Statistica Sinica **31** (2021), 1935-1959
doi:https://doi.org/10.5705/ss.202018.0379

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, accurate inferences with low computational costs are the key to improving scalability. When the sample size and dimensionality are both large, naively applying de-biasing to derive confidence intervals can be computationally inefficient or infeasible, because the de-biasing procedure increases the computational cost by an order of magnitude compared with that of the initial penalized estimation. Therefore, we suggest a split and conquer approach to improve the scalability of the de-biasing procedure, and show that the length of the established confidence interval is asymptotically the same as that using all of the data. Moreover, we demonstrate a significant improvement in the largest split size by separating the initial estimation and the relaxed projection steps, indicating that the sample sizes needed for these two steps with statistical guarantees are different. We propose a refined inference procedure to address the inflation issue in the finite sample performance when the split size becomes large. Lastly, numerical studies demonstrate the computational advantage and theoretical guarantee of our new methodology.

Key words and phrases: Big data, confidence intervals, de-biased estimator, divide and conquer, large split sizes, scalability.

1. Introduction

The rapid development of technologies and devices has made it easier than ever to generate large-scale data sets in areas such as meteorology, genomics, and economics, which are referred to as big data problems (Fan, Han and Liu (2014)). As a result, high-dimensional sparse modeling, applied when the number of variables can be larger than the sample size, has become a popular area of statistical research. When the sample size and dimensionality are both large, naively applying existing high-dimensional inference methods to large amounts of data can be computationally inefficient, or even infeasible. Thus, it is appealing to develop scalable methodologies that take advantage of huge data sets for accurate inferences with low computational costs.

Corresponding author: Jiarui Zhang and Yang Li, International Institute of Finance, University of Science and Technology of China, Hefei, Anhui 230026, China. E-mail: zjrt46@mail.ustc.edu.cn (Zhang), tjly@mail.ustc.edu.cn (Li).

As a reliable tool for producing meaningful and interpretable models, sparse modeling via regularization has shown its strengths in handling 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, Ritov and Tsybakov (2009), Fan, Richard and Wu (2009), Lv and Fan (2009), Zhang (2010), Sun and Zhang (2012), Fan and Lv (2013), Zheng, Fan and Lv (2014), Song and Liang (2015), Kong, Zheng and Lv (2016), Weng, Feng and Qiao (2017), and Hao, Feng and Zhang (2018). Based on sparse regularized estimators, statistical inferences such as hypothesis testing and confidence intervals can be made when the asymptotic distributions of the pilot estimators are derived. See, for instance, Lockhart et al. (2014) and Lee et al. (2016) for inferences using model selection and the Lasso (Tibshirani (1996)), and Javanmard and Montanari (2014), van de Geer et al. (2014), and Zhang and Zhang (2014) for inferences with de-biasing of the penalized estimators. We focus on improving the scalability of the methods in a big data setting, because the de-biasing procedure increases the computational cost by an order of magnitude close to the dimensionality compared with that of the initial penalized estimation, causing computational bottlenecks in large-scale applications.

A natural and efficient way to deal with big data problems is to use data splitting, where the entire data set is split into subsamples, and then the estimators obtained from each subsample are aggregated. This divide and conquer idea has been widely used to solve various kinds of problems (Fan, Guo and Hao (2012); Decrouez and Hall (2014); Kleiner et al. (2014); Mackey, Talwalkar and Jordan (2015); Shang and Cheng (2015); Zhang, Duchi and Wainwright (2015); Xu, Zhang and Li (2016); Zhao, Cheng and Liu (2016); Shang and Cheng (2017); Lian and Fan (2018)), with benefits such as robustness and enhanced stability in addition to the computational advantage being 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 penalized estimation approach that retains desired statistical properties and be more resistant to false model selections. Lee et al. (2017) devised a one-shot approach for a distributed sparse regression that averages the de-biased Lasso estimators and is 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 the presence of large-scale high-dimensional data remains largely unexplored. In general, deriving confidence intervals is not as flexible as hypothesis testing, because test statistics may not be inverted to pilot estimators. Even if we use the divide and conquer algorithm, how to preserve asymptotically equivalent statistical accuracy and efficiency for the full sample procedure is unclear. Moreover, neither the theoretical results nor the empirical performance of existing high-dimensional split and conquer inference methods 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 of big data applications. In this study, we provide answers to the aforementioned questions by introducing a new methodology for a scalable inference with partitioned data for deriving high-dimensional confidence intervals. The proposed method randomly splits the whole data set into subsamples of equal size, and generates a de-biased estimator for each subsample. Then, it constructs confidence intervals based on the bagging estimator that aggregates the estimators from all the subsamples.

