DIMACS Technical Report 2012-01
January 2012

A Split-and-Conquer Approach for Analysis of Extraordinarily Large Data

by

Xueying Chen and Minge Xie [1]

Department of Statistics and Biostatistics, Rutgers University
Piscataway, NJ 08854

[1]The research is supported in part by the National Science Foundation Grants DMS0915139, SES0851521 and DMS1107012 and by the Department of Homeland Security Grant 2008-DN-077-ARI012-02.

ABSTRACT

If there are extraordinarily large data, too large to fit into a single computer or too expensive to perform a computationally intensive data analysis, what should we do? To deal with this problem, we propose in this paper a *split-and-conquer* approach and illustrate it using a computationally intensive penalized regression method, along with a theoretical support. Consider a regression setting of generalized linear models with n observations and p covariates, in which n is extraordinarily large and p is either bounded or goes to ∞ at a certain rate of n. We propose to split the data of size n into K subsets of size $O(n/K)$. For each subset of data, we perform a penalized regression analysis and the results from each of the K subsets are then combined to obtain an overall result. We show that the combined overall result still retains the model selection consistency and asymptotic normality under mild conditions. When K is less than $O(n^{1/5})$, we also show that the combined result is asymptotically equivalent to the corresponding analysis result of using the entire data all together, assuming that there were a super computer that could carry out such an analysis. In addition, when a computational intensive algorithm is used in the sense that its computing expense is at the order of $O(n^a)$, $a > 1$, we show that the split-and-conquer approach can reduce computing time and computer memory requirement. Furthermore, the split-and-conquer approach involves a random splitting and a systemic combining. We establish an upper bound for the expected number of falsely selected variables and a lower bound for the expected number for truly selected variables. We also demonstrate that, from the splitting and combining, the approach has an inherent advantage of being more resistant to false model selections caused by spurious correlations. The proposed methodology is demonstrated numerically using both simulation and real data examples.

Keywords: Generalized linear models, Information combining, Large data analysis, Penalized regression

1 Introduction

Consider a generalized linear model:

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

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

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

where $\theta_i = \boldsymbol{x}_i'\boldsymbol{\beta}, i = 1, \ldots, n$. The log-likelihood function $\log f(\boldsymbol{y}; \boldsymbol{X}, \boldsymbol{\beta})$ is then given by

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

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

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

Here, $\boldsymbol{y}$ is a $n \times 1$ response vector, $\boldsymbol{X}$ is a $n \times p$ matrix; ρ is the penalty function with tuning parameter λ_a. The superscript a refers to the result obtained by analyzing *all* data simultaneously. Depending on the choice of penalty function $\rho(\boldsymbol{\beta}; \lambda_a)$, we have bridge regression (Frank and Friedman, 1993), LASSO estimator (Tibshirani, 1996; Chen et al., 2001), LARS algorithm (Efron et al., 2004), SCAD estimator (Fan and Li, 2001) and MCP estimators (Zhang, 2010), among others. In this paper, we focus on the settings used in Fan and Lv (2011) which includes a class of the most commonly used penalty functions to date. On these settings, Fan and Lv (2011) show that the penalized estimators under the generalized linear models (3) have good asymptotic properties, such as model selection consistency and asymptotic normality etc., under some regularity conditions.

In this paper, we propose a split-and-conquer approach for the situation that n is extraordinarily large, too large to perform the aforementioned penalized regression using a single computer or available computing resources to us. In this case, we split the whole

dataset into K subsets of smaller sample sizes. Each subset is then analyzed separately, provided that such an analysis can be performed on the smaller subsets. A set of K results are obtained. Subsequently, the K results are combined to obtain a final result.

The idea of this split-and-conquer approach is simple and straightforward. Its essence can be easily illustrated using a simple special case of the regular Gaussian linear regression where we have finite p and non-sparse $\boldsymbol{\beta}$. In particular, the least square estimator using entire data all together in this case is

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

When we split the dataset into K pieces, the least square estimator obtained from the k^{th} subset is $\hat{\boldsymbol{\beta}}_k = (\boldsymbol{X}_k'\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k'\boldsymbol{y}_k$, where $\boldsymbol{X}_k$ is the design matrix and $\boldsymbol{y}_k$ is the response vector for data in the k^{th} subset. These K least square estimators can be combined, using the inverse of $\hat{\boldsymbol{\beta}}_k$'s variance $S_k \stackrel{\mathrm{d}}{=} \boldsymbol{X}_k'\boldsymbol{X}_k$ as their combining weights, to form a new estimator

$$\hat{\boldsymbol{\beta}}^{(c)} \;=\; (\sum_{k=1}^{K}\boldsymbol{X}_k'\boldsymbol{X}_k)^{-1}\sum_{k=1}^{K}\{(\boldsymbol{X}_k'\boldsymbol{X}_k)\hat{\boldsymbol{\beta}}_k\} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sum_{k=1}^{K}\boldsymbol{X}_k'\boldsymbol{y}_k = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

This combined new estimator $\hat{\boldsymbol{\beta}}^{(c)}$ is identical to $\hat{\boldsymbol{\beta}}^{(a)}$, the least square estimator from analyzing the entire data all together. Thus, we do not lose any information through the split-and-conquer approach. For penalized estimators and under generalized linear models, the results are not so straightforward. Our goal in this paper is to investigate whether we have any similar results to support the split-and-conquer approach under generalized linear models and for penalized estimators. We also investigate whether there are any special properties and benefits for more complex settings beyond this simple case of the least squared estimation with a small fixed p.

The answers to our questions are affirmative. 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 entire data all together, provided that a penalty function discussed in Fan and Lv (2011) is used. The combined estimator can keep the sparsity property and is model selection consistent as long as the penalized estimators from the imposed penalty function are sparse and model selection consistent. When asymptotic normality is attainable, the combined estimator does not lose any efficiency through the split-and-conquer process, in the sense that it has the same asymptotic variance as the penalized estimator using entire data all together. In other words, although the combined estimator

may not be exactly the same as the one using complete data, it is as asymptotically efficient and asymptotically equivalent as the overall penalized estimator analyzing the entire data.

We study the choice of the number of subsets K. The number K should be relatively large so that each subset can by analyzed using computing resources available to us. But K cannot be too large either, because each subset should contain enough data to provide a meaningful estimator for the unknown regression parameter $\boldsymbol{\beta}$. When K is chosen at the order $O(n^{\delta})$, $0 \leq \delta \leq 1/5$, we demonstrate that the combined estimator has the desired properties mentioned in the previous paragraph.

Furthermore, when a computational intensive algorithm with computing expenses at the order of $O(n^a)$, $a > 1$, is used, we show using a simple calculation, as well as demonstrate using numerical examples, that the split-and-conquer approach can release the computing burden in the sense of reducing computing time and computer memory requirement. Consider a simple example of linear regression with L_1 norm penalty function. Even the LARS (Efron et al., 2004) algorithm, which is considered as a fast and efficient algorithm to solve the LASSO problem, requires $O(n^3)$ computations when $p \geq n$ and the computing time can be costly when both n and p are extraordinarily large. When generalized regression models or other more complicated penalty functions are used, the computing cost increases tremendously. In these cases, the split-and-conquer approach provides a feasible way to perform penalized regressions that can reduce both computer memory requirement and computing time.

The split-and-conquer approach involves random splitting that can introduce random errors. But the combining step provides a chance to average them out. In a penalized regression, improvements in model selection can be expected through a majority voting in the combining step. As a result, we are able to establish an upper bound for the expected number of falsely selected variables and a lower bound for the expected number of truly selected variables. Several studies have noticed that averaging over independent observations can reduce the impact of random errors. For example, Fan et al. (2010) propose refitted cross-validation to attenuate false correlations among the random errors and explanatory variables that they call spurious variables. Meinshausen and Buhlmann (2010) introduce stability selection which is a combination of subsampling and model selection algorithms. They get an exact error control bound because the data from subsampling are independent. Similarly, the split-and-conquer approach provides resistance to selection errors caused by spurious correlations and keeps a large amount of variables that are in the true model at the same time.

