Statistica Sinica **24** (2014), 1655-1684
doi:http://dx.doi.org/10.5705/ss.2013.088

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

Xueying Chen and Min-ge Xie

*Rutgers University*

*Abstract:* If there are datasets, too large to fit into a single computer or too expensive for a computationally intensive data analysis, what should we do? We propose a *split-and-conquer* approach and illustrate it using several computationally intensive penalized regression methods, along with a theoretical support. We show that the split-and-conquer approach can substantially reduce computing time and computer memory requirements. The proposed methodology is illustrated numerically using both simulation and data examples.

*Key words and phrases:* Big data, combining results from independent analyses, distributed computing, generalized linear models, large sample theory, penalized regression.

## 1. Introduction

Consider the 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. The sample size $n$ and the number of parameters $p$ are potentially very large. We assume that, given $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, the conditional distribution of $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is the canonical exponential distribution

$$f(\boldsymbol{y}; \boldsymbol{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} f_0(y_i; \theta_i) = \prod_{i=1}^{n} \left\{ c(y_i) \exp\left[ \frac{y_i \theta_i - b(\theta_i)}{\phi} \right] \right\}, \tag{1.1}$$

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

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X}) = \frac{\boldsymbol{y}^T \boldsymbol{X}\boldsymbol{\beta} - \mathbf{1}^T \boldsymbol{b}(\boldsymbol{X}\boldsymbol{\beta})}{n}, \tag{1.2}$$

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 $b(\cdot)$ has a second derivative. If $p$ is large (or grows with $n$) and $\boldsymbol{\beta}$ is sparse, a penalized likelihood estimator is often used in the general form

$$\hat{\boldsymbol{\beta}}^{(a)} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ \frac{\ell(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X})}{n} - \rho(\boldsymbol{\beta}; \lambda_a) \right\}. \tag{1.3}$$

Here, $\rho$ is the penalty function with tuning parameter $\lambda_a$. To distinguish the estimator obtained from our approach, we use the superscript $^{(a)}$ to indicate the estimator is obtained by analyzing the *entire* data $(\boldsymbol{y}, \boldsymbol{X})$. Depending on the choice of $\rho(\boldsymbol{\beta}; \lambda_a)$, we have the LASSO estimator (Tibshirani (1996); Chen, Donoho, and Saunders (2001)), the LARS algorithm (Efron et al. (2004)), the SCAD estimator (Fan and Li (2001)) and the MCP estimators (Zhang (2010)), among others.

We consider datasets 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 dataset into $K$ subsets; each subset is then to be analyzed separately. A set of $K$ results are obtained, to be combined to obtain a final result. Our task is to investigate whether the combined overall result can be as good as the result obtained from analyzing the entire dataset and, if conditions are needed, what they are. We assume that the same method (including software) is used to analyze each subset data as well as the entire data, if we would have enough computing power to do so.

We focus our developments on a general penalized regression setting considered in the review article of Fan and Lv (2011), that covers almost all commonly used penalty functions in penalized regression practice, the LASSO, SCAD, MCP, and others. In their setting, Fan and Lv (2011) show that penalized estimators under (1.3) have such good asymptotic properties as model selection consistency and asymptotic normality. We investigate here whether the combined result from our method retains these desired properties and, if so, under what conditions. We assume that each subset contains enough data to provide a meaningful inference for the unknown model parameters. This requirement of large enough subset data might be relaxed under some special situations yet often requires extra effort.

For the penalized estimators (1.3) and model (1.1), 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. The combined estimator is model selection consistent as long as the penalized estimators are model selection consistent. When asymptotic normality is attainable, the combined estimator has the same asymptotic variance as the penalized estimator using the entire data. The price that we need to pay is to require to have a slightly stronger assumption on the design matrix, a little larger coefficient signals and/or a slower growth rate of $p$.

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). With our 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, that are consistent with those reported in the literature. Note that, Fan, Guo, and Hao (2010) propose refitted cross-validation to attenuate serious correlations among the random errors and explanatory variables; Meinshausen and Buhlmann (2010) introduce a stability selection and an exact error control bound through a combination of subsampling and model selection algorithms; Shah and Samworth (2013) propose a variant of stability selection with improved error control property. Similarly, the split-and-conquer approach provides a resistance to selection errors caused by spurious correlations and it keeps a large number of variables that are in the true model at the same time. This control on the selected variables is not typically available for conventional penalized estimators on analyzing the entire data.

The split-and-conquer approach can substantially reduce computing time and memory requirements. For instance, in 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) report that the LARS algorithm requires $O(n^a)$ with $a > 1$ computations when $p \geq n$. The computing time can be costly when both $n$ and $p$ are large. We show, mathematically and numerically, that our approach with LARS can save up to $(1 - 1/K^{(a-1)})\%$ computing time, where $K$ is the number of splits. This result holds under a general setting. We provide 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 be found in the computer sciences community under the name of parallel and distributed computing (see, e.g., Andrews (2000)). Most of this research focuses on such aspects as accessing a shared memory, exchanging information between processors, or identifying parallel components within an algorithm, e.g., Ahmed et al. (2012). More recently, there is research on the performance of combined results. Mackey, Talwalkar, and Jordan (2011) propose a divide-and-conquer method for matrix factorization, that partitions a large-scale matrix into submatrices. Zhang, Duchi, and Wainwright (2013) provide a divide-and-combine method for kernel ridge regression, that divides a dataset into several subsets. See also Zhang, Duchi, and Wainwright (2012), Agarwal and Duchi (2012); Ahmed et al. (2012); Duchi, Agarwal, and Wainwright (2012). Here we use a weighted combination method and study the statistical performance and computing issues of the proposed method, alongside the results from our method and the corresponding results using the entire dataset. We treat statistical issues such as convergence and efficient estimation, and we provided discussions on computing time when computationally intensive approaches are involved.

The rest of this article is organized as follows. Section 2 takes a split-and-conquer approach to 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 utility of the methodology using both simulation studies and a 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 $p$ is large and the true parameter, $\boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_p^0)^T$, is sparse. Suppose the dataset of size $n$ is divided into $K$ subsets, and that 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

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

Corresponding to (1.3), the penalized estimator for the $k$th subset is

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

where $\rho(\boldsymbol{\beta}; \lambda_k)$ is the penalty function with tuning parameter $\lambda_k$. For simplicity, and 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)** $\rho(t; \lambda)$ 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 $\lambda$.

As noted by Fan and Lv (2011), (PC) covers most commonly used penalty functions, including the $\mathrm{L}_1$ penalty, SCAD, and MCP, among others.

Under the setup, the penalized estimator $\hat{\boldsymbol{\beta}}_k$ has the sparsity property; 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 $\hat{\boldsymbol{\beta}}_k$. For any indices set $S$, denote by $\hat{\boldsymbol{\beta}}_{k,S}$ a $|S| \times 1$ vector formed by the elements of $\hat{\boldsymbol{\beta}}_k$ whose indices are in $S$, so $\hat{\boldsymbol{\beta}}_{k,\hat{\mathcal{A}}_k}$ 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, the $\hat{\mathcal{A}}_k$ can differ, and the $K$ vectors $\hat{\boldsymbol{\beta}}_{k,\hat{\mathcal{A}}_k}$, $k = 1, \ldots, K$, may have different lengths.