The main contributions of this study are fourfold. First, we develop a new partitioned approach that substantially increases the computing speed of deriving confidence intervals in high dimensions. We prove that the length of the established confidence interval is asymptotically the same as that using all of the data, which means the information loss due to the divide and conquer procedure is negligible. Second, we demonstrate a significant improvement in the upper bound on the split size, which becomes the square of that in Battey et al. (2018), by separating the initial estimation and the relaxed projection steps. Thus the sample sizes needed in order for these two steps to enjoy statistical guarantees are different. Lastly, we propose a refined inference procedure to address the inflation issue in the finite sample performance when the split size indeed becomes large. Numerical studies show that the suggested methodology is communication efficient and can be more robust and resistant to heavy-tailedness and outliers.

The rest of the paper is organized as follows. Section 2 presents the proposed methodology of scalable inference with partitioned data in a big data settings. We provide confidence intervals based on the partitioned approach and the desired statistical guarantees in Section 3. The computational advantage and theoretical properties are demonstrated empirically using simulation studies in Section 4 and real data analyses in Section 5. 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 inferences with partitioned data to using high-dimensional linear regression models. Let $\mathbf{y} = (y_1, \ldots, y_n)^T$ be the n-dimensional response vector, and let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ be 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 $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a p-dimensional regression coefficient vector, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ 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:

$$\sum_{j=1}^{p} \min\left\{\frac{|\beta_j|}{\sigma\lambda_{\mathrm{univ}}}, 1\right\} \leq s, \tag{2.2}$$

where $\lambda_{\mathrm{univ}} = \sqrt{(2/n)\log p}$. This condition is weaker than the direct sparsity assumption $\|\boldsymbol{\beta}\|_0 \leq s$, because it allows for a large number of nonzero coefficients, as long as their magnitudes are small. We focus on the big data settings in which 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 study, we adopt the low dimensional projection estimator (LDPE) proposed in Zhang and Zhang (2014) to illustrate our partitioned approach, owing to the appealing property that the LDPE does not require the minimum signal strength condition.

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

$$\{\widehat{\boldsymbol{\beta}}^{(\mathrm{init})}, \widehat{\sigma}\} = \underset{\mathbf{b}\in\mathbb{R}^p,\, \sigma>0}{\operatorname{argmin}} \left\{\frac{\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2}{2\sigma n} + \frac{\sigma}{2} + \lambda_0\|\mathbf{b}\|_1\right\},$$

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

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

for $1 \le j \le p$, where $\mathbf{z}_j^T(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(\mathrm{init})})/\mathbf{z}_j^T\mathbf{x}_j$ is the de-biasing term, and $\mathbf{z}_j$ is a relaxed orthogonalization of $\mathbf{x}_j$ against the 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 $\mathbf{X}_{-j} = (\mathbf{x}_k : k \ne j)$ with the weighted L_1-penalty. That is,

$$\begin{aligned} \mathbf{z}_j &= \mathbf{x}_j - \mathbf{X}_{-j}\widehat{\boldsymbol{\gamma}}_j, \\ \{\widehat{\boldsymbol{\gamma}}_j, \widehat{\sigma}_j\} &= \underset{\mathbf{b}\in\mathbb{R}^{p-1},\sigma\in\mathbb{R}^+}{\operatorname{argmin}} \left\{ \frac{\|\mathbf{x}_j - \mathbf{X}_{-j}\mathbf{b}\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_0 \sum_{k\ne j} \frac{\|\mathbf{x}_k\|_2}{\sqrt{n}}|b_k| \right\}, \end{aligned} \tag{2.4}$$

