargo screening at U.S. Port-of-Entries (POEs). Section 5 provides further comments.

2. Split-and-conquer for penalized regressions

Suppose the number of parameters p can be very large and the true parameter, denoted by $\boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_p^0)^T$, is sparse. Assume the entire dataset of size n is divided into K subsets and the k^{th} subset has n_k observations: $(\boldsymbol{x}_{k,i}, y_{k,i})$, $i = 1, \ldots, n_k$. Write $\boldsymbol{y}_k = (y_{k,1}, \ldots, y_{k,n_k})^T$ and $\boldsymbol{X}_k = (\boldsymbol{x}_{k,1}^T, \ldots,\, \boldsymbol{x}_{k,n_k}^T)^T$, The log-likelihood function for the k^{th} subset, for $k = 1, \ldots, K$, is

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}_k, \boldsymbol{X}_k) = [\boldsymbol{y}_k^T \boldsymbol{X}_k \boldsymbol{\beta} - \mathbf{1}^T \boldsymbol{b}(\boldsymbol{X}_k \boldsymbol{\beta})]/n_k.$$

Corresponding to (3), the penalized estimator for the k^{th} subset is:

$$\hat{\boldsymbol{\beta}}_k = \operatorname{argmax}_{\boldsymbol{\beta}} \left\{ \ell(\boldsymbol{\beta}; \boldsymbol{y}_k, \boldsymbol{X}_k)/n_k - \rho(\boldsymbol{\beta}; \lambda_k) \right\},$$

where $\rho(\boldsymbol{\beta}; \lambda_k)$ is the penalty function with tuning parameter λ_k. For simplicity and also following Fan and Lv (2011), we write $\rho(\boldsymbol{\beta}; \lambda_k) = \sum_{j=1}^p \rho(\beta_j; \lambda_k)$ and assume that the penalty function $\rho(\beta_j; \lambda_k)$ satisfy the following condition:

- **(PC)** *Assume $\rho(t; \lambda)$ is increasing and concave in $t \in [0, \infty)$, and has a continuous derivative $\rho'(t; \lambda)$ with $\rho'(0+; \lambda) > 0$. In addition, $\rho'(t; \lambda)/\lambda$ is increasing in $\lambda \in [0, \infty)$ and $\rho'(0+; \lambda)/\lambda$ is independent of λ.*

As noted by Fan and Lv (2011), Condition (PC) covers most commonly used penalty functions, including L_1 penalty, SCAD, MCP, among others.

Under the setup, the penalized estimator $\hat{\boldsymbol{\beta}}_k$ has the so-called sparsity property (i.e., with many zero entries); see, e.g., Fan and Lv (2011). Denote by $\hat{\mathcal{A}}_k = \{j : \hat{\beta}_{k,j} \neq 0\}$ the set of non-zero elements of $\hat{\boldsymbol{\beta}}_k$. Also, for any indices set S, denote by $\hat{\boldsymbol{\beta}}_{k,S}$ a $|S| \times 1$ vector that is formed by the elements of $\hat{\boldsymbol{\beta}}_k$ whose indices are in S. Thus, $\hat{\boldsymbol{\beta}}_{k,\hat{\mathcal{A}}_k}$ is the sub-vector that contains only the non-zero elements of $\hat{\boldsymbol{\beta}}_k$. Since each $\hat{\boldsymbol{\beta}}_k$ is estimated from a different subset of data, $\hat{\mathcal{A}}_k$ can be different from one to another and the K vectors $\hat{\boldsymbol{\beta}}_{k,\hat{\mathcal{A}}_k}$, $k = 1, \ldots, K$, may have different lengths.

In order to obtain a combined estimator of $\boldsymbol{\beta}$ from $\hat{\boldsymbol{\beta}}_k$'s that retains good performance, we use a majority voting method. There are two considerations. First, the combined estimator, say $\hat{\boldsymbol{\beta}}^{(c)}$, should be based on the K estimators $\hat{\boldsymbol{\beta}}_k$'s, and a variable not in any of $\hat{\mathcal{A}}_k = \{j : \hat{\boldsymbol{\beta}}_{k,j} \neq 0\}$ should not be included by the set $\hat{\mathcal{A}}^{(c)} = \{j : \hat{\boldsymbol{\beta}}_j^{(c)} \neq 0\}$ for the combined estimator. Second, $\hat{\mathcal{A}}_k$ are subject to errors and there may be mismatches between the set $\hat{\mathcal{A}}_k$ from the analysis of the k^{th} subset and the true nonzero set, say $\mathcal{A} \overset{\mathrm{d}}{=} \{j : \beta_j^0 \neq 0\}$. In our majority voting method, we define

$$\hat{\mathcal{A}}^{(c)} \overset{\mathrm{d}}{=} \left\{ j : \sum_{k=1}^{K} \mathbf{I}(\hat{\beta}_{k,j} \neq 0) > w \right\} \tag{4}$$

as the set of selected variables of the combined estimator, where $w \in [0, K)$ is a prespecified threshold and $\mathbf{I}$ is the indicator function. From (4), it is clear that $\hat{\mathcal{A}}^{(c)} \subset \bigcup_{k=1}^{K} \hat{\mathcal{A}}_k$. Also, when the numbers of elements in $\hat{\mathcal{A}}_k$ (denoted by $|\hat{\mathcal{A}}_k|$) are small and the K sets $\hat{\mathcal{A}}_k$'s have many common elements, the numbers of elements in $\hat{\mathcal{A}}^{(c)}$ (denoted by $|\hat{\mathcal{A}}^{(c)}|$) is much smaller than p. In an extreme case in which the threshold $w \geq K - 1$, the majority voting set $\hat{\mathcal{A}}^{(c)}$ contains only the variables that are selected by all subset analyses. In the other extreme case in which the threshold $w = 0$, $\hat{\mathcal{A}}^{(c)}$ contains the variables that are selected by one or more subset analyses. The theoretical development presented later in Section 3.1 suggests that the choice of the threshold, as long as it is a fixed constant, does not affect the asymptotic results. But, in practice with given data and a fixed n, it can impact the numerical performance of the proposed approach. See further discussions in Section 3.2 and also numerical studies in Section 4.

It is also possible to extend the simple majority voting method in (4) to a weighted majority voting method to accommodate possible discrepancies among the K subsets of data (e.g., sample size or other non-random patterns). To keep our steps simple, we use in this paper the simple majority voting to determine the estimated number of selected variables and use a weighted scheme (to be shown below) to combine the penalized estimators from the K subset data.

We first introduce the following notations. For any $n \times 1$ vector of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$, we define an $n \times 1$ vector and an $n \times n$ matrix

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = (\mu(\theta_1), \ldots, \mu(\theta_n))^T \quad \text{and} \quad \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathrm{diag}(\sigma(\theta_1), \ldots, \sigma(\theta_n)),$$

respectively, where $\mu(\theta) = \partial b(\theta)/\partial \theta$ and $\sigma(\theta) = \partial^2 b(\theta)/\partial^2 \theta$. Also, let S be an index subset of $\{1, \ldots, n\}$ with $|S|$ elements $S = \{i_1, \ldots, i_{|S|}\}$. For any $|S| \times 1$ subvector of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}_S = (\theta_{i_1}, \ldots, \theta_{i_{|S|}})^T$, with a little abuse of notation, we define an $|S| \times 1$ vector and an $|S| \times |S|$ matrix

$$\boldsymbol{\mu}(\boldsymbol{\theta}_S) = (\mu(\theta_{i_1}), \ldots, \mu(\theta_{i_{|S|}}))^T \text{ and } \boldsymbol{\Sigma}(\boldsymbol{\theta}_S) = \operatorname{diag}(\sigma(\theta_{i_1}), \ldots, \sigma(\theta_{i_{|S|}})),$$

respectively.

Let $\hat{\boldsymbol{\beta}}_{k,\hat{\mathcal{A}}^{(c)}}$ be the sub-vector of $\hat{\boldsymbol{\beta}}_k$ confined by the majority voting set $\hat{\mathcal{A}}^{(c)}$ of (4). Also, let $\boldsymbol{E} = \operatorname{diag}(v_1, \ldots, v_p)$ be the $p \times p$ voting matrix with $v_j = 1$ if $\sum_{k=1}^{K} \mathrm{I}(\hat{\beta}_{k,j} \neq 0) > w$ and 0 otherwise, and let $\boldsymbol{A} = \boldsymbol{E}_{\hat{\mathcal{A}}^{(c)}}$ be the $p \times |\hat{\mathcal{A}}^{(c)}|$ selection matrix. Here, for any index subset S of $\{1, \ldots, p\}$, $\boldsymbol{E}_S$ stands for an $p \times |S|$ submatrix of $\boldsymbol{E}$ formed by columns whose indices are in S. Our combined estimator is defined as a weighted average of $\hat{\boldsymbol{\beta}}_{k,\hat{\mathcal{A}}^{(c)}}$, $k = 1, \ldots, K$:

$$\hat{\boldsymbol{\beta}}^{(c)} \overset{\mathrm{d}}{=} \boldsymbol{A} \left(\sum_{k=1}^{K} \boldsymbol{A}^T \{\boldsymbol{X}_k^T \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k) \boldsymbol{X}_k\} \boldsymbol{A} \right)^{-1} \sum_{k=1}^{K} \boldsymbol{A}^T \{\boldsymbol{X}_k^T \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k) \boldsymbol{X}_k\} \boldsymbol{A} \hat{\boldsymbol{\beta}}_{k,\hat{\mathcal{A}}^{(c)}}, \quad (5)$$

where $\hat{\boldsymbol{\theta}}_k = \boldsymbol{X}_k \hat{\boldsymbol{\beta}}_k$. Generally speaking, with a smaller sample size in each subset, we expect the estimator from each subset has a slower convergence rate. But the summation over the K pieces in (5) and the set of weights used in the combining can help boost up the estimation power and efficiency. As a result, we can show in the next section that $\hat{\boldsymbol{\beta}}^{(c)}$ is asymptotically equivalent to the corresponding estimator using the entire data, and more.

The majority voting idea discussed in this section is closely connected with the developments by Meinshausen and Buhlmann (2010) and Shah and Samworth (2012) on stability selection. For example, we may view the quantity $\sum_{k=1}^{K} \mathbf{I}(\hat{\beta}_{k,j} \neq 0)/K$ in (4) as a variant version of $\hat{\Pi}_j^{\lambda}$, the probability of variable j to be selected with tuning parameter λ, used in Meinshausen and Buhlmann (2010). However, the goal of Meinshausen and Buhlmann (2010) and Shah and Samworth (2012) is to develop stable penalized estimators, which is different from ours. Although our development also cares about performance and stable estimation, the main focus is to investigate whether we can analyze extremely large data by splitting the task; thus, computational feasibility is the forefront of our development. Different from Meinshausen and Buhlmann (2010) and Shah and Samworth (2012) (which are often computationally infeasible for extraordinarily large data due to multiple rounds of subsampling and calculation), the proposed majority voting approach in this paper only requires a single-round calculation for each data point and each subset has much less size than n when K is large. This difference help improve computing efficiency and reduce memory requirement which in turn increase the feasibility of handling extraordinarily large data. Also, unlike Meinshausen and Buhlmann (2010) and Shah and Samworth (2012) in which the same tuning parameter λ is used for all subsets, the tuning parameter λ_k in each subset analysis is chosen by a criterion, e.g. AIC, BIC or cross-validation, independently of others. This allows us to defer our task of coordinating the analyses from subset data to the last combination step, avoiding to place additional constraints on penalized regressions in the subset data analysis step. Finally, our development allows $K \to \infty$, as $n \to \infty$. It subsumes the situation discussed in Meinshausen and Buhlmann (2010) or Shah and Samworth (2012) in which K is always finite (e.g., $K = 2$).

In the proposed split and conquer approach, a solution path can be obtained for every subset. If a solution path is needed for the combined estimator, we can fix the tuning parameter at a grid and compute the combined estimator at each grid value to form a regularization path for the combined estimator.

3. Theoretical results

In this section, we investigate the asymptotic properties of the combined estimator $\hat{\boldsymbol{\beta}}^{(c)}$ defined in (5), and compare it with the penalized estimator $\hat{\boldsymbol{\beta}}^{(a)}$ defined in (3) which is from analyzing the entire dataset all at once.

3.1. Sign consistency

We show in this subsection that the combined estimator is sign consistent that each component of the combined estimator has the same sign as its true value.

Denote by $\boldsymbol{\theta}^0 = \boldsymbol{X}\boldsymbol{\beta}^0$, $\boldsymbol{\theta}_k^0 = \boldsymbol{X}_k\boldsymbol{\beta}^0$ and the minimal signal $\beta_* = \min\{|\beta_j^0| : \beta_j^0 \neq 0\}$. Let $\overline{\mathcal{A}}$ be the complement of the true nonzero set $\mathcal{A} = \{j : \beta_j^0 \neq 0\}$. Also, for any index set S, let $\boldsymbol{X}_S$ stand for an $n \times |S|$ submatrix of $\boldsymbol{X}$ formed by the columns whose indices are in S. Similarly, let $\boldsymbol{X}_{k,S}$ stand for an $n_k \times |S|$ submatrix of $\boldsymbol{X}_k$ formed by the columns whose indices are in S. In order to obtain model selection consistency of the combined estimator, we need to impose some regularity conditions on the design matrix. Specifically, let $\{b_{s,K}\}$ be a diverging sequence of positive numbers that depends on s and K. We extend the regularity conditions on design matrix imposed in Fan and Lv (2011) on the entire dataset to each subset (beside the entire data):

$$\mathbf{A1}\quad \begin{array}{l} \|\{\boldsymbol{X}_{\mathcal{A}}^T\boldsymbol{\Sigma}(\boldsymbol{\theta}^0)\boldsymbol{X}_{\mathcal{A}}\}^{-1}\|_\infty = O(b_{s,K}n^{-1}), \\ \|\{\boldsymbol{X}_{k,\mathcal{A}}^T\boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}\|_\infty = O(b_{s,K}n_k^{-1}); \\ \|\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T\boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}\{\boldsymbol{X}_{k,\mathcal{A}}^T\boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}\|_\infty = O(n_k^\alpha), \\ \max_{\boldsymbol{\delta}\in\mathcal{N}_0, 1\le j\le s} \lambda_{\max}[\nabla^2\{\boldsymbol{x}_{k,j}^T\boldsymbol{\mu}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})\}] = O(n_k), \end{array}$$

where $\alpha \in [0, 1/2]$, $\mathcal{N}_0 = \{\boldsymbol{\delta} \in \Re^s : \|\boldsymbol{\delta} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_\infty \le \beta_*/2\}$, and the operation ∇^2 is defined as $\nabla^2\gamma(\boldsymbol{\delta}) = \frac{\partial^2}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^T}\gamma(\boldsymbol{\delta})$ for any scalar function $\gamma(\boldsymbol{\delta})$ of an $s \times 1$ vector $\boldsymbol{\delta}$.