To obtain a combined estimator of $\boldsymbol{\beta}$ from $\hat{\boldsymbol{\beta}}_k$'s we use a majority voting method. There are two considerations in this. The combined estimator, say $\hat{\boldsymbol{\beta}}^{(c)}$, is based on $\hat{\boldsymbol{\beta}}_k$'s, and a variable not in any of $\hat{\mathcal{A}}_k = \{j : \hat{\boldsymbol{\beta}}_{k,j} \neq 0\}$ is not included in $\hat{\mathcal{A}}^{(c)} = \{j : \hat{\boldsymbol{\beta}}_j^{(c)} \neq 0\}$ for the combined estimator. As the $\hat{\mathcal{A}}_k$ are subject to errors, there may be mismatches between the set $\hat{\mathcal{A}}_k$ and the true nonzero set $\mathcal{A} \overset{\mathrm{d}}{=} \{j : \beta_j^0 \neq 0\}$. We take

$$\hat{\mathcal{A}}^{(c)} \overset{\mathrm{d}}{=} \left\{ j : \sum_{k=1}^{K} \mathbf{I}(\hat{\beta}_{k,j} \neq 0) > w \right\} \tag{2.1}$$

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 (2.1), it is clear that $\hat{\mathcal{A}}^{(c)} \subset \bigcup_{k=1}^{K} \hat{\mathcal{A}}_k$. Also, when the number of elements in $\hat{\mathcal{A}}_k$ (denoted by $|\hat{\mathcal{A}}_k|$) is small and the $K$ sets $\hat{\mathcal{A}}_k$'s have many common elements, the number $|\hat{\mathcal{A}}^{(c)}|$ of $\hat{\mathcal{A}}^{(c)}$ is much smaller than $p$. At the extremes, $\hat{\mathcal{A}}^{(c)}$ contains only the variables that are selected by all subset analyses, and $\hat{\mathcal{A}}^{(c)}$ contains the variables that are selected by one or more subset analyses. The theoretical development suggests that the choice of a fixed threshold, does not affect the asymptotic results, though in practice it can impact the numerical performance of the proposed approach.

It is possible to extend the simple majority voting method in (2.1) 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). For simplicity, we use the simple majority voting to determine the estimated number of selected variables and use a weighted scheme to combine the penalized estimators from the $K$ subset data.

For an $n \times 1$ vector of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$, let

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

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

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

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 (2.1). Take $\boldsymbol{E} = \operatorname{diag}(v_1, \ldots, v_p)$ to 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, as a weighted average of $\hat{\boldsymbol{\beta}}_{k,\hat{\mathcal{A}}^{(c)}}$, $k = 1, \ldots, K$, is

$$\hat{\boldsymbol{\beta}}^{(c)} \overset{\mathrm{d}}{=} \boldsymbol{A}\Big(\sum_{k=1}^{K} \boldsymbol{A}^T\{\boldsymbol{X}_k^T \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_k\}\boldsymbol{A}\Big)^{-1} \sum_{k=1}^{K} \boldsymbol{A}^T\{\boldsymbol{X}_k^T \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_k\}\boldsymbol{A}\hat{\boldsymbol{\beta}}_{k,\hat{\mathcal{A}}^{(c)}}, \quad (2.2)$$

where $\hat{\boldsymbol{\theta}}_k = \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_k$. The summation over the $K$ terms in (2.2) and the set of weights used in the combining can boost estimation power and efficiency. As a result, we can show that $\hat{\boldsymbol{\beta}}^{(c)}$ is asymptotically equivalent to the corresponding estimator using the full dataset, and more.

The majority voting idea discussed is closely connected with the developments in Meinshausen and Buhlmann (2010) and Shah and Samworth (2013) on stability selection. Their goal is to develop stable penalized estimators, while ours is to investigate whether we can analyze extremely large data by splitting the task. Here, computational feasibility is in the forefront of our development. Unlike Meinshausen and Buhlmann (2010) and Shah and Samworth (2013) in which the same tuning parameter $\lambda$ is used for all subsets, our $\lambda_k$'s are chosen independently. This allows us to defer our task of coordinating the analyses from subset data to the last combination step. Our development allows $K \to \infty$, as $n \to \infty$, whereas Meinshausen and Buhlmann (2010) and Shah and Samworth (2013) have $K$ finite, e.g., $K = 2$.

A solution path is obtained for every subset in the proposed split and conquer approach. 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 $\hat{\boldsymbol{\beta}}^{(c)}$ of (2.2), and compare it with the penalized estimator $\hat{\boldsymbol{\beta}}^{(a)}$ of (1.3).

### 3.1. Sign consistency

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

Let $\boldsymbol{\theta}^0 = \boldsymbol{X}\boldsymbol{\beta}^0$, $\boldsymbol{\theta}_k^0 = \boldsymbol{X}_k\boldsymbol{\beta}^0$, and $\beta_* = \min\{|\beta_j^0| : \beta_j^0 \neq 0\}$. Let $\overline{\mathcal{A}}$ be the complement of the true nonzero set $\mathcal{A} = \{j : \beta_j^0 \neq 0\}$. For any index set $S$, let $\boldsymbol{X}_S$ be the $n \times |S|$ submatrix of $\boldsymbol{X}$ formed by the columns whose indices are in $S$, and $\boldsymbol{X}_{k,S}$ be the $n_k \times |S|$ submatrix of $\boldsymbol{X}_k$ formed by the columns whose indices are in $S$. We need some regularity conditions on the design matrix. Let

$\{b_{s,K}\}$ be a diverging sequence of positive numbers that depends on $s$ and $K$. We extend the regularity conditions in Fan and Lv (2011) on the entire dataset to each subset:

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

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 $\nabla^2$ 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 **A1** is minor. For instance, in the linear regression model, we have $\boldsymbol{\Sigma}(\boldsymbol{\theta}^0) = \boldsymbol{I}_n$ and $\boldsymbol{\mu}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta}) = \boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta}$. Then **A1** is implied by

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

Since $\{b_{s,K}\}$ is a diverging sequence, the first two constraints in **A1** are weaker than the first two in $\mathbf{A1}_{(G)}$. Condition $\mathbf{A1}_{(G)}$ match with those discussed in the literature under both the settings of fixed matrix design (Fan and Lv (2011)) and of Gaussian random matrix design (Wainwright (2009)). In non-Gaussian generalized linear models, the variance matrices $\boldsymbol{\Sigma}(\boldsymbol{\theta}^0)$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0)$ typically involve the covariates through the linear predictors $\boldsymbol{\theta}^0$ and $\boldsymbol{\theta}_k^0$. If we take the smallest diagonal element of $\boldsymbol{\Sigma}(\boldsymbol{\theta}^0)$, say $h_n = \min_{1\le i\le n}\sigma(\theta_i^0)$, as bounded below by a small constant or $h_n^{-1} = O(b_{s,K})$, then Condition **A1** is implied by $\mathbf{A1}_{(G)}$. For example, consider a Poisson model, where $\boldsymbol{\Sigma}(\boldsymbol{\theta}^0) = \text{diag}\{\exp(\theta_i^0)\}$ 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 $\boldsymbol{I}_{|\mathcal{A}|}$, if the divergence sequence $b_{s,K}$ in **A1** is such that $\big[1 - \Phi\{\log(b_{s,K})\big/\|\boldsymbol{\beta}_{\mathcal{A}}^0\|_2\}\big]^n \to 1$, then by a direct calculation we have $h_n^{-1} = O_p(b_{s,K})$, and thus $\mathbf{A1}_{(G)}$ implies **A1**. See the discussions in Zhang and Huang (2008) and Wainwright (2009).

Following Zhang and Huang (2008), to obtain a slightly stronger sign consistency result than that of Fan and Lv (2011) (when $K = 1$), we introduce 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

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

These sequences are related to the error tolerance level under the design of **A1**; 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 require that the tuning parameter $\lambda_k$ satisfies:

$$\textbf{A3}\ \begin{cases} b_{s,K}\rho'(\frac{\beta_*}{2};\lambda_k) = o(\beta_*), \\ \max\limits_{\boldsymbol{\delta}\in\mathcal{N}_0} \kappa(\rho(\cdot;\lambda_k);\boldsymbol{\delta}) = o(\tau_{0,k}), \\ \|\boldsymbol{X}^T_{k,\overline{\mathcal{A}}}\boldsymbol{\Sigma}(\boldsymbol{\theta}^0_k)\boldsymbol{X}_{k,\mathcal{A}}\{\boldsymbol{X}^T_{k,\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\theta}^0_k)\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}\|_\infty \le \dfrac{C\rho'(0+;\lambda_k)}{\rho'(\beta_*/2;\lambda_k)}, \end{cases}$$

where $\kappa(\rho(\cdot;\lambda_k);\boldsymbol{\delta}) = \lim_{\epsilon\to 0+}\max_{1\le j\le s}\sup_{t_1<t_2\in(\delta_j-\epsilon,|\delta_j|+\epsilon)} -[\rho'(t_2;\lambda_k)-\rho'(t_1;\lambda_k)]/(t_2-t_1)$, $\tau_{0,k} = \min_{\boldsymbol{\delta}\in\mathcal{N}_0}\lambda_{\min}[n_k^{-1}\boldsymbol{X}^T_{k,\mathcal{A}}\Sigma(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})\boldsymbol{X}_{k,\mathcal{A}}]$, and $C$ is a positive constant with $C\in(0,1)$. The proof of the following is in the Appendix.