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

The LDPE has been shown to enjoy an asymptotic normal distribution. To gain insight into this, we consider the following decomposition:

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

where the first term is normally distributed, and the second term shows the approximation error. Let

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

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

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

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

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

Furthermore, a 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. Thus the proposed methodology focuses on improving the speed of calculating the relaxed projection vectors through data splitting, and uses 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, is demonstrated in Theorem 1. In cases where the initial estimator is infeasible owing to an extraordinarily large sample size or different locations, we suggest using the split and conquer approach of Chen and Xie (2014) to generate regularized initial estimators.

Our methodology for scalable inferences with partitioned data begins by splitting the entire data set into subsamples of equal size. Then it generates a de-biased estimator for each subsample using the same initial estimator. Finally, the de-biased estimators from all of the subsamples are aggregated using the mean in each coordinate. 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 the de-biased estimator $\widehat{\boldsymbol{\beta}}^{(l)} = (\widehat{\beta}_1^{(l)}, \ldots, \widehat{\beta}_p^{(l)})$ for each subsample by applying the bias correction idea of the LDPE to the initial estimator $\widehat{\boldsymbol{\beta}}^{(\text{init})}$ using the vectors $\mathbf{z}_j^{(l)}$ and $\mathbf{x}_j^{(l)}$. Finally, the mean bagging estimator $\widehat{\boldsymbol{\beta}}^{(\text{mean})} = K^{-1}\sum_{l=1}^{K}\widehat{\boldsymbol{\beta}}^{(l)}$ averages the de-biased estimators over all subsamples. That is,

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

We derive confidence intervals based on this mean bagging estimator in Section 3 by breaking the communication barriers between 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 popular LARS algorithm (Efron et al. (2004)) used to generate 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. Therefore, the computational cost of the LDPE, which needs about p Lasso solutions, is around $O(n^3p)$. Using our partitioned approach, the computational cost is $K \cdot O(\widetilde{n}^3p) = O(K^{-2}n^3p)$, representing a reduction by a factor of K^{-2} compared with the 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. Several methods use 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 is proposed for a penalized estimation of the regression coefficients under extraordinarily large data, where the combined estimator is shown to be asymptotically equivalent to the estimator analyzing all of the data, and is more robust in terms of variable selection. Theoretical and numerical analyses demonstrate similar asymptotic efficiency and robustness when deriving confidence intervals using our partitioned approach. The other two works, Lee et al. (2017) and Battey et al. (2018), are more related to ours, because both use de-biased estimators in each subsample. Because Lee et al. (2017) focuses mainly on estimation accuracy in a distributed sparse regression, we compare our work with that of Battey et al. (2018), who considered hypothesis testing using split and conquer approaches.

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

3. Theoretical Properties

Because our partitioned approach focuses on speeding up 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* $\{\widehat{\boldsymbol{\beta}}^{(\text{init})}, \widehat{\sigma}\}$ *satisfies that*

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

for some positive constants C_1 *and* 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 standard 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. When 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), whichthey show achieves similar asymptotic estimation accuracy.

In the fixed design setting, the confidence intervals of the LDPE are provided in Zhang and Zhang (2014). Here, we focus on the random design case with $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_n \otimes \boldsymbol{\Sigma})$ to analyze some key quantities such as $\|\mathbf{z}_j\|_2$, τ_j, and η_j in a probabilistic sense, and to derive 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* $\boldsymbol{\Sigma}$ *are bounded within the interval* $[M_*, M^*]$*, for some positive constants* M_* *and* M^**, and tha the rows of* $\boldsymbol{\Sigma}^{-1}$ *are sparse, with* $\max_i \sum_{j=1}^p I\{\boldsymbol{\Sigma}_{ij}^{-1} \neq 0\} \leq \sqrt{s}$*, 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 $\lim_{n\to\infty} \tau_j n^{1/2} = \boldsymbol{\Sigma}_{jj}^{-1/2}$ *hold simultaneously with probability at least* $1 - o(p^{-\delta+1})$*.*