The rest of this article is organized as follows. Section 2 proposes a split-and-conquer approach and a combined estimator under the generalized linear regression models. Section 3 studies theoretical properties of the combined estimator and explores issues related to computing and error bound controls. Section 4 illustrates the results using simulations and real data from an application of cargo screening in the U.S. Port-of-Entries (POEs) practices. Section 5 provides further discussions.

2 Split-and-conquer for penalized regressions

Suppose $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters that lies in the parameter space Ω and the true parameter, denoted by $\boldsymbol{\beta}^0$, is sparse. Let us divide the whole dataset into K subsets and the k^{th} subset has n_k observations: $(\boldsymbol{x}_{k,i}, y_{k,i})$, $i = 1, \ldots, n_k$. For the k^{th} subset, the log-likelihood function is

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}_k, \boldsymbol{X}_k) = [\boldsymbol{y}_k' \boldsymbol{X}_k \boldsymbol{\beta} - \mathbf{1}' \boldsymbol{b}(\boldsymbol{X}_k \boldsymbol{\beta})]/n_k, \quad k = 1, \ldots, K$$

where $\boldsymbol{y}_k = (y_{k,1}, \ldots, y_{k,n_k})'$ is a $n_k \times 1$ response vector, $\boldsymbol{X}_k = (\boldsymbol{x}_{k,1}', \ldots, \boldsymbol{x}_{k,n_k}')'$ is a $n_k \times p$ matrix. Corresponding to (3), the penalized estimator for the k^{th} subset is:

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

where ρ is the penalty function with tuning parameter λ_k. Suppose $\rho(\boldsymbol{\beta}; \lambda_k)$ is one of the penalty functions studied in Fan and Lv (2011). From Fan and Lv (2011), $\hat{\boldsymbol{\beta}}_k$ has the sparsity property with many zero entries.

Let us denote by $\hat{\mathcal{A}}_k = \{j : \hat{\beta}_{k,j} \neq 0\}$ the set of selected variables for $\hat{\boldsymbol{\beta}}_k$. Also, for any indices set S, denote by $\hat{\boldsymbol{\beta}}_{k,S}$ a $|S| \times 1$ vector that is formed by the elements of $\hat{\boldsymbol{\beta}}_k$ whose indices are in S. Thus, $\hat{\boldsymbol{\beta}}_{k,\hat{\mathcal{A}}_k}$ is the sub-vector that contains only the non-zero elements of $\hat{\boldsymbol{\beta}}_k$. Note that, since each $\hat{\boldsymbol{\beta}}_k$ is estimated from different data, $\hat{\mathcal{A}}_k$ can be different from one to another and the K vectors $\hat{\boldsymbol{\beta}}_{k,\hat{\mathcal{A}}_k}$ may have different lengths.

In order to obtain a combined estimator of $\boldsymbol{\beta}$ from $\hat{\boldsymbol{\beta}}_k$'s that retains good performance on model selection consistency, we use a majority voting method. This majority voting method is applied based on two considerations. First, the combined estimator should be formed based on $\hat{\boldsymbol{\beta}}_k$'s. A variable that is not selected in any of $\hat{\mathcal{A}}_k$ should also be excluded by the combined estimator. On the other hand, $\hat{\mathcal{A}}_k$ are subject to selection errors because only a portion of data is analyzed and the penalized likelihood estimator does not guarantee the perfect selection. In particular, $\hat{\mathcal{A}}_k$ from the analysis of the k^{th} subset may contain variables that are not in the true nonzero set $\mathcal{A} \overset{\text{d}}{=} \{j : \beta_j^0 \neq 0\}$ and

some variables in $\mathcal{A}$ may be missed in $\hat{\mathcal{A}}_k$. In our majority voting method, we define $\hat{\mathcal{A}}^{(c)} \stackrel{\text{d}}{=} \{j : v_j \neq 0\}$ as the set of selected variables of the combined estimator, where $\boldsymbol{v}(w) = (v_1, \ldots, v_p)'$ is an $p \times 1$ indicator such that

$$v_j = \begin{cases} 1 & \sum_{k=1}^{K} \mathbf{I}(\hat{\beta}_{k,j} \neq 0) > w \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

with $w \in [0, K)$ being a prespecified threshold and $\mathbf{I}$ being the indicator function. In one extreme case with $w = K-1$, only variables that are selected by all penalized estimators from the subsets are selected in $\hat{\mathcal{A}}^{(c)}$. In the other extreme case with $w = 0$, variables that are selected by at least one penalized estimator from the subsets are selected in $\hat{\mathcal{A}}^{(c)}$. According to (4), $\hat{\mathcal{A}}^{(c)}$ is a subset of $\bigcup_{k=1}^{K} \hat{\mathcal{A}}_k$. When the numbers of elements $|\hat{\mathcal{A}}_k|, k = 1, \ldots, K$ are small and the sets have lots of common elements, $|\hat{\mathcal{A}}^{(c)}|$ can be much smaller than p.

We introduce the following notations. For any $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, define

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = (\mu(\theta_1), \ldots, \mu(\theta_n))' \text{ and } \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \text{diag}(\sigma(\theta_1), \ldots, \sigma(\theta_n)),$$

where $\mu(\theta) = \partial b(\theta)/\partial\theta$ and $\sigma(\theta) = \partial^2 b(\theta)/\partial^2\theta$. We also define weight matrices

$$\boldsymbol{S}_k \stackrel{\text{d}}{=} \boldsymbol{X}_k' \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k) \boldsymbol{X}_k, \tag{5}$$

where $\hat{\boldsymbol{\theta}}_k = \boldsymbol{X}_k \hat{\boldsymbol{\beta}}_k$. The weight matrix $\boldsymbol{S}_k$ comes from the second order condition of the penalized likelihood function. It is approximately the inverse of the covariance matrix of $\hat{\boldsymbol{\beta}}_k$.

We propose to use the following combined estimator, which is the weighted average of $\hat{\boldsymbol{\beta}}_{k,\hat{\mathcal{A}}^{(c)}}$, $k = 1, \ldots, K$:

$$\hat{\boldsymbol{\beta}}^{(c)} \stackrel{\text{d}}{=} \boldsymbol{A} \left(\sum_{k=1}^{K} \boldsymbol{A}' \boldsymbol{S}_k \boldsymbol{A} \right)^{-1} \sum_{k=1}^{K} \boldsymbol{A}' \boldsymbol{S}_k \boldsymbol{A} \hat{\boldsymbol{\beta}}_{k,\hat{\mathcal{A}}^{(c)}}, \tag{6}$$

where $\boldsymbol{E} = \text{diag}(\boldsymbol{v}(w))$ and $\boldsymbol{A} = \boldsymbol{E}_{\hat{\mathcal{A}}^{(c)}}$. Here, for any set S, $\boldsymbol{E}_S$ stands for an $p \times |S|$ submatrix of $\boldsymbol{E}$ formed by columns whose indices are in S.

The derivation of the combined estimator is similar to that discussed under the framework of combining confidence distributions (c.f., Singh et al. (2005); Xie et al. (2011); Liu (2011)), although the number of parameters considered in those cases is fixed and finite. We will not repeat the detailed derivations in this paper. Instead, we will directly show that the combined estimator presented in (6) is model selection consistent and asymptotically equivalent to the penalized estimator directly using the entire data all together.

3 Theoretical results

In this section, we investigate the asymptotic properties of the combined estimator and compare it with the penalized estimator $\hat{\boldsymbol{\beta}}^{(a)}$ that is obtained from the entire dataset as defined in (3). We also provide nonasymptotic bounds for the expected number of false selected variables and the expected number of truly selected variables through the split-and-conquer approach. We assume in our development that the number of parameters $p = p_n$ can potentially increase with the number of observations n.

3.1 Model selection consistency

We first show that the combined estimator is model selection consistent given that the penalized estimator obtained from each subset is consistent. We show in this subsection that the combined estimator converges under the L_∞ and L_2 norms.