**Theorem 1.** *Suppose the sample size of the kth subset is $n_k = O(n/K)$, $k = 1,\ldots,K$, and that $\max_{1\le k\le K} n_k/\min_{1\le k\le K} n_k = O(1)$. If* A1–A3 *are satisfied and $s = o(\min\{(\beta_* b_{s,K})^{-1}, \beta_*^{-2}(K/n)^{\alpha}\})$, then with probability at least*

$$1 - 2Ks\exp\Big\{-\frac{v^2_{n,K}}{nK}\Big\} - 2K(p-s)\exp\Big\{-\frac{u^2_{n,K}}{nK}\Big\}, \tag{3.1}$$

*the combined estimator is sign consistent with $\|\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_\infty \le \beta_*/2$ 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 (3.1) goes to 1. To ensure later, we require $\log(Ks) = o(\min\{nb^{-2}_{s,K}\beta^2_*/K,\ n^{1-2\alpha}K^{2\alpha}\})$ 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 $K=1$. Thus, when we increase the number of splits, we impose a stronger constraint on the growth rate of $p$ to ensure that each subset contain enough data to provide a sign consistent estimator for the unknown model parameters.

Consider the 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 **A2** holds and $s = o(\min\{(\beta_* b_{s,K})^{-1}, \beta_*^{-2}(K/n)^{\alpha}\})$; A proof is provided in the Appendix. When $K=1$, these are the conditions 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 $K=1$, Theorem 1 is a slightly stronger result than that reported in Fan and Lv (2011) for the $K=1$ case.

### 3.2. Oracle property

We show here that, after we strengthen some of the regularity conditions, our combined estimator has an oracle property under the $L_2$ norm, that the combined estimator converges at rate of $O(\sqrt{s/n})$ under the $L_2$ norm. We also show that it is asymptotically normal with the same variance as the penalized estimator using all the data.

We impose regularity conditions on the design matrices that are the same as Condition 4 of Fan and Lv (2011), when $K=1$.

$$\mathbf{A4} \quad \begin{cases} \min\limits_{\boldsymbol{\delta}\in\mathcal{N}_\tau} \lambda_{\min}(\boldsymbol{X}_{k,\mathcal{A}}^T \boldsymbol{\Sigma}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})\boldsymbol{X}_{k,\mathcal{A}}) \geq c n_k, \\ tr(\boldsymbol{X}_{k,\mathcal{A}}^T \boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}) = O(s n_k), \\ \|\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T \boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}\|_{2,\infty} = O(n_k), \\ \max\limits_{\boldsymbol{\delta}\in\mathcal{N}_0, 1\leq j\leq s} \lambda_{\max}[\nabla^2\{\boldsymbol{x}_{k,j}^T \boldsymbol{\mu}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta}_{\mathcal{A}})\}] = O(n_k), \end{cases}$$

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

Again **A4** is minor, implied by

$$\mathbf{A4}_{(G)} \begin{cases} \min\limits_{\boldsymbol{\delta}\in\mathcal{N}_\tau} \lambda_{\min}(\boldsymbol{X}_{k,\mathcal{A}}^T \boldsymbol{X}_{k,\mathcal{A}}) \geq O_p(n_k), \\ tr(\boldsymbol{X}_{k,\mathcal{A}}^T \boldsymbol{X}_{k,\mathcal{A}}) = O_p(s n_k), \\ \|\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T \boldsymbol{X}_{k,\mathcal{A}}\|_{2,\infty} = O_p(n_k). \end{cases}$$

The first part of $\mathbf{A4}_{(G)}$ matches conditions discussed in the literature under the settings of fixed matrix design ( Fan and Lv (2011)) and Gaussian random matrix design (Marcenko and Pastur (1967); Takemura and Sheena (2005)). The third part of $\mathbf{A4}_{(G)}$ is minor, since $\|\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T \boldsymbol{X}_{k,\mathcal{A}}\|_{2,\infty} \leq \|\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T \boldsymbol{X}_{k,\mathcal{A}}\|_\infty$ and we can approximate the order of the $L_\infty$ norm. More generally, 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 $\lambda_{\min}(\boldsymbol{X}_{k,\mathcal{A}}^T \boldsymbol{\Sigma}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})\boldsymbol{X}_{k,\mathcal{A}}) \geq h_n \lambda_{\min}(\boldsymbol{X}_{k,\mathcal{A}}^T \boldsymbol{X}_{k,\mathcal{A}})$. In addition, by direct calculation, $tr(\boldsymbol{X}_{k,\mathcal{A}}^T \boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}) = \sum_{j=1}^s \sum_{i=1}^{n_k} x_{ij}^2 b''(\theta_i^0)$. It follows that, when $x_{ij}$ is fixed or a random variable such that such that $x_{ij} = O_p(1)$, $\mathbf{A4}_{(G)}$ and thus also Condition **A4** are satisfied.

Similar to **A3**, we impose a condition on the tuning parameter $\lambda_k$:

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

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

To ensure asymptotic normality, we impose a Lindeberg-type condition:

$$\textbf{A6} \quad \begin{cases} \max\limits_{i=1,\ldots,n} E|y_i - b'(\theta_i^0)|^3 = O(1), \\ \sum\limits_{i=1}^{n}(\boldsymbol{z}_i^T\boldsymbol{B}^{-1}\boldsymbol{z}_i)^{3/2} \to 0 \quad \text{as} \quad n\to\infty, \end{cases}$$

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

A proof of the following can be found in the Appendix.

**Theorem 2.** *Suppose the sample size of the kth 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* A4–A5 *are satisfied and that* $b_*/\sqrt{Ks/n}\to\infty$.

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

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

$$\boldsymbol{D}[\boldsymbol{X}_{\mathcal{A}}^T\boldsymbol{\Sigma}(\boldsymbol{\theta}^0)\boldsymbol{X}_{\mathcal{A}}]^{1/2}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0) \xrightarrow{\text{D}} N(\boldsymbol{0},\phi\boldsymbol{G}). \tag{3.2}$$

The limiting distribution of $\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)}$ in (3.2) is that of $\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(a)}$ in Fan and Lv (2011), where the entire dataset is analyzed. Thus, the combined estimator $\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)}$ is asymptotically as efficient as $\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(a)}$. Together with the fact that both estimators are model selection consistent, the combined estimator $\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)}$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(a)}$.

The signal strength and sparsity assumptions in Theorem 2 depend on the number of splits. When $K = O(1)$, asymptotic equivalence holds without any additional requirements on either, but with $K$ going to infinity, we pay the price that stronger conditions are needed. If stronger signal strength is a concern in a specific problem, the two-stage estimation approach of Zhang and Zhang (2014) can perhaps be used to weaken the requirement. The strengthened conditions ensure that each subset contains enough data to provide a meaningful inference for the unknown model parameters.

### 3.3. Error control

We provide an upper bound of the expected number of falsely selected variables and a lower bound of 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 $\bar{s}_k = E(|\hat{\mathcal{A}}_k|)$ be the average number of selected variables of the penalized estimator from the $k$th subset.