Furthermore, if Condition 1 *holds with* $C_1 C_j s \sqrt{(2/n)\log(p)\log(p/\epsilon)} \leq \epsilon_n'$*,*

$C_2 s(2/n)\log(p/\epsilon) \le \epsilon''_n$, *and* $\max(\epsilon'_n, \epsilon''_n) \to 0$ *as* $n \to \infty$, *then for sufficiently large* n, *the LDPE in* (2.3) *satisfies*

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

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

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

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

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

Let $\tau_j^{(l)}$ and $\eta_j^{(l)}$ be the noise and bias factors, respectively 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 a subsample size of $\widetilde{n} = n/K$.

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*

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

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

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

(B) *(Refined inference). Let $\omega_j^{(l)} = \widetilde{\tau}_j^{-1}\tau_j^{(l)}$ and $K_j = [\sum_{l=1}^{K}(\omega_j^{(l)})^2]^{-1}K^2$. Thus, $K_j \in [K, c_j^* K]$ 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*

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

Moreover, if $\sqrt{K_j}\epsilon'_n \to 0$, we have

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

Theorem 1 establishes the confidence intervals based on the mean bagging estimator of the suggested partitioned approach by breaking the communication barriers between the subsamples. Given the confidence intervals in (3.3), the statistical accuracy of our partitioned approach is asymptotically equivalent to that using the full sample, because $K^{-1/2}\widetilde{\tau}_j \asymp K^{-1/2}\widetilde{n}^{-1/2} = n^{-1/2} \asymp \tau_j$ by Theorem 1 and Proposition 1, and the limits of $K^{-1/2}\widetilde{\tau}_j n^{1/2}$ and $\tau_j n^{1/2}$ are both $\boldsymbol{\Sigma}_{jj}^{-1/2}$. This means that the lengths of the 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 Previously.

Furthermore, Theorem 1 provides the theoretical upper bound on the largest split size of $K = o(ns^{-2}\log^{-2} p)$, given 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 inferences using each subsample, in the sense of the minimax error bound for the initial Lasso estimator (Raskutti, Wainwright and Yu (2011)). We can improve the bound of K because we use a different partitioned inference strategy, where the de-biasing procedure is implemented using different subsamples to improve the computational efficiency, and 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 improved further.

Although the theoretical upper bound allows for large split sizes, the confidence intervals (3.3) yield higher coverage probabilities, in gereral, than the preassigned level in terms of the finite sample performance when the split size indeed becomes large, as shown in Section 4. Therefore, the statistical accuracy is insufficient,because the lengths of confidence intervals are longer than expected. This issue mainly results from the inflation of the overall noise factor $\widetilde{\tau}_j = \max_{1\leq l\leq K} \tau_j^{(l)}$, the magnitude of which can be larger than $\widetilde{n}^{-1/2}$ when K is large, owing to randomness in the subsamples. In this case, the order of $K^{-1/2}\widetilde{\tau}_j$ deviates from $n^{-1/2}$, leading to an overestimation of the variance.

To address this inflation issue and to take full advantage of the large theoretical split sizes with statistical accuracy, we propose a refined inference procedure in Part (**B**) of Theorem 1 that considers the noise factor $\tau_j^{(l)}$ of every subsample to adjust the length of the 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 because it takes the average rather than the maximum of the noise factors. Because K_j and K differ only by a constant, the asymptotic efficiency and upper bound on the split size for Part (**A**) also apply to Part (**B**). We show that this refined procedure maintains statistical accuracy, even under very large split sizes, making it applicable to large-scale applications with massive data sets.