The constraint imposed on the design matrices by Condition **A1** is minor. For instance, in the linear regression model, we have $\boldsymbol{\Sigma}(\boldsymbol{\theta}^0) = \boldsymbol{I}_n$ the identity matrix and $\boldsymbol{\mu}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta}) = \boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta}$. In this case, Condition **A1** can be implied by

$$\mathbf{A1}_{(G)}\quad \begin{array}{l} \|\{\boldsymbol{X}_{k,\mathcal{A}}^T\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}\|_\infty = O_p(n_k^{-1}), \\ \|\{\boldsymbol{X}_{\mathcal{A}}^T\boldsymbol{X}_{\mathcal{A}}\}^{-1}\|_\infty = O_p(n^{-1}), \\ \|\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T\boldsymbol{X}_{k,\mathcal{A}}\{\boldsymbol{X}_{k,\mathcal{A}}^T\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}\|_\infty = O(n_k^\alpha). \end{array}$$

Since $\{b_{s,K}\}$ is a diverging sequence, the first two equations in Condition **A1** are in fact weaker than the first two in Condition $\mathbf{A1}_{(G)}$. Condition $\mathbf{A1}_{(G)}$ match with those discussed in the literature under both the settings of fixed matrix design (cf., Fan and Lv (2011), pages 5470-5471) and Gaussian random matrix design (cf., Wainwright (2009), Lemma 5). More generally in non-Gaussian generalized linear models, the variance matrices $\mathbf{\Sigma}(\boldsymbol{\theta}^0)$ and $\mathbf{\Sigma}(\boldsymbol{\theta}_k^0)$ typically involve the covariates through the linear predictors $\boldsymbol{\theta}^0$ and $\boldsymbol{\theta}_k^0$, respectively. If we assume or can show that the smallest diagonal element of $\mathbf{\Sigma}(\boldsymbol{\theta}^0)$, say $h_n = \min_{1\le i\le n}\sigma(\theta_i^0)$, is bounded below by a small constant or $h_n^{-1} = O(b_{s,K})$, then Condition **A1** can also be implied by $\mathbf{A1}_{(G)}$. For example, consider a Poisson model, where $\mathbf{\Sigma}(\boldsymbol{\theta}^0) = \mathrm{diag}\{\exp(\theta_i^0)\}$ with $\theta_i^{(0)} = \boldsymbol{x}_i^T\boldsymbol{\beta}^{(0)} = \boldsymbol{x}_{i,\mathcal{A}}^T\boldsymbol{\beta}_{\mathcal{A}}^{(0)}$. In the fixed design matrix case, we only need to impose that $\theta_i^{(0)} = \boldsymbol{x}_{i,\mathcal{A}}^T\boldsymbol{\beta}_{\mathcal{A}}^{(0)}$ is bounded below away from $-\infty$ or is a sequence tending to $-\infty$ slower than $O(\log(b_{s,K}))$. In the random design matrix case with $\boldsymbol{x}_{i,\mathcal{A}}$ being i.i.d Gaussian vectors with mean 0 and variance $\boldsymbol{I}_{|\mathcal{A}|}$, suppose that the divergence sequence $b_{s,K}$ in **A1** is selected such that $\left[1-\Phi\{\log(b_{s,K})/\|\boldsymbol{\beta}_{\mathcal{A}}^0\|_2\}\right]^n \to 1$, then by a direct calculation we have $h_n^{-1} = O_p(b_{s,K})$ and thus Condition $\mathbf{A1}_{(G)}$ implies Condition **A1**. See also discussions in Zhang and Huang (2008) and Wainwright (2009).

Following Zhang and Huang (2008) and also to obtain a slightly stronger sign consistency result than that of Fan and Lv (2011) (when $K=1$), we introduce two diverging sequences $v_{n,K}$ and $u_{n,K}$ that depend on the total sample size n and the number of subsets K, and assume that

$$\mathbf{A2} \qquad v_{n,K} = o(\min\{n_k K b_{s,K}^{-1}\beta_*, n^{1-\alpha}K^{\alpha}\}) \text{ and } u_{n,K} = o(n).$$

These two sequences $v_{n,k}$ and $u_{n,k}$ are related to the error tolerance level under the design of Condition **A1**. Specifically, the probability of obtaining the correct signs of nonzero variables increases with $v_{n,k}$ and the probability of excluding variables with zero coefficients increases with $u_{n,k}$.

We also require that the tuning parameter λ_k satisfy the following conditions:

$$\begin{array}{ll} & b_{s,K}\rho'(\beta_*/2;\lambda_k) = o(\beta_*), \\ \mathbf{A3} & \max_{\boldsymbol{\delta}\in\mathcal{N}_0}\kappa(\rho(\cdot;\lambda_k);\boldsymbol{\delta}) = o(\tau_{0,k}), \\ & \|\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T\mathbf{\Sigma}(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}\{\boldsymbol{X}_{k,\mathcal{A}}^T\mathbf{\Sigma}(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}\|_\infty \le C\rho'(0+;\lambda_k)/\rho'(\beta_*/2;\lambda_k), \end{array}$$

where $\kappa(\rho(\cdot;\lambda_k);\boldsymbol{\delta}) = \lim_{\epsilon\to 0+}\max_{1\le j\le s}\sup_{t_1<t_2\in(\delta_j-\epsilon,|\delta_j|+\epsilon)} -[\rho'(t_2;\lambda_k)-\rho'(t_1;\lambda_k)]/(t_2-t_1)$, $\tau_{0,k} = \min_{\boldsymbol{\delta}\in\mathcal{N}_0}\lambda_{\min}[n_k^{-1}\boldsymbol{X}_{k,\mathcal{A}}^T\Sigma(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})\boldsymbol{X}_{k,\mathcal{A}}]$ and C is a positive constant with $C\in(0,1)$. We have the following theorem, and a proof is provided in Appendix.

THEOREM 1. *Suppose the sample size of the k^{th} subset $n_k = O(n/K)$, $k = 1,\ldots,K$, and $\max_{1\le k\le K} n_k/\min_{1\le k\le K} n_k = O(1)$. Assume that the regularity conditions A1 - A3 are satisfied and $s = o(\min\{(\beta_* b_{s,K})^{-1}, \beta_*^{-2}(K/n)^{\alpha}\})$. Then, with probability at least*

$$1 - 2Ks\exp\{-v_{n,K}^2/(nK)\} - 2K(p-s)\exp\{-u_{n,K}^2/(nK)\}, \tag{6}$$

the combined estimator is sign consistent, i.e. $\operatorname{sgn}(\hat{\boldsymbol{\beta}}^{(c)}) = \operatorname{sgn}(\boldsymbol{\beta}^0)$. *More specifically, we have* $\|\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_\infty \leq \beta_*/2$ *and* $\hat{\boldsymbol{\beta}}^{(c)}_{\overline{\mathcal{A}}} = 0$.

Theorem 1 suggests that the combined estimator $\hat{\boldsymbol{\beta}}^{(c)}$ is sign consistent under some regularity conditions and when the probability in (6) goes to 1. To ensure the probability in (6) goes to 1, we need to require that $\log(Ks) = o(\min\{nb_{s,K}^{-2}\beta_*^2/K,\ n^{1-2\alpha}K^{2\alpha}\})$ and $\log(Kp) = o(n/K)$. The latter requirement suggests that the growth rate of p needs to be controlled by $e^{n/K - \log(K)}$. This rate decreases in K and it is e^n when $K = 1$. So when we increase the number of splits, we impose a stronger constraint on the growth rate of p. This strengthened constraint ensures that each subset contain enough data to provide a sign consistent estimator for the unknown model parameters.

Let us consider now a special case with $\beta_* = O(n^{-\gamma}\log n)$, $\gamma \in (0, 1/2]$, the signal strength imposed in Fan and Lv (2011), and assume $s = O(n^{\alpha_0})$ with $\alpha_0 \in (0, \min(\gamma, 2\gamma - \alpha))$. Let $b_{s,K} = o(\min\{K^{-1/2}n^{1/2-\gamma}\sqrt{\log n}, s^{-1}n^{\gamma}/\log n\})$ and $K = o\{\min(n^{1-2\gamma}\log n, n^{\alpha_1})\}$. If we choose $v_{n,K} = \sqrt{Kn\log n}$ and $u_{n,k} = K^{1/2}n^{1-\alpha_1}(\log n)^{1/2}$ with $\alpha_1 = \min(1/2, 2\gamma - \alpha_0) - \alpha$, then we can show that Condition **A2** holds and $s = o(\min\{(\beta_* b_{s,K})^{-1}, \beta_*^{-2}(K/n)^{\alpha}\})$; A proof of this claim is provided in the Appendix. When $K = 1$, the conditions specified above in this paragraph are the same as those imposed in Fan and Lv (2011). A basic calculation in this special case leads us to require the growth rate of p be controlled by $e^{n^{1-2\alpha_1}/K}$, which becomes $e^{n^{1-2\alpha_1}}$ when $K = 1$. This rate $e^{n^{1-2\alpha_1}}$ is also reported in Fan and Lv (2011). Since the upper bound rate for p that is implied by Theorem 1 is e^n when $K = 1$, as described in the previous paragraph, Theorem 1 provides a slightly stronger result than that reported in Fan and Lv (2011) for the $K = 1$ case.

3.2. Oracle property

In this subsection, we show that, after we strengthen some of the regularity conditions, our combined estimator can also have an oracle property under the L_2 norm. Specifically, we show that the combined estimator can converge at the regular order of $O(\sqrt{s/n})$ under the L_2 norm. We also show that the combined estimator obtains asymptotic normality with the same variance as the penalized estimator using the entire data all at once. These results fully establish the asymptotic equivalence between the combined estimator and the penalized estimator using the entire data all at once.

In this subsection, we impose the following regularity conditions on the design matrices. They are the same as Condition 4 of Fan and Lv (2011) when $K = 1$.

$$\textbf{A4}\qquad \begin{array}{l} \min_{\boldsymbol{\delta}\in\mathcal{N}_\tau} \lambda_{\min}(\boldsymbol{X}_{k,\mathcal{A}}^T\boldsymbol{\Sigma}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})\boldsymbol{X}_{k,\mathcal{A}}) \geq cn_k, \\ tr(\boldsymbol{X}_{k,\mathcal{A}}^T\boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}) = O(sn_k), \\ \|\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T\boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}\|_{2,\infty} = O(n_k), \\ \max_{\boldsymbol{\delta}\in\mathcal{N}_0, 1\leq j\leq s} \lambda_{\max}[\nabla^2\{\boldsymbol{x}_{k,j}^T\boldsymbol{\mu}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta}_{\mathcal{A}})\}] = O(n_k), \end{array}$$

where c is some positive constant, $\|A\|_{2,\infty} = \max_{\|\boldsymbol{v}\|_2=1} \|A\boldsymbol{v}\|_\infty$ and $\boldsymbol{\delta} \in \mathcal{N}_\tau = \{\boldsymbol{\delta} \in \Re^s : \|\boldsymbol{\delta} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_2 \le \tau\sqrt{Ks/n}\}$ for any given positive constant τ.

The constraint imposed by Condition **A4** is also minor. Similar to Condition **A1**, in the linear regression model, Condition **A4** can be implied by

$$\mathbf{A4}_{(G)} \quad \begin{array}{l} \min_{\boldsymbol{\delta}\in\mathcal{N}_\tau} \lambda_{\min}(\boldsymbol{X}^T_{k,\mathcal{A}}\boldsymbol{X}_{k,\mathcal{A}}) \ge O_p(n_k), \\ tr(\boldsymbol{X}^T_{k,\mathcal{A}}\boldsymbol{X}_{k,\mathcal{A}}) = O_p(sn_k), \\ \|\boldsymbol{X}^T_{k,\bar{\mathcal{A}}}\boldsymbol{X}_{k,\mathcal{A}}\|_{2,\infty} = O_p(n_k). \end{array}$$

The first equation in condition $\mathbf{A4}_{(G)}$ matches with those discussed in the literature under both the settings of fixed matrix design (cf., Fan and Lv (2011), pages 5472) and Gaussian random matrix design (cf., Marcenko and Pastur (1967), theorem 1; Takemura and Sheena (2005), Lemma 1). The requirement in the third equation is also minor, since $\|\boldsymbol{X}^T_{k,\bar{\mathcal{A}}}\boldsymbol{X}_{k,\mathcal{A}}\|_{2,\infty} \le \|\boldsymbol{X}^T_{k,\bar{\mathcal{A}}}\boldsymbol{X}_{k,\mathcal{A}}\|_\infty$ and we can approximate the order from L_∞ norm. More generally and in non-Gaussian generalized linear models, we can bound the smallest eigenvalue of $\boldsymbol{\Sigma}(\boldsymbol{\theta}^0)$ by $h_n > 0$, similar to $\mathbf{A1}_{(G)}$. If $h_n = O(1)$, we have $\lambda_{\min}(\boldsymbol{X}^T_{k,\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})\boldsymbol{X}_{k,\mathcal{A}}) \ge \{h_n + o_p(1)\}\lambda_{\min}(\boldsymbol{X}^T_{k,\mathcal{A}}\boldsymbol{X}_{k,\mathcal{A}})$. In addition, by direct calculation: $tr(\boldsymbol{X}^T_{k,\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\theta}^0_k)\boldsymbol{X}_{k,\mathcal{A}}) = \sum_{j=1}^s \sum_{i=1}^{n_k} x^2_{ij}\sigma(\theta^0_i) \ge h_n \sum_{j=1}^s \sum_{i=1}^{n_k} x^2_{ij}$. It follows that, when either x_{ij} is fixed or a random variable such that $x_{ij} = O_p(1)$, Condition $\mathbf{A4}_{(G)}$ and thus also Condition **A4** are satisfied.

Similar to **A3**, we also impose a condition on the tuning parameter λ_k:

$$\mathbf{A5} \quad \max_{\boldsymbol{\delta}\in\mathcal{N}_\tau} \kappa(\rho(\cdot;\lambda_k);\boldsymbol{\delta}) = o(\tau_{1,k}),\ \rho'(\beta_*/2;\lambda_k) = O(n^{-1/2}),$$

where $\tau_{1,k} = \min_{\boldsymbol{\delta}\in\mathcal{N}_\tau} \lambda_{\min}[n_k^{-1}\boldsymbol{X}^T_{k,\mathcal{A}}\Sigma(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})\boldsymbol{X}_{k,\mathcal{A}}]$.

Finally, to ensure asymptotic normality, we also impose a Lindeberg-type of condition:

$$\mathbf{A6} \quad \begin{array}{l} \max_{i=1,\ldots,n} E|y_i - b'(\theta^0_i)|^3 = O(1), \\ \sum_{i=1}^n (\boldsymbol{z}_i^T\boldsymbol{B}^{-1}\boldsymbol{z}_i)^{3/2} \to 0 \ \text{ as } \ n \to \infty, \end{array}$$

where $\boldsymbol{B} = \boldsymbol{X}^T_{\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\theta}^0)\boldsymbol{X}_{\mathcal{A}}$ and $\boldsymbol{X}_{\mathcal{A}} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^T$.

The equivalence results are stated in the following theorem. A proof of the theorem can be found in the Appendix.

THEOREM 2. *Suppose the sample size of the k^{th} subset $n_k = O(n/K)$, $k = 1, \ldots, K$, and $\max_{1\le k\le K} n_k / \min_{1\le k\le K} n_k = O(1)$. Assume that the regularity Conditions A4- A5 are satisfied and $b_*/\sqrt{Ks/n} \to \infty$.*

(i) Assume that $Ks = o(\sqrt{n})$. We have, with probability approaching 1, $\hat{\boldsymbol{\beta}}^{(c)}_{\bar{\mathcal{A}}} = 0$ as $n \to \infty$ and $\|\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_2 = O(\sqrt{s/n})$.