A similar result is provided by Fan, Samworth, and Wu (2009) and Meinshausen and Buhlmann (2010), both of which only considered the special case of $K = 2$.

**Theorem 3.** *Assume the distributions 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* $k = 1, \ldots, K$*, and that* $E(|\mathcal{A} \cap \hat{\mathcal{A}}_k|)/E(|\overline{\mathcal{A}} \cap \hat{\mathcal{A}}_k|) \geq |\mathcal{A}|/|\overline{\mathcal{A}}|$*, for the set of selected variables* $\hat{\mathcal{A}}_k$ *of any penalized estimator. If* $w \geq s^*K/p - 1$*, then for the combined estimator* $\hat{\boldsymbol{\beta}}^{(c)}$*,*

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

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

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

Here $s^*$ and $s_*$ depend on the choice of the threshold $w$. If $w = K - 1$, the combined estimator only selects the variables that are selected in all $K$ subsets, and the expected number of falsely selected variables is bounded above by $(s^*)^K/p^{K-1}$. If $s^*$ is bounded by $c^{1/K}p^{1-1/K}$ for a constant $c$, the expected number of falsely selected variables is bounded by the constant $c$. In sparse models, $s^*$ is usually small and so is $c$. If $w = 0$ the combined estimator selects any variables that are selected in one or more subsets. Then the lower bound for the expected number of truly selected variables achieves the true number of non-zero set $|\mathcal{A}|$, but the upper bound for the expected number of false selected variables can be very loose, up to $|\overline{\mathcal{A}}|$.

There is a trade-off between the upper and lower bounds in Theorem 3 for the choice of $w$. 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

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

The following provides some detail.

**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* $23n^3/3 + 71n/6 + 8n^2(p - 31/16) + 12np$*.*

The following result specifies mild conditions under which the number of computing steps needed in our approach is always less than that of a direct use of the LARS algorithm on the entire data.

**Theorem 4.** *In the setting of 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 split-and-conquer approach has fewer computing steps than that of a direct use of a LARS algorithm on the full dataset.*

The statement in Theorem 4 is a conservative one and our numerical study in Section 4.1 suggests that on average the computing time saved by the split-and-conquer approach is quite significant. Figure 1 demonstrates how the average computing time changes for different $n$, $p$, and $K$ using the LARS algorithm.

Calculations of computing savings by average computing steps can be obtained for 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. A similar finding in a computational intensive robust multivariate scale estimation (where $p$ is fixed) was reported in Section 5.3 of Singh, Xie, and Strawderman (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 that the computing effort of the combination is ignorable. Then the split-and-conquer approach needs* $O(n^a p^b / K^{a-1})$ *steps and using the split-and-conquer approach results in a computing saving on the order of* $K^{a-1}$ *times.*

That the computing effort in the combination is negligible is often satisfied in our context. We use majority voting to determine the number of non-zero coefficients and then use a weighted linear combination formula to combine the $K$ estimators. There are roughly $Ks$ non-zero coefficients across all $K$ subsets, and a computing order of $O(Ks)$ is often enough to identify them. Weighing depends on the number of non-zero coefficients $s$ and $K$, and the computing order is $O(Ks + Ks^2 + s^3)$, where the highest order $O(s^3)$ is for the inversion of the roughly the size $s \times s$ matrix (Golub and Van Loan (1983); Trefethen and Bau III (1997)). As $n \to \infty$, this computation is often negligible compared to the order $O(n^a p^b)$.

For LARS algorithm, a reviewer pointed out an alternative approach that directly applies a split-and-conquer method to the calculation of the sample covariance matrix instead of the LARS estimator. In the LASSO and LARS setting,

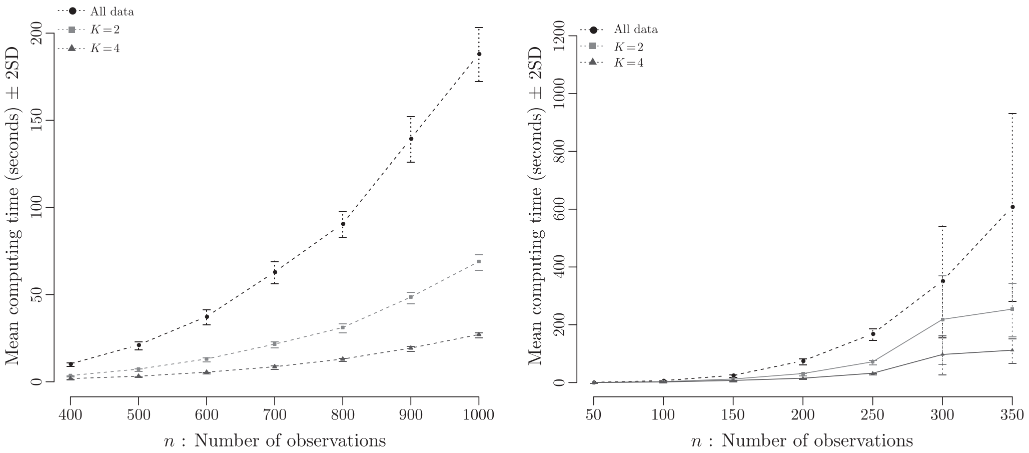

Figure 1. Computing time comparison for different $K$ using LARS algorithm: Mean ± 2Standard Deviation (SD) over 100 replications. Figure 1 (a) [left] is for $p = 2n$ with $n = 400, 500, 600, 700, 800, 900, 1,000$. Figure 1 (b) [right] is for $p = 100n$ with $n = 50, 100, 150, 200, 250, 300, 350$.

the Gram matrix $\boldsymbol{X}^T\boldsymbol{X}$ is the only sufficient statistic; when $p < n$, it is independent of $n$. We can use a parallel computing approach to obtain the overall Gram matrix $\boldsymbol{X}^T\boldsymbol{X}$, and then feed it into the LARS solver. This approach can also effectively handle big data problems involving the LASSO/LARS method when $p < n$, but when $p > n$, the LARS algorithm fits at most $n$ variables. Since the inversion of the Gram matrix is the most costly computing part in the LARS algorithm, the alternative approach often does not save significant computing time, even when $p > n$.

## 4. Numerical Studies

We use several numerical studies, using both simulation and real data, to illustrate the performance of the proposed split-and-conquer approach. We also compare the combined estimators with their corresponding penalized estimators obtained using the entire dataset, 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 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 development is illustrated using the $L_1$, SCAD and MCP penalty functions.

### 4.1. Linear regression with $L_1$ norm penalty

Here the response variable $\boldsymbol{y}$ follows a Guassian linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$, where $\varepsilon$ are IID $N(0, 1)$ errors and the explanatory variables $\boldsymbol{X}$ are generated from a $N(0, \boldsymbol{I})$ distribution with $\boldsymbol{I}$ the identity matrix. In our 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 contained $s = \lfloor\sqrt{p}\rfloor$ nonzero coefficients whose true values were around $\sqrt{2K\log(p)/n}$. To get the LASSO estimators using the $L_1$ norm penalty, the LARS algorithm (Efron et al., 2004) was applied and BIC criterion was used for selecting the tuning parameter.

We repeated our simulation 100 times under each setting of $(n, p)$. For the final estimators, we recorded the means of computing time and the numbers of selected nonzero coefficients. To demonstrate the error control property, we calculated model selection sensitivity (the number of truly selected variables divided by the true model size) and model selection specificity (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 dataset was used to get the LASSO estimator. To examine the performance of the combined estimator, we took $K = 2, 4, 6$ and $w = 1, \ldots, \lfloor K/2\rfloor$.

According to Table 1, all estimators selected some noise variables in addition to the true $s$ nonzero variables, consistent with the performance of LASSO-type estimators. When $K = 4$ or 6 and $w = 2$ or 3, the model selection specificities increase a lot, which indicates that the combined estimator is more efficient in removing noise and spurious variables from the selected models. At each given setting of $(n, p, K)$, with the increase of the threshold $w$, model selection sensitivity decreases as specificities increases.

