Statistica Sinica Preprint No: SS-2018-0379	
Title	Partitioned Approach for High-dimensional Confidence Intervals with Large Split Sizes
Manuscript ID	SS-2018-0379
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0379
Complete List of Authors	Zemin Zheng, Jiarui Zhang, Yang Li and Yaohua Wu
Corresponding Author	Jiarui Zhang
E-mail	zjrt46@mail.ustc.edu.cn
Notice: Accepted version subject to English editing.	

Partitioned Approach for High-dimensional Confidence Intervals with Large Split Sizes

Zemin Zheng, Jiarui Zhang, Yang Li, and Yaohua Wu

University of Science and Technology of China

Abstract: With the availability of massive data sets, making accurate inference with low computational cost is the key to improving scalability. When both the sample size and the dimensionality are large, naively applying the de-biasing idea to derive confidence intervals can be computationally inefficient or infeasible as the de-biasing procedure increases the computational cost by an order of magnitude compared with the initial penalized estimation. Therefore, we suggest a split and conquer approach to ameliorate the scalability in the de-biasing procedure and show that the length of the established confidence interval is asymptotically the same as that using the data all at once. Moreover, a significant improvement in the largest split size is demonstrated by separating the initial estimation and the relaxed projection steps, which reveals that the sample sizes needed for these two steps with statistical guarantees are different. Last but not least, a refined inference procedure is proposed to address the inflation issue in finite sample performances when the split size indeed gets large. Both computational advantage and theoretical guarantee of our new methodology are evidenced by numerical studies.

Key words and phrases: Big data, Confidence intervals; Scalability, De-biased estimator, Divide and conquer, Large split sizes.

1. Introduction

Due to the highly developed technologies and devices, it is easier than ever to generate large-scale data sets in many areas including meteorology, genomics, and economics, which are referred to as big data problems (Fan et al., 2014). High-dimensional sparse modeling, which explores the situation where the number of variables can be larger than the sample size, has also become an intensely popular area in the research of statistics over the past decade. When both the scales of the sample size and the dimensionality are large, naively applying the existing high-dimensional inference methods to extract knowledge from large amounts of data can be computationally inefficient or even infeasible. Thus, it is appealing to develop scalable methodologies to make advantage of the huge data sets for accurate inference with low computational cost.

As a solid tool to produce meaningful and interpretable models, sparse modeling via regularization has shown its strengths in handling the high-dimensional data sets. See, for example, Tibshirani (1996); Fan and Li (2001); Zou and Hastie (2005); Zou (2006); Candes and Tao (2007); Liu and Wu (2007); Zou and Li (2008); Bickel et al. (2009); Fan et al. (2009); Lv and Fan (2009); Zhang (2010); Sun and Zhang (2012); Fan and Lv (2013); Zheng, et al. (2014); Song and Liang (2015); Kong et al. (2016); Weng et al. (2017); Hao et al. (2018), among many others. Based on the sparse regularized estimators, statistical inferences such as hypothesis testing and confidence intervals can be made when asymptotic distributions of the pilot estimators are derived. See, for instance, Lockhart et al. (2014) and Lee et al. (2016) for inference with model selection via the Lasso (Tibshirani, 1996), and Javanmard and Montanari (2014), van de Geer et al. (2014), and Zhang and Zhang (2014) for inference through de-biasing the penalized estimators. We will focus on ameliorating the scalability of the latter type of inference methods in the big data settings since the de-biasing procedure increases the computational cost by an order of magnitude close to the dimensionality compared with the initial penalized estimation, which yields the computational bottleneck for large-scale applications.

A natural and efficient way to deal with big data problems is through data splitting, that is, splitting the entire data set into subsamples and then aggregating the estimators obtained from each subsample. In fact, this divide and conquer idea has been widely used to solve various kinds of problems (Fan et al., 2012; Decrouez and Hall, 2014; Kleiner et al., 2014; Mackey et al., 2015; Shang and Cheng, 2015; Zhang et al., 2015; Xu et al., 2016; Zhao et al., 2016; Shang and Cheng, 2017; Lian and Fan, 2018), where extra benefits such as robustness and enhanced stability besides the computational advantage were demonstrated. For high-dimensional regression models, divide and conquer methods also play important roles in the analysis of extraordinarily large data sets. For instance, Chen and Xie (2014) developed a split and conquer approach for penalized estimation, which was proved to retain desired statistical properties and be more resistant to false model selections. A one-shot approach to distributed sparse regression that averages the de-biased Lasso estimators was devised in Lee et al. (2017) and shown to converge at the same rate as the Lasso. Battey et al. (2018) proposed divide and conquer Wald and score statistics for hypothesis testing based on the de-biasing procedures in Javanmard and Montanari (2014) and van de Geer et al. (2014), respectively.

Despite the fast growing literature, how to construct confidence intervals in presence of large-scale high-dimensional data remains largely unexplored. Generally speaking, deriving confidence intervals is not that flexible as hypothesis testing since test statistics may not be inverted to pilot estimators. Even if we utilize the divide and conquer algorithm, how to preserve asymptotically equivalent statistical accuracy and efficiency to the full sample procedure is unclear. Moreover, both theoretical results and empirical performances in existing high-dimensional split and conquer inference methods do not allow for large split sizes. It would be interesting to study whether the largest split size with a statistical guarantee can be improved both theoretically and empirically to enhance the scalability in big data applications. In this paper, we intend to provide some answers to the aforementioned questions by introducing a new methodology of scalable inference with partitioned data for deriving high-dimensional confidence intervals. It randomly splits the whole data set into subsamples of equal size and generates debiased estimator from each subsample, then constructing confidence intervals based on the bagging estimator that aggregates all estimators from different subsamples.

The main contributions of this paper are fourfold. First of all, a new partitioned approach is developed to substantially increase the computing speed for deriving confidence intervals in high dimensions. We prove that the length of the established confidence interval is asymptotically the same as that using the data all at once, which means that the information loss due to the divide and conquer procedure is negligible. Second, a significant improvement in the upper bound on the split sizes, which becomes a square of that in Battey et al. (2018), is demonstrated by separating the initial estimation and the relaxed projection steps. It reveals that the sample sizes needed for these two steps to enjoy statistical guarantee are indeed different. Last but not least, a refined inference procedure is proposed to address the inflation issue in finite sample performances when the split size indeed gets large. Numerical studies show that the suggested methodology is communication efficient and can be more robust and resistent to heavy-tailedness and outliers.

The rest of the paper is organized as follows. Section 2 presents the new methodology of scalable inference with partitioned data in the big data settings. We provide confidence intervals based on the partitioned approach as well as the desired statistical guarantees in Section 3. The computational advantage and theoretical properties are backed up empirically through simulation studies in Section 4 and real data analyses in Section 5, respectively. Section 6 concludes with extensions and potential future research. All technical details are relegated to the Supplementary Material.

2. Scalable inference with partitioned data

We illustrate scalable inference with partitioned data through high-dimensional linear regression models. Denote by $\mathbf{y} = (y_1, \cdots, y_n)^T$ the n-dimensional response vector and $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_p)$ an $n \times p$ random design matrix with p covariates. Assume that the rows of $\mathbf{X}$ are independent and normally distributed with mean zero and covariance matrix $\mathbf{\Sigma}$, that is, $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{\Sigma})$. Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is a p-dimensional regression coefficient vector and $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^T \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$ is an n-dimensional error vector independent of $\mathbf{X}$. The true regression coefficient vector $\boldsymbol{\beta}$ is assumed to satisfy the following capped L_1 sparsity condition

$$\sum_{j=1}^{p} \min\{|\beta_j|/(\sigma\lambda_{\text{univ}}), 1\} \leq s \tag{2.2}$$

with $\lambda_{\text{univ}} = \sqrt{(2/n)\log p}$. This condition is weaker than the direct sparsity assumption $\|\boldsymbol{\beta}\|_0 \leq s$ as it allows for a large number of nonzero coefficients as long as their magnitudes are small. We focus on the big data settings where both the sample size n and the number of covariates p diverge, satisfying $\log(p) = o(n)$ and n can be extremely large.

1. Low dimensional projection estimator. As mentioned in Section 1, there are several de-biasing based inference methods for constructing confidence intervals in high dimensions. In this paper, we adopt the low dimensional projection estimator (LDPE) proposed in Zhang and Zhang (2014) to illustrate our partitioned approach due to the appealing property that LDPE does not require the minimum signal strength condition.