(ii) Suppose further Condition A6 holds and assume that $K^{2/3}s = o(n^{1/3})$ and $\rho'(\beta_*/2;\lambda_k) = o(s^{-1/2}n^{-1/2})$. If $\boldsymbol{D}$ is a $q \times s$ matrix such that $\boldsymbol{D}\boldsymbol{D}^T \to \boldsymbol{G}$, where $\boldsymbol{G}$ is a $q \times q$ symmetric positive definite matrix, we have

$$\boldsymbol{D}[\boldsymbol{X}^T_{\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\theta}^0)\boldsymbol{X}_{\mathcal{A}}]^{1/2}(\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) \xrightarrow{\mathrm{D}} N(\boldsymbol{0}, \phi\boldsymbol{G}). \qquad (7)$$

The limiting distribution of $\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}}$ reported in (7) is exactly the same as that of $\hat{\boldsymbol{\beta}}^{(a)}_{\mathcal{A}}$ reported in Fan and Lv (2011), where the entire data is analyzed all at once. Thus, the combined estimator $\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}}$ is asymptotically as efficient as $\hat{\boldsymbol{\beta}}^{(a)}_{\mathcal{A}}$. Together with the fact that both estimators are model selection consistent, the combined estimator $\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}}$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}^{(a)}_{\mathcal{A}}$.

The requirements of the signal strength and the sparsity assumption in Theorem 2 depends on the number of splits. In the special case when $K = O(1)$ the asymptotic equivalence result holds without any additional requirements on either the signal strength or the sparsity assumption. But when we allows many splits with K going to infinity, we pay the price of imposing stronger conditions since some conditions such as minimal signal strength typically depend on the sample size and each subset has a smaller sample size. Specifically, in Theorem 2, we require that β_* need to be larger than $O(\sqrt{Ks/n})$ to ensure the sign consistency for the penalized estimator of each subset. So, compared with the signal strength $O(\sqrt{s/n})$ required by the penalized estimator using the entire dataset, the combined estimator needs larger coefficients to entail L_2 norm consistency. Regardless of the stronger conditions, the combined estimator remains asymptotically equivalent to the penalized estimator using the entire dataset. If stronger signal strength is a concern in a specific problem, two-stage estimation approach such as Zhang and Zhang (2011) may be used to weaken the requirement. In addition, the number of non-zero coefficients s has to be at the order of $O(n^{1/3}/K^{2/3})$ which is smaller than $O(n^{1/3})$ required when analyzing the entire dataset all at once. These strengthened conditions ensure that each subset contains enough data to provide a meaningful inference for the unknown model parameters. See Section 5 for further discussions.

3.3. Error control

In this subsection, we provide an upper bound for the expected number of falsely selected variables and a lower bound for the expected number of truly selected variables. In Theorem 3 below, $s^* = \sup_k \bar{s}_k$ and $s_* = \inf_k \bar{s}_k$ where $\bar{s}_k = E(|\hat{\mathcal{A}}_k|)$ is the average number of selected variables of the penalized estimator from the k^{th} subset. A similar result is provided by Fan et al. (2009) and Meinshausen and Buhlmann (2010), both of which only considered the special case of $K = 2$.

Theorem 3. *Assume the distribution of* $\{\mathbf{1}_{j\in\hat{\mathcal{A}}_k} : j \in \mathcal{A}\}$ *and* $\{\mathbf{1}_{j\in\hat{\mathcal{A}}_k} : j \in \overline{\mathcal{A}}\}$ *are exchangeable for all* $k = 1, \ldots, K$. *Also, assume the penalized estimators used are not worse than random guessing, i.e.* $E(|\mathcal{A} \cap \hat{\mathcal{A}}_k|)/E(|\overline{\mathcal{A}} \cap \hat{\mathcal{A}}_k|) \geq |\mathcal{A}|/|\overline{\mathcal{A}}|$, *for the set of selected variables* $\hat{\mathcal{A}}_k$ *of any penalized estimator. If* $w \geq s^*K/p - 1$, *then we have, for the combined estimator* $\hat{\boldsymbol{\beta}}^{(c)}$,

(i) the expected number of false selected variables has an upper bound: $E(|\overline{\mathcal{A}} \cap \hat{\mathcal{A}}^{(c)}|) \leq |\overline{\mathcal{A}}|\{1 - F(w|K, s^*/p)\}$,

(ii) the expected number of truly selected variables has a lower bound: $E(|\mathcal{A} \cap \hat{\mathcal{A}}^{(c)}|) \geq |\mathcal{A}|\{1 - F(w|K, s_*/p)\}$,

where $F(\cdot|m, q)$ *is the cumulative distribution function of binomial distribution with* m *trials and success probability* q.

Both s^* and s_* depend on the choice of the threshold w. In one extreme case with the threshold $w = K - 1$, the combined estimator only selects the variables that are selected in all K subsets. In this case, the expected number of falsely selected variables is bounded above by $(s^*)^K/p^{K-1}$. If further s^* is bounded by $c^{1/K}p^{1-1/K}$ for a constant c, the expected number of falsely selected variables is then bounded by the constant c. In sparse models, s^* is usually small and so is c. Therefore, the combined estimator controls the model selection error in a foreseeable way. In the opposite extreme case with the threshold $w = 0$, the combined estimator selects any variables that are selected in one or more subsets. In this case, the lower bound for the expected number of truly selected variables is tight, achieving the true number of non-zero set $|\mathcal{A}|$. But, in this latter case, the upper bound for the expected number of false selected variables can be very loose, up to $|\overline{\mathcal{A}}|$ the number of variables in the entire noise set.

There is a trade-off between the upper and lower bounds in Theorem 3 for the choice of w. A larger w typically gives us a smaller upper bound of the expected number of false selected variables but a smaller lower bound of the expected number of truly selected variables. A smaller w typically gives us a larger lower bound of the expected number of truly selected variables but a larger upper bound of the expected number of false selected variable. This intuition may help guide us to select a threshold w, depending the specific needs of applications for instance whether we prefer to control the sensitivity or specificity of model selections. In our numerical studies in Section 4, we use $w = K/2$, which appears to provide a good balance when s^* is smaller than $p/2$.

3.4. Computing issues

The split-and-conquer approach can significantly save computing time when the analysis involved is computationally intensive. In this subsection, we first study in detail the computing steps of LASSO estimators using the LARS algorithm (Efron et al., 2004) when $p > n$, and provide conditions under which the split and conquer approach is always computationally faster. We use the LARS algorithm as an illustrative example, since it has trackable computing steps and is also one of the most well known and studied methods. We then provide a calculation of average computing orders under general settings, which covers computationally intensive algorithms at the order of $O(n^a p^b)$, $a > 1$ and $b \geq 0$.

The following lemma is a more detailed restatement of computing related issues discussed in Section 7 of Efron et al. (2004), in which we spell out the exact computing steps in the LARS algorithm.

Lemma 1. *Suppose a LARS algorithm is applied to a data set with* n *observations and* p *variables where* $p \geq n$. *Then, the number of computing steps in* *the algorithm is greater than* $5n^3/3 + 23n/6 + 4n^2(p - 7/8) + 6np$ *but less than* $23n^3/3 + 71n/6 + 8n^2(p - 31/16) + 12np$.

Based on the upper and lower bounds stated in Lemma 1, the following theorem specifies a set of mild conditions under which the number of computing steps needed in the proposed split-and-conquer approach is always less than that of a directly use of the LARS algorithm on the entire data.

THEOREM 4. *Under the assumption in lemma 1, suppose* $p \geq 2$ *and the dataset is split into* K *subsets of size* $n_k = O(n/K)$*, for* $k = 1, \ldots, K$*, with* $\max_{1\leq k\leq K} n_k / \min_{1\leq k\leq K} n_k = O(1)$. *If* $K \geq 3$*,* $n \geq 4(4+3p)/\{1 + 8p(1 - 2/K) + 31/K - 7\}$ *and the computing effort of the combination is ignorable, the proposed split and conquer approach always has less computing steps than that of a direct use of a LARS algorithm on the entire data all together.*

The statement in Theorem 4 is conservative since, to ensure that the statement always holds, the conclusion is reached based on a comparison of the worst case scenario in the split and conquer approach against the best case scenario of using the LARS algorithm on the entire data. Our numerical study later in Section 4.1 suggests that on average the computing time saved by the split-and-conquer approach is quite significant. For instance, Figure 1 demonstrates how the average computing time changes for different n, p and K using the LARS algorithm. It indicates that the average computing time decreases dramatically for the split and conquer approach compared with that required for analyzing the entire dataset all at once. In Figure 1(a), the number of parameters p and the total sample size n are at the same scale ($p = 2n$). In Figure 1(b), the number of parameters p is much larger than the total sample size n ($p = 100n$).

Based on Lemma 1, the computing steps of the LARS algorithm are at the order of $O(n^2p)$, when p is much bigger than n. Thus, when $n_k = O(n/K)$ and the computing effort of the combination is ignorable, the split and conquer approach only needs $K \times O((n/K)^2 p) = 1/K \times O(n^2p)$ steps. This means that the proposed split and conquer approach, on average, can result in a computing saving by the order of $(1 - 1/K)100\%$. [Remark: In Efron et al. (2004) (page 444), the computing order of a LARS algorithm was reported as $O(n^3)$ instead of $O(n^2p)$ when $p > n$. We think this is a typographical error. If the order of $O(n^3)$ were correct, the computing saving by a split and conquer approach would be by the order of $(1 - 1/K^2)100\%$.] When p is at the same order as n, the computing order of a LARS algorithm is $O(n^2p + n^3) = O(n^3)$. Thus, when $n_k = O(n/K)$ and the computing effort of the combination is ignorable, the split-and-conquer approach, on average, saves computing time roughly by the order $(1 - 1/K^2)100\%$.

The above calculation of computing saving can be extended to any statistical procedure that requires $O(n^a p^b)$ computing steps, for any $a > 1$ and $b \geq 0$. We have the following statement, which provides an intuitive interpretation on how much computing time, on average, can be saved. A similar result in a computational intensive robust multivariate scale estimation (where p is fixed) was reported in Section 5.3 of Singh et al. (2005).

THEOREM 5. *Assume a statistical procedure requires $O(n^a p^b)$ computing steps, $a > 1$ and $b \geq 0$, when sample size is n. Suppose the dataset is split into K subsets with almost equal sample size $n_k = O(n/K)$ and the computing effort of the combination is ignorable. Then, the split-and-conquer approach only needs $K \times O((n/K)^a p^b)$, that is $O(n^a p^b / K^{a-1})$, steps. Thus, using the split-and-conquer approach results in a computing saving by the order of K^{a-1} times.*

In Theorems 4 and 5, we assume that the computing effort in the combination is negligible. This assumption is often satisfied in our context. Specifically, the combined estimator in (5) are calculated in two parts: we first use a majority voting to determine the number of non-zero coefficients and then use a weighted linear combination formula to combine the K estimators from the subsets. In the first part, there are roughly Ks non-zero coefficients across all K subsets, and a computing up to the order of $O(Ks)$ is often enough to identify these Ks variables of non-zero coefficients. In the second part the computing order is also depending on the number of non-zero coefficients s and K, and the total computing order is $O(Ks + Ks^2 + s^3)$ where the highest order $O(s^3)$ is for the inversion of the size (roughly) $s \times s$ matrix (Golub and Van Loan, 1983; Trefethen and Bau III, 1997). As $n \to \infty$, the computation in the combination is often negligible in many examples, compared to the oder $O(n^a p^b)$.

In the special example of the LARS algorithm, a reviewer pointed out an interesting alternative approach which directly applies a split and conquer method in the calculation of sample covariance matrix instead of the LARS estimator. Under the LASSO and LARS setting, the Gram matrix $\boldsymbol{X}^T \boldsymbol{X}$ is the only sufficient statistic. When $p < n$, it is independent of n. We can simply use a parallel computing approach to obtained the overall Gram matrix $\boldsymbol{X}^T \boldsymbol{X}$, and then feed it into the LARS solver. This alternative approach can effectively handle big data problems involving the LASSO/LARS method when $p < n$. However, when $p > n$ (the case considered in this subsection), the LARS algorithm fits at most n variables. It is undesirable to directly feed the overall Gram matrix to the LARS solver. We also received a warning message in the LARS solver in R (R package *lars*) when we tried to use this alternative approach to re-analyze the data of Figure 1(b). Furthermore, since the inversion of the Gram matrix is the most costly computing part in the LARS algorithm (cf. the proof of Lemma 1), the alternative approach often does not significantly save computing time, even in the case when $p > n$. In summary, when $p < n$, both the proposed and the alternative approaches can be used, and they produce asymptotically equivalent results. When $p > n$, we suggest to use our proposed split and conquer approach. It can save computing time and be easily implemented by using the standard software to analyze each subset data and then combining them via a simple majority voting and a weighted linear sum.

4. Numerical studies

In this section, we provide several numerical studies, using both simulation and real data, to illustrate the performance of the proposed split-and-conquer ap-

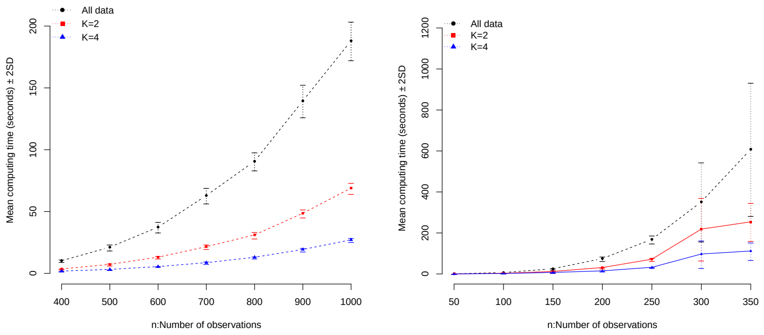

Fig. 1. Computing time comparison for different K using LARS algorithm: Mean ± 2Standard Deviation (SD) over 100 replications. Figure 1 (a) [left] is for $p = 2n$ with $n = 400, 500, 600, 700, 800, 900, 1000$. Figure 1 (b) [right] is for $p = 100n$ with $n = 50, 100, 150, 200, 250, 300, 350$.

proach. We also compare the combined estimators with their corresponding penalized estimators obtained using the entire data all at once, whenever the computing of the latter approach does not reach the limits of the computer used in our project (a W35653 20GHz, 2G RAM workstation using R 2.13.1 under Windows 7). We focus on two models, the Gaussian linear regression model and the logistic model, with different choices of sample size n, number of parameters p and true model size s (the number of nonzero regression parameters). The development is illustrated using the L_1, SCAD and MCP penalty functions, three widely used penalties in the literature.

4.1. Linear regression with L_1 norm penalty

This subsection focuses on the setting of a linear regression and the L_1 norm penalty. In particular, the response variable $\boldsymbol{y}$ follows a Guassian linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon,$$

where ε are IID $N(0, 1)$ errors and the explanatory variables $\boldsymbol{X}$ are generated from a $N(0, \boldsymbol{I})$ distribution with $\boldsymbol{I}$ being identity matrix. In our simulation study, four sample settings of (n, p), with $n \leq p$, are considered (see Table 1). The true model $\mathcal{A} = \{j : \beta_j^{(0)} \neq 0\}$ in each setting contains $s = \lfloor \sqrt{p} \rfloor$ nonzero coefficients whose true values are around $\sqrt{2K \log(p)/n}$. To get the LASSO estimators using the L_1 norm penalty, the LARS algorithm (Efron et al., 2004) is applied and BIC criterion is used for selecting the tuning parameter.