Computing time decreases drastically as $K$ increases, as seen in Column 5 of Table 1. The time savings reported are between $(1 - 1/K^2)100\%$ and $(1-1/K)100\%$, perhaps because $n$ and $p$ are roughly the same in Table 1. To further study computing savings in the LARS algorithm, we performed additional simulations with $p = 2n$ for $n = 400-1,000$, and with $p = 100n$ for $n = 50-350$, this under the same linear regression set up. We considered $K = 1, 2, 4$. The average computing times (with standard errors) over 100 repetitions are plotted in Figure 1. Savings appear to be between $(1 - 1/K^2)100\%$ and $(1 - 1/K)100\%$ in Figure 1(a), and roughly $(1 - 1/K)100\%$ in Figure 1(b).

### 4.2. Generalized linear model with SCAD and MCP penalties

The SCAD and MCP estimators are obtained based on non-concave penalized likelihood functions and compared with the LASSO estimators, they often select a tighter model and fewer noise variables. We considered the SCAD and MCP estimators under both the linear regression and logistic models.

For the linear regression case, the response variable $\boldsymbol{y}$ is $\boldsymbol{y} = \boldsymbol{X\beta} + \varepsilon$, where $\varepsilon$ are IID $N(0, 1)$ errors. For the logistic regression case, the response variable $\boldsymbol{y}$ follows the Bernoulli distribution with success probability $p(\boldsymbol{X\beta}) = e^{\boldsymbol{X\beta}}/(1+$

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

| Simulation setting | | | | Computing time (in second) | Model selection | | | specificity (in %) |
|---|---|---|---|---|---|---|---|---|
| $n$ | $p$ | $s$ | $K$ | | $w$ | # selected variables | sensitivity (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 | 1,000 | 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) |
| 1,000 | 1,000 | 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) |

$e^{\boldsymbol{X\beta}}$). In our simulations, we considered two settings to generate the design matrix $\boldsymbol{X}$: a set of $p$ variables were generated as $N(0, \boldsymbol{I})$; a set of $p$ variables were generated as $N(0, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}(i, j) = 0.6^{|i-j|}$.

Sample sizes were $n = 10,000$ and $n = 100,000$. For the linear regression, $p = 1,000$ and for the logistic model $p = 200$. In all cases, the true model had $s = 30$ nonzero coefficients (with values around 0.4). In order to get the SCAD and MCP estimators, the NCVREG algorithm (Breheny and Huang (2011)) was applied and a 10-fold cross-validation was used to select the tuning parameters.

The simulation was repeated 100 times. We recorded the computing time and the number of selected variables, and calculated model selection sensitivity and specificity. In addition, the MSE (mean squared error) was calculated in the linear regression case, and the misclassification rate with 0.5 as threshold was reported in the logistic regression case. The results are in Table 2.

Computing times were reduced through the split-and-conquer procedure under all settings. For the SCAD and MCP penalties, the proposed split-and-conquer approach reduced computing time drastically in the linear regression setting (with $K = 10$), using about 1/10 of the time when the explanatory variables were independent and about 1/3 of when the explanatory variables were correlated. For the logistic model with $K = 5$, the average saving was a little less. When the explanatory variables were independent, the combined estimator needed about half of the time compared to directly performing the analysis on the full dataset. When the explanatory variables were correlated, the combined estimator by the proposed method saved up to 25% time. With $n = 100,000$, we were not able to perform the SCAD or the MCP regression on the full dataset due to computer memory limitations, but we obtained estimators using the split-and-conquer procedure with results reported in Table 2.

According to Table 2, the SCAD estimators performed similarly to the MCP estimators. In all cases, the combined estimators had good model selection results with high model selection sensitivity and specificity that were similar to those of the penalized estimators analyzing the full dataset. This is held for the MSE's in linear regression settings, and for misclassification rates in logistic regression settings.

Figure 2 presents several sets of side-by-side boxplots to compare the estimates $\hat{\boldsymbol{\beta}}^{(c)}$ with the penalized estimates $\hat{\boldsymbol{\beta}}^{(a)}$, when both are available in the settings of Table 2. From the boxplots, we can see that the combined estimates had almost the same mean and spread as the estimates obtained using full dataset.

### 4.3. Numerical analysis on POEs manifest data

After the 911 terrorist attack, substantial efforts were made to devise strategies for inspecting containers coming through the US POEs to intersect 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 potentially illegitimate activities is a small but important part of layered defenses for 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 obtained a set of manifest data that contain all shipping records coming through the POEs across the US in February, 2009. The goal

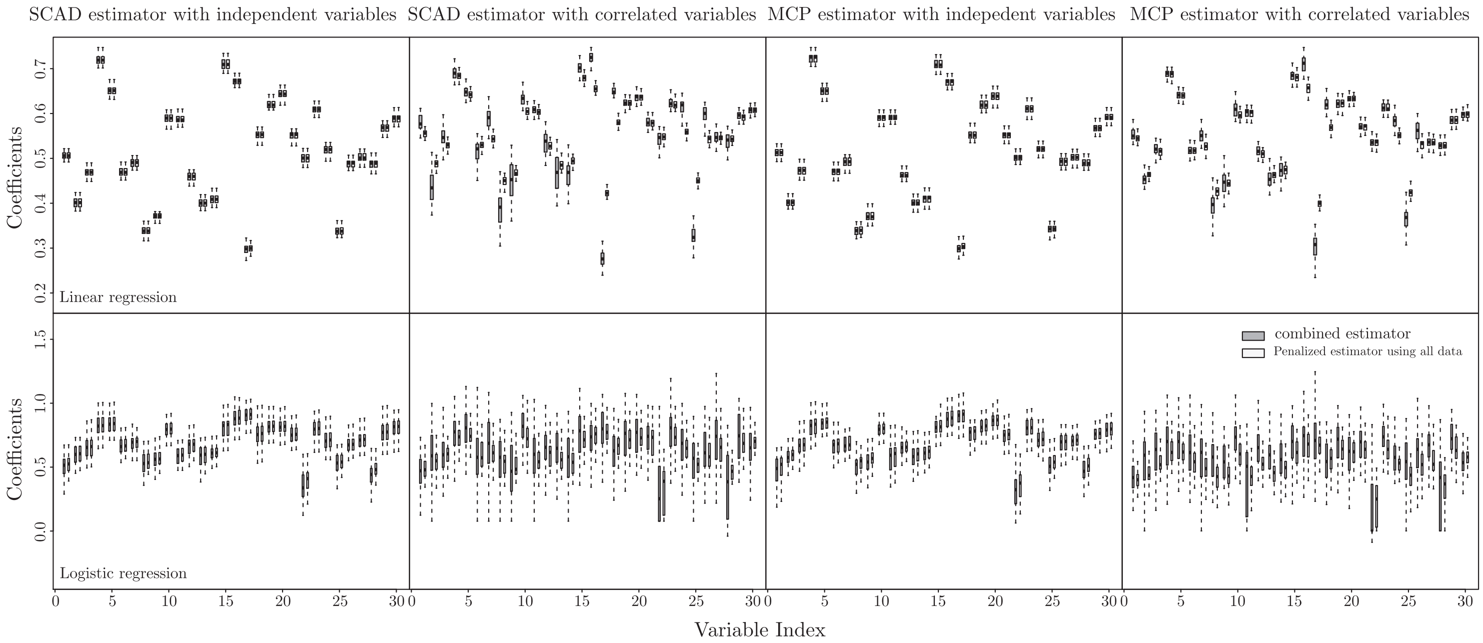

Figure 2.Comparison of parameter estimation for the combined estimator and the penalized estimator using all data. Box plots of estimation for variables in the true model. Grey: the combined estimator; White: the estimator using all data. Top panels: Linear regression; bottom panels: Logistic regression.

