The Annals of Statistics
2021, Vol. 49, No. 5, 2851–2869
https://doi.org/10.1214/21-AOS2062

DISTRIBUTED STATISTICAL INFERENCE FOR MASSIVE DATA

BY SONG XI CHEN[1] AND LIUHUA PENG[2]

[1]*Guanghua School of Management and Center for Statistical Science, Peking University, songxichen@pku.edu.cn*
[2]*School of Mathematics and Statistics, University of Melbourne, liuhua.peng@unimelb.edu.au*

This paper considers distributed statistical inference for general symmetric statistics in the context of massive data with efficient computation. Estimation efficiency and asymptotic distributions of the distributed statistics are provided, which reveal different results between the nondegenerate and degenerate cases, and show the number of the data subsets plays an important role. Two distributed bootstrap methods are proposed and analyzed to approximation the underlying distribution of the distributed statistics with improved computation efficiency over existing methods. The accuracy of the distributional approximation by the bootstrap are studied theoretically. One of the methods, the pseudo-distributed bootstrap, is particularly attractive if the number of datasets is large as it directly resamples the subset-based statistics, assumes less stringent conditions and its performance can be improved by studentization.

1. Introduction. Massive data with rapidly increasing size are encountered in many scientific fields that call needs for new statistical analysis. Not only the size of the data is an issue, but also that the data are often stored in multiple locations. This implies that statistical procedures formulated on the the entire data have to involve data communication between different storage facilities, which are expensive and slow down the computation, and sometimes are impossible in some cases due to privacy concerns that prevent data sharing between different data locations.

Two strains of methods have been developed to deal with the challenges with massive data. One is the "split-and-conquer" (SaC) method considered in Lin and Xi (2010), Zhang, Duchi and Wainwright (2013), Chen and Xie (2014), Volgushev, Chao and Cheng (2019) and Battey et al. (2018). From the estimation point of view, SaC partitions the entire data into subsets of smaller sizes, performs the estimation on each subset and then aggregates to form the final estimator. SaC has been used in different settings, for instance the M-estimation by Zhang, Duchi and Wainwright (2013) and the generalized linear models by Chen and Xie (2014); see also Volgushev, Chao and Cheng (2019) and Battey et al. (2018).

The other strain of the methods makes the bootstrap adaptive to massive data for obtaining the standard errors or confidence intervals for statistical inference. As the bootstrap resampling of the entire dataset is not feasible for massive data, Kleiner et al. (2014) introduced the bag of little bootstrap (BLB) that incorporates subsampling and the m out of n bootstrap to assess the variation of estimators for various inference purposes. Sengupta, Volgushev and Shao (2016) proposed the subsampled double bootstrap (SDB) that combines the BLB with a fast double bootstrap (Davidson and MacKinnon (2002), Chang and Hall (2015)) that has computational advantages over the BLB. Both the BLB and SDB's core idea is to construct bootstrap resamples that match to the entire data. Their computational efficiency relies on that the estimator of interest admitting a weighted subsample representation. However, for

Received August 2020; revised January 2021.
MSC2020 subject classifications. Primary 62G09; secondary 62G20.
Key words and phrases. Distributed bootstrap, distributed statistics, massive data, pseudo-distributed bootstrap.

estimators without such weighted subsample representation, the BLB and SDB can be computationally much involved.

We consider distributed statistical inference for a broader class of symmetric statistics (Lai and Wang (1993)), which encompass both the linear and nonlinear statistics, and cover both nondegenerate and degenerate cases. Although we use the standard SaC formulation, our study reveals new findings. For the degenerate case, the distributed formulation no longer attains the same efficiency as the full sample statistic, which differs from the nondegenerate case. The asymptotic distributions of the distributed statistics under the degeneracy are, respectively, the weighted χ^2 and Gaussian, depending on K, the number of data blocks, being finite or divergent. Higher order agreement between the distributions of the distributed statistic and the full sample statistic is established, which shows the roles played by K.

We propose two bootstrap algorithms to approximate the distribution of the distributed statistic: the distributed bootstrap (DB) and the pseudo-distributed bootstrap (PDB), which have advantages over the BLB and SDB. The DB does resampling within each subset, which is distributive and makes it well suitable for the distributed formulation. The PDB directly resamples the subset-based statistics, which offers further computation saving over the DB, BLB and SDB, while requiring K being large. Furthermore, the PDB works under less stringent conditions for both the nondegenerate and degenerate cases, and its performance can be improved by studentization, which inherits a property of the conventional bootstrap.