First of all, we need an initial penalized estimator $\widehat{\boldsymbol{\beta}}^{(\text{init})} = (\widehat{\beta}_1^{(\text{init})}, \cdots, \widehat{\beta}_p^{(\text{init})})^T$, which can be generated from the scaled Lasso (Sun and Zhang, 2012) given by

$$\{\widehat{\boldsymbol{\beta}}^{(\text{init})}, \widehat{\sigma}\} = \mathop{\arg\min}_{\mathbf{b}\in\mathbb{R}^p,\, \sigma>0} \left\{ \frac{\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2}{2\sigma n} + \frac{\sigma}{2} + \lambda_0\|\mathbf{b}\|_1 \right\},$$

where λ_0 is a universal regularization parameter independent of the noise level. As a self-bias correction from the initial estimator, the LDPE $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \cdots, \widehat{\beta}_p)^T$ is then defined through each coordinate as

$$\widehat{\beta}_j = \widehat{\beta}_j^{(\text{init})} + \mathbf{z}_j^T(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(\text{init})})/\mathbf{z}_j^T\mathbf{x}_j \tag{2.3}$$

for $1 \le j \le p$, where $\mathbf{z}_j^T(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(\text{init})})/\mathbf{z}_j^T\mathbf{x}_j$ is the de-biasing term and $\mathbf{z}_j$ is a relaxed orthogonalization of $\mathbf{x}_j$ against other covariate vectors. For simple tuning, $\mathbf{z}_j$ can be obtained as the residual of the scaled Lasso regression for $\mathbf{x}_j$ against $\mathbf{X}_{-j} = (\mathbf{x}_k : k \neq j)$ with the weighted L_1-penalty. That is,

$$\mathbf{z}_j = \mathbf{x}_j - \mathbf{X}_{-j}\widehat{\boldsymbol{\gamma}}_j, \tag{2.4}$$
$$\{\widehat{\boldsymbol{\gamma}}_j, \widehat{\sigma}_j\} = \operatorname*{arg\,min}_{\mathbf{b}\in\mathbb{R}^{p-1},\sigma\in\mathbb{R}^+} \left\{ \frac{\|\mathbf{x}_j - \mathbf{X}_{-j}\mathbf{b}\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_0 \sum_{k\neq j} \frac{\|\mathbf{x}_k\|_2}{\sqrt{n}} |b_k| \right\},$$

where the vector $\mathbf{b}$ is indexed by $\{k : 1 \le k \le p, k \neq j\}$.

The LDPE was proved to enjoy asymptotic normal distributions. To gain insight into this, we can look at the following decomposition

$$\widehat{\beta}_j - \beta_j = \frac{\mathbf{z}_j^T\boldsymbol{\varepsilon}}{\mathbf{z}_j^T\mathbf{x}_j} + \frac{\sum_{k\neq j} \mathbf{z}_j^T\mathbf{x}_k(\beta_k - \widehat{\beta}_k^{(\text{init})})}{\mathbf{z}_j^T\mathbf{x}_j},$$

where the first term is normally distributed and the second term shows the approximation error. Denote by

$$\tau_j = \|\mathbf{z}_j\|_2/|\mathbf{z}_j^T\mathbf{x}_j| \text{ and } \eta_j = \max_{k\neq j} |\mathbf{z}_j^T\mathbf{x}_k|/\|\mathbf{z}_j\|_2. \tag{2.5}$$

Then τ_j is the noise factor relative to the standard deviation of the asymptotic distribution and η_j is the bias factor controlling the approximation error by

$$|\sum_{k\neq j} \mathbf{z}_j^T\mathbf{x}_k(\beta_k - \widehat{\beta}_k^{(\text{init})})| \le (\max_{k\neq j} |\mathbf{z}_j^T\mathbf{x}_k|)\|\widehat{\boldsymbol{\beta}}^{(\text{init})} - \boldsymbol{\beta}\|_1 = \eta_j\|\mathbf{z}_j\|_2\|\widehat{\boldsymbol{\beta}}^{(\text{init})} - \boldsymbol{\beta}\|_1.$$

It shows that the roles of the initial estimator $\widehat{\boldsymbol{\beta}}^{(\text{init})}$ and the relaxed projection vectors $\mathbf{z}_j$ are relatively independent, which motivates us to separate the initial estimation and relaxed projection steps in the partitioned approach.

2. Scalable inference with partitioned data. For bias correction based high dimensional inference methods such as LDPE, the computational bottleneck comes from the de-biasing step instead of the initial estimation since the initial Lasso type estimator is a linear programming problem that can be efficiently solved and implemented by packages such as '*lars*' and '*glmnet*', while the de-biasing step generally requires intensive computing. For instance, the construction of all relaxed projection vectors $\mathbf{z}_j$ requires p times of Lasso type solutions in LDPE, which takes the majority of computational cost in high dimensions.

Furthermore, since larger sample size provides better accuracy in controlling the approximation error and the role of the initial estimator is relatively independent of the projection vectors, our new methodology will focus on improving the speed of calculating the relaxed projection vectors through data splitting and utilize the initial estimator generated by the full sample. The extra benefit of this strategy on the largest possible split size subject to a statistical guarantee will be demonstrated in Theorem 1. In cases where the initial estimator is infeasible due to extraordinarily large sample sizes or different locations, we suggest the split and conquer approach in Chen and Xie (2014) for generating regularized initial estimators.

Our new methodology of scalable inference with partitioned data begins with s-plitting the entire data set into subsamples of equal size, then generating de-biased estimator in each subsample with the same initial estimator, and finally all de-biased estimators from different subsamples will be aggregated through the mean in each co-ordinate. Specifically, we divide the whole data set of size n into K groups of size

$\widetilde{n} = n/K$, and generate relaxed projection vectors $\mathbf{z}_j^{(l)}$ from the corresponding predictors $\mathbf{x}_j^{(l)}$ in the lth subsample for $1 \leq l \leq K$. Then we obtain de-biased estimator $\widehat{\boldsymbol{\beta}}^{(l)} = (\widehat{\beta}_1^{(l)}, \cdots, \widehat{\beta}_p^{(l)})$ from each subsample by applying the bias correction idea of LDPE to the initial estimator $\widehat{\boldsymbol{\beta}}^{(\mathrm{init})}$ using vectors $\mathbf{z}_j^{(l)}$ and $\mathbf{x}_j^{(l)}$. The mean bagging estimator $\widehat{\boldsymbol{\beta}}^{(\mathrm{mean})} = K^{-1}\sum_{l=1}^{K}\widehat{\boldsymbol{\beta}}^{(l)}$ finally averages the de-biased estimators over all subsamples. That is,

$$\widehat{\beta}_j^{(\mathrm{mean})} = \widehat{\beta}_j^{(\mathrm{init})} + K^{-1}\sum_{l=1}^{K}(\mathbf{z}_j^{(l)})^T(\mathbf{y}^{(l)} - \mathbf{X}^{(l)}\widehat{\boldsymbol{\beta}}^{(\mathrm{init})})/(\mathbf{z}_j^{(l)})^T\mathbf{x}_j^{(l)}.$$

We will derive confidence intervals based on this mean bagging estimator in Section 3 by breaking the communication barriers between different subsamples.

From a practical point of view, the proposed partitioned approach can significantly reduce the computational cost. As discussed in Chen and Xie (2014), the popularly used LARS algorithm (Efron et al., 2004) for generating the Lasso solution takes computing steps of $O(n^2p + n^3)$, which is around $O(n^3)$ when the sample size n is at least of the same order as p. So the computational cost of LDPE which needs about p times of Lasso solutions will be around $O(n^3p)$. Through our partitioned approach, the computational cost will be $K \cdot O(\widetilde{n}^3p) = O(K^{-2}n^3p)$, reduced by a factor of K^{-2} compared with LDPE using the entire sample. In fact, for any algorithm of de-biased estimator that requires $O(n^ap^b)$ computing steps with some constants $a > 1$ (nonlinear in n) and $b > 0$, the partitioned approach can improve the computing speed by K^{a-1} times in the same device and K^a times if K devices are employed simultaneously, when the computational cost of the bagging procedure is negligible.

3. Comparison with existing works. There are several methods utilizing the split and conquer framework in high-dimensional regression models including Chen and Xie (2014), Lee et al. (2017), and Battey et al. (2018), which are closely related to our work. In Chen and Xie (2014), a divide and conquer approach was proposed for penalized estimation of the regression coefficients under extraordinarily large data, where the combined estimator was shown to be asymptotically equivalent to the estimator analyzing the entire data all at once and more robust in variable selection. Similar asymptotic efficiency and robustness in deriving confidence intervals will be demonstrated for our partitioned approach through theories and numerical analyses. The other two works, Lee et al. (2017) and Battey et al. (2018), are indeed more related to ours since both of them utilized de-biased estimators in each subsample. As Lee et al. (2017) mainly focused on estimation accuracy in distributed sparse regression, we compare our work with Battey et al. (2018) which considered hypothesis testing via split and conquer approaches as follows.