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 a 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*

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

(B) *If, in addition, Condition* 1 *holds with* $\max_{j\in S}\sqrt{c_j^*}C_jC_1s\sqrt{(2/n)\log(p)\log(p/\epsilon)}$ $\leq\epsilon_n'$, *then for any* $t\geq(1+\max_{j\in S}\sqrt{K_j}\epsilon_n')/(1-\epsilon_n'')$, *we have*

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

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

The bootstrap-assisted procedure starts by 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 that introduced in Zhang and Cheng (2017). The bootstrap critical value is given by $c_G(\alpha)=\inf\{t\in\mathbb{R}:P\left(W_G\leq t|(\mathbf{y},\mathbf{X})\right)\geq 1-\alpha\}$. 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* $(\log(pn))^7/n\leq C_3n^{-c_3}$ *hold, for some positive constants* C_3 *and* c_3. *Then for any* $G\subseteq\{1,2,\ldots,p\}$, *we have*

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

Theorem 3 establishes the theoretical guarantee of constructing simultaneous confidence intervals using the proposed bootstrap-assisted procedure, which explicitly accounts for the effect of $|G|$, given 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, because $K^{-1/2}\widetilde{\tau}_j$ is of similar magnitude to τ_j.

Lastly, similarly to Zhang and Zhang (2014), the de-biased mean bagging estimator can also be used for variable selection and estimation for the entire regression coefficient vector after a simple soft thresholding: that is,

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

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 the variable selection and estimation properties listed below, showing 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 $\widehat{\boldsymbol{\beta}}^{(\mathrm{t})} = (\widehat{\beta}_1^{(\mathrm{t})},\ldots,\widehat{\beta}_p^{(\mathrm{t})})^T$ *be the soft thresholded mean bagging estimator with these* $\widehat{t}_j$. *Then, for any given* $\mathbf{X}$, *there exists an event* Ω_n *with* $P\{\Omega_n\} \ge 1-3\epsilon$, *such that*

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

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

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

4. Simulation Studies

In this section, we investigate the finite sample performance of the two versions of the scalable inference with partitioned data (denoted by IPAD and R-IPAD, respectively) listed in Theorem 1 with different split sizes K, and compare it with that of the LDPE using the full data set. We adopt similar high-dimensional settings to those in Zhang and Zhang (2014), where $(n, p) = (600, 1000)$ in the first example, and $(n, p) = (2000, 3000)$ with higher dimensionality in the second example. Each simulation consists of 100 replications. In every replication, we generate the data set $(\mathbf{X}, \mathbf{y})$ from the linear regression model in (2.1), where the rows of $\mathbf{X}$ are independent and identically distributed (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

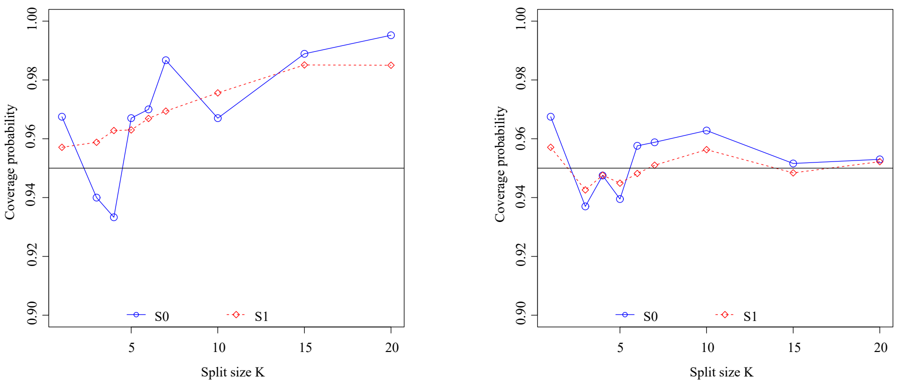
In this simulation study, the true regression coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ satisfies $\beta_j = 3\lambda_{\text{univ}}$ with $\lambda_{\text{univ}} = \sqrt{(2/n)\log p}$, for $j = 200, 400, 600, 800, 1000$, and $\beta_j = 3\lambda_{\text{univ}}/j$ for all other j. It is a mixture of strong and weak signals without any zero coefficients, originally designed in Zhang and Zhang (2014). By setting $\alpha = 0.05$, we aim to achieve confidence intervals with 95% coverage probability for each coefficient. For convenience, let $S_0 = \{\beta_j : \beta_j = 3\lambda_{\text{univ}}, j = 1, 2, \ldots, p\}$ and $S_1 = \{\beta_j : \beta_j = 3\lambda_{\text{univ}}/j, j = 1, 2, \ldots, p\}$ be 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 using different methods, where $K = 1$ corresponds to the LDPE without partitioning the data.