We repeat our simulation 100 times under each setting of (n, p). For the final overall estimators, we record the mean of computing time and the number of selected nonzero coefficients. To demonstrate the error control property, we

Table 1. Comparison of the combined estimator and the complete estimator (with standard deviation in the parenthesis)

Simulation setting					Model selection			
n	p	s	K	Computing time (in second)	w	# selected variables	sensitivity (in %)	specificity (in %)
500	500	22	1	41.55 (5.37)	-	36.01 (5.87)	100 (0)	97.07 (1.23)
			2	5.77 (0.51)	1	66.56 (9.72)	100 (0)	90.68 (2.03)
			4	2.74 (0.46)	1	157.55 (16.05)	100 (0)	71.64 (3.36)
					2	37.49 (5.10)	98.68 (2.53)	96.70 (1.06)
			6	1.71 (0.22)	1	221.92 (14.25)	99.73 (1.08)	58.16 (2.93)
					2	63.96 (7.63)	96.73 (3.66)	91.07 (1.58)
					3	24.11 (2.61)	86.73 (7.33)	98.95 (0.42)
500	800	28	1	24.72 (1.77)	-	48.32 (6.60)	100 (0)	97.37 (0.86)
			2	8.10 (0.60)	1	102.75 (11.84)	99.93 (0.50)	90.31 (1.53)
			4	3.60 (0.38)	1	240.06 (14.99)	99.18 (1.89)	72.50 (1.94)
					2	50.96 (5.92)	92.29 (5.39)	96.75 (0.74)
			6	2.52 (0.27)	1	294.60 (11.50)	97.18 (2.93)	65.36 (1.50)
					2	69.31 (6.84)	83.50 (7.28)	94.05 (0.90)
					3	20.66 (3.34)	58.46 (9.71)	99.44 (0.27)
500	1000	31	1	28.06 (1.85)	-	59.09 (7.80)	100 (0)	97.20 (0.81)
			2	10.04 (0.60)	1	135.72 (16.58)	99.81 (1.32)	89.28 (1.71)
			4	4.48 (0.41)	1	284.18 (15.89)	97.03 (2.68)	73.85 (1.64)
					2	54.13 (5.80)	83.53 (6.86)	97.17 (0.56)
			6	2.92 (0.27)	1	325.83 (10.94)	93.19 (4.34)	69.42 (1.15)
					2	64.46 (5.84)	70.31 (7.58)	95.67 (0.63)
					3	16.60 (3.16)	41.88 (7.77)	99.67 (0.19)
1000	1000	31	1	393.10 (46.82)	-	47.86 (6.54)	100 (0)	98.36 (0.68)
			2	57.30 (2.87)	1	83.51 (12.31)	98.36 (0.68)	94.68 (1.27)
			4	20.21 (2.24)	1	217.77 (18.11)	100 (0)	80.81 (1.87)
					2	46.53 (4.72)	99.87 (0.62)	98.50 (0.49)
			6	12.66 (1.63)	1	381.51 (21.69)	99.94 (0.44)	63.89 (2.24)
					2	94.18 (8.31)	99.81 (0.75)	93.57 (0.86)
					3	37.51 (3.13)	97.59 (2.62)	99.35 (0.30)

also calculate model selection sensitivity and model selection specificity. Here, model selection sensitivity is defined as the number of truly selected variables divided by the true model size, and model selection specificity is defined as the number of truly removed variables divided by the number of noise variables. The simulation results are shown in Table 1. In Table 1, $K = 1$ means the entire data are used all at once to get the LASSO estimator; otherwise, the combined estimator proposed in this paper is used. To examine the performance of the combined estimator, we have tried different values of K and w, where $K = 2, 4, 6$ and $w = 1, \dots, \lfloor K/2 \rfloor$.

According to Table 1, all estimators select some noise variables in addition to the true s nonzero variables. This is consistent with a known performance of the LASSO-type estimators that they usually intend to include more variables than desired in model selections. When $K \geq 2$ and $w = [K/2]$, the combined estimator shows the benefit of error controls through the split-and-conquer approach. In particular, when $K = 4$ or 6 and $w = 2$ or 3, the model selection specificities increase a lot. This indicates that the combined estimator is more efficient in removing noise and spurious variables from the selected models. Also, at each given setting of (n, p, K), with the increase of the threshold w, the model selection sensitivity decreases but its specificities increases. This is also consistent with the message on trade-offs of two types errors discussed in Section 3.3.

Moreover, the computing time decreases drastically when K increases, as indicated reported in column 4 of Table 1. The time savings reported are between $(1 - 1/K^2)100\%$ and $(1 - 1/K)100\%$. We speculate this may be because n and p are roughly the same in the cases considered in Table 1. In such cases, both the two leading terms $O(1/n^3)$ and $O(1/(n^2 p))$ described in Lemma 1 contribute significantly to the overall saveings of the computing time, resulting to the percentages between $(1 - 1/K^2)100\%$ and $(1 - 1/K)100\%$. To further study computing savings in the LARS algorithm, we have performed additional simulations with (a) $p = 2n$ for $n = 400 - 1000$ and (b) $p = 100n$ for $n = 50 - 350$, under the same linear regression set up. In the simulations we have considered $K = 1, 2, 4$. The average computing times (with standard errors) over 100 repetitions are plotted in Figure 1. As shown in the figure, the savings appear to be between $(1 - 1/K^2)100\%$ and $(1 - 1/K)100\%$ in Figure 1(a) and roughly around $(1 - 1/K)100\%$ in Figure 1(b).

4.2. Generalized linear model with SCAD and MCP penalties

The SCAD and MCP estimators are two commonly used estimators that are obtained based on non-concave penalized likelihood functions. Compared with the LASSO estimators, they often select a tighter model and fewer noise variables. We consider in this subsection both the SCAD and MCP estimators under both the linear regression and logistic models.

For the linear regression case, the response variable $\boldsymbol{y}$ follows the model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$, where ε are IID $N(0, 1)$ errors. For the logistic regression case, the response variable $\boldsymbol{y}$ follows the Bernoulli distribution with the success probability $p(\boldsymbol{X}\boldsymbol{\beta}) = e^{\boldsymbol{X}\boldsymbol{\beta}}/(1 + e^{\boldsymbol{X}\boldsymbol{\beta}})$. In our simulations, we consider two settings to generate the design matrix $\boldsymbol{X}$: one is for independent variables and the other is for correlated variables.

1. Independent variables: a set of p variables are generated from a $N(0, \boldsymbol{I})$ distribution, where $\boldsymbol{I}$ is identity matrix.

2. Correlated variables: a set of p variables are generated from a $N(0, \boldsymbol{\Sigma})$ distribution, where $\boldsymbol{\Sigma}(i, j) = 0.6^{|i-j|}$ is the covariance matrix.

Two settings of sample sizes are considered $n = 10000$ and $n = 100000$. In the linear regression, the number of parameters $p = 1000$ and in the logistic model, the number of parameters $p = 200$. In all cases, the true model contains $s = 30$ nonzero coefficients (with values around 0.4) and the true model size $s = 30$

is relatively small compared to p and n. In order to get the SCAD and MCP estimators, the NCVREG algorithm (Breheny and Huang, 2011) is applied and a 10-fold cross-validation is used to select the tuning parameters.

The simulation is repeated 100 times. Similarly as in Example 1, we record the computing time and the number of selected variables and calculate model selection sensitivity and specificity. In addition, the MSE (mean squared error) is calculated in the linear regression case and the misclassification rate with 0.5 as threshold is reported in the logistic regression case. The results are displayed in Table 2. In the table, $K = 1$ refers to analysis of an entire dataset all at once. When $K > 1$, the proposed split-and-conquer approach is applied.

The computing times are reduced through the split-and-conquer procedure under all settings, although we do not have a theoretical calculation of the order of computing steps, thus the computing savings, in the algorithms to perform SCAD and MCP penalized regressions. For both the SCAD and MCP penalties, the proposed split-and-conquer approach can reduce the computing time drastically in the linear regression setting (with $K = 10$), using about 1/10 of time when the explanatory variables are independent and about 1/3 of when the explanatory variables are correlated. Under the setup of the logistic model (with $K = 5$), the average saving is a little less. When the explanatory variables are independent, the combined estimator needs about half of the time compared to directly performing the same analysis on the entire data all together. When the explanatory variables are correlated, the combined estimator by the proposed method can save up to 25% time compared to directly performing the same analysis on the entire data all at once. When the sample size $n = 100000$, we are not able to perform either the SCAD or the MCP regression on the entire data all together due to computer memory limitations. But we can obtain the combined estimators using the split-and-conquer procedure; see the results reported in Table 2.

According to Table 2, the SCAD estimators perform similar to the MCP estimators. In all cases, the combined estimators have good model selection results with high model selection sensitivity and specificity that are similar to those of the penalized estimators analyzing the entire data all together. Moreover, under the linear regression setting, the combined estimator have similar MSEs to the corresponding penalized estimators using the entire data all together. Under the logistic regression settings, the misclassification rates of the combined estimators are also close to those of the corresponding penalized estimators using the entire data all together.

Figure 2 presents several sets of side-by-side boxplots to compare the combined estimates $\hat{\boldsymbol{\beta}}^{(c)}$ by the proposed split-and-conquer method with the penalized estimates $\hat{\boldsymbol{\beta}}^{(a)}$ by analyzing the entire data all together, when both are available in the settings of Table 2. The top panel of four figures are for the linear regression settings and the bottom panel of four figures are for logistic regression settings. Within each figure, estimates of nonzero $\boldsymbol{\beta} \in \mathcal{A} = \{j : \beta_j^0 \neq 0\}$ are plotted, with $\hat{\boldsymbol{\beta}}^{(c)}$ arranged on the left (colored brown) and $\hat{\boldsymbol{\beta}}^{(a)}$ on the right

Table 2. Comparison of the combined estimates and the complete estimates (with standard deviation in the parenthesis); Here, $s = 30$ under all settings.

Part I: Linear regression								
Simulation setting					Model selection			
Design matrix	n	p	K	Computing time (in second)	# selected variables	sensitivity (in %)	specificity (in %)	MSE
SCAD: Linear regression								
Independent	10000	1000	1	815.27 (77.98)	34.58 (9.81)	100 (0)	99.53 (1.01)	1.00 (0.01)
			10	104.96 (9.55)	30 (0)	100 (0)	100 (0)	1.00 (0.01)
Correlated	10000	1000	1	755.4 (157.56)	34.00 (12.22)	96.00 (19.79)	99.46 (1.02)	0.96 (0.20)
			10	289.17 (61.03)	28.72 (6.13)	95.87 (19.78)	100 (0)	1.00 (0.01)
Independent	100000	1000	1	- (-)	- (-)	- (-)	- (-)	- (-)
			100	1136.70 (74.65)	30 (0)	100 (0)	100 (0)	1.00 (0.01)
Correlated	100000	1000	1	- (-)	- (-)	- (-)	- (-)	- (-)
			100	3074.53 (25.01)	30 (0)	100 (0)	100 (0)	1.06 (0.01)
MCP: Linear regression								
Independent	10000	1000	1	2243.45 (155.82)	34.58 (9.81)	100 (0)	99.79 (0.41)	1.00 (0.01)
			10	163.72 (12.95)	30 (0)	100 (0)	100 (0)	1.00 (0.01)
Correlated	10000	1000	1	1244.73 (80.86)	31.92 (5.69)	100 (0)	99.80 (0.59)	0.99 (0.01)
			10	442.14 (42.42)	29.98 (0.14)	99.93 (0.47)	100 (0)	1.01 (0.02)
Independent	100000	1000	1	- (-)	- (-)	- (-)	- (-)	- (-)
			100	1565.54 (132.38)	30 (0)	100 (0)	100 (0)	1.00 (0.01)
Correlated	100000	1000	1	- (-)	- (-)	- (-)	- (-)	- (-)
			100	4256.52 (215.60)	30 (0)	100 (0)	100 (0)	1.02 (0.01)
Part II: Logistic regression								
Simulation setting					Model selection			
Design matrix	n	p	K	Computing time (in second)	# selected variables	sensitivity (in %)	specificity (in %)	Misclassificaton rate (in %)
SCAD: Logistic regression								
Independent	10000	200	1	198.85 (5.88)	35.54 (5.71)	100 (0)	96.74 (3.36)	17.32 (0.40)
			5	116.49 (2.78)	31.70 (1.33)	100 (0)	99.00 (0.78)	17.40 (0.38)
Correlated	10000	200	1	463.61 (20.16)	38.18 (5.58)	99.33 (1.35)	95.02 (3.15)	9.90 (0.29)
			5	359.29 (7.94)	32.38 (2.42)	96.07 (2.75)	97.84 (1.27)	10.10 (0.26)
Independent	100000	200	1	- (-)	- (-)	- (-)	- (-)	- (-)
			20	1352.14 (76.2)	30 (0)	100 (0)	100 (0)	17.38 (0.12)
Correlated	100000	200	1	- (-)	- (-)	- (-)	- (-)	- (-)
			20	4014.48 (284.69)	29.97 (0.2)	99.87 (0.67)	100 (0)	9.96 (0.09)
MCP: Logistic regression								
Independent	10000	200	1	201.46 (6.74)	31.8 (2.77)	100 (0)	98.94 (1.63)	17.31 (0.34)
			5	118.85 (3.17)	30.24 (0.62)	99.87 (0.66)	99.84 (0.34)	17.38 (0.35)
Correlated	10000	200	1	582.182 (59.02)	35.48 (4.22)	98.73 (1.89)	96.55 (2.27)	9.84 (0.33)
			5	557.43 (22.7)	28.7 (1.63)	92.93 (3.85)	99.52 (0.60)	10.17 (0.32)
Independent	100000	200	1	- (-)	- (-)	- (-)	- (-)	- (-)
			20	1301.95 (63.27)	30 (0)	100 (0)	100 (0)	17.34 (0.13)
Correlated	100000	200	1	- (-)	- (-)	- (-)	- (-)	- (-)
			20	4485.9 (186.29)	29.58 (0.50)	98.60 (1.66)	100 (0)	10.00 (0.09)

(colored yellow) within each pair of boxplots. From these boxplots, we can clearly see that the combined estimates have almost the same mean and spread as the estimates obtained using entire data all at once. In the logistic model case, the estimation of covariance matrix can influence the combined estimates. We use here the maximum likelihood estimates of covariance matrices based on only the selected variables in $\hat{\mathcal{A}}$ to get the weight matrices.