In Battey et al. (2018), a divide and conquer Wald statistic that aggregates the Wald statistics from different subsamples through the mean was proposed for hypothesis testing in high dimensions, where its asymptotic inferential efficiency was proved by showing that the mean bagging estimator has the same statistical error as the full sample de-biased estimator. In contrast, we will establish confidence intervals through the partitioned approach and show the equivalence in asymptotic efficiency by proving that the lengths of confidence intervals are the same. This is a more concrete result as the length of confidence interval takes both bias and variance into consideration. Moreover, by separating the initial estimation and the relaxed projection steps, we will demonstrate in Theorem 1 that the theoretical upper bound on the split size will be $K = o(ns^{-2}\log^{-2} p)$, which is a significant improvement compared with the largest split size $K = o(n^{1/2}s^{-1}\log^{-1} p)$ in Battey et al. (2018). It implies that the sample sizes needed for initial estimation and relaxed projection to enjoy statistical guarantees are indeed different.

3. Theoretical properties

As our partitioned approach focuses on speeding the de-biasing step, we impose the same assumption on the initial estimator as that in Zhang and Zhang (2014).

Condition 1. Assume that the initial estimator $\{\widehat{\boldsymbol{\beta}}^{(\mathrm{init})}, \widehat{\sigma}\}$ satisfies that

$$P\Big\{\|\widehat{\boldsymbol{\beta}}^{(\mathrm{init})} - \boldsymbol{\beta}\|_1 \geq C_1 s\sigma^* \sqrt{(2/n)\log(p/\epsilon)}\Big\} \leq \epsilon,$$
$$P\Big\{|\widehat{\sigma}/\sigma^* - 1| \geq C_2 s(2/n)\log(p/\epsilon)\Big\} \leq \epsilon$$

for some positive constants C_1, C_2 and any ϵ satisfying $\alpha_0/p^2 \leq \epsilon \leq 1$, where $\alpha_0 \in (0,1)$ is a preassigned constant and $\sigma^* = \|\boldsymbol{\varepsilon}\|_2/\sqrt{n}$ is the oracle estimate of the noise stand deviation σ.

Condition 1 characterizes the estimation accuracy of the initial estimator, which has been shown to hold for various regularized estimators including the scaled Lasso under both fixed and random design settings with mild regularity conditions. In situations where the data sets are located in different areas, we can use the initial estimator based on the divide and conquer approach proposed in Chen and Xie (2014), which was shown to achieve similar asymptotic estimation accuracy.

In the fixed design setting, the confidence intervals of LDPE were already provided in Zhang and Zhang (2014). Here we focus on the random design case with $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{\Sigma})$ to facilitate the analysis of some key quantities such as $\|\mathbf{z}_j\|_2$, τ_j, and η_j in a probabilistic sense, and derive the confidence intervals based on the LDPE using the full sample of size n.

Proposition 1. *Suppose that* $s = o(\sqrt{n}/\log p)$ *and* $\lambda_0 = (1+\varepsilon)\sqrt{2\delta\log(p)/n}$ *for some* $\delta \geq 1$ *and* $\varepsilon > 0$ *in (2.4). Assume in addition that the eigenvalues of* $\mathbf{\Sigma}$ *are bounded within the interval* $[M_*, M^*]$ *for some positive constants* M_* *and* M^**, and the rows of* $\mathbf{\Sigma}^{-1}$ *are sparse with* $\max_i \sum_{j=1}^{p} I\{\mathbf{\Sigma}_{ij}^{-1} \neq 0\} \leq \sqrt{s}$*, where* $I\{\cdot\}$ *is the indicator function. Then for sufficiently large* n*, there exist positive constants* c_j*,* $\widetilde{c}_j$*, and* C_j *such that*

$$\widetilde{c}_j n^{-1/2} \leq \tau_j \leq c_j n^{-1/2}, \ \eta_j \leq C_j\sqrt{\log(p)}, \tag{3.1}$$

and $\lim_{n\to\infty} \tau_j n^{1/2} = \mathbf{\Sigma}_{jj}^{-1/2}$ *hold simultaneously with probability at least* $1 - o(p^{-\delta+1})$.

Furthermore, if Condition 1 holds with $C_1 C_j s\sqrt{(2/n)\log(p)\log(p/\epsilon)} \leq \epsilon_n'$, $C_2 s(2/n)\log(p/\epsilon) \leq \epsilon_n''$*, and* $\max(\epsilon_n', \epsilon_n'') \to 0$ *as* $n \to \infty$*, then for sufficiently large* n*, the LDPE in (2.3) satisfies*

$$P(|\widehat{\beta}_j - \beta_j| \geq \tau_j\widehat{\sigma}t) \leq 2\Phi_{n-1}\big\{[-(1-\epsilon_n'')t + \epsilon_n'] \cdot \sqrt{1-n^{-1}}\big\} + 2\epsilon + o(p^{-\delta+1})$$

for any $t \geq (1+\epsilon_n')/(1-\epsilon_n'')$*, where* $\Phi_n(t)$ *is the Student t-distribution function with* n *degrees of freedom. By setting* $n \to \infty$*, we get*

$$\lim_{n\to\infty} P\big\{|\widehat{\beta}_j - \beta_j| \leq \tau_j\widehat{\sigma}\Phi^{-1}(1-\alpha/2)\big\} = 1 - \alpha, \tag{3.2}$$

where $\Phi(t)$ *is the normal distribution function.*

Both the conditions and confidence intervals in Proposition 1 are very similar to that of Zhang and Zhang (2014) under fix design settings, but we also provide quantitative analysis for the bias and noise factors η_j and τ_j from a probabilistic point of view. Based on the established confidence intervals (3.2), the noise factor τ_j is a key quantity for determining the statistical accuracy (the length of confidence interval for a preassigned α), which is of the order $n^{-1/2}$ in view of (3.1), denoted as $\tau_j \asymp n^{-1/2}$. We will compare the statistical accuracy of confidence intervals achieved by the partitioned approach with that of using the the entire sample given in Proposition 1.

Denote by $\tau_j^{(l)}$ and $\eta_j^{(l)}$ the noise and bias factors of the lth subsample, $\widetilde{\tau}_j = \max_{1\le l\le K} \tau_j^{(l)}$, and $\widetilde{\eta}_j = \max_{1\le l\le K} \eta_j^{(l)}$. The following theorem provides the confidence intervals based on the bagging estimator $\widehat{\boldsymbol{\beta}}^{(\mathrm{mean})}$ which takes the mean of $\widehat{\boldsymbol{\beta}}^{(l)}$ through each coordinate in the proposed partitioned approach with subsample size $\widetilde{n} = n/K$.

Theorem 1. *Suppose that $s = o(\sqrt{\widetilde{n}}/\log p)$, $\lambda_0 = (1+\varepsilon)\sqrt{2\delta\log(p)/\widetilde{n}}$ for some $\delta > 1$ and $\varepsilon > 0$, and both the initial estimator and $\boldsymbol{\Sigma}$ satisfy the same conditions as in Proposition 1. Then the following statements hold.*

*(**A**) (Asymptotic efficiency). For any $t \ge (1+\sqrt{K})\epsilon_n'/(1-\epsilon_n'')$, with sufficiently large n, the bagging estimator $\widehat{\boldsymbol{\beta}}^{(\mathrm{mean})}$ satisfies*

$$P(\sqrt{K}|\widehat{\beta}_j^{(\mathrm{mean})} - \beta_j| \ge \widetilde{\tau}_j\widehat{\sigma}t) \le 2\Phi_{n-1}[-(1-\epsilon_n'')t + \sqrt{K}\epsilon_n'] + 2\epsilon + o(Kp^{-\delta+1}),$$

where $\widetilde{\tau}_j \asymp \widetilde{n}^{-1/2}$ with probability at least $1 - o(Kp^{-\delta+1})$. Furthermore, if $\sqrt{K}\epsilon_n' \to 0$, we have

$$\lim_{n\to\infty} P\big\{|\widehat{\beta}_j^{(\mathrm{mean})} - \beta_j| \le K^{-1/2}\widetilde{\tau}_j\widehat{\sigma}\Phi^{-1}(1-\alpha/2)\big\} = 1-\alpha. \tag{3.3}$$