The coverage probabilities of IPAD match well with the preassigned level when the split size K is relatively small (less than five), and start approaching one when K gets larger. Thus, the length of confidence interval is longer than needed, and so that the statistical accuracy decreases. This can be seen directly from the average lengths of the confidence intervals for a typical strong signal β_{200} and a weak signal β_{201} in Table 2.This agrees with our previous theoretical analysis; that is, when K gets large, the randomness in the K groups inflates the overall noise factor $\widetilde{\tau}_j$, which is the maximum of the individual noise factors $\tau_j^{(l)}$ over K groups.

On the other hand, this inflation issue is solved by the refined inference R-IPAD given the corresponding results in Tables 1 and 2 and Figure 1 (b), because the coverage probability stays around the preassigned 95 percent, and the length of the confidence interval maintains the same level. Note that even if the split

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

Table 2. Average lengths of confidence intervals for β_{200} and β_{201} using 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

size K is as large as 20, such that each subgroup contains only 30 samples, R-IPAD still works well in terms of both the coverage probability and the statistical

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

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

accuracy. This makes the proposed partitioned approach scalable for analyzing massive data sets with large splits. Of course, the split size should not keep increasing because we need a sufficient sample size in each subgroup to provide relatively accurate estimates.

Furthermore, the computational cost has been significantly reduced after partitioning the data in Table 3 and Figure 3 (a), where the average system running time for each replication is about two hours, using all samples at once. Furthermore, it decreases dramatically after splitting the data, with a running time of just four minutes when the split size is equal to 20. This statistical analysis was performed on a usual PC with an Intel Core i7-7700 CPU (3.60 GHz) and 8 GB RAM. In addition, parallel computing was employed in the computation of the relaxed projection residual vectors $\mathbf{z}_j$ and the 100 simulation replications, where the computation tasks were divided into seven cores using the R package "snowfall," confirming our aforementioned computational advantage using a fair comparison on a single computing device. The computational advantage can be enhanced further by using multiple PCs to analyze different subsamples.

Lastly, we present the coverage probabilities, average lengths of the confidence intervals, and average system running times for the simultaneous confidence intervals of the coefficients β_j in S_0 and all β_j, using the proposed bootstrap-assisted procedure with the preassigned 95% coverage probability over 100 replications in Table 4. The results show that, the coverage probabilities stay around

Table 5. Average coverage probabilities using 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

0.95 over different split sizes, and the average lengths of the confidence intervals are also very stable, demonstrating the validity of the proposed method. Moreover, significant improvements in the running times are evident as the split size becomes larger.

4.2. Simulation example 2

In this example, we increase both the dimensionality and the sample size to $p = 3{,}000$ and $n = 2{,}000$, such that a usual PC finds it difficult to implement the LDPE without partitioning the data, owing to the significant computational cost. Therefore our statistical analysis starts with a split size $K = 5$ and ends with $K = 40$. The true regression coefficient vector $\boldsymbol{\beta}$ takes strong signals of the same magnitude as those 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 as before. Table 5 and Figure 2 summarize the results for the average coverage probabilities using different methods and split sizes; a similar conclusion to that in Section 4.1 can be drawn. The IPAD method works well when the split size K is no larger than 10, but begins to lose statistical accuracy after K becomes larger, owing to the inflation issue. However, R-IPAD maintains good performance under different split sizes, as shown in Tables 5 and 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 with a stable length of confidence intervals. At the same time, a significant improvement in computing speed is evident Table 7 and Figure 3 (b). The performance of the simultaneous confidence intervals is also similar to that in the first example, as shown in Table 8.