Denote by $\boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_{p_n}^0)$ the true parameter. Also, denote by $\mathcal{A} = \{i : \beta_i^0 \neq 0\}$ the true nonzero set and denote by $\mathcal{B}$ its complement $\mathcal{A}^c$, the set of noise variables. We write the half of the minimal signal as $d_n = 2^{-1} \min\{|\beta_j^0| : \beta_j^0 \neq 0\}$. For any indices set S, $\boldsymbol{X}_S$ stands for an $n \times |S|$ submatrix of $\boldsymbol{X}$ formed by columns with indices in S. Similarly, $\boldsymbol{X}_{k,S}$ stands for an $n_k \times |S|$ submatrix of $\boldsymbol{X}_k$ formed by columns with indices in S.

In order to obtain model selection consistency of the combined estimator, we require certain regularity conditions on the penalty functions and the design matrix. The regular conditions 1-5 listed in Appendix A are adapted from Fan and Lv (2011). Condition 1 requires that the increasing concave penalty function has continuous derivative and Condition 2 provides the upper bound of the tail probability of the response variables. In Conditions 3 and 5, we assume weak correlations between $\boldsymbol{X}_{\mathcal{B}}$ and $\boldsymbol{X}_{\mathcal{A}}$. Here, we require that each subset with sample size n_k satisfies these conditions, which is the only difference from Fan and Lv (2011). In Conditions 4 and 6, we assume that the minimal signal is large enough for detection. Also, p_n can not increase faster than $o(e^{n_k})$ and the model size $s_n = |\mathcal{A}|$ may increase at a rate of $o(n_k)$. Under these conditions, following Fan and Lv (2011), all K penalized estimators obtained from the subsets are model selection consistent and converge under the L_∞ or L_2 norm. The following theorem concerns the consistency of the combined estimator $\hat{\boldsymbol{\beta}}^{(c)}$ under the L_∞ and L_2 norms.

Theorem 1 *Suppose the regularity Conditions 1-2 in Appendix A are satisfied. Assume the dataset is divided into $K = O(n^\delta)$, $0 \leq \delta \leq 1/2$ subsets and $n_k = O(n/K)$.*

(i) If the regularity Conditions 3-4 in Appendix A are satisfied, $s_n = o(n^{1-\delta})$, $\log p_n = O(n^{(1-\delta)(1-2\alpha)})$ *and in addition*

$$\max_{\boldsymbol{\delta}\in\mathcal{N}_0=\{\boldsymbol{\delta}\in\Re^{s_n}:\|\boldsymbol{\delta}-\boldsymbol{\beta}^0_{\mathcal{A}}\|_\infty\le d_n\}} \|[\boldsymbol{X}'_{\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{X}_{\mathcal{A}}\boldsymbol{\delta})\boldsymbol{X}_{\mathcal{A}}]^{-1}\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})\boldsymbol{X}_{k,\mathcal{A}}\|_\infty = O(n_k/n), \quad (7)$$

then we have, with probability approaching 1, $\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{B}} = 0$ *as* $n \to \infty$ *and* $\|\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_\infty = O(n^{-\gamma(1-\delta)}\log n)$*, for some* $0 < \gamma < 1/2$.

(ii) If the regularity Conditions 5-6 in Appendix A are satisfied, $s_n = o(n^{1-\delta})$, $\log p_n = O(n^{(1-\delta)(1-2\alpha)})$ *and in addition*

$$\lambda_{\max}[\{\boldsymbol{X}'_{\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{X}_{\mathcal{A}}\boldsymbol{\delta})\boldsymbol{X}_{\mathcal{A}}\}^{-1}\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})\boldsymbol{X}_{k,\mathcal{A}}] = O(n_k/n), \quad (8)$$

for $\boldsymbol{\delta} \in \mathcal{N}_0 = \{\boldsymbol{\delta} \in \Re^{s_n} : \|\boldsymbol{\delta} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_\infty \le d_n\}$*, then we have, with probability approaching 1,* $\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{B}} = 0$ *as* $n \to \infty$ *and* $\|\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_2 = O(\sqrt{s_n/n^{1-\delta}})$.

The regularity conditions in Theorem 1 are mild, and they include the same class of widely used penalty functions discussed in Fan and Lv (2011). According to Theorem 1, the combined estimator is model selection consistent as long as the penalized estimator for each subset $\hat{\boldsymbol{\beta}}_k$ is model selection consistent. The additional conditions in (7) and (8), beyond those adapted from Fan and Lv (2011), are minor. They essentially require the penalized estimator obtained from the k^{th} subset contributes to the combined estimator proportionally to the subset's sample size n_k. Asymptotically, the same limiting model would be obtained whichever fixed $w \in [0, K)$ is chosen.

Note that the sample size of each subset n_k is smaller than n, the consistency rate under L_∞ loss is changed to $\|\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_\infty = O(n_k^{-\gamma}\log n_k) = O(n^{-\gamma(1-\delta)}\log n)$ rather than $O(n^{-\gamma}\log n)$ in Fan and Lv (2011). Thus, the combined estimator has slower consistency rate than the penalized estimator using all data under L_∞ norm because n_k may go to infinity at a slower rate compared with n, unless $\delta = 0$ or equivalently K is a constant. Similarly, under the L_2 norm, the consistency rate of the combined estimator is $O(\sqrt{s_n/n^{1-\delta}})$ rather than $O(\sqrt{s_n/n})$, unless K is a constant. This is not surprising, since, without any further assumptions, the combined estimator typically converges at the rate of its individual components. In the special case when $\delta = 0$, or K is a constant, the combined estimator converges at the same rate as the penalized estimator directly using entire data all together.

3.2 Oracle property

The usual penalized likelihood estimators possess oracle property, with a better rate of model selection consistency and asymptotic normality, if we strengthen the regularity conditions; see, e.g., Fan and Lv (2011). In this section, we show that our combined estimator also has such an oracle property when the oracle property is attainable for the penalized estimator obtained from each subset.

Before obtaining the asymptotic normality, we first show that, under the L_2 norm, the combined estimator is able to achieve the $\sqrt{s_n/n}$ convergence rate that is the same as the penalized estimator using entire data all together, if we further constrain $s_n = o(n_k^{1/3})$ and $K \leq O(s_n)$. This is a stronger consistency result than that in Theorem 1 (ii). Then, we show that the combined estimator obtains asymptotic normality with the same variance as the penalized estimator using entire data all together. Therefore, we fully establish the asymptotic equivalence between the combined estimator and the penalized estimator using entire data all together. The results are stated in the following theorem.

Theorem 2 *Suppose the regularity Conditions 1-2, 5-6 in Appendix A are satisfied. Assume the dataset is divided into $K = O(n^\delta)$ subsets and $n_k = O(n/K)$.*

(i) Suppose the regularity Condition 7 in Appendix is satisfied, $s_n = O(n^{(1-\delta)/3})$ and $\log p_n = O(n^{(1-\delta)(1-2\alpha)})$. If $0 \leq \delta \leq 1/4$, we have, with probability approaching 1, $\hat{\boldsymbol{\beta}}_{\mathcal{B}}^{(c)} = 0$ as $n \to \infty$ and $\|\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_2 = O(\sqrt{s_n/n})$.

(ii) Suppose the regularity conditions 8-9 in Appendix A are satisfied, $s_n = o(n^{(1-2\delta)/3})$ and $\boldsymbol{D}$ is a $q \times s_n$ matrix such that $\boldsymbol{D}\boldsymbol{D}' \to \boldsymbol{G}$, $\boldsymbol{G}$ is a $q \times q$ symmetric positive definite matrix. If $0 \leq \delta \leq 1/5$, we have

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

Fan and Lv (2011) show that $\boldsymbol{D}[\boldsymbol{X}_{\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\theta}^0)\boldsymbol{X}_{\mathcal{A}}]^{1/2}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(a)} - \boldsymbol{\beta}_{\mathcal{A}}^0) \xrightarrow{\mathrm{D}} N(\boldsymbol{0}, \phi\boldsymbol{G})$. From Theorem 2 (ii), the combined estimator $\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)}$ has the same limiting normal distribution as $\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(a)}$. Thus, the combined estimator is as asymptotically efficient as the penalized estimator $\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(a)}$ which is obtained using the entire data all together. Together with the fact that both estimators are model selection consistent, the combined estimator is asymptotically equivalent to the penalized estimator analyzing the entire data all together.