(**B**) (Refined inference). Denote by $\omega_j^{(l)} = \widetilde{\tau}_j^{-1}\tau_j^{(l)}$ and $K_j = [\sum_{l=1}^K (\omega_j^{(l)})^2]^{-1}K^2$. We have $K_j \in [K, c_j^* K]$ holds with probability at least $1 - o(Kp^{-\delta+1})$ for some constant $c_j^* \geq 1$. Then for any $t \geq (1 + \sqrt{K_j}\epsilon_n')/(1-\epsilon_n'')$, with sufficiently large n, the bagging estimator $\widehat{\boldsymbol{\beta}}^{(\text{mean})}$ satisfies

$$P(\sqrt{K_j}|\widehat{\beta}_j^{(\text{mean})} - \beta_j| \geq \widetilde{\tau}_j\widehat{\sigma}t) \leq 2\Phi_{n-1}[-(1-\epsilon_n'')t + \sqrt{K_j}\epsilon_n'] + 2\epsilon + o(Kp^{-\delta+1}).$$

Moreover, if $\sqrt{K_j}\epsilon_n' \to 0$, we get

$$\lim_{n\to\infty} P\big\{|\widehat{\beta}_j^{(\text{mean})} - \beta_j| \leq K_j^{-1/2}\widetilde{\tau}_j\widehat{\sigma}\Phi^{-1}(1-\alpha/2)\big\} = 1-\alpha. \tag{3.4}$$

Theorem 1 establishes the confidence intervals based on the mean bagging estimator of the suggested partitioned approach by breaking the communication barriers between different subsamples. In view of confidence intervals (3.3), the statistical accuracy of our partitioned approach is asymptotically equivalent to that of using the sample all at once since $K^{-1/2}\widetilde{\tau}_j \asymp K^{-1/2}\widetilde{n}^{-1/2} = n^{-1/2} \asymp \tau_j$ by Theorem 1 and Proposition 1 and both the limits of $K^{-1/2}\widetilde{\tau}_j n^{1/2}$ and $\tau_j n^{1/2}$ are $\boldsymbol{\Sigma}_{jj}^{-1/2}$. It means that the lengths of confidence intervals are the same in the asymptotic sense given a preassigned level α. For finite samples, the tail probability is inflated from $o(p^{-\delta+1})$ to $o(Kp^{-\delta+1})$, but the partitioned approach saves the computational cost by about K^2 times as discussed before.

Furthermore, Theorem 1 provides the theoretical upper bound on the largest split size with $K = o(ns^{-2}\log^{-2} p)$ in view of the constraints that the validity of Theorem 1 relies on $\sqrt{K}\epsilon_n' = o(1)$ with $\epsilon_n' \geq C_1 C_j s\sqrt{(2/n)\log(p)\log(p/\epsilon)}$. Compared with the theoretical largest split size $K = o(n^{1/2}s^{-1}\log^{-1} p)$ in Battey et al. (2018) for statistical inference, our new approach allows for much larger split sizes by separating the initial estimation and the relaxed projection steps. It also implies that the sample size needed in the relaxed projection procedure with a statistical guarantee is smaller than that needed in the initial estimation. In fact, the upper bound $K = o(n^{1/2}s^{-1}\log^{-1}p)$ of the number of partitions in Battey et al. (2018) is the sharp one for valid inference with each subsample, in the sense of the minimax error bound for the initial Lasso estimator (Raskutti et al., 2011). The reason why we can improve the bound of K is that a different partitioned inference strategy is used, where the de-biasing procedure is implemented through different subsamples to improve the computational efficiency while the initial estimator is computed based on the full data set. If some other penalty function beyond the Lasso is adopted to reduce the bias of the initial estimator, the theoretical upper bound of the number of partitions may be further improved.

Although the theoretical upper bound allows for large split sizes, the confidence intervals (3.3) generally yield higher coverage probabilities than the preassigned level in finite sample performance when the split size indeed gets large, which will be demonstrated in Section 4, meaning insufficient statistical accuracy as the lengths of confidence intervals are longer than expected. This issue mainly results from the inflation of the overall noise factor $\widetilde{\tau}_j = \max_{1\leq l\leq K}\tau_j^{(l)}$, whose magnitude can be larger than $\widetilde{n}^{-1/2}$ when K is large due to the randomness in different subsamples. Then the order of $K^{-1/2}\widetilde{\tau}_j$ will deviate from $n^{-1/2}$, leading to overestimation of the variance.

To address this inflation issue and take full advantage of the theoretical large split sizes with statistical accuracy, we propose a refined inference procedure in Part (**B**) of Theorem 1 that takes the noise factor $\tau_j^{(l)}$ of every subsample into consideration to adjust the length of confidence intervals. The corresponding noise factor $K_j^{-1/2}\widetilde{\tau}_j = K^{-1/2}\sqrt{\sum_{l=1}^{K}(\tau_j^{(l)})^2/K} \asymp K^{-1/2}\widetilde{n}^{-1/2} = n^{-1/2}$ in the refined confidence intervals (3.4) is more accurate under finite samples since it takes the average of the noise factors instead of the maximum of them. As K_j and K only differ by a constant, the asymptotic efficiency and upper bound on the split size for Part (**A**) also apply to Part (**B**). We will show that this refined procedure maintains statistical accuracy even under very large split sizes, adapting it to large-scale application with massive datasets.

Based on the asymptotic normality of the bagging estimator established in Theorem 1, we immediately have the following simultaneous confidence intervals for multiple coefficients β_j.

Theorem 2. *For any subset $S \subset \{j : 1 \leq j \leq p\}$ with finite number of elements $|S|$, under the assumptions of Theorem 1, we have the following statements.*

*(**A**) If in addition Condition 1 holds with $\max_{j\in S} C_j C_1 s\sqrt{(2/n)\log(p)\log(p/\epsilon)} \leq \epsilon_n'$, then for any $t \geq (1+\sqrt{K}\epsilon_n')/(1-\epsilon_n'')$, the bagging estimator $\widehat{\boldsymbol{\beta}}^{(\mathrm{mean})}$ satisfies*

$$P(\max_{j\in S}\sqrt{K}|\widehat{\beta}_j^{(\mathrm{mean})} - \beta_j|/\widetilde{\tau}_j \geq \widehat{\sigma}t) \leq |S|\cdot 2\Phi_{n-1}[-(1-\epsilon_n'')t + \sqrt{K}\epsilon_n'] + 2\epsilon + o(Kp^{-\delta+1}).$$

*(**B**) If in addition Condition 1 holds with $\max_{j\in S}\sqrt{c_j^*}C_j C_1 s\sqrt{(2/n)\log(p)\log(p/\epsilon)} \leq \epsilon_n'$, then for any $t \geq (1+\max_{j\in S}\sqrt{K_j}\epsilon_n')/(1-\epsilon_n'')$, we have*

$$P(\max_{j\in S}\sqrt{K_j}|\widehat{\beta}_j^{(\mathrm{mean})} - \beta_j|/\widetilde{\tau}_j \geq \widehat{\sigma}t) \leq \sum_{j\in S} 2\Phi_{n-1}[-(1-\epsilon_n'')t + \sqrt{K_j}\epsilon_n'] + 2\epsilon + o(Kp^{-\delta+1}).$$

Theorem 2 provides the simultaneous confidence intervals corresponding to the two parts of Theorem 1 through Bonferroni adjustments, which mainly works for a finite number of coefficients. For statistical inference on a large number of coefficients, we then suggest a bootstrap-assisted procedure similar to that in Zhang and Cheng (2017) based on the mean bagging estimator. It facilitates simultaneous inference under the split and conquer framework for an arbitrary subset $G \subseteq \{1, 2, \cdots, p\}$, where $|G|$ is allowed to grow as fast as p.

The bootstrap-assisted procedure starts with generating a sequence of random variables $\{e_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} N(0,1)$. Then the multiplier bootstrap statistic is defined as

$$W_G = \max_{j \in G} \left| \frac{\sqrt{n}}{K} \sum_{l=1}^{K} \sum_{i=1}^{\widetilde{n}} \frac{z_{i,j}^{(l)} \widehat{\sigma} e_i^{(l)}}{(\mathbf{z}_j^{(l)})^T \mathbf{x}_j^{(l)}} \right|,$$

where $e^{(l)}$ is the corresponding lth subsample of $\{e_i\}_{i=1}^n$ and $z_{i,j}^{(l)}$ is the ith entry of $\mathbf{z}_j^{(l)}$. When there is no splitting, that is, $K = 1$, the above statistic reduces to the one introduced in Zhang and Cheng (2017). The bootstrap critical value is given by $c_G(\alpha) = \inf\{t \in \mathbb{R} : P\left(W_G \le t | (\mathbf{y}, \mathbf{X})\right) \ge 1 - \alpha\}$. We have the following theorem guaranteeing the validity of the proposed procedure.