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 | 10,000 | 1,000 | 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 | 10,000 | 1,000 | 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 | 100,000 | 1,000 | 1 | - (-) | - (-) | - (-) | - (-) | - (-) |
| | | | 100 | 1136.70 (74.65) | 30 (0) | 100 (0) | 100 (0) | 1.00 (0.01) |
| Correlated | 100,000 | 1,000 | 1 | - (-) | - (-) | - (-) | - (-) | - (-) |
| | | | 100 | 3074.53 (25.01) | 30 (0) | 100 (0) | 100 (0) | 1.06 (0.01) |
| MCP: Linear regression | | | | | | | | |
| Independent | 10,000 | 1,000 | 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 | 10,000 | 1,000 | 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 | 100,000 | 1,000 | 1 | - (-) | - (-) | - (-) | - (-) | - (-) |
| | | | 100 | 1565.54 (132.38) | 30 (0) | 100 (0) | 100 (0) | 1.00 (0.01) |
| Correlated | 100,000 | 1,000 | 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 | 10,000 | 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 | 10,000 | 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 | 100,000 | 200 | 1 | - (-) | - (-) | - (-) | - (-) | - (-) |
| | | | 20 | 1352.14 (76.2) | 30 (0) | 100 (0) | 100 (0) | 17.38 (0.12) |
| Correlated | 100,000 | 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 | 10,000 | 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 | 10,000 | 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 | 100,000 | 200 | 1 | - (-) | - (-) | - (-) | - (-) | - (-) |
| | | | 20 | 1301.95 (63.27) | 30 (0) | 100 (0) | 100 (0) | 17.34 (0.13) |
| Correlated | 100,000 | 200 | 1 | - (-) | - (-) | - (-) | - (-) | - (-) |
| | | | 20 | 4485.9 (186.29) | 29.58 (0.50) | 98.60 (1.66) | 100 (0) | 10.00 (0.09) |

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

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 |

shipments. In the project, a logistic regression model was 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 is much redundant information, we used 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 are categorical and there are $p = 213$ variables in total. There are also text fields but we do not consider any semantic analysis or text mining here.

Due to the amount of traffic and a large number of entry sites, it is a challenge to analyze the full dataset on a single computer. This 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 experts in the field suggest that 1% to 10% of cargo containers need further inspections. With the assistance of field experts, 22 potentially influential shipment characteristics were selected used in a logistic model to generate 1% to 10% high-risk shipments. Our task was then to test whether a penalized regression technique could identify these 22 characteristics among all shipment features recorded in the manifest data. Our computer could do a SCAD penalized regression on single-day data, but not week-long data. Thus we performed the SCAD penalized regression on each day's data and combined the daily estimators to obtain 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 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. From Table 4, the split-and conquer approach has identified the most influential variables in the manifest data. In particular, the combined estimates have both high model selection sensitivity and

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) |

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) |

specificity, while the daily estimates either select more noise variables or exclude more influential variables. The combined estimates are more stable than the daily estimates. The combined estimator has a slightly smaller misclassification rate.

The Sporting variable in the category of Container Contents is left out in the model selections of all daily analyses and the split-and-conquer approach. All other variables with non-zero coefficients are recovered by the split-and-conquer approach.

## 5. Discussions

One important step in the split-and-conquer approach is the combination, it depends on the desired statistical procedure. According to Singh, Xie, and Strawderman (2005), Xie, Singh, Strawderman (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 (e.g., any settings where a likelihood or penalized likelihood or estimating equation method applies), as well as to problems of hypothesis testings and confidence intervals.

## Acknowledgements

This research is partly supported by research grants from NSF (DMS1107012, DMS0915139, SES0851521). It is developed from a chapter of the first author's thesis. The authors thank Professor Naisyin Wang, an 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.

## Appendix

### Proof of Theorem 1

We state two lemmas without proofs. The first is Proposition 4 of Fan and Lv (2011), and the second 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 an $n$-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 $\epsilon \in (0, \infty)$,

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

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

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

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

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

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

$$\boldsymbol{X}_{k,\mathcal{A}}^T\boldsymbol{y}_k - \boldsymbol{X}_{k,\mathcal{A}}^T\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}_k) - n_k\bar{\rho}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}; \lambda_k) = 0, \tag{A.1}$$

$$n_k^{-1}\|\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T[\boldsymbol{y}_k - \boldsymbol{\mu}(\hat{\boldsymbol{\theta}}_k)]\|_\infty < \rho'(0+; \lambda_k), \tag{A.2}$$

$$\lambda_{\min}[\boldsymbol{X}_{k,\mathcal{A}}^T\Sigma(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}] > n_k\kappa(\rho; \hat{\boldsymbol{\beta}}_{k,\mathcal{A}}), \tag{A.3}$$

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

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

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

For any $\hat{\boldsymbol{\beta}}_k = (\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}^T, 0)^T$ with $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} \in \mathcal{N}_0 = \{\boldsymbol{\delta} \in \Re^s : \|\boldsymbol{\delta} - \boldsymbol{\beta}_\mathcal{A}^0\|_\infty \le \beta_*/2\}$, (A.1) 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 $\boldsymbol{r}_{k,\mathcal{A}} = (r_{k1}, \dots, r_{ks})^T$ 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 $j = 1, \dots, s$. It follows that

$$\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}_\mathcal{A}^0 = \{\boldsymbol{X}_{k,\mathcal{A}}^T\Sigma(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}\big[\boldsymbol{X}_{k,\mathcal{A}}^T\{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)\} - n_k\bar{\rho}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}; \lambda_k) - \boldsymbol{r}_{k,\mathcal{A}}\big]. \tag{A.4}$$

Thus,

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_\infty &\le \|\{\boldsymbol{X}^T_{k,\mathcal{A}}\Sigma(\boldsymbol{\theta}^0_k)\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}\|_\infty\big\{\|\boldsymbol{\xi}_{k,\mathcal{A}}\|_\infty + \|\boldsymbol{\eta}_{k,\mathcal{A}}\|_\infty + \|\boldsymbol{r}_{k,\mathcal{A}}\|_\infty\big\}\\ &\le O(b_{s,K}n_k^{-1})\big\{c_1^{-1/2}\frac{v_{n,K}}{K} + n_k\rho'(\frac{\beta_*}{2};\lambda_k) + O(n_k)\times\beta_*^2 s\big\}\\ &= O(c_1^{-1/2}b_{s,K}\frac{v_{n,K}}{n_kK} + b_{s,K}\rho'(\frac{\beta_*}{2};\lambda_k) + b_{s,K}\beta_*^2 s) = o(\beta_*),\end{aligned}$$

where $\boldsymbol{\xi}_{k,\mathcal{A}} = \boldsymbol{X}^T_{k,\mathcal{A}}\{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}^0_k)\}$ 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 $E_{1k}$, A1, and that the concavity of $\rho$. The last line holds by A2 and A3.

By Miranda's existence theorem (e.g., Vrahatis (1989)), there exists a solution $\hat{\boldsymbol{\beta}}_k = (\hat{\boldsymbol{\beta}}^T_{k,\mathcal{A}}, 0)^T$, with $\hat{\beta}_{k,\mathcal{A}} \in \mathcal{N}_0$, to (A.1).