The conditions required in Theorem 2 are mostly adapted from Theorem 4 of Fan and Lv (2011), but some of them required on the subset level. In addition, we require $\lambda_{\max}[\boldsymbol{X}'_{k,\mathcal{A}}\{\boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0) - \boldsymbol{\Sigma}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})\}\boldsymbol{X}_{k,\mathcal{A}}]$ to be smaller than certain rate. This is because

when calculate $\hat{\boldsymbol{\beta}}^{(c)}$, estimated covariance matrix $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)$ is used rather than the true value $\boldsymbol{\Sigma}(\boldsymbol{\theta}^0)$. To obtain the oracle properties, the estimated weight matrices have to be close to the true ones. This requirement is usually satisfied in practice because $\hat{\boldsymbol{\beta}}_k$ is close to $\boldsymbol{\beta}^0$, which we have already known, and the pug-in estimator $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)$ usually is a good estimator of $\boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0)$. In the simple case of linear regressions $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the identity matrix not dependent on $\hat{\boldsymbol{\beta}}_k$; thus, $\{\boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0) - \boldsymbol{\Sigma}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})\}$ is always 0. Also, we can always choose K that is small enough to accomplish this condition, such as $K = O(1)$. In particular, at least in the special case with $K = O(1)$, the combined estimator achieves asymptotic equivalence under the same conditions of Fan and Lv (2011).

3.3 Computing issues

In this subsection, we discuss potential computing savings through the split-and-conquer approach. We have the following simple proposition for a computational demanding procedure.

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

Proposition 1 provides an intuitive interpretation on how much computing time can be saved. In the numerical example of a relatively simple Gaussian regression with L1 penalty in Section 4.1, the exact order of the computing saving stated in the proposition is achieved using the LARS algorithm. However, under more complex situations with generalized linear models and more complicated algorithms such as those studied in Section 4.2, the computing saving is less than what would be predicted by the simple proposition, although the computing time is still reduced in a great amount in all those examples. The complexity of an algorithm and its computing time are associated with its computing paths in search for a numerical solution of the optimization. Cross-validations used for selecting the tuning parameter in the penalized likelihood function add another degree of complexity to the problem. We speculate that the computing time for analysis of the K subset is different, sometimes substantially, from one to another in these more complex situations. This makes a prediction of computing savings a much harder task.

Although we can not use the simple proposition to calculate computing savings in the more complex cases, it still provides an intuition that can help us understand why the split-and-conquer approach can reduce computing time. In the simple setting of Section 4.1 where the LARS algorithm is used in a Gaussian regression with a L1 penalty, the proposition can in fact be used to calculate the computing savings; see Section 4.1 later. A similar finding in a computational intensive robust multivariate scale estimation is also reported in Section 5.3 of Singh et al. (2005).

3.4 Error control

Since the observations are independent and the splitting is random, the majority voting proposed in our approach enables us to find an upper bound of the expected number of falsely selected variables and a lower bound of the expected number of truly selected variables for the combined estimator. Let $\bar{s}_k = E(|\hat{\mathcal{A}}_k|)$ be the average number of selected variables of the penalized estimator from the k^{th} subset. Theorem 3 below provides an upper bound of the expected number of falsely selected variables and a lower bound of the expected number of truly selected variables, both of which depend on the choice of the threshold w in the proposed majority voting method. A similar result is also provided by Meinshausen and Buhlmann (2010) who only consider the $K = 2$ situation and only give a lower bound of the expected number of false selected variables.

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

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

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

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

In an extreme case with $w = K - 1$, the combined estimator selects a variable when it is selected by all penalized estimators from the K subsets. Then, the upper bound for

the expectation of selected noise variables is $(s^*)^K/p^{K-1}$. Usually, it is hard to get s^*. However, as long as s^* is bounded by $c^{1/K}p^{1-1/K}$, the average number of noise variables is bounded by c, where c is constant. In sparse models, s^* is usually small and so is c. Therefore, the combined estimator controls the model selection error in a foreseeable way. In another extreme case with $w = 0$, the combined estimator selects a variable when it is selected by at least a penalized estimator from the K subsets. In this case, the lower bound for the expected number of truly selected variables is tight, achieving the true number of non-zero set $|\mathcal{A}|$. However, in this latter case, the upper bound for the expected number of false selected variables is very loose, up to $|\mathcal{B}|$ the number of the entire noise set.

Indeed, there is a trade off between the upper and lower bounds in Theorem 3 for the choice of w. A larger w typically gives us a smaller upper bound of the expected number of false selected variable as well as a smaller lower bound of the expected number of truly selected variables. A smaller w typically gives us a larger upper bound of the expected number of false selected variable as well as a larger lower bound of the expected number of truly selected variables. We use $w = K/2$ in our numerical studies in Section 4. It appears to be able to provide a good balance between selecting nonzero coefficients in the true model and excluding noise variables, provided that s^* is smaller than half of p. Our numerical studies show that when $w = K/2$, the combined estimators select very few noise variables while keep most variables in the true model. In fact, when $w \in [K/3, K/2]$, variables that are selected by the split-and-conquer are often the same or very similar. Our empirical experience seems to suggest that model selection results are not very sensitive to the choice of $w \in [K/3, K/2]$.

4 Numerical studies

In this section, we provide 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 data all together, whenever the latter approach can be performed and does not reach the limits of our computer. The L_1 norm, SCAD and MCP, three of the most widely used penalty functions in the literature, are used in our illustration. We focus on two models, the Guassian linear regression model and the logistic model, with different choices of sample size n, number of parameters p and true model size s (the number of nonzero regression parameters). All analyses are performed on a W35653 20GHz,

2G(RAM) workstation using R 2.13.1 under Windows 7.

4.1 Linear regression with L_1 norm penalty

We consider in this subsection a simple case with a linear regression and the L_1 norm penalty to demonstrate the properties of the combined estimator. In particular, the response variable $\boldsymbol{y}$ follows a Guassian linear model

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

where ε are IID $N(0,1)$ errors and the explanatory variables $\boldsymbol{X}$ are generated from a $N(0, \boldsymbol{I})$ distribution with $\boldsymbol{I}$ being identity matrix. In our simulation study, we generate $p = 1000$ variables and the true model $\mathcal{A}^0$ contains $s = 30$ nonzero coefficients. The total sample size is picked to be $n = 1000$ which is equal to p. To get the LASSO estimators using the L_1 norm penalty, the LARS algorithm (Efron et al., 2004) is applied and a 10-fold cross-validation is used for selecting the tuning parameter. When $p \geq n$, the computing order of the LARS is $O(n^3)$ that is computationally intensive.

We repeat our simulation 100 times. For the final overall estimators, we record the mean of computing time and the number of selected nonzero coefficients. To demonstrate the error control property, we also calculate model selection sensitivity and model selection specificity. Here, model selection sensitivity is defined as the number of truly selected variables divided by the true model size, and model selection specificity is defined as the number of truly removed variables divided by the number of noise variables. The simulation results are shown in Table 1. In Table 1, $K = 1$ means the entire dataset is used to get the LASSO estimator; otherwise, the combined estimator proposed in this paper is used.

According to Table 1, all estimators select some noise variables in addition to the true $s = 30$ nonzero variables. This is consistent with a known performance of the LASSO-type estimators that they usually intend to include more variables than desired in model selections. When K gets larger, the combined estimator shows the benefit of error controls through the split-and-conquer approach. In particular, when $K = 4$ and 6, the model selection specificities increase a lot. This indicates that the combined estimator is more efficient in removing noise and spurious variables from the selected models. Moreover, the computing time is decreasing when K is increasing. Since the computing order for the LARS algorithm is $O(n^3)$ when $p \geq n$, Proposition 1 in Section 3.3 suggests that the split-and-conquer approach can save computing time by the order

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

Simulation setting					Model selection		
Design matrix	n	p	K	Computing time (in second)	# selected variables	sensitivity (in %)	specificity (in %)
Independent variables	10000	1000	1	1929.99 (91.46)	151.36 (24.53)	100 (0)	87.49 (2.53)
			2	433.87 (14.34)	210.16 (42.50)	100 (0)	81.43 (4.38)
			4	140.39 (13.39)	97.64 (15.80)	100 (0)	93.03 (1.63)
			6	90.69 (3.23)	47.92(6.85)	100 (0)	98.15 (0.71)

of K^2. This is exactly the order achieved in this simulation study, as indicated in column 5 in the middle of Table 1.