Both simulation examples illustrate the statistical accuracy and computa-

Table 6. Average lengths of confidence intervals for β_{500} and β_{501} using 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

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 the 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

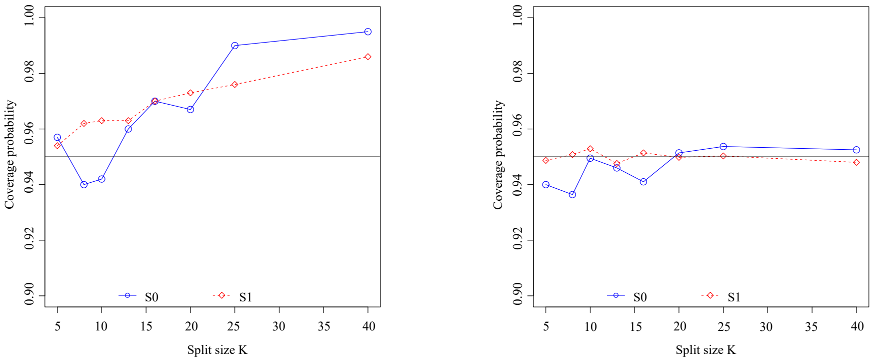
tional advantage of constructing confidence intervals using R-IPAD, even under very large split sizes. We focus on this refined inference procedure in our analysis of real data sets.

5. Real-Data Analyses

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

5.1. Application to student performance data

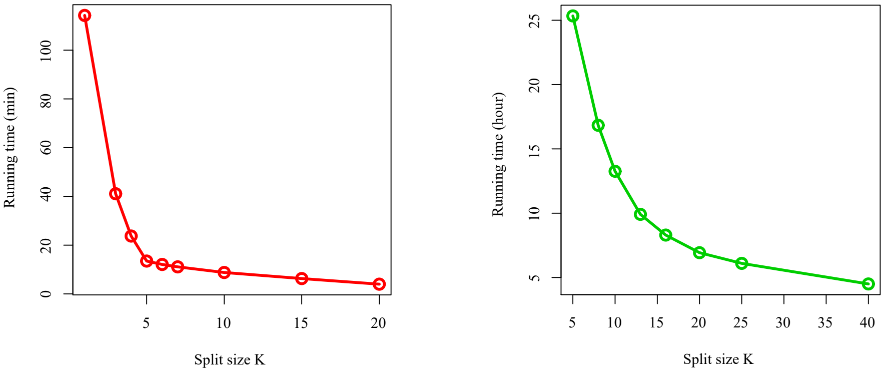
This data set was studied in Cortez and Silva (2008) to evaluate students' performance in two Portuguese public schools, and is available from the UCI Machine



(a) IPAD

(b) R-IPAD

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



(a) $n = 600, p = 1{,}000$

(b) $n = 2{,}000, p = 3{,}000$

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

Learning Repository (`https://archive.ics.uci.edu/ml/datasets/Student+Performance`). It consists of 32 predictive variables, including the 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 28 students with a zero final grade,the final sample size was $n = 357$. Moreover, we added interactions between each pair of 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.

Table 9. Model sizes, system running times, and confidence intervals of the three most significant coefficients by the 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	$(-0.45, -0.13)$	$(-0.43, -0.12)$	$(-0.40, -0.11)$	$(-0.38, -0.11)$	$(-0.36, -0.08)$

Similarly 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 shows the sizes of the selected models, system running times, and confidence intervals of the three most significant coefficients in terms of the R-IPAD p-values over different split sizes, where $K = 1$ corresponds to the 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 using popular sparse modeling methods, including the Lasso, minimax concave penalty (Zhang (2010)), and smoothly clipped absolute deviations penalty (Fan and Li (2001)), and tuned using cross-validation. From Table 9, the sizes of the selected models and the confidence intervals of the three most significant coefficients are all around the same level over different split sizes, which demonstrates the statistical accuracy of R-IPAD. Furthermore, there is a significant improvement in the computing speed when the split size increases.