4.3. Numerical analysis on POEs manifest data

After the 911 terrorist attack, substantial efforts have been made in devising strategies for inspecting containers coming through the US POEs every day to interdict illicit nuclear and chemical materials. Manifest data, compiled from the custom forms submitted by merchants or shipping companies, are collected by the US custom offices and the Department of Homeland Security (DHS). Analysis of the manifest data to flag out potentially illegitimate activities is a small but

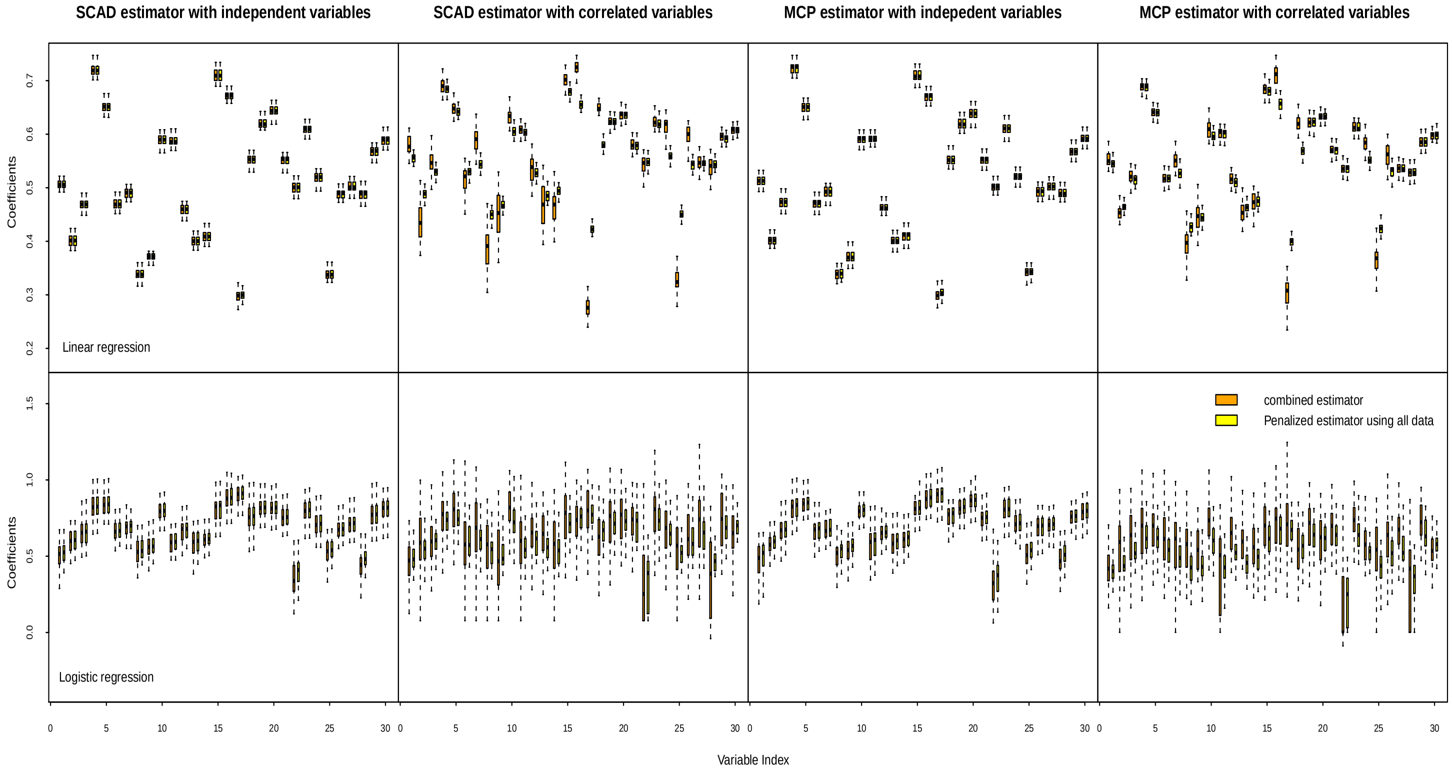

Fig. 2. Comparison of parameter estimation for the combined estimator and the penalized estimator using all data. Box plots of estimation for variables in the true model. Orange: the combined estimator; Yellow: the estimator using all data. Top panels: Linear regression; bottom panels: Logistic regression

Table 3. Manifest data: Dictionary of Variables

Variables	Number of Categories	Definition
$\boldsymbol{X}_1$	9	Vessel Country Code
$\boldsymbol{X}_2$	69	Voyage Number
$\boldsymbol{X}_3$	9	dp of Unlading
$\boldsymbol{X}_4$	14	Foreign Port Lading
$\boldsymbol{X}_5$	68	Foreign Port
$\boldsymbol{X}_6$	35	Inbond Entry Type
$\boldsymbol{X}_7$	17	Container Cotents

important part of layered defenses for the national security. In a nuclear detection project sponsored by the Command, Control, and Interoperability Center for Advanced Data Analysis (CCICADA), a Department of Homeland Security (DHS) Center of Excellence, we obtain a set of manifest data that contain all shipping records coming through the POEs across the US in February, 2009. The goal is to make quantitative evaluations of the manifest data and to develop an effective risk scoring approach that can be used to assist the assessment of future shipments. In the project, a logistic regression model has been used to enhance the effectiveness of the real-time inspection system with binary response variable indicating high-risk shipments. Since not all information collected in the manifest data are relevant to risk scoring and there are also many redundant information, we use the SCAD penalized regression to evaluate the importance of these variables. Table 3 provides the definition and a description of some variables contained in the manifest data. Most of these variables are categorical and dummy variables for each categorical variable are created which results in $p = 213$ variables in total. There are also text fields that can potentially lead to a much larger p. But to simply our discussion and without loss of our focus, we only illustrate the proposed split-and-conquer approach using this $p = 213$ variables and we do not consider any semantic analysis and text mining approaches in this paper.

Due to the enormous size of traffic and a large number of entry sites, it is a challenge to us to analyze the entire data all at once on a single computer. For instance, there are 164721 shipments in one week from February 20, 2009 to February 26, 2009. A computer with 2 GB memory and 3.2GHz CPU fails to perform the SCAD penalized regression on the one-week data. Even though we can solicit a high-performance super computer from our university by purchasing its computing hours, it still takes time to carry out the task and also has additional cost. Relying on a super computer to do research is inefficient in practice, especially we also need to constantly update the models over the months and years. This difficulty has motivated our research to propose the split-and-conquer approach.

Because of security concerns, the record of high-risk shipments is not provided in the project. But we have been told by experts in the filed to use the rate 1% to 10% of cargo containers that need further inspections in the context of inspections of drugs and other illicit materials. We turn to a simulation to

Table 4. Comparison of the combined weekly estimator and daily estimators (standard deviation in the parenthesis)

	Model selection			
	# of selected variables	Sensitivity (in %)	Specificity (in %)	Misclassification rate (in %)
Week (Combined)	21.06 (0.38)	95.25 (0.09)	99.95 (0.14)	3.97 (0.05)
Mon	32.66 (4.00)	92.53 (0.36)	94.2 (1.78)	3.99 (0.05)
Tues	29.18 (3.07)	95.4 (0.05)	96.14 (1.44)	3.98 (0.05)
Wed	9.22 (4.58)	23.13 (1.2	98.05 (1.18)	3.99 (0.05)
Thur	10.86 (4.6)	27.73 (1.08)	97.76 (1.28)	3.98 (0.05)
Fri	25.6 (2.09)	95.45 (0)	97.83 (0.98)	4.00 (0.05)
Sat	29.76 (3.47)	95 (0.14)	95.82 (1.61)	3.98 (0.05)
Sun	30.6 (3.31)	95.1 (0.12)	95.44 (1.57)	3.99 (0.05)

generate 1% to 10% high-risk shipments among the given manifest data. In particular, with assistance of two field experts, potentially influential shipment characteristics are selected (c.f., the 22 variables listed in the first column of Table 5) and we then use these characteristics and a logistic model to generate 1% to 10% high-risk shipments. Our task now becomes to test whether a penalized regression technique can be used to identify these 22 characteristics (pretending we do not know these 22 variables are important) among all shipment features recorded in the manifest data. Our computer can perform the SCAD penalized regression on a single-day data but it runs out of memery when we try to analyze a week-long data all at once. To overcome the difficulty, we apply the proposed split-and-conquer method to analyze the week-long manifest data. Specifically, we perform the SCAD penalized regression on everyday's data and combine the seven daily estimators together to obtain an overall combined estimator.

Tables 4 contains the values of model selection sensitivity, model selection specificity and misclassification rate and Table 5 reports the average estimates of the non-zero parameters from 100 replications, based on the split-and-conquer approach as well as the SCAD penalized regression using the data of a single day. The true model used in this study has $s = 22$ non-zero parameters, corresponding to a subset of 22 dummy variables from three categories: Vessel Country Code, Foreign Port Landing and Container Contents. Clearly, the split-and-conquer approach succeeds in performing the penalized logistic regression analysis on the whole week manifest data. As we can see from Table 4, the split-and conquer approach has identified most influential variables in the manifest data. In particular, the combined estimates have both high model selection sensitivity and specificity. On a contrast, the daily estimates either select more noise variables or exclude more influential variables. Also, the combined estimates are more stable (small standard errors in the average model size (0.38), sensitivity (0.09) and specificity (0.14)) than daily estimates. The combined estimator has a slightly smaller misclassification rate, comparable to all other estimators which is around but slightly less than 4%.

In terms of parameter estimation, as indicated in Table 5, the combined estimates also have smaller variances than the penalized estimates that only use

Table 5. Manifest data analysis through split-and-conquer approach

	Week	Daily estimation						
Categories	(Combined)	Mon	Tues	Wed	Thur	Fri	Sat	Sun
				Vessel country code				
PA	0.33(0.06)	0.2(0.17)	0.36(0.15)	0.07(0.14)	0.14(0.14)	0.46(0.07)	0.41(0.16)	0.4(0.14)
LR	1.78(0.07)	1.7(0.22)	1.75(0.19)	0.8(0.39)	1.64(0.16)	1.78(0.16)	1.75(0.17)	1.73(0.13)
DE	0.26(0.06)	0.22(0.17)	0.39(0.16)	0.01(0.06)	0.02(0.11)	0.47(0.11)	0.32(0.19)	0.31(0.2)
				Foreign port lading				
570	1.54(0.05)	1.59(0.15)	1.56(0.13)	0.92(0.35)	1.36(0.33)	1.53(0.08)	1.58(0.17)	1.53(0.12)
582	0.9(0.07)	1(0.23)	1.1(0.14)	0.26(0.21)	0.36(0.23)	0.84(0.17)	0.92(0.26)	0.63(0.25)
580	1.13(0.06)	1.39(0.17)	0.85(0.23)	0.03(0.09)	0.45(0.29)	1.33(0.1)	0.72(0.23)	1.27(0.14)
				Container contents				
Material	1.31(0.1)	1.98(0.24)	2.03(0.18)	0.12(0.27)	0.1(0.22)	2.06(0.17)	2(0.23)	1.97(0.24)
Animals	0.05(0.11)	0.27(0.21)	0.74(0.28)	0(0)	0(0)	0.63(0.21)	0.47(0.24)	0.46(0.25)
Entertainment	1.04(0.15)	1.55(0.36)	1.75(0.32)	0.03(0.12)	0.03(0.14)	1.85(0.23)	1.48(0.31)	1.56(0.33)
Industry	0.76(0.1)	1.39(0.25)	1.5(0.19)	0.03(0.22)	0.01(0.1)	1.55(0.18)	1.43(0.2)	1.44(0.18)
Cloth	0.65(0.08)	1.31(0.17)	1.37(0.12)	0.03(0.19)	0.02(0.13)	1.4(0.1)	1.32(0.17)	1.3(0.15)
Electro	0.44(0.13)	1.02(0.37)	1.09(0.28)	0.01(0.12)	0.01(0.12)	1.38(0.26)	0.91(0.26)	1.02(0.28)
Food	0.7(0.08)	1.41(0.14)	1.4(0.15)	0.02(0.17)	0.05(0.19)	1.46(0.11)	1.36(0.14)	1.34(0.12)
Furniture	1.34(0.11)	2.01(0.25)	2.09(0.22)	0.08(0.24)	0.12(0.23)	2.14(0.18)	2.01(0.26)	1.95(0.22)
Hardware	0.24(0.07)	0.88(0.18)	0.94(0.14)	0.01(0.1)	0(0.03)	0.97(0.1)	0.87(0.17)	0.9(0.15)
Health	0.53(0.09)	1.18(0.15)	1.23(0.13)	0.02(0.14)	0.01(0.12)	1.25(0.1)	1.19(0.15)	1.18(0.13)
Home	1.18(0.1)	1.91(0.24)	1.91(0.19)	0.09(0.26)	0.03(0.16)	1.95(0.15)	1.87(0.2)	1.83(0.2)
Motor	0.28(0.14)	0.89(0.3)	1.01(0.32)	0.03(0.25)	0.01(0.1)	1.19(0.29)	1.18(0.37)	1(0.33)
Media	0.98(0.11)	1.69(0.23)	1.75(0.26)	0.03(0.14)	0.02(0.13)	1.79(0.2)	1.47(0.29)	1.46(0.28)
Office	-0.17(0.13)	0.24(0.25)	0.55(0.26)	0.01(0.06)	0(0)	0.55(0.25)	0.4(0.25)	0.54(0.29)
Sporting	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
Mature	0.45(0.08)	1.15(0.13)	1.17(0.13)	0.02(0.15)	0.01(0.1)	1.23(0.1)	1.14(0.14)	1.14(0.11)

daily data. The Sporting variable in the category of Container Contents is left out in the model selections by all daily analyses and also the split-and-conquer approach. All other 19 variables with non-zero coefficients are recovered by the split-and-conquer approach. In general, by incorporating one-week information, the split-and-conquer approach provides more reliable results with better performance than any of the daily analysis.

5. Discussions

We propose in this paper a split and conquer methodology for analysis of big data. Specifically, an entire dataset is randomly split into non-overlapped small subsets. Each subset is analyzed separately using a desired statistical procedure. Then, the results from all subsets are combined together to provide a final overall statistical inference that contains information from the entire dataset. We demonstrate the proposed split and conquer approach using several penalized regression approaches that are widely used in the analysis of high-dimensional data. It is shown, both theoretically and numerically, that the result obtained from the split and conquer approach is asymptotically equivalent to the result of analyzing the entire data all at once under some mild conditions.

The split and conquer approach provides an applicable way to analyze big data using available resources. The approach is very general and can have many applications. As the entire data are split into smaller pieces, each subset requires a smaller storage space and computer memory when we perform our statistical analysis. Moreover, we have shown that the split and conquer approach needs less computing time when the desired statistical method is computationally intensive. Even in the case in which the desired statistical method is efficient, a reduced computing time can be expected operationally because we now can analyze different subsets at the same time using different computers. Also, in the case when the regular non-split version and the proposed split and conquer method both can be applied, if computing method is intensive, it may also be beneficial to use the split and conquer method to save computing time, although some stronger assumptions are needed. This computing improvement is very useful in many practical applications.

One important step in the split-and-conquer approach is the combination. The specific combination method to be used depends on the desired statistical procedure. As illustrated by penalized regressions in this paper, the properly weighted and linearly combined estimator is asymptotically equivalent to the one from analyzing the entire data all together although the combined estimator requires slightly stronger conditions. According to Singh et al. (2005), Xie et al. (2011) and Liu (2012), equivalent combined statistics or asymptotic efficiency are achievable for many other models. The proposed split and conquer approach can be easily extended to other problem settings as well as problems beyond point estimations including those using hypothesis testings and confidence intervals.

As a general methodology and for easy implementation in practice, we assume in this development that the intended statistical method and computing algorithm for the entire data is also used for each subset. No extra effort or a difference algorithm is required. We also assume that we can get a meaningful estimator from each subset. With a smaller sample size in each subset, the assumption may impose sometimes stronger conditions for analysis. We argue that this might not be a big concern in many applications involving big data. For instance, in the example of regular linear and regression and least squares estimator, in order to obtain meaningful least squares estimators from the subsets, we need assume a lightly strong condition that $\boldsymbol{X}_k^T\boldsymbol{X}_k$ is invertible. Although depending on the specific practice, this is often not an issue especially when we have big data. The theoretical developments in Sections 3.1 and 3.2 are in the same nature in this aspect. In some specific cases, it is possible to obtain the same conclusion without any stronger conditions, although to achieve this goal extra analytical and computing efforts are often required and it also depends on the specific problem involved.

Acknowledgements

This research is partly supported by research grants from NSF (DMS1107012, DMS0915139, SES0851521). The authors thank Professor Naisyin Wang, the AE and two reviewers for their constructive suggestions. They also thank Professors Cun-hui Zhang and Hui Zou for helpful discussions on several technical aspects of penalized regressions.

6. Appendix

6.1. Proof of Theorem 1

Before we prove Theorem 1, we state following two lemmas without proofs. Lemma A1 is Proposition 4 of Fan and Lv (2011) and Lemma A2 is a restatement of Theorem 1 of Fan and Lv (2011) but on analysis of a subset of data.