4.2 Generalized linear model with SCAD and MCP penalties

The SCAD and MCP estimators are two commonly used estimators that are obtained based on non-concave penalized likelihood functions. They often have a better performance than the LASSO estimators, in terms of selecting a tighter model and fewer noise variables. We consider in this subsection both the SCAD and MCP estimators under both the linear regression and logistic models.

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

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

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

We consider two settings of sample sizes: $n = 10000$ that is large but not too large and $n = 100000$ that is very large. In the linear regression, the number of parameters $p = 1000$ and in the logistic model, the number of parameters $p = 200$. In all cases, the true model contains $s = 30$ nonzero coefficients. The true model size $s = 30$ is

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

The simulation is repeated 100 times. Similarly as in example 1, we record the computing time and the number of selected variables and calculate model selection sensitivity and specificity. In addition, the MSE (mean squared error) is calculated in the linear regression case and the misclassification rate with 0.5 as threshold is reported in the logistic regression case. The results are displayed in Table 2. In the table, $K = 1$ refers to the entire data is used all together with no splitting. For any $K > 1$, the proposed split-and-conquer approach is applied.

According to Table 2, the SCAD estimators performs similar to the MCP estimators. In either case, the combined estimator has good model selection results with high model selection sensitivity and specificity that are similar to those of the penalized estimator using entire data all together. Moreover, in the linear regression case, the combined estimator has a similar MSE to that of the penalized estimator using entire data all together. In the logistic regression case, the misclassification rate of the combined estimator is also close to that of the penalized estimator using entire data all together.

The computing time is reduced through the split-and-conquer procedure, although we cannot calculate the exact order of computing savings in these complicated settings. For both the SCAD and MCP penalties, the proposed split-and-conquer approach can reduce the computing time by almost 10 times on average in the linear regression setting. For the logistic model, the average saving is a little less. When the explanatory variables are independent, the combined estimator needs about half of the time compared to directly performing the same analysis on the entire data all together. When the explanatory variables are correlated, the combined estimator by the proposed method can save up to 25% time compared to directly performing the same analysis on the entire data all together. When the sample size $n = 100000$, we are not able to perform either the SCAD or the MCP regression on the entire data all together due to computer memory limitations. However, the combined estimators can still be obtained using the split-and-conquer procedure.

We also compare the values of the combined estimators and the penalized estimators analyzing entire data all together in all the settings of Table 2 when both are available; see Figure 1. For the linear regression case, the boxplots of the $\boldsymbol{\beta}$ estimation in the true model $\mathcal{A} = \{j : \beta_j^0 \neq 0\}$ are plotted in the top panels. We can see that the estimation of the combined estimators has the similar mean and spread to those of the

estimators using entire data all together. For the logistic regression, the boxplots of the $\boldsymbol{\beta}$ estimation in the true model are plotted in the bottom panels. In the logistic model case, the estimation of covariance matrix can influence the combined estimator. We use the maximum likelihood estimator based on only the selected variables in $\hat{\mathcal{A}}$ to get the weight matrix. Again, the combined estimators using the proposed split-and conquer approach perform similarly to the penalized estimators using entire data all together.

4.3 Numerical analysis on POEs manifest data

In this section, we study a set of manifest data collected at the US Port of Entries (POEs) to demonstrate an application of the split-and-conquer approach. To counter potential terrorists' threats, substantial efforts have been made in devising strategies for inspecting containers coming through the US POEs every day to interdict illicit nuclear and chemical materials. Manifest data, compiled from the custom forms submitted by merchants or shipping companies, are collected by the US custom offices and the Department of Homeland Security (DHS). Analysis of the manifest data is a part of effort to build up layered defenses for the national security. In a nuclear detection project sponsored by the Command, Control, and Interoperability Center for Advanced Data Analysis (CCICADA), a Department of Homeland Security (DHS) Center of Excellence, we obtain a set of manifest data that contain all shipping records coming through the POEs across the US in February, 2009. The goal is to make quantitative evaluations of the manifest data and to develop an effective risk scoring approach that can be used to assess future shipments. In our project, a logistic regression model has been used to enhance the effectiveness of the real-time inspection system with binary response variable indicating high-risk shipments. Since not all information collected in the manifest data are relevant to risk scoring and there are also many redundant information, we need to determine the effects of different sources of information in the manifest data and penalized regression provides a way to evaluate the importance of these variables. Table 3 provides the definition and a description of some variables contained in the manifest data. Most of these variables are categorical and dummy variables for each categorical variable are created which results in $p = 213$ variables in total. There are also text fields that can potentially lead to a much larger p. To simply our discussion and without loss of our focus, we only illustrate the proposed split-and-conquer approach using this $p = 213$ variables and we do not consider any semantic analysis and text mining approaches in this paper.

Practical issues and challenges exist in carrying out this important task. Due to

Table 2: Comparison of the combined estimator and the complete estimator (standard deviation in the parenthesis)

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

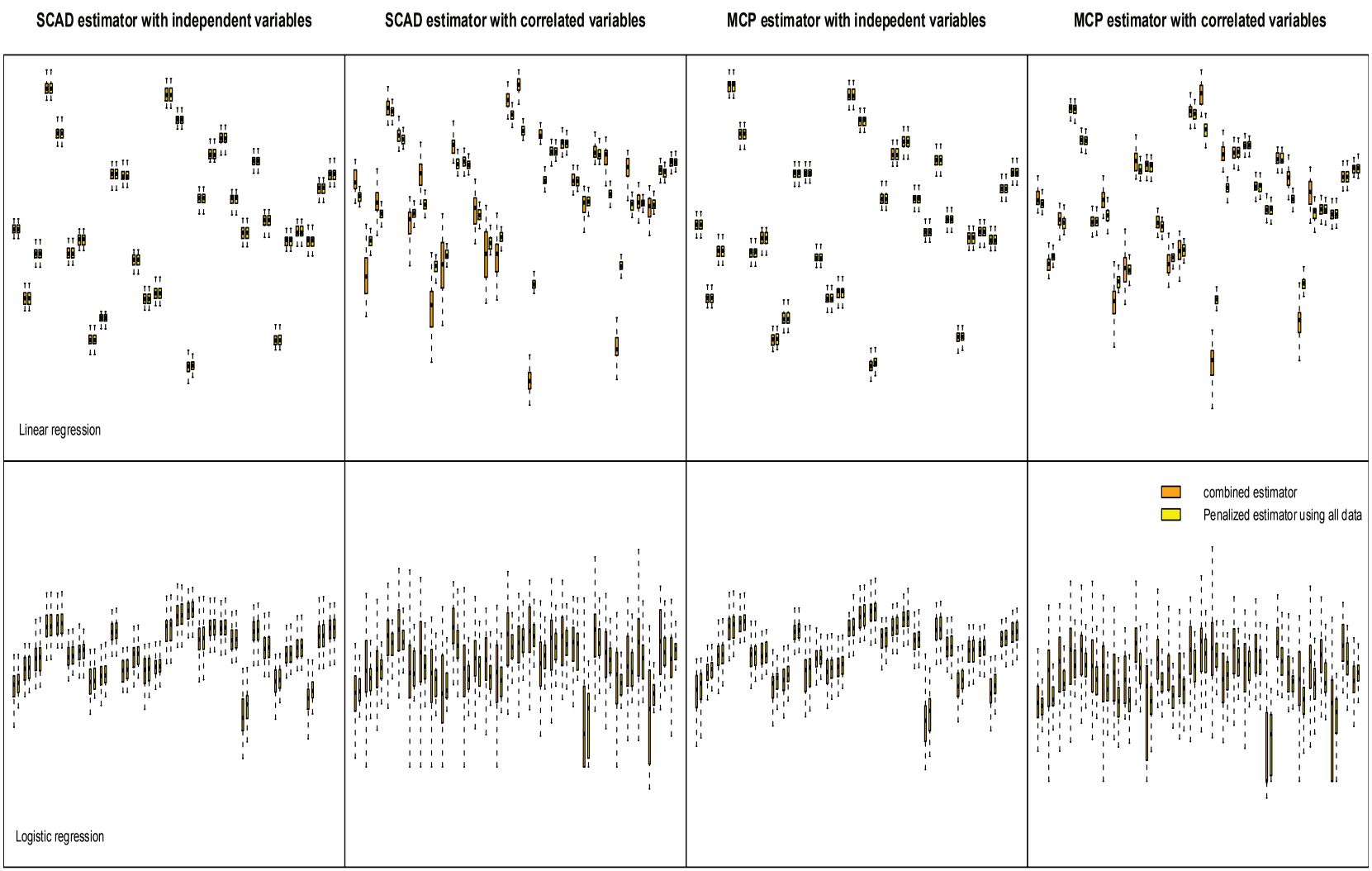

Figure 1: Comparison of parameter estimation for the combined estimator and the penalized estimator using all data. Box plots of estimation for variables in the true model. Orange: the combined estimator; Yellow: the estimator using complete data. Top panels: Linear regression; bottom panels: Logistic regression