The paper is organized as follows. The efficiency and asymptotic distributions of the distributed statistics are established in Section 3. The two distributed bootstrap procedures are studied in Sections 4 and 5. Section 6 provides numerical verification to the theoretical results. Proofs, technical details, discussion on computational complexity and extra numerical studies including a real data analysis are in the supplementary materials (SM) (Chen and Peng (2021)).

2. Distributed symmetric statistics. Let $\mathfrak{X}_N = \{X_1, \ldots, X_N\}$ be a sequence of independent random vectors taking values in a measurable space $(\mathcal{X}, \mathcal{B})$ with a common distribution F. A symmetric statistic $T_N = T(\mathfrak{X}_N)$ for a parameter $\theta = \theta(F)$ is invariant under data permutations, that admits a general nonlinear expansion

$$T_N = \theta + N^{-1}\sum_{i=1}^{N}\alpha(X_i; F) + N^{-2}\sum_{1\le i<j\le N}\beta(X_i, X_j; F) + R_N, \tag{2.1}$$

where $\alpha(x; F)$ and $\beta(x, y; F)$ are known functions, and $R_N = R(\mathfrak{X}_N; F)$ is a remainder term. We assume the following conditions on α and β, and two sets of conditions on R_N.

CONDITION C1. (i) *The functions $\alpha(x; F)$ and $\beta(x, y; F)$, depending on F, are known measurable functions of x and y, satisfying* $\mathrm{E}\{\alpha(X_1; F)\} = 0$ *and* $\mathrm{Var}\{\alpha(X_1; F)\} = \sigma_\alpha^2 \in [0, \infty)$, *and $\beta(x, y; F)$ being symmetric in x and y such that* $\mathrm{E}\{\beta(X_1, X_2; F)|X_1\} = 0$ *and* $\mathrm{Var}\{\beta(X_1, X_2; F)\} = \sigma_\beta^2 \in [0, \infty)$. (ii) $\mathrm{E}|\alpha(X_1; F)|^{2+\delta} < \infty$ *and* $\mathrm{E}|\beta(X_1, X_2; F)|^{2+\delta'} < \infty$ *for some constants $0 \le \delta \le 1$ and $0 \le \delta' \le 1$.* (iii) *The distribution of $\alpha(X_1; F)$ is nonlattice and* $\mathrm{E}|\alpha(X_1; F)|^3 < \infty$.

CONDITION C2. (i) $\mathrm{E}(R_N) = b_1 N^{-\tau_1} + o(N^{-\tau_1})$ *and* $\mathrm{Var}(R_N) = O(N^{-\tau_2})$ *for some $b_1 \ne 0$, $\tau_1 \ge 1$ and $\tau_2 > 1$.* (ii) $\mathrm{P}(|R_N| \ge CN^{-\tau_3}) = o(N^{-\tau_4})$ *for some positive constant C, $\tau_3 > 1/2$ and $\tau_4 \ge 0$.*

The statistic T_N encompasses a wide class of statistics, for instance the U- and L-statistics, and the M-estimator (Jing and Wang (2010), Lahiri (1994), Lai and Wang (1993)). For a U-statistic, $\alpha(X_i; F)$ and $\beta(X_i, X_j; F)$ are determined by the first- and second-order terms

of the Hoeffding's decomposition (Serfling (1980)), and R_N satisfies Condition C2(i) with $\mathrm{E}(R_N) = 0$ and $\mathrm{Var}(R_N) = O(N^{-3})$. The linear term involving $\alpha(X_i; F)$ can vanish as the degenerate U-statistics. When the influence function of the M-estimator is twice differentiable and its second derivative is Lipschitz continuous, the M-estimator can be expressed as (2.1) with explicit α and β, and $R_N = O_p(N^{-1})$ satisfies Condition C2(ii) with $\tau_3 = 1$ and $\tau_4 = 1/2$ (Lahiri (1994), or Lemma S2.5 in the Supplementary Material). However, when the influence function is not smooth enough, the M-estimator may not be expanded to the second-order $\beta(X_i, X_j; F)$ term. In this case, we can absorb the quadratic term into R_N that satisfies Condition C2(ii) (He and Shao (1996), Volgushev, Chao and Cheng (2019)), which is often $O(N^{-3/4})$ almost surely.