By a Taylor expansion,

$$\begin{aligned}\boldsymbol{X}^T_{k,\overline{\mathcal{A}}}\{\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}_{k,\mathcal{A}}) - \boldsymbol{\mu}(\boldsymbol{\theta}^0_k)\} &= \boldsymbol{X}^T_{k,\overline{\mathcal{A}}}\Sigma(\boldsymbol{\theta}^0_k)\boldsymbol{X}_{k,\mathcal{A}}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) + \boldsymbol{w}_{k,\overline{\mathcal{A}}}\\ &= \boldsymbol{X}^T_{k,\overline{\mathcal{A}}}\Sigma(\boldsymbol{\theta}^0_k)\boldsymbol{X}_{k,\mathcal{A}}\{\boldsymbol{X}^T_{k,\mathcal{A}}\Sigma(\boldsymbol{\theta}^0_k)\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}(\boldsymbol{\xi}_{k,\mathcal{A}} - \boldsymbol{\eta}_{k,\mathcal{A}} - \boldsymbol{r}_{k,\mathcal{A}}) + \boldsymbol{w}_{k,\overline{\mathcal{A}}}, \quad \text{(A.5)}\end{aligned}$$

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

$$\begin{aligned}n_k^{-1}\|\boldsymbol{X}^T_{k,\overline{\mathcal{A}}}[\boldsymbol{y}_k - \boldsymbol{\mu}(\hat{\boldsymbol{\theta}}_k)]\|_\infty &\le n_k^{-1}\|\boldsymbol{\xi}_{k,\overline{\mathcal{A}}}\|_\infty + \|\boldsymbol{X}^T_{k,\overline{\mathcal{A}}}\{\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}_{k,\mathcal{A}}) - \boldsymbol{\mu}(\boldsymbol{\theta}^0_k)\}\|_\infty\\ &\le n_k^{-1}\|\boldsymbol{\xi}_{k,\overline{\mathcal{A}}}\|_\infty + n_k^{-1}\|\boldsymbol{X}^T_{k,\overline{\mathcal{A}}}\Sigma(\boldsymbol{\theta}^0_k)\boldsymbol{X}_{k,\mathcal{A}}\{\boldsymbol{X}^T_{k,\mathcal{A}}\Sigma(\boldsymbol{\theta}^0_k)\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}\|_\infty\\ &\quad\times\big\{\|\boldsymbol{\xi}_{k,\mathcal{A}}\|_\infty + \|\boldsymbol{\eta}_{k,\mathcal{A}}\|_\infty + \|\boldsymbol{r}_{k,\mathcal{A}}\|_\infty\big\} + n_k^{-1}\|\boldsymbol{w}_{k,\overline{\mathcal{A}}}\|_\infty\\ &= c_1^{-1/2}\frac{u_{n,K}}{n_kK} + O(n_k^{\alpha-1}\frac{v_{n,K}}{K}) + O(n_k^\alpha s\beta_*^2) + C\rho'(0+;\lambda_k) + O(s\beta_*^2)\\ &= o(1) + C\rho'(0+;\lambda_k) < \rho'(0+;\lambda_k).\end{aligned}$$

Here, $\boldsymbol{\xi}_{k,\overline{\mathcal{A}}} = \boldsymbol{X}^T_{k,\overline{\mathcal{A}}}[\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}^0_k)]$. So (A.1) and (A.2) hold for $\hat{\beta}_k = (\hat{\boldsymbol{\beta}}^T_{k,\mathcal{A}}, 0)^T$. Since $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} \in \mathcal{N}_0$, by A3 (A.3) is satisfied. Thus, by Lemma A2, $\hat{\beta}_k = (\hat{\boldsymbol{\beta}}^T_{k,\mathcal{A}}, 0)^T$ is the solution in the $k$th subset, for $k = 1,\ldots,K$.

Here $\hat{\beta}_k$ is evaluated under the event $E = \cap_{k=1}^K(E_{1k}\cap E_{2k})\}$ which is an intersection over all $k = 1,2,\ldots,K$, and the event $E$ holds with probability 1 as $n\to\infty$. When $n$ is large enough, we have $\hat{\mathcal{A}}_k = \mathcal{A}$ for all subsets and $\hat{\mathcal{A}}^{(c)} = \mathcal{A}$. In this case, $\hat{\boldsymbol{\beta}}^{(c)}_{\overline{\mathcal{A}}} = 0$, and $\boldsymbol{X}_k\boldsymbol{A} = \boldsymbol{X}_{k,\mathcal{A}}$ where $\boldsymbol{A} = \boldsymbol{E}_{\hat{\mathcal{A}}^{(c)}}$ is the selection matrix defined in (2.2). It follows from (2.2), (A.4), and also $\boldsymbol{X}^T_{k,\mathcal{A}}\Sigma(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}} = \boldsymbol{X}^T_{k,\mathcal{A}}\Sigma(\boldsymbol{\theta}^0_k)\boldsymbol{X}_{k,\mathcal{A}} + o_p(1)$, that, uniformlly

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

Based on A1 and A2,

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

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

### A.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 $\hat{\boldsymbol{\beta}}_k = (\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}^T, 0)^T$ is a consistent estimator of $\boldsymbol{\beta}$ for each $k$ and obtain an asymptotic expansion of $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}$. We then use the asymptotic expansions of $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}$ and also the fact that $\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)}$ is weighted sum of $\hat{\boldsymbol{\beta}}_k$ to obtain the desired results.

**Proof of Theorem 2** (i) Let $u_{n,K}$ be a divergent sequence depending on the total sample size $n$ and the number of subsets $K$ such that $u_{n,K} = o(n)$ and $pK\exp\{-u_{n,k}^2/(nK)\} = o(1)$. Consider events $E_{2k} = \{\|\boldsymbol{X}_{k,\overline{\mathcal{A}}}^T\{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)\}\|_\infty \le c_1^{-1/2}u_{n,K}/K\}$, for $k = 1,\ldots,K$. From Lemma A1, we have that $\mathrm{P}\{\cap_{k=1}^K E_{2k}\} \ge 1-2K(p-s)\exp\{-u_{n,K}^2/(nK)\}$. So, the event $E_a = \cap_{k=1}^K E_{2k}$ holds in probability 1.

First, let us constrain the parameter space to the subspace $\{\boldsymbol{\beta} : \boldsymbol{\beta}_{\overline{\mathcal{A}}} = 0\}$ and also define $\mathcal{N}_\tau = \{\boldsymbol{\delta} \in \Re^s : \|\boldsymbol{\delta} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_2 \le \sqrt{Ks/n}\tau\}$ 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 $k$, let $F_k = \{Q_k(\boldsymbol{\beta}_{\mathcal{A}}^0) > \max_{\boldsymbol{\delta}\in\partial\mathcal{N}_\tau} Q_k(\boldsymbol{\delta})\}$, 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 a 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 $\boldsymbol{\delta}_k^*$ an $s \times 1$ vector on the segment joining $\boldsymbol{\delta}$ and $\boldsymbol{\beta}_{\mathcal{A}}^0$.

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

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

$$\mathrm{P}(F_k) \ge \mathrm{P}(\|\boldsymbol{v}_k\|_2^2 < \frac{c^2Ks\tau^2}{16n}) \ge 1 - 16n\frac{E\|\boldsymbol{v}_k\|_2^2}{c^2Ks\tau^2} \ge 1 - O(\tau^{-2}).$$

Since this holds for any (arbitrarily large) constant $\tau > 0$ and $Q_k(\boldsymbol{\delta})$ is a continuous injective function, there exists a $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} \in \mathcal{N}_\tau$ 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.

Constrain the parameter $\boldsymbol{\beta}$ to the subspace $\{\boldsymbol{\beta} : \boldsymbol{\beta}_{\overline{\mathcal{A}}} = 0\}$. We can show that, under Condition A4 and A5 and also $n_k = O(n/K)$,

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

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

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

Thus, when $n$ is large enough, (A.2) holds. Since A5 also implies (A.3), 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 $k$th subset data, for $k = 1,\ldots,K$.

All statements about $\hat{\beta}_k$ are evaluated under the event $\cap_{k=1}^K E_{2k}$ and $\cap_{k=1}^K E_{2k}$ holds with probability 1, as $n \to \infty$. When $n$ is large enough, we have $\hat{\mathcal{A}}_k = \mathcal{A}$ for all subsets, and $\hat{\mathcal{A}}^{(c)} = \mathcal{A}$. 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 (2.2). Since $\hat{\boldsymbol{\beta}}_{k,\overline{\mathcal{A}}} = 0$ for all $k$, we immediately have $\hat{\boldsymbol{\beta}}_{\overline{\mathcal{A}}}^{(c)} = 0$.