Table 3: Manifest data: Dictionary of Variables

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

the enormous size of traffic and a large number of entry sites, it is impossible for us to analyze the whole data simultaneously on a single computer. For instance, there are 164721 shipments in one week from February 20, 2009 to February 26, 2009. A computer with 2 GB memory and 3.2GHz CPU fails to perform the SCAD penalized regression on the one-week data. Even if high-performance computer is available, it will takes a long time to carry out the task and this is very inefficient in practice, especially we may need to constantly update the models over the months and years. Nevertheless, we can solve this problem by applying the split-and-conquer approach with the assumption that the underlying regression model stay more or less the same over a short period of time of one week or one month.

Because of security concerns, the indicator of high-risk shipments are not accessible to us, but we have been told to use the rate 1% to 10% of cargo containers that need further inspections in the context of inspections of drugs and other illicit materials. To illustrate our approach, we turn to a simulation to generate the risk scores based on the given manifest data. In particular, potential influential characteristics are selected to generate the risk scores using logistic models. Then, we perform the SCAD penalized regression on everyday's data and combine the seven daily estimators together to obtain an overall combined estimator. Note that, due to the computing limitations of our personal computer, we are not able to perform the SCAD analysis on the whole week of data all together.

The results from the split-and-conquer approach are displayed in Tables 4 and 5, in which we report the model selection sensitivity, model selection specificity, misclassifi-

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 %)	Misclassificaton 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)

cation rate and 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 $s = 22$ non-zero parameters are from three categorical variables: Vessel Country Code, Foreign Port Landing and Container Contents. Clearly, the split-and-conquer approach succeeds in performing the penalized logistic regression analysis on the whole week manifest data. As we can see from Table 4, the split-and conquer approach has identified most influential variables in the manifest data. In particular, the combined estimator has both high model selection sensitivity and specificity. On a contrast, the daily estimators either selects many more noise variables or excludes many influential variables. Also, the combined estimator is more stable than daily estimators because it has much smaller variances in the values of the average model size, model selection sensitivity and specificity. Although the combined estimator has a slightly smaller misclassification rate, all estimators have more or less the similar misclassification rates, which are on average slightly less than 4%.

In terms of estimation, as in Table 5, the combined estimator also has smaller variance than the penalized estimators that only use daily data. For the categories Animals and Office in Container Contents, some of the daily estimators fails to select them and they are not significant in the combined estimators. Also, the Sporting variable is left out in the model by all the estimators. But all other 19 variables are found by the combined

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)

estimator. The same performance can not be achieved by any of the penalized estimators using only daily data. By incorporating one-week information, the split-and-conquer approach provides more reliable results with better performance than any of the daily analysis.

5 Discussions

We propose in this paper a split-and-conquer methodology for analysis of extraordinarily large data. The split-and-conquer approach contains two operational steps. Firstly, the entire dataset is randomly split into non-overlapped small subsets, and each subset is analyzed separately using desired statistical procedures. Then, the results from all subsets are combined together and provide a final overall statistical inference that contains information from the entire dataset. We demonstrate the split-and-conquer approach for penalized regression models that are widely used in the analysis high-dimensional data.

The split-and-conquer approach provides an applicable way to analyze extraordinarily large datasets. The approach is very general and can have a lot of applications. As the entire dataset is split into smaller pieces, each subset requires a smaller storage space and computer memory when we perform our statistical analysis. Moreover, we have shown that the split-and-conquer approach needs less computing time when the desired statistical method is computationally intensive. Even in the case in which the desired statistical method is efficient, a reduced computing time can be expected operationally because we now can analyze different subsets at the same time using different computers. This computing improvement is very useful in many practical applications.

One important step in the split-and-conquer approach is the combination. We have demonstrated in our settings that the combined results obtained from the subsets do not cause any bias or efficiency loss, asymptotically. The specific combination method to be used depends on the desired statistical procedure. As illustrated by penalized regressions in this paper, the properly weighted and linearly combined estimator is asymptotically equivalent to the one from analyzing the entire data all together. According to Singh et al. (2005), Xie et al. (2011) and Liu (2011), equivalent combined statistics or asymptotic efficiency are achievable for many other models. The proposed split-and-conquer approach can be easily extended to other problem settings as well as problems beyond point estimations including those using hypothesis testings and confidence intervals.

6 Appendix

6.1 Appendix A

Condition 1 $\rho(t;\lambda)$ is increasing and concave in $t \in [0,\infty)$, and has a continuous derivative $\rho'(t;\lambda)$ with $\rho'(0+;\lambda) > 0$. In addition, $\rho'(0+;\lambda)$ is increasing in $\lambda \in [0,\infty)$ and $\rho'(0+;\lambda)$ is independent of λ.

Condition 2 Let $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ be the n-dimensional independent random response vector and $\boldsymbol{a} \in \Re^n$. Assume

$$\mathrm{P}(|\boldsymbol{a}'\boldsymbol{Y} - \boldsymbol{a}'\boldsymbol{\mu}(\boldsymbol{\theta}^0)| > \|\boldsymbol{a}\|_2\epsilon) \le \psi(\epsilon)$$

where $\psi(\epsilon) = 2e^{-c_1\epsilon^2}$ and c_1 is a constant, see Fan and Lv (2011) for details about c_1.

Condition 3 The sub-design matrix $\boldsymbol{X}_k$, $k = 1, \ldots, K$ satisfies

$$\|[\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})\boldsymbol{X}_{k,\mathcal{A}}]^{-1}\|_\infty = O(b_s n_k^{-1}),$$

$$\|\boldsymbol{X}'_{k,\mathcal{B}}\boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}[\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}]^{-1}\|_\infty \le \min\{C\rho'(0+)/\rho'(d_n), O(n_k^{\alpha_1})\},$$

$$\max_{\boldsymbol{\delta}\in\mathcal{N}_0}\max_{j=1,\ldots,p_n} \lambda_{\max}[\boldsymbol{X}'_{k,\mathcal{A}}\mathrm{diag}\{|\boldsymbol{x}_{k,j}| \circ |\boldsymbol{\mu}''(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})|\}\boldsymbol{X}_{k,\mathcal{A}}] = O(n_k),$$

where $C \in (0,1)$, $\alpha_1 \in [0, 1/2]$ and $\mathcal{N}_0 = \{\boldsymbol{\delta} \in \Re^{s_n} : \|\boldsymbol{\delta} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_\infty \le d_n\}$. Here, b_s is associated with the non-sparsity size $s_n = O(n_k^{\alpha_0})$; see Fan and Lv (2011) for a detailed definition. The derivative is taken componentwise.