5.2. Application to the PCR data set

In this second example, we compare R-IPAD with the LDPE on a PCR data set. This set was originally studied in Lan et al. (2006). It examines the genetics of two inbred mouse populations, and comprises $n = 60$ samples, with 29 males and 31 females. The expression levels of 22,575 genes were measured. Following Song and Liang (2015) and Kong, Zheng and Lv (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 the $p = 2{,}000$ genes with 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 performing 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	$(-0.002, \quad 0.475)$	$(0.097, \quad 0.477)$	$(0.158, \quad 0.492)$

We identified the significant predictors in the same way as in Section 5.1 at the $\alpha = 0.005$ level. This is stricter because hundreds of genes are selected if we keep $\alpha = 0.05$ after the de-biasing step, owing to the large residual errors of the prediction based on the initial estimator. The split sizes for R-IPAD were two and three. We did not split the data into more groups because there are only 60 samples in total. Nevertheless, because the dimensionality is high, there is a significant improvement in computing speed, as shown in the system running times in Table 10. The selected models varied womewhat over different split sizes because the p-values of some selected genes were on the boundary. However, 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. Note that the significant gene "1438937_x_at" identified by R-IPAD fell into the rejection boundary of the LDPE in terms of its confidence intervals. However, this gene was the only significant one reported in Song and Liang (2015), and is shared by other five popular variable selection approaches. This verifies the robustness of R-IPAD in presence of heavy-tailedness and outliers, owing to the split and conquer procedure.

6. Discussion

We have proposed a new methodology for scalable inferences with partitioned data for big data applications. To the best of our knowledge, this study is one of the first attempts to derive high-dimensional confidence intervals using a split and conquer framework. Compared with inferences using the LDPE without splitting the data, the proposed method improves the computational speed by about K^2 times in a single computing device, and by about K^3 times if multiple devices are employed simultaneously. We prove theoretically that the length of the confidence intervals constructed using 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. Lastly, we suggest a bootstrap-assisted procedure for simultaneous inferences on a large number of coefficients. The results of our 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.

In addition to the mean bagging estimator, we can also adopt 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. Furthermore, we believe that the idea of deriving high-dimensional confidence intervals using a partitioned approach can be applied to other models, such as generalized linear models, to reduce the computational cost of the de-biasing step. 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 this research, and are left to future research.

Supplementary Material

Proofs of the theoretical results are available in the online Supplementary Material.

Acknowledgments

The authors thank the editors and referees for their valuable comments. This work was supported by the National Natural Science Foundation of China(grants-72071187, 11971457, 11601501, 11471302, 11671374, 71731010, and 71921001), Fundamental Research Funds for the Central Universities-WK3470000017, WK2 040000027, and Wk2040160028, and University of Hong Kong, Zhejiang Institute of Research and Innovation Seed Fund. Part of this work was completed 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 Jiarui Zhang and Yang Li are co-corresponding authors.

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. In *Proceedings of 5th Annual Future Business Technology Conference*, 5–12. Porto.

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. et al. (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. J. 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.

Zemin Zheng

International Institute of Finance, The School of Management, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China.

E-mail: zhengzm@ustc.edu.cn

Jiarui Zhang

International Institute of Finance, The School of Management, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China.

E-mail: zjrt46@mail.ustc.edu.cn

Yang Li

International Institute of Finance, The School of Management, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China.

E-mail: tjly@mail.ustc.edu.cn

Yaohua Wu

International Institute of Finance, The School of Management, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China.

E-mail: wuyh@ustc.edu.cn

(Received September 2018; accepted February 2020)