For any $\hat{\boldsymbol{\beta}}_k = (\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}^T, 0)^T$ with $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} \in \mathcal{N}_\tau = \{\boldsymbol{\delta} \in \Re^s : \|\boldsymbol{\delta} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_2 \le \sqrt{Ks/n\tau}\}$, we can obtain by Taylor expansion the same expression of $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}$ as in (A.4). Since $\|n_k\bar{\rho}(\boldsymbol{\beta}_{\mathcal{A}}^0; \lambda_k)\|_2 \le n_k\sqrt{s}\rho'(\beta_*/2; \lambda_k) = O(\sqrt{ns}/K)$ and $\|\boldsymbol{r}_{k\mathcal{A}}\|_2 = \sqrt{s}O(n_k)O_p(Ks/n) = O_p(s^{3/2})$, it follows that

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

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

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} = \boldsymbol{\beta}_{\mathcal{A}}^0 &+ \Big[\sum_{k=1}^K \boldsymbol{X}_{k,\mathcal{A}}^T\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}\Big]^{-1}\\ &\times\Bigg\{\sum_{k=1}^K\{1 + o_p(1)\}\Bigg[\boldsymbol{X}_{k,\mathcal{A}}^T\{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)\} + O_p(\frac{\sqrt{ns}}{K})\Bigg]\Bigg\}.\end{aligned}\tag{A.6}$$

Since $\lambda_{\min}\big(\sum_{k=1}^K \boldsymbol{X}_{k,\mathcal{A}}^T\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}\big) \ge \sum_{k=1}^K \lambda_{\min}\big(\boldsymbol{X}_{k,\mathcal{A}}^T\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}\big)$, by A4 we have that $\lambda_{\max}\big([\sum_{k=1}^K \boldsymbol{X}_{k,\mathcal{A}}^T\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}]^{-1}\big) = O_p(n^{-1})$. In addition, $E\|\boldsymbol{X}_{\mathcal{A}}^T[\boldsymbol{y}-\mu(\boldsymbol{\theta}_k^0)]\|_2^2 \le \phi tr(\boldsymbol{X}_{\mathcal{A}}^T\Sigma(\boldsymbol{\theta}^0)\boldsymbol{X}_{\mathcal{A}}) = \phi\sum_{k=1}^K tr(\boldsymbol{X}_{k,\mathcal{A}}^T\Sigma(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}) = O(sn)$ by A4. It follows $\|\boldsymbol{X}_{\mathcal{A}}^T[\boldsymbol{y} - \mu(\boldsymbol{\theta}^0)]\|_2^2 = O_p(ns)$. Thus, by (A.6),

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

(ii) Under the 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 (A.6) is in fact $o_p(\sqrt{n}/K)$. By (A.6) with this modification, we have

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

From Condition A6,

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

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

### A.3. Proof of Theorem 3

**Proof of Theorem 3.** We show that $\text{P}(j \in \hat{\mathcal{A}}_k) \le \bar{s}_k/p$, $j \in \overline{\mathcal{A}}$, and $\text{P}(j \in \hat{\mathcal{A}}_k) \ge \bar{s}_k/p$ 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 $E(|\mathcal{A} \cap \hat{\mathcal{A}}^{(c)}|) = \sum_{j \in \mathcal{A}} \mathrm{P}(j \in \hat{\mathcal{A}}^{(c)}) \geq |\mathcal{A}|(1 - F(w|K, s_*/p))$.

## A.4. Proof of Lemma 1 and Theorem 4

**Proof of Lemma 1.** We 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 by *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 $U$,$G_A = U^TU$. 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 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 $Y_{11}$ are known from the previous loop, we can update $G_A^{-1}$ as followings: $U_{12} = Y_{11}^TG_{12}, U_{22} = \sqrt{G_{22} - U_{12}^TU_{12}}, Y_{22} = U_{22}^{-1}, Y_{12} = -Y_{11}U_{12}Y_{22}$, and compute $G_{11}^{-1} + Y_{12}Y_{12}^T$, $Y_{12}Y_{22}^T$ and $Y_{22}Y_{22}^T$. Thus, $comp[2] = 8|A|^2 - 10|A| + 7 + (2|A| - 1)n$.

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

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

Each time dropping variable occurs, it add and 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 is $6n^2 + 4n(p-3) + 6p + 8$ computing steps each time and the solution path has $n$ times downsize. The computing steps for the worst case is $23n^3/3 + 71n/6 + 8n^2(p - 31/16) + 12np$.

**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 $\sum_{k=1}^{K} 5n_k^3/3 + 23n_k/6 + 4n_k^2(p - 7/8) + 6n_kp$. Since $\sum_{k=1}^{K} n_k = n$, $5n^3/3 + 23n/6 + 4n^2(p - 7/8) + 6np \geq \sum_{k=1}^{K} 5n_k^3/3 + 23n_k/6 + 4n_k^2(p - 7/8) + 6n_kp$. 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 fewer computing steps than the best case for the 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 the 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 then 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.

### A.5. Verification of A2 in the special case described in the last paragraph of Section 3.1

**Proof.** The proof is straightforward. We compute the order of each term in 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, $v_{n,K}b_{s,K}/(n_kK\beta_*) = o(\sqrt{Kn\log(n)}b_{s,K}/(n^{1-\gamma}\log(n))) = o(1)$ and $v_{n,K}/(n_k^{1-\alpha}K) = o(\sqrt{\log(n)}K^{1/2-\alpha}/n^{1/2-\alpha}) = o(1)$. Finally, $u_{n,K}/(n_kK) = O(\sqrt{K\log(n)}/n^{\alpha_1}) = o(1)$.

## References

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

Ahmed, A., Aly, M., Das, A., Smola, A. J. and Anastasakos, T. (2012). Web-scale multi-task feature selection for behavioral targeting. Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 1737-1741.

Andrews, G. R. (2000). *Foundations of Multithreaded, Parallel, and Distributed Programming.* Addison-Wesley.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Statist.* **5**, 232-253.

Chen, S. S., Donoho, D. L. and Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM Rev.* 129-159.

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

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407-451.

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

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J. and Lv, J. (2011). *Non-concave penalized likelihood with NP-dimensionality. IEEE transaction on Information Theory* **57**, 5467-5484.

Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *J. Mach. Learn. Res.* **10**, 2013-2038.

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

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

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

Meinshausen, N. and Buhlmann, P. (2010). Stability selection. *J. Roy. Statist. Soc. Ser. B* **72**, 417-473.

Shah, R. and Samworth, R. J. (2013). Variable selection with error control: Another look at stability selection. *J. Roy. Statist. Soc. Ser. B* **75**, 55-80.

Singh, K. and Xie, M. and Strawderman, W. E. (2005). Combining information from independent sources through confidence distributions. *Ann. Statist.* **33**, 159-183.

Takemura, A. and Sheena, Y. (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. *J. Multivariate Anal.* **94**, 271-299.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Trefethen, L. N. and Bau III, D. (1997). *Numerical Linear Algebra.* SIAM.

Vrahatis, M. N. (1989). A short proof and a generalization of Miranda's existence theorem. *Proc. Amer. Math. Soc*, **107**, 701-703.

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

Xie, M. and Singh, K. and Strawderman, W. E. (2011). Confidence distributions and a unifying framework for meta-analysis. *J. Amer. Statist. Assoc.* **106**, 320-333.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68**, 49-67.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894-942.

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

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

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* **36**, 1567-1594.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low-dimensional parameters in high-dimensional linear models. *J. Roy. Statist. Soc. Ser. B* **76**, 217-242.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67**, 301-320.

Department of Statistics, Rutgers University, Piscataway, NJ 08854, USA.

E-mail: xueychen@stat.rutgers.edu

Department of Statistics, Rutgers University, Piscataway, NJ 08854, USA.

E-mail: mxie@stat.rutgers.edu

(Received April 2013; accepted December 2013)