Condition 4

For some $\gamma \in (0, 1/2]$, assume that $d_n \ge n_k^{-\gamma}\log n_k$ and $b_s = o\{\min(n_k^{1/2-\gamma})\sqrt{\log n_k}, s_n^{-1}n_k^{\gamma}/\log n_k\}$. In addition, assume $\rho'(d_n;\lambda_k) = o(b_s^{-1}n_k^{-\gamma}\log n_k)$, $\lambda_k \gg n_k^{-\alpha}(\log n_k)^2$, and $\lambda_k\kappa_0 = o(\tau_{k0})$, where $\alpha = \min(1/2, 2\gamma - \alpha_0) - \alpha_1$, $\tau_{k0} = \max_{\boldsymbol{\delta}\in\mathcal{N}_0}\lambda_{\min}[n_k^{-1}\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})\boldsymbol{X}_{k,\mathcal{A}}]$. Assume that $\max_{j=1,\ldots,p_n}\|\boldsymbol{x}_{k,j}\|_\infty = o(n_k^{\alpha}/\sqrt{\log n_k})$ if the responses are unbounded.

Condition 5 The sub-design matrix $\boldsymbol{X}_k$, $k = 1, \ldots, K$ satisfies

$$\min_{\boldsymbol{\delta}\in\mathcal{N}_0}\lambda_{\min}[\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})\boldsymbol{X}_{k,\mathcal{A}}] \ge cn_k, \quad \mathrm{tr}[\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}] = O(s_n n_k)$$

$$\|\boldsymbol{X}'_{k,\mathcal{B}}\boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}\|_{2,\infty} = O(n_k)$$

$$\max_{\boldsymbol{\delta}\in\mathcal{N}_0}\max_{j=1,\ldots,p_n} \lambda_{\max}[\boldsymbol{X}'_{k,\mathcal{A}}\mathrm{diag}\{|\boldsymbol{x}_{k,j}| \circ |\boldsymbol{\mu}''(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})|\}\boldsymbol{X}_{k,\mathcal{A}}] = O(n_k)$$

where $\mathcal{N}_0 = \{\boldsymbol{\delta} \in \Re^{s_n} : \|\boldsymbol{\delta} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_\infty \leq d_n\}$, c is some positive constant, and $\|\boldsymbol{B}\|_{2,\infty} = \max_{\|\boldsymbol{v}\|_2=1} \|\boldsymbol{B}\boldsymbol{v}\|_\infty$.

Condition 6 Assume that $d_n \gg \lambda_k \gg \max\{\sqrt{s_n/n_k}, n_k^{(\alpha-1)/2}(\log n_k)^{1/2}\}$, $\rho'(d_n; \lambda_k) = O(n_k^{-1/2}K^{-1/2})$; and $\lambda_k\kappa_0 = o(1)$, where $\kappa_0 = \max_{\boldsymbol{\delta}\in\mathcal{N}_0} \kappa(\rho; \boldsymbol{\delta})$. In addition assume that $\max_{j=1}^{p_n} \|\boldsymbol{x}_{k,j}\|_\infty = o(n_k^{(1-\alpha)/2}/\sqrt{\log n_k})$ if the responses are unbounded.

Condition 7

$$\lambda_{\max}[\boldsymbol{X}'_{k,\mathcal{A}}\{\boldsymbol{\Sigma}(\boldsymbol{\theta}^0_k) - \boldsymbol{\Sigma}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})\}\boldsymbol{X}_{k,\mathcal{A}}] = O(n_k/\sqrt{s_n}) \tag{9}$$

where $\boldsymbol{\delta}$ is in $\mathcal{N}_0 = \{\boldsymbol{\delta} \in \Re^{s_n} : \|\boldsymbol{\delta} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_\infty \leq d_n\}$.

Condition 8

$$\lambda_{\max}[\boldsymbol{X}'_{k,\mathcal{A}}\{\boldsymbol{\Sigma}(\boldsymbol{\theta}^0_k) - \boldsymbol{\Sigma}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})\}\boldsymbol{X}_{k,\mathcal{A}}] = o(n_k/\sqrt{s_nK}) \tag{10}$$

where $\boldsymbol{\delta}$ is in $\mathcal{N}_0 = \{\boldsymbol{\delta} \in \Re^{s_n} : \|\boldsymbol{\delta} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_\infty \leq d_n\}$.

Condition 9 Assume that $\rho'(d_n; \lambda_k) = o(s_n^{-1/2}n^{-1/2})$, $\max_{i=1,\ldots,n} E|y_i - b'(\theta^0_i)|^3 = O(1)$ and $\sum_{i=1}^n (\boldsymbol{z}_i'\boldsymbol{B}^{-1}\boldsymbol{z}_i)^{3/2} \to 0$ as $n \to \infty$, where $\boldsymbol{B} = \boldsymbol{X}'_{\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\theta}^0)\boldsymbol{X}_{\mathcal{A}}$ and $\boldsymbol{X}_{\mathcal{A}} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)'$.

6.2 Appendix B

Proof of Theorem 1 Proof: We first prove part (1). According to Fan and Lv (2011), under the regularity conditions, when $n_k \to \infty$, $k = 1, \ldots, K$, with probability 1, we have $\hat{\boldsymbol{\beta}}_{k,\mathcal{B}} = 0$,$\boldsymbol{S}_k = \text{diag}(\boldsymbol{S}_{k,\mathcal{A}}, 0)$ and $\boldsymbol{A} = \text{diag}(\text{I}_{s_n}, 0)$. Therefore,

$$\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{B}} = 0$$

In addition, the definition of $\hat{\boldsymbol{\beta}}^{(c)}$ shows that

$$\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}} = [\sum_{k=1}^K \boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}]^{-1}[\sum_{k=1}^K \boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}})]$$

.

Since $\|\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_\infty = O(n_k^{-\gamma}\log n_k)$ and (7) in conditions 4, we have

$$\begin{aligned}
\|\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_\infty &= \|[\sum_{k=1}^K \boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}]^{-1}[\sum_{k=1}^K \boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}})]\|_\infty \\
\leq & \sum_{k=1}^K\{\|[\sum_{k=1}^K \boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}]^{-1}\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}\|_\infty\|\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_\infty\} \\
= & \sum_{k=1}^K\{\|[\boldsymbol{X}'_{\mathcal{A}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\boldsymbol{X}_{\mathcal{A}}]^{-1}\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}\|_\infty\|\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_\infty\} \\
= & \sum_{k=1}^K O(n_k/n)O(n_k^{-\gamma}\log n_k) \\
= & O(n_k^{-\gamma}\log n) = O(n^{-\gamma(1-\delta)}\log n)
\end{aligned}$$

The proof of part (2) is similar. Since $\|\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_2 = O(\sqrt{s_n/n_k})$ (Fan and Lv, 2011) and (8) in condition 6, it gives us

$$\begin{aligned}
\|\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_2 \leq & \|\sum_{k=1}^K[\{\sum_{k=1}^K \boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}})]\|_2 \\
\leq & \sum_{k=1}^K \lambda_{\max}[\{\sum_{k=1}^K \boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}]\|\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_2 \\
\leq & \sum_{k=1}^K O(n_k/n)O(\sqrt{s_n/n_k}) = O(\sqrt{s_n/n_k}) = O(\sqrt{s_n/n^{1-\delta}})
\end{aligned}$$

Proof of Theorem 2 We first prove part (1). Constrained on the subspace $\{\boldsymbol{\beta} : \boldsymbol{\beta}_{\mathcal{B}} = 0\}$, we take Taylor expansion of the penalized likelihood function at $\boldsymbol{\beta}^0_{\mathcal{A}}$. Since $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}$ is local maximum and $\|\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_2 = O(\sqrt{s_n/n_k})$,

$$\begin{aligned}
& n_k^{-1}\boldsymbol{X}'_{k,\mathcal{A}}[\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}^0_k)] - n_k^{-1}\boldsymbol{X}'_{k,\mathcal{A}}[\boldsymbol{\Sigma}(\boldsymbol{\theta}^0_k) - \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)]\boldsymbol{X}_{k,\mathcal{A}}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) \qquad (11) \\
- & n_k^{-1}\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) - \bar{\rho}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}};\lambda_k) + O_p(s_n^{3/2}n_k^{-1}) = 0 \qquad (12)
\end{aligned}$$

From (9) in condition 7, we have

$$\|n_k^{-1}\boldsymbol{X}'_{k,\mathcal{A}}[\boldsymbol{\Sigma}(\boldsymbol{\theta}^0_k) - \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)]\boldsymbol{X}_{k,\mathcal{A}}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}})\|_2 \leq O(1/\sqrt{s_n})\|\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_2 = O(1/\sqrt{n_k}) \qquad (13)$$