Lemma A 1. *Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ be the n-dimensional independent random response vector and $\boldsymbol{a} \in R^n$. Then*

a) If $Y_1, \ldots, Y_n$ are bounded in $[c, d]$ for some $c, d \in R$, then for any $\epsilon \in (0, \infty)$

$$\mathrm{P}(|\boldsymbol{a}^T\boldsymbol{Y} - \boldsymbol{a}^T\boldsymbol{\mu}(\boldsymbol{\theta}^0)| > \epsilon) \leq 2\exp[-2\epsilon^2/(\|\boldsymbol{a}\|_2^2(d-c)^2)].$$

b) If $Y_1, \ldots, Y_n$ are unbounded and there exist some $M, v_0 \in (0, \infty)$ such that

$$\max_{i=1,\ldots,n} E\{\exp[(Y_i - b'(\theta_i^0))/M] - 1 - |Y_i - b'(\theta_i^0)|/M\}M^2 \leq v_0/2$$

with $\boldsymbol{\theta}^0 = (\theta_i^0, \ldots, \theta_n^0)$, then for any $\epsilon \in (0, \infty)$

$$\mathrm{P}(|\boldsymbol{a}^T\boldsymbol{Y} - \boldsymbol{a}^T\boldsymbol{\mu}(\boldsymbol{\theta}^0)| > \epsilon) \leq 2\exp[-\epsilon^2/(2\|\boldsymbol{a}\|_2^2 v_0 + \|\boldsymbol{a}\|_\infty M\epsilon)].$$

Lemma A 2. *A vector $(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}, 0)$ is a strict local maximizer, if*

$$\boldsymbol{X}_{k,\mathcal{A}}^T\boldsymbol{y}_k - \boldsymbol{X}_{k,\mathcal{A}}^T\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}_k) - n_k\bar{\rho}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}; \lambda_k) = 0, \tag{8}$$

$$n_k^{-1}\|\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T[\boldsymbol{y}_k - \boldsymbol{\mu}(\hat{\boldsymbol{\theta}}_k)]\|_\infty < \rho'(0+; \lambda_k), \tag{9}$$

$$\lambda_{\min}[\boldsymbol{X}_{k,\mathcal{A}}^T\Sigma(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}] > n_k\kappa(\rho; \hat{\boldsymbol{\beta}}_{k,\mathcal{A}}), \tag{10}$$

where $\bar{\rho}(\hat{\beta}_{k,\mathcal{A}}; \lambda_k) = (\rho'(\hat{\beta}_{k,j}; \lambda_k), (k, j) \in \mathcal{A})$.

Proof of Theorem 1. For $k = 1, \ldots, K$, let us define events $E_{1k} = \{\|\boldsymbol{X}_{k,\mathcal{A}}^T\{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)\}\|_\infty \leq c_1^{-1/2}v_{n,K}/K\}$ and $E_{2k} = \{\|\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T\{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)\}\|_\infty \leq c_1^{-1/2}u_{n,K}/K\}$, where $c_1 = 2/(d-c)^2$ for the case of bounded responses and $c_1 = 1/(2v_0 + 2M)$ for the case unbounded responses. From Lemma A1, we have

$$\begin{aligned}
\mathrm{P}\{\cap_{k=1}^K(E_{1k} \cap E_{2k})\} &\geq 1 - \sum_{k=1}^K \mathrm{P}(E_{1k}^c) - \sum_{k=1}^K \mathrm{P}(E_{2k}^c) \\
&\geq 1 - \sum_{k=1}^K\sum_{j=1}^s \mathrm{P}(|\boldsymbol{x}_{k,j}^T\boldsymbol{y}_k - \boldsymbol{x}_{k,j}^T\boldsymbol{\mu}(\boldsymbol{\theta}_k^0)| > c_1^{-1/2}v_{n,K}/K) \\
&\quad - \sum_{k=1}^K\sum_{j=s+1}^p \mathrm{P}(|\boldsymbol{x}_{k,j}^T\boldsymbol{y}_k - \boldsymbol{x}_{k,j}^T\boldsymbol{\mu}(\boldsymbol{\theta}_k^0)| > c_1^{-1/2}u_{n,K}/K) \\
&\geq 1 - 2Ks\exp\{-v_{n,K}^2/(nK)\} - 2K(p-s)\exp\{-u_{n,K}^2/(nK)\}.
\end{aligned}$$

Thus, the event $E = \cap_{k=1}^{K}(E_{1k} \cap E_{2k})$ holds in probability 1, provided that both $Ks \exp\{-v_{n,K}^2/(nK)\} \to 0$ and $K(p-s)\exp\{-u_{n,K}^2/(nK)\} \to 0$, as $n \to \infty$.

Now, for any $\hat{\boldsymbol{\beta}}_k = (\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}^T, 0)^T$ with $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} \in \mathcal{N}_0 = \{\boldsymbol{\delta} \in \Re^s : \|\boldsymbol{\delta} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_\infty \le \beta_*/2\}$, equation (8) can be re-written as

$$\boldsymbol{X}_{k,\mathcal{A}}^T\{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)\} - \boldsymbol{X}_{k,\mathcal{A}}^T \Sigma(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}(\hat{\beta}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0) - n_k\bar{\rho}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}};\lambda_k) - \boldsymbol{r}_{k,\mathcal{A}} = 0,$$

where $\boldsymbol{r}_{k,\mathcal{A}} = (r_{k1},\ldots,r_{ks})^T$ with $r_{kj} = (\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0)^T \nabla^2 \gamma_{k,j}(\tilde{\boldsymbol{\delta}}_j)(\hat{\beta}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0)$, $\gamma_{k,j}(\boldsymbol{\delta}) = \boldsymbol{x}_{k,j}^T\boldsymbol{\mu}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})$ and $\tilde{\boldsymbol{\delta}}_j$ being an s-dimensional vector on the segment between $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}$ and $\boldsymbol{\beta}_{\mathcal{A}}^0$, for $j = 1,\ldots,s$. It follows that

$$\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0 = \{\boldsymbol{X}_{k,\mathcal{A}}^T\Sigma(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}\big[\boldsymbol{X}_{k,\mathcal{A}}^T\{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)\} - n_k\bar{\rho}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}};\lambda_k) - \boldsymbol{r}_{k,\mathcal{A}}\big]. \quad (11)$$

Thus,

$$\begin{aligned}
\|\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_\infty &\le \|\{\boldsymbol{X}_{k,\mathcal{A}}^T\Sigma(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}\|_\infty\big\{\|\boldsymbol{\xi}_{k,\mathcal{A}}\|_\infty + \|\boldsymbol{\eta}_{k,\mathcal{A}}\|_\infty + \|\boldsymbol{r}_{k,\mathcal{A}}\|_\infty\big\} \\
&\le O(b_{s,K}n_k^{-1})\{c_1^{-1/2}v_{n,K}/K + n_k\rho'(\beta_*/2;\lambda_k) + O(n_k) \times \beta_*^2 s\} \\
&= O(c_1^{-1/2}b_{s,K}v_{n,K}/(n_kK) + b_{s,K}\rho'(\beta_*/2;\lambda_k) + b_{s,K}\beta_*^2 s) = o(\beta_*),
\end{aligned}$$

where $\boldsymbol{\xi}_{k,\mathcal{A}} = \boldsymbol{X}_{k,\mathcal{A}}^T\{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)\}$ and $\boldsymbol{\eta}_{k,\mathcal{A}} = n_k\bar{\rho}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}};\lambda_k)$. The second inequality holds by the definition of event E_{1k}, Condition A1 and that the the penalty function ρ is concave. The order of last line holds by Conditions A2 and A3.

Therefore, by Miranda's existence theorem (e.g. Vrahatis (1989)), there exists a solution $\hat{\boldsymbol{\beta}}_k = (\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}^T, 0)^T$ with $\hat{\beta}_{k,\mathcal{A}} \in \mathcal{N}_0$ for equation (8).

In addition, by Taylor expansion,

$$\begin{aligned}
\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T\{\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}_{k,\mathcal{A}}) - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)\} &= \boldsymbol{X}_{k,\overline{\mathcal{A}}}^T\Sigma(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0) + \boldsymbol{w}_{k,\overline{\mathcal{A}}} \qquad (12) \\
&= \boldsymbol{X}_{k,\overline{\mathcal{A}}}^T\Sigma(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}\{\boldsymbol{X}_{k,\mathcal{A}}^T\Sigma(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}(\boldsymbol{\xi}_{k,\mathcal{A}} - \boldsymbol{\eta}_{k,\mathcal{A}} - \boldsymbol{r}_{k,\mathcal{A}}) + \boldsymbol{w}_{k,\overline{\mathcal{A}}},
\end{aligned}$$

where $\boldsymbol{w}_{k,\overline{\mathcal{A}}} = (w_{k,s+1},\ldots,w_{kp})$ and $w_{kj} = (\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0)^T\nabla^2\gamma_{kj}(\boldsymbol{\delta}_j)(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0)$ with $\gamma_{kj}(\boldsymbol{\delta}) = \boldsymbol{x}_{k,j}^T\boldsymbol{\mu}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})$ for $j \in \overline{\mathcal{A}}$ and some $s \times 1$ vector $\tilde{\boldsymbol{\delta}}_j$ on the segment $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}$ and $\boldsymbol{\beta}_{\mathcal{A}}^0$. Similar to $\boldsymbol{r}_{k,\mathcal{A}}$, $\|\boldsymbol{w}_{k,\overline{\mathcal{A}}}\|_\infty = O(n_k s\beta_*^2)$. Thus, under the event $E_{1k} \cap E_{2k}$, by the last condition in A3 and also conditions A1 and A2,

$$\begin{aligned}
&n_k^{-1}\|\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T[\boldsymbol{y}_k - \boldsymbol{\mu}(\hat{\boldsymbol{\theta}}_k)]\|_\infty \le n_k^{-1}\|\boldsymbol{\xi}_{k,\overline{\mathcal{A}}}\|_\infty + \|\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T\{\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}_{k,\mathcal{A}}) - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)\}\|_\infty \\
&\le n_k^{-1}\|\boldsymbol{\xi}_{k,\overline{\mathcal{A}}}\|_\infty + n_k^{-1}\|\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T\Sigma(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}\{\boldsymbol{X}_{k,\mathcal{A}}^T\Sigma(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}\|_\infty\big\{\|\boldsymbol{\xi}_{k,\mathcal{A}}\|_\infty + \\
&\qquad \|\boldsymbol{\eta}_{k,\mathcal{A}}\|_\infty + \|\boldsymbol{r}_{k,\mathcal{A}}\|_\infty\big\} + n_k^{-1}\|\boldsymbol{w}_{k,\overline{\mathcal{A}}}\|_\infty \\
&= c_1^{-1/2}u_{n,K}/(n_kK) + O(n_k^{\alpha-1}v_{n,K}/K) + O(n_k^{\alpha}s\beta_*^2) + C\rho'(0+;\lambda_k) + O(s\beta_*^2) \\
&= o(1) + C\rho'(0+;\lambda_k) < \rho'(0+;\lambda_k).
\end{aligned}$$

Here, $\boldsymbol{\xi}_{k,\overline{\mathcal{A}}} = \boldsymbol{X}_{k,\overline{\mathcal{A}}}^T[\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)]$. So, both (8) and (9) hold for $\hat{\beta}_k = (\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}^T, 0)^T$. In addition, since $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} \in \mathcal{N}_0$ and by A3, equation (10) is satisfied. Thus, by Lemma A2, $\hat{\beta}_k = (\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}^T, 0)^T$ is the our solution in the kth subset, for $k = 1,\ldots,K$.

Finally, note that the above $\hat{\beta}_k$ is evaluated under the event $E = \cap_{k=1}^K (E_{1k} \cap E_{2k})\}$ which is an intersection over all $k = 1, 2, \ldots, K$. Also, the event E holds in probability 1, as $n \to \infty$. So, the above convergence statements regarding $\hat{\beta}_k$ are uniform over all K subsets. Specifically, when n is large enough, we have $\hat{\mathcal{A}}_k = \mathcal{A}$ for all subsets and also $\hat{\mathcal{A}}^{(c)} = \mathcal{A}$. In this case, $\hat{\boldsymbol{\beta}}^{(c)}_{\bar{\mathcal{A}}} = 0$ and also $\boldsymbol{X}_k \boldsymbol{A}$ $= \boldsymbol{X}_{k,\mathcal{A}}$ where $\boldsymbol{A} = \boldsymbol{E}_{\hat{\mathcal{A}}^{(c)}}$ is the selection matrix defined in (5). It follows from (5), (11) and also $\boldsymbol{X}^T_{k,\mathcal{A}} \Sigma(\hat{\boldsymbol{\theta}}_k) \boldsymbol{X}_{k,\mathcal{A}} = \boldsymbol{X}^T_{k,\mathcal{A}} \Sigma(\boldsymbol{\theta}^0_k) \boldsymbol{X}_{k,\mathcal{A}} + o_p(1)$, uniformly, that

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}} &= \boldsymbol{\beta}^0_{\mathcal{A}} + \Big(\sum_{k=1}^K \boldsymbol{X}^T_{k,\mathcal{A}} \Sigma(\hat{\boldsymbol{\theta}}_k) \boldsymbol{X}_{k,\mathcal{A}} \Big)^{-1} \Big[\\
&\qquad \sum_{k=1}^K \{\boldsymbol{X}^T_{k,\mathcal{A}} \Sigma(\hat{\boldsymbol{\theta}}_k) \boldsymbol{X}_{k,\mathcal{A}}\} \{\boldsymbol{X}^T_{k,\mathcal{A}} \Sigma(\boldsymbol{\theta}^0_k) \boldsymbol{X}_{k,\mathcal{A}}\}^{-1} \{\boldsymbol{\xi}_{k,\mathcal{A}} - \boldsymbol{\eta}_{k,\mathcal{A}} + \boldsymbol{r}_{k,\mathcal{A}}\} \Big] \\
&= \boldsymbol{\beta}^0_{\mathcal{A}} + \big\{ \boldsymbol{X}^T_{\mathcal{A}} \Sigma(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}} + o_p(1) \big\}^{-1} \Big\{ \sum_{k=1}^K \big[I + \\
&\qquad o_p(1) \{\boldsymbol{X}_{k,\mathcal{A}} \Sigma(\boldsymbol{\theta}^0_k) \boldsymbol{X}_{k,\mathcal{A}}\}^{-1} \big] (\boldsymbol{\xi}_{k,\mathcal{A}} - \boldsymbol{\eta}_{k,\mathcal{A}} - \boldsymbol{r}_{k,\mathcal{A}}) \Big\}.
\end{aligned}$$

Based on conditions A1 and A2,

$$\begin{aligned}
\|\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_\infty &\le O(1) \|(\boldsymbol{X}^T_{\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}})^{-1}\|_\infty \| \sum_{k=1}^K (\boldsymbol{\xi}_{k,\mathcal{A}} - \boldsymbol{\eta}_{k,\mathcal{A}} - \boldsymbol{r}_{k,\mathcal{A}}) \|_\infty \\
&= O\big(b_{s,K} v_{n,K}/n\big) + O\big(b_{s,K} \sum_{k=1}^K n_k \rho'(\beta_*/2; \lambda_k)/n\big) + O\big(b_{s,K} \beta_*^2 s\big) = o(\beta_*).
\end{aligned}$$

Thus, $\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}} \in \mathcal{N}_0$.

6.2. Proof of Theorem 2

To achieve the convergence rate of $\sqrt{s/n}$ for the $\boldsymbol{\beta}$ estimators under the L_2 norm and also to show the property of asymptotic normality, we first show that $\hat{\boldsymbol{\beta}}_k = (\hat{\boldsymbol{\beta}}^T_{k,\mathcal{A}}, 0)^T$ is a consistent estimator of $\boldsymbol{\beta}$ for each k and obtain an asymptotic expansion of $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}$. To do so, we adopt the same method used by Fan and Lv (2011) in the proofs of their Theorems 3 and 4, in which an injective mapping from a $\sqrt{Ks/n}$-ball under L_2 norm is utilized instead of using Milanda's existing theorem. We then use the asymptotic expansions of $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}$ and also the fact that $\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}}$ is weighted sum of $\hat{\boldsymbol{\beta}}_k$ to obtain the desired results.