Theorem 3. *Under the same conditions as those in Theorem 1 and suppose that* $s^2(\log(p))^3/\widetilde{n} = o(1)$, $s(\log(p\widetilde{n}))^3(\log(\widetilde{n}))^2/\widetilde{n} = o(1)$, *and* $(\log(pn))^7/n \le C_3 n^{-c_3}$ *hold for some positive constants* C_3 *and* c_3*. Then we have for any* $G \subseteq \{1, 2, \cdots, p\}$,

$$\sup_{\alpha \in (0,1)} \left| P\left(\max_{j \in G} \sqrt{n} \left| \widehat{\beta}_j^{(\text{mean})} - \beta_j \right| > c_G(\alpha) \right) - \alpha \right| = o(1).$$

Theorem 3 establishes the theoretical guarantee of constructing simultaneous confidence intervals via the proposed bootstrap-assisted procedure, which explicitly accounts for the effect of $|G|$ in view of the dependence of $c_G(\alpha)$ on the set G. The additional dimensionality constraints are imposed to control the estimation errors, which are very similar to those in Zhang and Cheng (2017). The statistical accuracy also remains the same in the asymptotic sense since $K^{-1/2}\widetilde{\tau}_j$ are around the same magnitudes as τ_j.

Last but not least, similar to Zhang and Zhang (2014), the de-biased mean bagging estimator can also be used for variable selection and estimation of the entire regression coefficient vector after a simple soft thresholding, that is,

$$\widehat{\beta}_j^{(\mathrm{t})} = \mathrm{sgn}(\widehat{\beta}_j^{(\mathrm{mean})})(|\widehat{\beta}_j^{(\mathrm{mean})}| - \widehat{t}_j)^+$$

with the selected model

$$\widehat{S}^{(\mathrm{t})} = \{j : |\widehat{\beta}_j^{(\mathrm{mean})}| > \widehat{t}_j\}$$

for some thresholds $\widehat{t}_j$. Then we have a parallel theorem guaranteeing variable selection and estimation properties listed below, which shows that the soft thresholded mean bagging estimator enjoys the same asymptotic efficiency in variable selection and estimation as that in Zhang and Zhang (2014).

Theorem 4. *Let $L_0 = \Phi^{-1}(1-\alpha/(2p))$, $\widetilde{t}_j = K^{-1/2}\widetilde{\tau}_j\sigma L_0$, and $\widehat{t}_j = (1+c_n)\,K^{-1/2}\widehat{\sigma}\widetilde{\tau}_j L_0$ with positive constants α and c_n. Assume that Condition 1 holds with $\max_{j\le p}\widetilde{\eta}_j C_1 s/\sqrt{n}$ $\le \epsilon_n'$ and*

$$P\left\{\frac{(\widehat{\sigma}/\sigma)\vee(\sigma/\widehat{\sigma}) - 1 + \epsilon_n'\sigma^*/(\widehat{\sigma}\wedge\sigma)}{1-(\widehat{\sigma}/\sigma-1)_+} > c_n\right\} \le 2\epsilon.$$

Let $\widehat{\boldsymbol{\beta}}^{(\mathrm{t})} = \left(\widehat{\beta}_1^{(\mathrm{t})},\ldots,\widehat{\beta}_p^{(\mathrm{t})}\right)^T$ be the soft thresholded mean bagging estimator with these $\widehat{t}_j$. Then for any given $\mathbf{X}$, there exist an event Ω_n with $P\{\Omega_n\} \ge 1-3\epsilon$ such that

$$E\left\|\widehat{\boldsymbol{\beta}}^{(\mathrm{t})} - \boldsymbol{\beta}\right\|_2^2 I_{\Omega_n} \le \sum_{j=1}^{p}\min\left\{\beta_j^2, \frac{1}{K^2}\sum_{l=1}^{K}(\tau_j^{(l)})^2\sigma^2\left(L_0^2\left(1+2c_n\right)^2+1\right)\right\} + \frac{\epsilon L_n\sigma^2}{pK}\sum_{j=1}^{p}\widetilde{\tau}_j^2,$$

where $L_n = 4/L_0^3 + 4c_n/L_0 + 12c_n^2 L_0$. *Furthermore, with at least probability* $1-\alpha-3\epsilon$,

$$\left\{j : |\beta_j| > (2+2c_n)\,\widetilde{t}_j\right\} \subseteq \widehat{S}^{(\mathrm{t})} \subseteq \left\{j : \beta_j \neq 0\right\}.$$

4. Simulation studies

In this section, we investigate the finite sample performances of the two versions of scalable inference with partitioned data (denoted by IPAD and R-IPAD) listed in Theorem 1 with different split sizes K, and compare it with LDPE using the entire data set. We adopt similar high-dimensional settings as that in Zhang and Zhang (2014), where $(n, p) = (600, 1000)$ in the first example and $(n, p) = (2000, 3000)$ of higher dimensionality in the second example. Each simulation experiment consists of 100 replications. In every replication, we generate the data set $(\mathbf{X}, \mathbf{y})$ from linear regression model (2.1), where the rows of $\mathbf{X}$ are i.i.d. $N(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = (\rho^{|j-k|})_{p\times p}$ for $\rho = 0.2$, and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}_n)$.

4.1 Simulation example 1

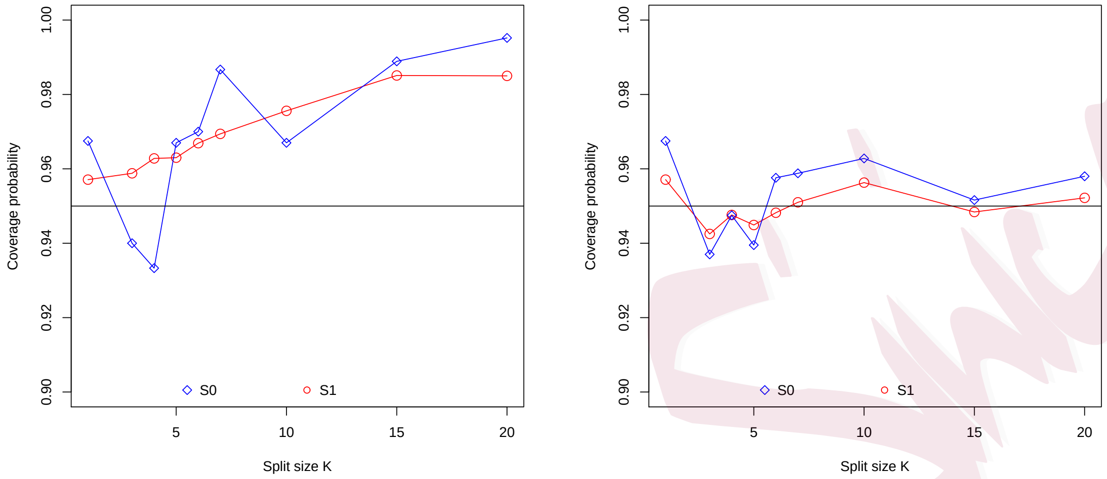
In this simulation study, the true regression coefficient vector $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ satisfies $\beta_j = 3\lambda_{\text{univ}}$ with $\lambda_{\text{univ}} = \sqrt{(2/n)\log p}$ for $j = 200, 400, 600, 800, 1000$ and $\beta_j = 3\lambda_{\text{univ}}/j$ for all the other j. It is a mixture of strong and weak signals without any zero coefficient, originally designed in Zhang and Zhang (2014). By setting $\alpha = 0.05$, we aimed at achieving confidence intervals with 95% coverage probability for each coefficient. For convenience, denote by $S_0 = \{\beta_j : \beta_j = 3\lambda_{\text{univ}}, j = 1, 2, \cdots, p\}$ and $S_1 = \{\beta_j : \beta_j = 3\lambda_{\text{univ}}/j, j = 1, 2, \cdots, p\}$ the sets of strong and weak signals, respectively. Table 1 and

Figure 1 summarize the average coverage probabilities for the coefficients in S_0, S_1, and all coefficients by different methods, where $K = 1$ corresponds to LDPE without partitioning the data.

In view of the results, the coverage probabilities of IPAD match well with the pre-assigned level when the split size K is relatively small (under 5) and start approaching one when K gets larger, meaning that the length of confidence interval is longer than needed so that the statistical accuracy decreases. This can be seen directly from the average lengths of confidence intervals for a typical strong signal β_{200} and a weak one β_{201} in Table 2. It is in accordance with our theoretical analysis before, that is, when K gets large, the randomness in K groups inflates the overall noise factor $\widetilde{\tau}_j$, which is the maximum of the individual noise factor $\tau_j^{(l)}$ over K groups.