Since $s_n = O(n^{(1-\delta)/3}) = O(n_k^{1/3})$ and $\rho'(d_n;\lambda_k) = o(s_n^{-1/2}n_k^{-1/2})$, together with (13), (11) gives

$$n_k^{-1}\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) = n_k^{-1}\boldsymbol{X}'_{k,\mathcal{A}}[\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}^0_k)] + O(1/\sqrt{n_k})$$

By the definition of $\hat{\boldsymbol{\beta}}^{(c)}$, we have

$$[\sum_{k=1}^{K} \boldsymbol{X}'_{k,\mathcal{A}} \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k) \boldsymbol{X}_{k,\mathcal{A}}](\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0) = \sum_{k=1}^{K} \boldsymbol{X}'_{k,\mathcal{A}}[\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)] + O(\sqrt{nK})$$

Therefore,

$$\begin{aligned} & [\sum_{k=1}^{K} \boldsymbol{X}'_{k,\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0) \boldsymbol{X}_{k,\mathcal{A}}](\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0) \\ = & \sum_{k=1}^{K} \boldsymbol{X}'_{k,\mathcal{A}}[\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)] + O(\sqrt{nK}) - [\sum_{k=1}^{K} \boldsymbol{X}_{k,\mathcal{A}}\{\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k) - \boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0)\}\boldsymbol{X}_{k,\mathcal{A}}](\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0) \end{aligned}$$

Since $\|\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_2 = O(\sqrt{s_n K/n})$ and (9) in condition 7, we have

$$[\sum_{k=1}^{K} \boldsymbol{X}'_{k,\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0) \boldsymbol{X}_{k,\mathcal{A}}](\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0) = \sum_{k=1}^{K} \boldsymbol{X}'_{k,\mathcal{A}}[\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)] + o(\sqrt{nK})$$

The above equation is equivalent to

$$\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0 = [\boldsymbol{X}'_{\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}}]^{-1} \boldsymbol{X}_{\mathcal{A}}[\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}^0)] + o(\sqrt{K/n})$$

Since $K = O(n^{\delta}), 0 \leq \delta \leq 1/4$ and $s_n = O(n_k^{1/3})$, we have $K \leq O(s_n)$ and

$$\begin{aligned} \|\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_2 &= \|[\boldsymbol{X}'_{\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}}]^{-1} \boldsymbol{X}_{\mathcal{A}}[\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}^0)]\|_2 + o(\sqrt{K/n}) \\ &= O(\sqrt{s_n/n}) + o(\sqrt{K/n}) = O(\sqrt{s_n/n}) \end{aligned}$$

Proof of (2). From (10) in condition 8, we have

$$\|n_k^{-1} \boldsymbol{X}'_{k,\mathcal{A}}[\boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0) - \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)] \boldsymbol{X}_{k,\mathcal{A}}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0)\|_2 \leq o(1/\sqrt{s_n K}) \|\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_2 = o(1/\sqrt{K n_k}) \quad (14)$$

Since $s_n = o(n_k^{1/3}/K^{1/3})$ and $\rho'(d_n; \lambda_k) = o(s_n^{-1/2} n_k^{-1/2} K^{-1/2})$, together with (14), (11) gives

$$n_k^{-1} \boldsymbol{X}'_{k,\mathcal{A}} \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k) \boldsymbol{X}_{k,\mathcal{A}}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0) = n_k^{-1} \boldsymbol{X}'_{k,\mathcal{A}}[\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)] + o(1/\sqrt{n_k K})$$

By the definition of $\hat{\boldsymbol{\beta}}^{(c)}$, we have

$$[\sum_{k=1}^{K} \boldsymbol{X}'_{k,\mathcal{A}} \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k) \boldsymbol{X}_{k,\mathcal{A}}](\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0) = \sum_{k=1}^{K} \boldsymbol{X}'_{k,\mathcal{A}}[\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)] + o(\sqrt{n})$$

Therefore,

$$
\begin{aligned}
& [\sum_{k=1}^{K} \boldsymbol{X}'_{k,\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0) \boldsymbol{X}_{k,\mathcal{A}}](\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0) \\
= & \sum_{k=1}^{K} \boldsymbol{X}'_{k,\mathcal{A}} [\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)] + o(\sqrt{n}) - [\sum_{k=1}^{K} \boldsymbol{X}_{k,\mathcal{A}} \{\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k) - \boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0)\} \boldsymbol{X}_{k,\mathcal{A}}](\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0)
\end{aligned}
$$

Since $\|\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_2 = O(\sqrt{s_n/n})$ and (9) in condition 8, we have

$$
[\sum_{k=1}^{K} \boldsymbol{X}'_{k,\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0) \boldsymbol{X}_{k,\mathcal{A}}](\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0) = \sum_{k=1}^{K} \boldsymbol{X}'_{k,\mathcal{A}} [\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)] + o(\sqrt{n})
$$

The above equation is equivalent to

$$
[\boldsymbol{X}'_{\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}}](\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0) = \boldsymbol{X}_{\mathcal{A}} [\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}^0)] + o(\sqrt{n})
$$

Thus,

$$
\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0 = [\boldsymbol{X}'_{\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}}]^{-1} \boldsymbol{X}_{\mathcal{A}} [\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}^0)] + o(1/\sqrt{n})
$$

In addition,

$$
\boldsymbol{D}[\boldsymbol{X}_{\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}}]^{1/2} (\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0) = \boldsymbol{D}[\boldsymbol{X}_{\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}}]^{-1/2} \boldsymbol{X}_{\mathcal{A}} [\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}^0)] + o(1)
$$

and under Condition 9, we have

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

This complete the proof.

Proof of Theorem 3

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

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

Using the exchangeability assumption, $\mathrm{P}(j \in \hat{\mathcal{A}}_k) = E(|\mathcal{B} \cap \hat{\mathcal{A}}_k|)/|\mathcal{B}|$, $j \in \mathcal{B}$ and $\mathrm{P}(j \in \hat{\mathcal{A}}_k) = E(|\mathcal{A} \cap \hat{\mathcal{A}}_k|)/|\mathcal{A}|$, $j \in \mathcal{A}$. Therefore, $\mathrm{P}(j \in \hat{\mathcal{A}}_k) \leq \bar{s}_k/p \leq s^*/p$, $j \in \mathcal{B}$ 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 $w \geq s^*K/p - 1$, $\mathrm{P}(j \in \hat{\mathcal{A}}) \leq 1 - F(w|K, s^*/p), j \in \mathcal{B}$ and $\mathrm{P}(j \in \hat{\mathcal{A}}) \geq 1 - F(w|K, s_*/p), j \in \mathcal{A}$. Therefore, $E(|\mathcal{B} \cap \hat{\mathcal{A}}|) = \sum_{j \in \mathcal{B}} \mathrm{P}(j \in \hat{\mathcal{A}}) \leq |\mathcal{B}|\{1 - F(w|K, s^*/p)\}$ and $E(|\mathcal{A} \cap \hat{\mathcal{A}}|) = \sum_{j \in \mathcal{A}} \mathrm{P}(j \in \hat{\mathcal{A}}) \geq |\mathcal{A}|(1 - F(w|K, s_*/p))$

Proof of Proposition 1: Each sub-sample has n/K observations, so the computing steps for the combined estimator is $O(K \cdot (n/K)^a) = O(n^\alpha / K^{a-1})$. The result follows immediately.

References

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

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

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least angle regression," *Annals of statistics*, 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," *Journal of the American Statistical Association*, 96, 1348–1360.

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

Frank, I. and Friedman, J. (1993), "A statistical view of some chemometrics regression tools," *Technometrics*, 35, 109–135.

Liu, D. (2011), "Combination of Confidence Distributions and an Efficient Approach for Meta-Analysis of Heterogeneous Studies," *Ph.D thesis*, Department of Statistics and Biostatistics, Rutgers University.

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

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

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

Xie, M., Singh, K., and Strawderman, W. (2011), "Confidence Distributions and a Unifying Framework for Meta-analysis," *Journal of the American Statistical Association*, 106, 320–333.

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