Proof of Theorem 2 (i) Let $u_{n,K}$ be a diverging sequence depending on the total sample size n and the number of subsets K such that $u_{n,K} = o(n)$ and $pK \exp\{-u^2_{n,k}/(nK)\} = o(1)$. Consider events $E_{2k} = \{\|\boldsymbol{X}^T_{k,\bar{\mathcal{A}}}\{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}^0_k)\}\|_\infty \le c_1^{-1/2} u_{n,K}/K\}$, for $k = 1, \ldots, K$. As in the proof of Theorem 1 and from Lemma

A1, we have that the probability $\mathrm{P}\{\cap_{k=1}^{K} E_{2k}\} \geq 1 - 2K(p-s)\exp\{-u_{n,K}^2/(nK)\}$. So, the event $E_a = \cap_{k=1}^{K} E_{2k}$ holds in probability 1.

First, let us constrain the parameter space on the subspace $\{\boldsymbol{\beta} : \boldsymbol{\beta}_{\overline{\mathcal{A}}} = 0\}$ and also define $\mathcal{N}_\tau = \{\boldsymbol{\delta} \in \Re^s : \|\boldsymbol{\delta} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_2 \leq \sqrt{Ks/n}\tau\}$, for any given constant $\tau > 0$. Since $\beta_* \gg \sqrt{Ks/n}$, we have that, when n is large enough, $\beta_*/2 > \sqrt{Ks/n}\tau$ and thus $\mathrm{sgn}(\boldsymbol{\delta}) = \mathrm{sgn}(\boldsymbol{\beta}_{\mathcal{A}}^0)$ for any $\boldsymbol{\delta} \in \mathcal{N}_\tau$.

For each k, we define an event $F_k = \{Q_k(\boldsymbol{\beta}_{\mathcal{A}}^0) > \max_{\boldsymbol{\delta} \in \partial \mathcal{N}_\tau} Q_k(\boldsymbol{\delta})\}$, where $Q_k(\boldsymbol{\delta}) = \ell(\boldsymbol{\delta}; \boldsymbol{y}_k, \boldsymbol{X}_{k,\mathcal{A}}) - \rho(\boldsymbol{\delta}; \lambda_k)$ is the penalized likelihood. By Taylor expansion, we have

$$Q_k(\boldsymbol{\delta}) - Q_k(\boldsymbol{\beta}_{\mathcal{A}}^0) = (\boldsymbol{\delta} - \boldsymbol{\beta}_{\mathcal{A}}^0)^T \boldsymbol{v}_k - (\boldsymbol{\delta} - \boldsymbol{\beta}_{\mathcal{A}}^0)^T V_k (\boldsymbol{\delta} - \boldsymbol{\beta}_{\mathcal{A}}^0),$$

where $\boldsymbol{v}_k = n_k^{-1} \boldsymbol{X}_{k,\mathcal{A}}^T [\boldsymbol{y}_k - \mu(\boldsymbol{\theta}_k^0)] - \bar{\rho}(\boldsymbol{\beta}_{\mathcal{A}}^0; \lambda_k)$ and $V_k = n_k^{-1} \boldsymbol{X}_{k,\mathcal{A}}^T \Sigma(\boldsymbol{\theta}_k^*) \boldsymbol{X}_{k,\mathcal{A}} + \mathrm{diag}(\kappa(\rho(\cdot; \lambda_k); \boldsymbol{\delta}_k^*))$ with $\boldsymbol{\theta}_k^* = \boldsymbol{X}_{k,\mathcal{A}} \boldsymbol{\delta}_k^*$ and $\boldsymbol{\delta}_k^*$ being an $s \times 1$ vector on the segment joining $\boldsymbol{\delta}$ and $\boldsymbol{\beta}_{\mathcal{A}}^0$.

By Conditions A4 and A5, we have $E\|\boldsymbol{v}_k\|_2^2 \leq n_k^{-2} \phi tr(\boldsymbol{X}_{k,\mathcal{A}}^T \Sigma(\boldsymbol{\theta}_k^0) \boldsymbol{X}_{k,\mathcal{A}}) + \|\bar{\rho}(\boldsymbol{\beta}_{\mathcal{A}}^0; \lambda_k)\|_2^2 \leq n_k^{-2} \phi tr(\boldsymbol{X}_{k,\mathcal{A}}^T \Sigma(\boldsymbol{\theta}_k^0) \boldsymbol{X}_{k,\mathcal{A}}) + s\rho'(\beta_*/2; \lambda_k)^2 = O(s n_k^{-1}) = O(Ks/n)$ and $\lambda_{\min}(V_k) \geq \tau_{1,k}\{1 - o(1)\} \geq c/2$. Therefore,

$$\max_{\boldsymbol{\delta} \in \partial \mathcal{N}_\tau} Q_k(\boldsymbol{\delta}) - Q_k(\boldsymbol{\beta}_{\mathcal{A}}^0) \leq \sqrt{Ks/n}\tau(\|\boldsymbol{v}_k\|_2 - c\sqrt{Ks/n}\tau/4),$$

and thus

$$\mathrm{P}(F_k) \geq \mathrm{P}(\|\boldsymbol{v}_k\|_2^2 < c^2 K s \tau^2/(16n)) \geq 1 - 16nE\|\boldsymbol{v}_k\|_2^2/(c^2 K s \tau^2) \geq 1 - O(\tau^{-2}).$$

Since the above result holds for any (arbitrarily large) constant $\tau > 0$ and $Q_k(\boldsymbol{\delta})$ is a continuous injective function, there exists a maximizer $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} \in \mathcal{N}_\tau$ (that maximizes $Q_k(\boldsymbol{\delta})$ for $\boldsymbol{\delta} \in \mathcal{N}_\tau$) and $\|\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_2 = O_p(\sqrt{Ks/n})$, in probability.

Now, let us constrain the parameter $\boldsymbol{\beta}$ on the subspace $\{\boldsymbol{\beta} : \boldsymbol{\beta}_{\overline{\mathcal{A}}} = 0\}$. Following the first expression in (12) by Taylor expansion, we can show that, under Condition A4 and A5 and also $n_k = O(n/K)$,

$$\begin{aligned} \|\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T\{\mu(\hat{\boldsymbol{\theta}}_k) - \mu(\boldsymbol{\theta}_k^0)\}\|_\infty &\leq \|\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T \Sigma(\boldsymbol{\theta}_k^0) \boldsymbol{X}_{k,\mathcal{A}} (\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0)\|_\infty + \|\boldsymbol{w}_{k,\overline{\mathcal{A}}}\|_\infty \\ &\leq O(n_k)\|\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_2 + O(n_k)\|\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_2 = O_p(\sqrt{ns/K}). \end{aligned}$$

Thus, under event $E_{2k} = \{\|\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T\{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)\}\|_\infty \leq c_1^{-1/2} u_{n,K}/K\}$, we have

$$\begin{aligned} \|n_k^{-1} \boldsymbol{X}_{k,\overline{\mathcal{A}}}^T\{\boldsymbol{y}_k - \{\mu(\hat{\boldsymbol{\theta}}_k)\}\|_\infty &= \|n_k^{-1}\left[\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T\{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)\} - \boldsymbol{X}_{k,\overline{\mathcal{A}}}^T\{\mu(\hat{\boldsymbol{\theta}}_k) - \mu(\boldsymbol{\theta}_k^0)\}\right]\|_\infty \\ &\leq n_k^{-1}[\|\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T\{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)\}\|_\infty + \|\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T\{\mu(\hat{\boldsymbol{\theta}}_k) - \mu(\boldsymbol{\theta}_k^0)\}\|_\infty] \\ &\leq c_1^{-1/2} u_{n,K}/(n_k K) + O_p(\sqrt{sK/n}) = o(1). \end{aligned}$$

Thus, when n is large enough, equation (9) holds. Since condition A5 also implies equation (10), we conclude based on Lemma A2 that $\hat{\beta}_k = (\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}^T, 0)^T$ with $\|\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_2 = O_p(\sqrt{Ks/n})$ is a local maximizer in the analysis of the kth subset data, for $k = 1, \ldots, K$.

Follow the argument in the proof of Theorem 1, the above statements about $\hat{\beta}_k$ hold uniformly (Specifically, all statements about $\hat{\beta}_k$ is evaluated under the event $\cap_{k=1}^{K} E_{2k}$ which is an intersection over all $k = 1, 2, \ldots, K$ and the event $\cap_{k=1}^{K} E_{2k}$ holds in probability 1, as $n \to \infty$. So, the above convergence statements regarding $\hat{\beta}_k$ are uniform over all K subsets.) In particular, when n is large enough, we have $\hat{\mathcal{A}}_k = \mathcal{A}$ for all subsets and also $\hat{\mathcal{A}}^{(c)} = \mathcal{A}$. In this case, $\boldsymbol{X}_k \boldsymbol{A} = \boldsymbol{X}_{k,\mathcal{A}}$ where $\boldsymbol{A} = \boldsymbol{E}_{\hat{\mathcal{A}}^{(c)}}$ is the selection matrix defined in (5). As in the proof of Theorem 1, since $\hat{\boldsymbol{\beta}}_{k,\overline{\mathcal{A}}} = 0$ for all k, we immediately have $\hat{\boldsymbol{\beta}}^{(c)}_{\overline{\mathcal{A}}} = 0$.

For any $\hat{\boldsymbol{\beta}}_k = (\hat{\boldsymbol{\beta}}^T_{k,\mathcal{A}}, 0)^T$ with $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} \in \mathcal{N}_\tau = \{\boldsymbol{\delta} \in \Re^s : \|\boldsymbol{\delta} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_2 \le \sqrt{Ks/n\tau}\}$, we can obtain by Taylor expansion the same expression of $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}$ as stated in equation (11). Since $\|n_k \bar{\rho}(\boldsymbol{\beta}^0_{\mathcal{A}}; \lambda_k)\|_2 \le n_k \sqrt{s} \rho'(\beta_*/2; \lambda_k) = O(\sqrt{ns}/K)$ and $\|\boldsymbol{r}_{k\mathcal{A}}\|_2 = \sqrt{s} O(n_k) O_p(Ks/n) = O_p(s^{3/2})$, it follows that

$$\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} = \boldsymbol{\beta}^0_{\mathcal{A}} + \{\boldsymbol{X}^T_{k,\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}^0_k) \boldsymbol{X}_{k,\mathcal{A}}\}^{-1} \left[\boldsymbol{X}^T_{k,\mathcal{A}}\{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}^0_k)\} + O_p(\max\{s^{3/2}, \sqrt{ns}/K\})\right]$$

Therefore, by the definition of $\hat{\boldsymbol{\beta}}^{(c)}$ in (5) and noting that $Ks = O(n^{1/2})$ and $\boldsymbol{X}^T_{k,\mathcal{A}} \Sigma(\hat{\boldsymbol{\theta}}_k) \boldsymbol{X}_{k,\mathcal{A}} = \boldsymbol{X}^T_{k,\mathcal{A}} \Sigma(\boldsymbol{\theta}^0_k) \boldsymbol{X}_{k,\mathcal{A}} + o_p(1)$, uniformly, we have

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}} = \boldsymbol{\beta}^0_{\mathcal{A}} + [\sum_{k=1}^{K} \boldsymbol{X}^T_{k,\mathcal{A}} \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k) \boldsymbol{X}_{k,\mathcal{A}}]^{-1} \Bigg\{ \\ \sum_{k=1}^{K} \{1 + o_p(1)\} \left[\boldsymbol{X}^T_{k,\mathcal{A}}\{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}^0_k)\} + O_p(\sqrt{ns}/K)\right] \Bigg\} \end{aligned} \tag{13}$$

Since $\lambda_{\min}\big(\sum_{k=1}^{K} \boldsymbol{X}^T_{k,\mathcal{A}} \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k) \boldsymbol{X}_{k,\mathcal{A}}\big) \ge \sum_{k=1}^{K} \lambda_{\min}\big(\boldsymbol{X}^T_{k,\mathcal{A}} \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k) \boldsymbol{X}_{k,\mathcal{A}}\big)$, by condition A4 we have that $\lambda_{\max}\big([\sum_{k=1}^{K} \boldsymbol{X}^T_{k,\mathcal{A}} \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k) \boldsymbol{X}_{k,\mathcal{A}}]^{-1}\big) = O_p(n^{-1})$. In addition, $E\|\boldsymbol{X}^T_{\mathcal{A}}[\boldsymbol{y} - \mu(\boldsymbol{\theta}^0_k)]\|^2_2 \le \phi tr(\boldsymbol{X}^T_{\mathcal{A}} \Sigma(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}}) = \phi \sum_{k=1}^{K} tr(\boldsymbol{X}^T_{k,\mathcal{A}} \Sigma(\boldsymbol{\theta}^0_k) \boldsymbol{X}_{k,\mathcal{A}}) = O(sn)$ by Condition A4. It follows $\|\boldsymbol{X}^T_{\mathcal{A}}[\boldsymbol{y} - \mu(\boldsymbol{\theta}^0)]\|^2_2 = O_p(ns)$. Thus, by (13),

$$\|\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_2 \le O_p(n^{-1})\left[\|\boldsymbol{X}^T_{\mathcal{A}}\{\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}^0)\}\|_2 + O_p(\sqrt{ns})\right] = O_p(\sqrt{s/n}).$$

(ii) Under the further assumption that $K^{2/3}s = o(n^{1/3})$ and $\rho'(\beta_*/2; \lambda_k) = o(s^{-1/2}n^{-1/2})$, the remaining term $O_p(\sqrt{ns}/K)$ in (13) is in fact $o_p(\sqrt{n}/K)$. By (13) with this modification, we have

$$\begin{aligned} \boldsymbol{D}[\boldsymbol{X}^T_{\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}}]^{1/2} (\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) = \boldsymbol{D}[\boldsymbol{X}^T_{\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}}]^{1/2} \big\{\boldsymbol{X}^T_{\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}} \\ + o_p(1)\big\}^{-1} \boldsymbol{X}_{\mathcal{A}} \big\{\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}^0)\big\} \big\{1 + o_p(1)\big\} + o_p(1). \end{aligned}$$

From Condition A6,

$$\boldsymbol{D}[\boldsymbol{X}_{\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}}]^{-1/2} \boldsymbol{X}_{\mathcal{A}}[\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}^0)] \xrightarrow{\mathrm{D}} N(\mathbf{0}, \phi \boldsymbol{G}).$$

Thus, the asymptotic normality result in (ii) holds.

6.3. Proof of Theorem 3

Proof of Theorem 3 We fist show that $\mathrm{P}(j \in \hat{\mathcal{A}}_k) \leq \bar{s}_k/p$, $j \in \overline{\mathcal{A}}$, and $\mathrm{P}(j \in \hat{\mathcal{A}}_k) \geq \bar{s}_k/p$, for $j \in \mathcal{A}$ and $k = 1, \ldots, K$.