Table 1: Average coverage probabilities by different methods and split sizes over 100 replications in Section 4.1 with $(n, p) = (600, 1000)$

Method	(LDPE)				IPAD				
	$K = 1$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$	$K = 10$	$K = 15$	$K = 20$
β_j (S_0)	0.9675	0.9400	0.9333	0.9672	0.9700	0.9867	0.9670	0.9889	0.9952
β_j (S_1)	0.9570	0.9589	0.9630	0.9629	0.9667	0.9692	0.9756	0.9851	0.9849
All β_j	0.9571	0.9588	0.9628	0.9630	0.9669	0.9694	0.9756	0.9851	0.9850
Method					R-IPAD				
		$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$	$K = 10$	$K = 15$	$K = 20$
β_j (S_0)		0.9370	0.9475	0.9395	0.9576	0.9588	0.9628	0.9516	0.9530
β_j (S_1)		0.9426	0.9476	0.9450	0.9481	0.9510	0.9563	0.9484	0.9523
All β_j		0.9425	0.9476	0.9451	0.9482	0.9510	0.9563	0.9484	0.9524



(a) IPAD

(b) R-IPAD

Figure 1: Average coverage probabilities by different methods and split sizes over 100 replications in Section 4.1 with $(n, p) = (600, 1000)$

On the other hand, this inflation issue has been solved by the refined inference R-IPAD in view of the corresponding results in Tables 1, 2 and Figure 1 (b) as the coverage probability keeps around the preassigned 95 percent and the length of confidence interval maintains at the same level. It is worth pointing out that even if the split size K is as large as 20 such that each subgroup contains only 30 samples, R-IPAD still works well in terms of both coverage probability and statistical accuracy. It makes the proposed partitioned approach scalable for analyzing massive data sets with large splits. Of course, the split size should not keep increasing since we need sufficient sample size in each subgroup to provide relatively accurate estimate.

Furthermore, the computational cost has been significantly reduced after partition-

Table 2: Average lengths of confidence intervals for β_{200} and β_{201} by different methods and split sizes over 100 replications in Section 4.1 with $(n, p) = (600, 1000)$

IPAD	$K = 1$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$	$K = 10$	$K = 15$	$K = 20$
β_{200}	0.1733	0.1806	0.1821	0.1885	0.1935	0.1928	0.2020	0.2206	0.2280
β_{201}	0.1793	0.1825	0.1869	0.1940	0.2003	0.2010	0.2064	0.2155	0.2243
R-IPAD		$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$	$K = 10$	$K = 15$	$K = 20$
β_{200}		0.1735	0.1728	0.1744	0.1747	0.1753	0.1746	0.1782	0.1795
β_{201}		0.1799	0.1787	0.1796	0.1804	0.1789	0.1773	0.1770	0.1765

Table 3: Average system running times over 100 replications in Section 4.1 with $(n, p) = (600, 1000)$

Split size	$K = 1$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$	$K = 10$	$K = 15$	$K = 20$
Time (mins)	114.0	41.1	24.0	13.5	12.1	11.1	8.8	6.3	4.0

ing the data in view of Table 3 and Figure 3 (a), where the average system running time for each replication takes about 2 hours by utilizing all samples at once, and dramatically decreases after splitting the data, ending up with only 4 minutes when the split size equals to 20. This statistical analysis was conducted by a usual PC with Intel Core i7-7700 CPU (3.60 GHz) and 8 GB RAM, and parallel computing was employed in the computation of the relaxed projection residual vectors $\mathbf{z}_j$ and the 100 simulation replications, where computation tasks were divided into 7 cores using R package 'snowfall'. It confirms our aforementioned computational advantage by a fair comparison on the single computing device. The computational advantage can be further enhanced if we use multiple PCs to analyze different subsamples.

Table 4: Coverage probabilities, average lengths of confidence intervals, and average system running times for simultaneous confidence intervals over 100 replications in Section 4.1 with $(n, p) = (600, 1000)$

Probability	$K=1$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=10$	$K=15$	$K=20$
β_j (S_0)	0.91	0.97	0.92	0.94	0.95	0.98	0.94	0.93	0.92
All β_j	0.94	0.96	0.95	0.91	0.93	0.93	0.94	0.97	0.98
Length	$K=1$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=10$	$K=15$	$K=20$
β_j (S_0)	0.2227	0.2241	0.2254	0.2300	0.2306	0.2286	0.2298	0.2273	0.2252
All β_j	0.3542	0.3567	0.3552	0.3560	0.3565	0.3558	0.3579	0.3567	0.3561
Time	$K=1$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=10$	$K=15$	$K=20$
mins	126.6	50.5	35.5	28.9	23.7	19.1	14.3	10.1	7.2

Last but not least, we present the coverage probabilities, average lengths of confidence intervals, and average system running times for simultaneous confidence intervals of coefficients β_j in S_0 and all β_j via the proposed bootstrap-assisted procedure with the preassigned 95% coverage probability over 100 replications in Table 4. In view of the results, the coverage probabilities keep around 0.95 over different split sizes and the average lengths of confidence intervals are also very stable, demonstrating the validity of the proposed method. Moreover, significant improvement in running times can also be seen as the split size gets larger.

4.2 Simulation example 2

In this example, we increase both the dimensionality and sample size to $p = 3000$ and $n = 2000$ such that a usual PC is difficult to implement LDPE without partitioning the data due to the huge computational cost. So our statistical analysis starts with split size $K = 5$ until $K = 40$. The true regression coefficient vector $\boldsymbol{\beta}$ takes the strong signals of the same magnitude as that in the first example for $j = 500, 1000, 1500, 2000, 2500, 3000$, and adopts the same pattern for the weak signals. The sets S_0 and S_1 are defined similarly as before. Table 5 and Figure 2 summarize the results of average coverage probabilities by different methods and split sizes, and a similar conclusion as that in Section 4.1 can be drawn. The IPAD method works well when the split size K is no larger than 10 and begins to lose statistical accuracy after K gets larger, due to the inflation issue. But R-IPAD maintains nice performances under different split sizes in view of the corresponding results in Tables 5, 6 and Figure 2. Even if the split size $K = 40$ with each subgroup containing only 50 samples, the coverage probability matches well with the preassigned level by a stable length of confidence intervals. At the same time, significant improvement in computing speed can be seen from Table 7 and figure 3 (b). The performance of simultaneous confidence intervals is also similar to that in the first example in view of the results in Table 8.

Table 5: Average coverage probabilities by different methods and split sizes over 100 replications in Section 4.2 with $(n, p) = (2000, 3000)$

IPAD	$K = 5$	$K = 8$	$K = 10$	$K = 13$	$K = 16$	$K = 20$	$K = 25$	$K = 40$
β_j (S_0)	0.9571	0.9403	0.9425	0.9601	0.9708	0.9679	0.9906	0.9952
β_j (S_1)	0.9545	0.9624	0.9631	0.9637	0.9708	0.9732	0.9766	0.9870
All β_j	0.9545	0.9624	0.9630	0.9637	0.9708	0.9732	0.9766	0.9869
R-IPAD	$K = 5$	$K = 8$	$K = 10$	$K = 13$	$K = 16$	$K = 20$	$K = 25$	$K = 40$
β_j (S_0)	0.9400	0.9364	0.9495	0.9460	0.9401	0.9514	0.9537	0.9525
β_j (S_1)	0.9487	0.9508	0.9529	0.9476	0.9514	0.9498	0.9504	0.9480
All β_j	0.9487	0.9508	0.9529	0.9476	0.9514	0.9498	0.9503	0.9480

Table 6: Average lengths of confidence intervals for β_{500} and β_{501} by different methods and split sizes over 100 replications in Section 4.2 with $(n, p) = (2000, 3000)$

IPAD	$K = 5$	$K = 8$	$K = 10$	$K = 13$	$K = 16$	$K = 20$	$K = 25$	$K = 40$
β_{500}	0.0938	0.0951	0.0969	0.0975	0.0989	0.0959	0.0978	0.1039
β_{501}	0.0921	0.0939	0.0958	0.0972	0.0994	0.1015	0.1046	0.1067
R-IPAD	$K = 5$	$K = 8$	$K = 10$	$K = 13$	$K = 16$	$K = 20$	$K = 25$	$K = 40$
β_{500}	0.0915	0.0916	0.0919	0.0921	0.0923	0.0922	0.0929	0.0945
β_{501}	0.0890	0.0889	0.0888	0.0893	0.0898	0.0901	0.0903	0.0900

Both two simulation examples illustrate the statistical accuracy and computational