Because $E(|\overline{\mathcal{A}} \cap \hat{\mathcal{A}}_k|) = E(|\hat{\mathcal{A}}_k|) - E(|\mathcal{A} \cap \hat{\mathcal{A}}_k|) = \bar{s}_k - E(|\mathcal{A} \cap \hat{\mathcal{A}}_k|)$ and $E(|\mathcal{A} \cap \hat{\mathcal{A}}_k|)/E(|\overline{\mathcal{A}} \cap \hat{\mathcal{A}}_k|) \geq |\mathcal{A}|/|\overline{\mathcal{A}}|$, we have $E(|\overline{\mathcal{A}} \cap \hat{\mathcal{A}}_k|) \leq \bar{s}_k/(1 + |\mathcal{A}|/|\overline{\mathcal{A}}|)$ and $E(|\mathcal{A} \cap \hat{\mathcal{A}}_k|) \geq \bar{s}_k/(1 + |\overline{\mathcal{A}}|/|\mathcal{A}|)$. Therefore, $E(|\overline{\mathcal{A}} \cap \hat{\mathcal{A}}_k|) \leq \bar{s}_k|\overline{\mathcal{A}}|/p$ and $E(|\mathcal{A} \cap \hat{\mathcal{A}}_k|) \geq \bar{s}_k|\mathcal{A}|/p$.

Using the exchangeability assumption, $\mathrm{P}(j \in \hat{\mathcal{A}}_k) = E(|\overline{\mathcal{A}} \cap \hat{\mathcal{A}}_k|)/|\overline{\mathcal{A}}|$, $j \in \overline{\mathcal{A}}$ and $\mathrm{P}(j \in \hat{\mathcal{A}}_k) = E(|\mathcal{A} \cap \hat{\mathcal{A}}_k|)/|\mathcal{A}|$, $j \in \mathcal{A}$. Therefore, $\mathrm{P}(j \in \hat{\mathcal{A}}_k) \leq \bar{s}_k/p \leq s^*/p$, $j \in \overline{\mathcal{A}}$ and $\mathrm{P}(j \in \hat{\mathcal{A}}_k) \geq \bar{s}_k/p \geq s_*/p$, $j \in \mathcal{A}$.

Since the observations in each subset are independent and $w \geq s^*K/p - 1$, $\mathrm{P}(j \in \hat{\mathcal{A}}^{(c)}) \leq 1 - F(w|K, s^*/p), j \in \overline{\mathcal{A}}$ and $\mathrm{P}(j \in \hat{\mathcal{A}}^{(c)}) \geq 1 - F(w|K, s_*/p), j \in \mathcal{A}$. Therefore, $E(|\overline{\mathcal{A}} \cap \hat{\mathcal{A}}^{(c)}|) = \sum_{j \in \overline{\mathcal{A}}} \mathrm{P}(j \in \hat{\mathcal{A}}^{(c)}) \leq |\overline{\mathcal{A}}|\{1 - F(w|K, s^*/p)\}$ and $E(|\mathcal{A} \cap \hat{\mathcal{A}}^{(c)}|) = \sum_{j \in \mathcal{A}} \mathrm{P}(j \in \hat{\mathcal{A}}^{(c)}) \geq |\mathcal{A}|(1 - F(w|K, s_*/p))$.

6.4. Proof of Lemma 1 and Theorem 4

Proof of Lemma 1: We first state the LARS algorithm for LASSO here:

- Initialize, let the active set which contains the variables with nonzero coefficients $A = \emptyset$, current mean estimate $\hat{\mu}_A = 0$, current coefficient $\hat{\boldsymbol{\beta}}_A = 0$ and step size $\gamma = 0$. Let $\boldsymbol{a} = 0$.
- Repeat the following steps until $|A| = n$.

  [1] Calculate the correlation between variables and the current residual $\hat{\boldsymbol{c}} = \boldsymbol{X}_{A^c}^T \boldsymbol{y} - \gamma \boldsymbol{a}$ and $\hat{C} = \max\{|\hat{c}_j|\}$, where $\hat{c}_j$ is the elements of $\hat{\boldsymbol{c}}$ for $j \in A$.

  [2] Let $A = \{j : |\hat{c}_j| = \hat{C}\}$ if $A = \emptyset$, $s_j = \mathrm{sgn}(\hat{c}_j)$ and $\boldsymbol{X}_A = (\ldots, s_j \boldsymbol{x}_j, \ldots)$, $j \in A$. Calculate the next moving direction $G_A = \boldsymbol{X}_A^T \boldsymbol{X}_A$, $Q_A = (\mathbf{1}_A^T G_A^T \mathbf{1}_A)^{-1/2}$ and $\boldsymbol{w}_A = Q_A G_A^{-1} \mathbf{1}_A = (\ldots, w_j, \ldots)$, $\boldsymbol{u}_A = \boldsymbol{X}_A \boldsymbol{w}_A$. Here $\mathbf{1}_A$ is a vector of size $|A|$ with all 1s.

  [3] Calculate the size of tuning parameter. Let $\hat{d}_j = s_j w_j$, $j \in A$ and $\boldsymbol{a} = \boldsymbol{X}_{A^c}^T \boldsymbol{u}_A = (\ldots, a_j, \ldots)$. Calculate $\gamma_j = -\hat{\beta}_j/\hat{d}_j$, $\tilde{\gamma} = \min_{\gamma_j > 0}(\gamma_j)$ and $\hat{\gamma} = \min_{j \in A^c}{}^{+}\{(\hat{C} - \hat{c}_j)/(Q_A - a_j), (\hat{C} - \hat{c}_j)/(Q_A + a_j)\}$, where $\min^{+}$ means that the minimum is taken over only positive components.

  [4] If $\tilde{\gamma} \leq \hat{\gamma}$, update $\hat{\mu}_A \leftarrow \hat{\mu}_A + \tilde{\gamma}\boldsymbol{u}_A$, $\hat{\beta}_j \leftarrow \hat{\beta}_j + \tilde{\gamma} s_j w_j$ $A \leftarrow A - \tilde{j}$ where $\tilde{j}$ is the index for which the minimizing index in obtaining $\tilde{\gamma}$, and $\gamma = \tilde{\gamma}$. If $\tilde{\gamma} > \hat{\gamma}$, update $\hat{\mu}_A \leftarrow \hat{\mu}_A + \hat{\gamma}\boldsymbol{u}_A$, $\hat{\beta}_j \leftarrow \hat{\beta}_j + \hat{\gamma} s_j w_j$, $A \leftarrow A + \tilde{j}$ where $\tilde{j}$ is the index for which the minimizing index in obtaining $\hat{\gamma}$ and $\gamma = \hat{\gamma}$.

Denote *comp*$[i]$ the computing steps at step i in each loop, $i = 1, 2, 3, 4$. Suppose linear search is used to find the maximum or minimum and schoolbook matrix multiplication algorithm is applied. We have *comp*$[1] = 2n(p - |A|)$.

In step [2], computing Q_A requires $|A|^2$ computing steps. When compute G_A^{-1}, Cholesky factorization is applied to update the inverse matrix. Details are given below. Get the block representation of G_A and the Cholesky factor of G_A, denoted by U, i.e., $G_A = U^T U$. Also denote the inverse matrix of U by $Y = U^{-1}$ and write $G_A = \begin{pmatrix} G_{11} & G_{12} \\ G_{12}^T & G_{22} \end{pmatrix}, U = \begin{pmatrix} U_{11} & U_{12} \\ U_{12}^T & U_{22} \end{pmatrix}$ and $Y = \begin{pmatrix} Y_{11} & Y_{12} \\ Y_{12}^T & Y_{22} \end{pmatrix}$, where G_{22} is a one-dimension matrix (a number) representing the newly added variable. Thus, $G_A^{-1} = \begin{pmatrix} Y_{11}Y_{11}^T + Y_{12}Y_{12}^T & Y_{12}Y_{22}^T \\ Y_{22}Y_{12}^T & Y_{22}Y_{22}^T \end{pmatrix}$, where $G_{11}^{-1} = Y_{11}Y_{11}^T$.

Since U_{11} and Y_{11} is known from the previous loop, we can update G_A^{-1} by the following equations: $U_{12} = Y_{11}^T G_{12}, U_{22} = \sqrt{G_{22} - U_{12}^T U_{12}}, Y_{22} = U_{22}^{-1}, Y_{12} = -Y_{11}U_{12}Y_{22}$, and compute $G_{11}^{-1} + Y_{12}Y_{12}^T$, $Y_{12}Y_{22}^T$ and $Y_{22}Y_{22}^T$. Thus, $comp[2] = 8|A|^2 - 10|A| + 7 + (2|A| - 1)n$.

In Step [3] and [4], we have $comp[3] = |A| + (2n-1)(p-|A|) + 2|A| + 7(p-|A|)$, and $comp[4] = 2|A| + 1$, respectively.

In all, one loop in LARS algorithm requires $8|A|^2 - 11|A| + (4n+6)p - 2n|A| + 8 - n$. Therefore, since $p \geq n$, at most n variables will be fitted and the LARS algorithm requires at least $\sum_{|A|=1}^{n} 8|A|^2 - 11|A| + (4n+6)p - 2n|A| + 8 - n = 5n^3/3 + 23n/6 + 4n^2(p - 7/8) + 6np$.

Each time dropping variable occurs, it will add additional $8|A|^2 - 11|A| + (4n+6)p - 2n|A| + 8 - n$ computing steps depending on the number of current active variables. The worst case would be $6n^2 + 4n(p-3) + 6p + 8$ computing steps each time and the solution path has n times downsize. The computing steps for the worst case would be $23n^3/3 + 71n/6 + 8n^2(p - 31/16) + 12np$.

Proof of Theorem 4: According to Lemma 1, as each sub-sample has n_k observations, for the best case, the computing steps for the combined estimator is $\sum_{k=1}^{K} 5n_k^3/3 + 23n_k/6 + 4n_k^2(p - 7/8) + 6n_k p$. Since $\sum_{k=1}^{K} n_k = n$, $5n^3/3 + 23n/6 + 4n^2(p - 7/8) + 6np \geq \sum_{k=1}^{K} 5n_k^3/3 + 23n_k/6 + 4n_k^2(p - 7/8) + 6n_k p$. The result follows immediately. Similarly, the combined estimator requires less computing steps for the worst case.

We only need to show that under the assumptions, the worst case for split-and-conquer approach requires less computing steps than the best case for LARS algorithm using the entire dataset. When $n_k = O(n_k)$, split-and-conquer approach requires at most $23n^3/(3K^2) + 71n/6 + 8n^2(p - 31/16)/K + 12np$ computing steps and LARS algorithm using the entire dataset needs at least $5n^3/3 + 23n/6 + 4n^2(p - 7/8) + 6np$ computing steps. It is equivalent to show that $\{5n^3/3 + 23n/6 + 4n^2(p-7/8) + 6np\} - \{23n^3/(3K^2) + 71n/6 + 8n^2(p-31/16)/K + 12np\} = (5 - 23/K^2)n^3/3 + \{4p(1-2/K) + (31/K - 7)/2\}n^2 - (8+6p)n \geq 0$.

When $K \geq 3$ and $p \geq 2$, we have $5 - 23/K^2 > 0$ and $4p(1-2/K) + (31/K - 7)/2 > 0$. Thus, when $n \geq 4(4+3p)/\{1 + 8p(1-2/K) + 31/K - 7\}$, we have $(5 - 23/K^2)n^2/3 + \{4p(1-2/K) + (31/K-7)/2\}n - (8+6p) > 0$. The result follows immediately.

6.5. *Verification of Condition* $\mathbf{A2}$ *in the special case described in the last paragraph of Section3.1*

Proof: The proof is straightforward. We compute the order of each term in Condition A2. First, $\beta_* s b_{s,K} = O(n^{-\gamma}\log(n) s b_{s,K}) = o(1)$ and $\beta_*(n_k^{\alpha}s)^{1/2} = o(\log(n)/n^{\gamma-\alpha/2-\alpha_0/2}) = o(1)$.

Then, $v_{n,K}b_{s,K}/(n_k K\beta_*) = o(\sqrt{Kn\log(n)}b_{s,K}/(n^{1-\gamma}\log(n))) = o(1)$ and $v_{n,K}/(n_k^{1-\alpha}K) = o(\sqrt{\log(n)}K^{1/2-\alpha}/n^{1/2-\alpha}) = o(1)$. Finally, $u_{n,K}/(n_k K) = O(\sqrt{K}\log(n)/n^{\alpha_1}) = o(1)$.

References

Agarwal, A. and J. C. Duchi (2012). Distributed delayed stochastic optimization. In *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pp. 5451–5452. IEEE.

Ahmed, A., M. Aly, A. Das, A. J. Smola, and T. Anastasakos (2012). Web-scale multi-task feature selection for behavioral targeting. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 1737–1741. ACM.

Andrews, G. (2000). *Foundations of multithreaded, parallel, and distributed programming*, Volume 1. Addison-Wesley.

Breheny, P. and J. Huang (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics 5*(1), 232–253.

Chen, S., D. Donoho, and M. Saunders (2001). Atomic decomposition by basis pursuit. *SIAM review*, 129–159.

Duchi, J. C., A. Agarwal, and M. J. Wainwright (2012). Dual averaging for distributed optimization: convergence analysis and network scaling. *Automatic Control, IEEE Transactions on 57*(3), 592–606.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of statistics 32*(2), 407–451.

Fan, J., S. Guo, and N. Hao (2010). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Arxiv preprint arXiv:1004.5178*.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fan, J. and J. Lv (2011). Non-concave penalized likelihood with np-dimensionality. *IEEE transaction on information theory 57*(8), 5467–5484.

Fan, J., R. Samworth, and Y. Wu (2009). Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research 10*, 2013–2038.

Golub, G. H. and C. F. Van Loan (1983). Matrix computations. *Johns Hopkins University, Press, Baltimore, MD, USA.*

Liu, D. (2012). Combination of confidence distributions and an efficient approach for meta-analysis of heterogeneous studies. *Ph.D thesis*, Department of Statistics and Biostatistics, Rutgers University.

Mackey, L., A. Talwalkar, and M. Jordan (2011). Divide-and-conquer matrix factorization. *arXiv preprint arXiv:1107.0789*.

Marcenko, V. A. and L. A. Pastur (1967). Distribution of eigenvalues for some sets of random matrices. *Sbornik: Mathematics 1*(4), 457–483.

Meinshausen, N. and P. Buhlmann (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(4), 417–473.

Shah, R. and R. Samworth (2012). Variable selection with error control: Another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) to appear.*

Singh, K., M. Xie, and W. Strawderman (2005). Combining information from independent sources through confidence distributions. *Annals of statistics*, 159–183.

Takemura, A. and Y. Sheena (2005). Distribution of eigenvalues and eigenvectors of wishart matrix when the population eigenvalues are infinitely dispersed and its application to minimax estimation of covariance matrix. *Journal of Multivariate Analysis 94*(2), 271–299.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Trefethen, L. N. and D. Bau III (1997). *Numerical linear algebra.* Number 50. Siam.

Vrahatis, M. (1989). A short proof and a generalization of mirandas existence theorem. In *Proc. Amer. Math. Soc*, Volume 107, pp. 701–703.

Wainwright, M. (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using 1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory 55*(5), 2183–2202.

Xie, M., K. Singh, and W. Strawderman (2011). Confidence distributions and a unifying framework for meta-analysis. *Journal of the American Statistical Association 106*(493), 320–333.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(1), 49–67.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics 38*(2), 894–942.

Zhang, C.-H. and J. Huang (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics 36*(4), 1567–1594.

Zhang, C.-H. and S. S. Zhang (2011). Confidence intervals for low-dimensional parameters in high-dimensional linear models. *arXiv preprint arXiv:1110.2563*.

Zhang, Y., J. C. Duchi, and M. J. Wainwright (2012). Communication-efficient algorithms for statistical optimization. In *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pp. 6792–6792. IEEE.

Zhang, Y., J. C. Duchi, and M. J. Wainwright (2013). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *arXiv preprint arXiv:1305.5029*.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 301–320.