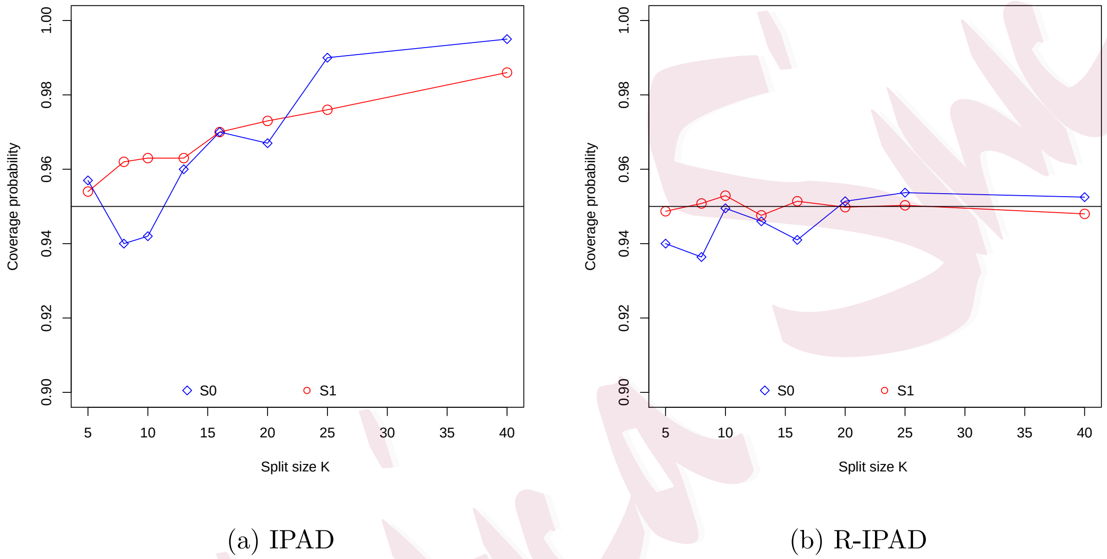
Table 7: Average system running times over 100 replications in Section 4.2 with $(n, p) = (2000, 3000)$

Split size	$K = 5$	$K = 8$	$K = 10$	$K = 13$	$K = 16$	$K = 20$	$K = 25$	$K = 40$
Time (hours)	26.4	16.8	13.3	9.9	8.3	6.9	6.1	4.9

Table 8: Coverage probabilities, average lengths of confidence intervals, and average system running times for simultaneous confidence intervals over 100 replications in Section 4.2 with $(n, p) = (2000, 3000)$

Probability	$K = 15$	$K = 20$	$K = 25$	$K = 30$	$K = 40$
β_j (S_0)	0.93	0.97	0.92	0.93	0.94
All β_j	0.96	0.96	0.95	0.95	0.97
Length	$K = 15$	$K = 20$	$K = 25$	$K = 30$	$K = 40$
β_j (S_0)	0.1254	0.1271	0.1270	0.1266	0.1276
All β_j	0.2026	0.2038	0.2047	0.2053	0.2067
Time	$K = 15$	$K = 20$	$K = 25$	$K = 30$	$K = 40$
hours	10.2	8.7	7.7	7.0	5.7

advantage in constructing confidence intervals by R-IPAD, even under very large split sizes. We will focus on this refined inference procedure in the following analysis of real data sets.



(a) IPAD

(b) R-IPAD

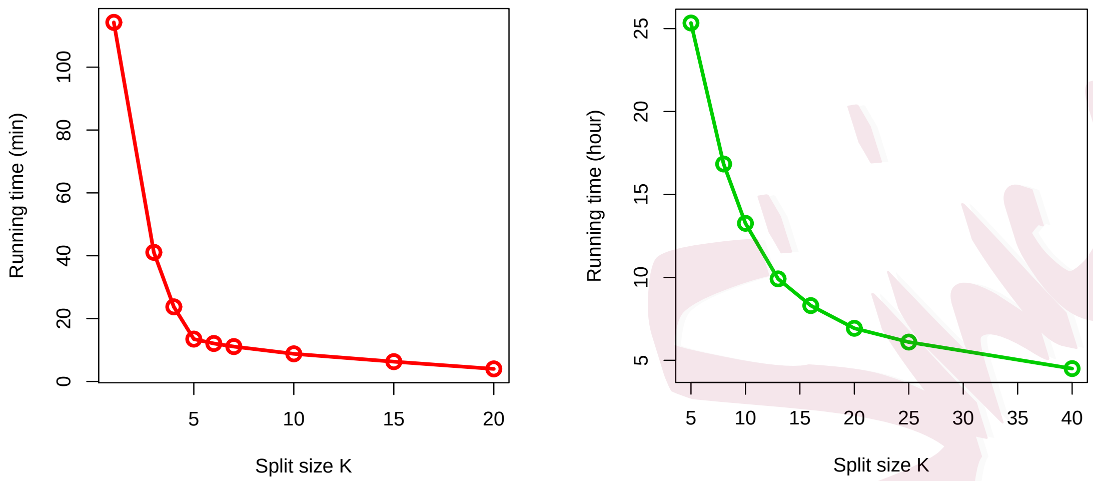
Figure 2: Average coverage probabilities by different methods and split sizes over 100 replications in Section 4.2 with $(n, p) = (2000, 3000)$

5. Real-data analyses

In this section, we will apply both LDPE and R-IPAD to two real data sets: the student performance data set and the polymerase chain reaction (PCR) data set.

5.1 Application to the student performance data

This data set was studied in Cortez and Silva (2008) to evaluate the students' performance in two Portuguese public schools, available at the UCI Machine Learning Repos-



(a) $n = 600, p = 1000$

(b) $n = 2000, p = 3000$

Figure 3: Average system running time over 100 replications under different settings in Section 4

itory (`https://archive.ics.uci.edu/ml/datasets/Student+Performance`). It consists of 32 predictive variables including studying times, first and second period grades, activities, health conditions and so on, obtained from 395 students through school reports and questionnaires. The response of interest is the final grade of the students. After removing the 28 students with zero final grade, we got $n = 357$ samples in total. Moreover, we added the interactions between each pair of the variables, resulting in $p = 528$ predictors. The predictors were standardized to have mean zero and L_2-norm $\sqrt{n}$ in each column, and the response was centralized to have mean zero.

Similar to Janková and van de Geer (2016), after constructing the confidence intervals for all coefficients, we identified the significant level at $\alpha = 0.05$, meaning their confidence intervals of 95% coverage probability did not contain zero. Table 9 concludes sizes of the selected models, system running times and confidence intervals of the three most significant coefficients in terms of p-values by R-IPAD over different split sizes, where $K = 1$ corresponds to LDPE. The most significant variables were 'Grade 2' (second period grade), 'V528' (interaction of 'Grade 1' and 'Grade 2'), and 'V456' (interaction of 'whether attending nursery school' and 'time spending on going out with friends'). They were also identified by popular sparse modeling methods including the Lasso, MCP (Zhang, 2010), and SCAD (Fan and Li, 2001), tuned by cross-validation. In view of Table 9, the sizes of the selected models and the confidence intervals of the three most significant coefficients were all around the same level over different split sizes, which demonstrates the statistical accuracy of R-IPAD. Furthermore, it can be seen that there is a significant improvement on the computing speed when the split size increases.

Table 9: Model sizes, system running times, and confidence intervals of the three most significant coefficients by LDPE and R-IPAD over different split sizes in Section 5.1

Split size	K=1	K=3	K=5	K=10	K=15
Model size	29	31	33	33	34
Time (mins)	60.2	21.7	18.9	7.3	4.5
Grade 2	$(2.35, 3.05)$	$(2.42, 2.97)$	$(2.44, 2.85)$	$(2.45, 2.75)$	$(2.44, 2.79)$
V528	$(0.26, 0.75)$	$(0.31, 0.75)$	$(0.35, 0.74)$	$(0.38, 0.72)$	$(0.40, 0.69)$
V456	$(-0.45, -0.13)$	$(-0.43, -0.12)$	$(-0.40, -0.11)$	$(-0.38, -0.11)$	$(-0.36, -0.08)$

5.2 Application to the PCR data set

In this second example, we compare R-IPAD with LDPE on a polymerase chain reaction data set. The PCR data set was originally studied in Lan et al. (2006). It examines the genetics of two inbred mouse populations and comprises of $n = 60$ samples with 29 males and 31 females. Expression levels of 22575 genes were measured. Following Song and Liang (2015) and Kong et al. (2016), we studied the linear relationship between the numbers of phosphoenolpyruvate carboxykinase (PEPCK), a phenotype measured by quantitative real-time PCR, and the gene expression levels. We picked $p = 2000$ genes having the highest marginal correlations with the PEPCK as predictors. The predictors were standardized to have mean zero and L_2-norm $\sqrt{n}$ in each column and the responses were centralized before we conducted the analysis.

Table 10: Model sizes, system running times, and confidence intervals for coefficients of significant genes by LDPE and R-IPAD over different split sizes in Section 5.2

Split size	K=1	K=2	K=3
Model size	18	24	22
Time (mins)	42.2	24.8	16.4
1438819_at	$(-0.389, -0.113)$	$(-0.398, -0.126)$	$(-0.418, -0.142)$
1460011_at	$(-0.317, -0.043)$	$(-0.341, -0.064)$	$(-0.343, -0.068)$
1438937_x_at	$(-0.002, 0.475)$	$(0.097, 0.477)$	$(0.158, 0.492)$

We identified the significant predictors in the same way as in Section 5.1 at the $\alpha = 0.005$ level, which is stricter since hundreds of genes would be selected if we keep $\alpha = 0.05$ after the de-biasing step due to the large residual errors of prediction based on the initial estimator. The split sizes for R-IPAD were 2 and 3. We did not split the data into more groups since there were only 60 samples in total. Nevertheless, since the dimensionality is high, there is a significant improvement in computing speed in view of the system running times from Table 10. The selected models varied a bit over different split sizes since the p-values of some selected genes were on the boundary, but the confidence intervals of the most significant genes were around the same level. See, for instance, the confidence intervals of the top two significant genes '1438819_at' and '1460011_at' over different split sizes in Table 10. An interesting thing is that the significant gene '1438937_x_at' identified by R-IPAD fell into the rejection boundary of LDPE in view of its confidence intervals. But this gene was the only significant one reported in Song and Liang (2015) and shared by other five popular variable selection approaches. It verifies the robustness of R-IPAD in presence of heavy-tailedness and outliers due to the split and conquer procedure.

6. Discussion

We have proposed a new methodology of scalable inference with partitioned data to adapt to big data applications. To the best of our knowledge, it is among the first attempts for deriving high-dimensional confidence intervals in a split and conquer framework. Compared with inference by LDPE without splitting the data, the suggested method improves the computational speed by about K^2 times in a single computing device and about K^3 times if multiple devices are employed simultaneously. Theoretically, we prove that the length of the confidence intervals constructed by the partitioned approach is asymptotically equivalent to that without splitting the data, along with a significantly larger upper bound on the split sizes. Moreover, a refined inference procedure is developed to address the inflation issue under finite samples and large split sizes. Last but not least, we suggest a bootstrap-assisted procedure for simultaneous inference on a large number of coefficients. Simulation studies are consistent with our theoretical results and real data analyses show that the proposed partitioned approach can be more robust and resistent to heavy-tailedness and outliers.

Besides the mean bagging estimator, we can also adopt some other bagging estimators such as the median to further enhance the robustness of the partitioned approach. It would be interesting to derive the theoretical properties and the noise factors based on other bagging estimators. We also believe that the idea of deriving high-dimensional confidence intervals through a partitioned approach can be applied to other models such as generalized linear models to reduce the computational cost of the de-biasing step in broader applications. Then the key questions are about the upper bound on the split sizes and how to develop an inference procedure with accurate confidence intervals under finite samples and large split sizes. These problems are beyond the scope of the current paper and will be interesting topics for future research.

Supplementary Materials

Proofs of the theoretical results are available in the Supplementary Materials.

Acknowledgements

The authors sincerely thank editors and referees for their valuable comments that helped improve the article substantially. This work was supported by National Natural Science Foundation of China Grants-11601501, 11471302, 11671374, 11971457 and 71731010, Fundamental Research Funds for the Central Universities-WK2040160028, Anhui Provincial Natural Science Foundation Grant-1708085QA02, and the University of Hong Kong, Zhejiang Institute of Research and Innovation Seed Fund. Part of this work is done when Jiarui Zhang was a visiting researcher at the University of Hong Kong, Zhejiang Institute of Research and Innovation. Zemin Zheng and Jiarui Zhang are the co-first authors and the corresponding author is Jiarui Zhang.

References

Battey, H., Fan, J., Liu, H., Lu, J. and Zhu, Z. (2018). Distributed estimation and inference with statistical guarantees. *Ann. Statist.* **46**, 1352–1382.

Bickel, P. J., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37**, 1705–1732.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35**, 2313–2351.

Chen, X. and Xie, M. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statist. Sinica* **24**, 1655–1684.

Cortez, P. and Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. *Proceedings of 5th Annual Future Business Technology Conference*, Porto, 5–12.

Decrouez, G. and Hall, P. (2014). Split sample methods for constructing confidence intervals for binomial and Poisson parameters. *J. R. Statist. Soc. Ser.* B **76**, 949–975.

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. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Statist. Soc. Ser.* B **74**, 37–65.

Fan, J., Han, F. and Liu, H. (2014). Challenges of big data analysis. *National Sci. Rev.* **1**, 293–314.

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., Richard, S. and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *J. Mach. Learn. Res.* **10**, 1829–1853.

Fan, Y. and Lv, J. (2013). Asymptotic equivalence of regularization methods in thresholded parameter space. *J. Amer. Statist. Assoc.* **108**, 1044–1061.

Hao, N., Feng, Y. and Zhang, H. H. (2018). Model selection for high dimensional quadratic regression via regularization. *J. Amer. Statist. Assoc.* **113**, 615–625.

Janková, J. and van de Geer, S. (2016). Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics* **9**, 1205–1229.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15**, 2869–2909.

Kleiner, A., Talwalkar, A., Sarkar, P. and Jordan, M. (2014). A scalable bootstrap for massive data. *J. R. Statist. Soc. Ser.* B **76**, 795–816.

Kong, Y., Zheng, Z. and Lv, J. (2016). The constrained Dantzig selector with enhanced consistency. *J. Mach. Learn. Res.* **17**, 1–22.

Lan, H., Chen, M., Flowers, J., Yandell, D., Mata, C., Mui, E., Flowers, M., Schueler, K., Manly, K., Williams, M., Kendziorski, O. and Antie, A. D. (2006). Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genetics* **57**, 53–63.

Lee, J., Liu, Q., Sun, Y. and Taylor, J. (2017). Communication-efficient distributed sparse regression. *J. Mach. Learn. Res.*, **18**, 1–30.

Lee, J., Sun, D., Sun, Y. and Taylor, J. (2016). Exact post-selection inference with the lasso. *Ann. Statist.* **44**, 907–927.

Lian, H. and Fan, Z. (2018). Divide-and-conquer for debiased l_1-norm support vector machine in ultra-high dimensions. *J. Mach. Learn. Res.* **18**, 1–26.

Liu, Y. and Wu, Y. (2007). Variable selection via a combination of the L_0 and L_1 penalties. *Journal of Computational and Graphical Statistics* **16**, 782–798.

Lockhart, R., Taylor, J., Tibshirani, R. and Tibshirani, R. (2014). A significance test for the lasso. *Ann. Statist.* **42**, 413–468.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37**, 3498–3528.

Mackey, L., Talwalkar, A. and Jordan M. (2015). Distributed matrix completion and robust factorization. *J. Mach. Learn. Res.* **16**, 913–960.

Raskutti, G., Wainwright, M. J. and Yu. B. (2011). Minimax rates of estimation for high-dimensional linear regression over l_q-balls. *IEEE Transactions on Information Theory.* **57**, 6976–6994.

Shang, Z. and Cheng, G. (2015). Nonparametric Bayesian aggregation for massive data. *J. Mach. Learn. Res.* **20**, 1–81.

Shang, Z. and Cheng, G. (2017). Computational limits of a distributed algorithm for smoothing spline. *J. Mach. Learn. Res.* **18**, 1–37.

Song, Q. and Liang, F. (2015). High-dimensional variable selection with reciprocal L_1-regularization. *J. Amer.*

Statist. Assoc. **110**, 1607–1620.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99**, 879–898.

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

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42**, 1166–1202.

Weng, H., Feng, Y. and Qiao, X. (2017). Regularization after retention in ultrahigh dimensional linear regression models. *Statist. Sinica.* **29**, 387–407.

Xu, C., Zhang, Y. and Li, R. (2016). On the feasibility of distributed kernel regression for big data. *IEEE Transactions on Knowledge and Data Engineering.* **28**, 3041–3052.

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

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

Zhang, X. and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. *J. Amer. Statist. Assoc.* **112**, 757–768.

Zhang, Y., Duchi, J. C. and Wainwright, M. J. (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.* **16**, 3299–3340.

Zhao, T., Cheng, G. and Liu, H. (2016). A partially linear framework for massive heterogeneous data. *Ann. Statist.* **44**, 1400–1437.

Zheng, Z., Fan, Y. and Lv, J. (2014). High-dimensional thresholded regression and shrinkage effect. *J. R. Statist. Soc. Ser.* B **76**, 627–649.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418–1429.

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

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36**, 1509–1566.

The School of Management, the School of Data Science, and International Institute of Finance, University of Science and Technology of China, Hefei 230026, China.

E-mail: zhengzm@ustc.edu.cn, zjrt46@mail.ustc.edu.cn, tjly@mail.ustc.edu.cn, wuyh@ustc.edu.cn