To improve computation of T_N, we divide the full data $\mathfrak{X}_N$ into K data blocks. Let $\mathfrak{X}_{N,K}^{(k)} = \{X_{k,1}, \ldots, X_{k,n_k}\}$ be the kth data block of size n_k, for $k = 1, \ldots, K$. Such division is naturally available when $\mathfrak{X}_N$ is stored over K storage facilities. Otherwise, the blocks can be attained by random splitting.

Let $T_{N,K}^{(k)} = \theta + n_k^{-1} \sum_{i=1}^{n_k} \alpha(X_{k,i}; F) + n_k^{-2} \sum_{1 \le i < j \le n_k} \beta(X_{k,i}, X_{k,j}; F) + R_{N,K}^{(k)}$ that mimics (2.1) on $\mathfrak{X}_{N,K}^{(k)}$, where $R_{N,K}^{(k)} = R(\mathfrak{X}_{N,K}^{(k)}; F)$ is the remainder term specific to the kth block.

By averaging the K blockwise statistics, the distributed statistic

$$T_{N,K} = N^{-1} \sum_{k=1}^{K} n_k T_{N,K}^{(k)}, \tag{2.2}$$

which can be expressed as

$$T_{N,K} = \theta + N^{-1} \sum_{i=1}^{N} \alpha(X_i; F) + N^{-1} \sum_{k=1}^{K} n_k^{-1} \sum_{1 \le i < j \le n_k} \beta(X_{k,i}, X_{k,j}; F) + R_{N,K},$$

where $R_{N,K} = N^{-1} \sum_{k=1}^{K} n_k R_{N,K}^{(k)}$. It is clear that the difference between $T_{N,K}$ and T_N occurs at the terms involving β and the remainders.

While the SaC formulation is not new, our analysis on the general symmetric statistics contain fresh results for both nondegenerate ($\sigma_\alpha^2 > 0$) and degenerate ($\sigma_\alpha^2 = 0$) cases, and two distributed bootstrap algorithms to approximate the distribution of $T_{N,K}$ which can be used for inference purposes.

3. Statistical efficiency and asymptotic distributions. Our study on the statistical efficiency and asymptotic distributions of $T_{N,K}$ relative to those of T_N requires the following conditions.

CONDITION C3. *There exist positive constants c_1 and c_2 such that $c_1 \le \inf_{k_1,k_2} n_{k_1}/n_{k_2} \le \sup_{k_1,k_2} n_{k_1}/n_{k_2} \le c_2$, and K can be either finite or diverging to infinity as long as $K/N \to 0$ as $N \to \infty$.*

CONDITION C4. (i) *If $\mathrm{E}(R_N) = b_1 N^{-\tau_1} + o(N^{-\tau_1})$ and $\mathrm{Var}(R_N) = O(N^{-\tau_2})$ for some $b_1 \ne 0$, $\tau_1 \ge 1$ and $\tau_2 > 1$, then $\mathrm{E}(R_{N,K}^{(k)}) = b_{1,k} n_k^{-\tau_1} + o(n_k^{-\tau_1})$ for some $b_{1,k} \ne 0$ and $\mathrm{Var}(R_{N,K}^{(k)}) = O(n_k^{-\tau_2})$, for $k = 1, \ldots, K$.*

(ii) *If $\mathrm{P}(|R_N| \ge C N^{-\tau_3}) = o(N^{-\tau_4})$ for some positive constant C, $\tau_3 > 1/2$ and $\tau_4 \ge 0$, then $\mathrm{P}(|R_{N,K}^{(k)}| \ge C_k n_k^{-\tau_3}) = o(n_k^{-\tau_4})$ for some positive constant C_k, for $k = 1, \ldots, K$.*

Condition C3 assumes that $\{n_k\}_{k=1}^K$ are of the same order and $K = o(N)$. Condition C4 prescribes that $R_{N,K}^{(k)}$ inherits the properties of R_N with C4(i) and (ii) in the forms of the moments and probability, respectively.

The next theorem gives the biases and variances of T_N and $T_{N,K}$.

THEOREM 3.1. *Under* C1(i), C2(i), C3, C4(i) *with* $\tau_2 > 2$, $\mathrm{Bias}(T_N) = b_1 N^{-\tau_1} + o(N^{-\tau_1})$ *and* $\mathrm{Bias}(T_{N,K}) = N^{-1}\sum_{k=1}^K b_{1,k} n_k^{1-\tau_1} + o(K^{\tau_1} N^{-\tau_1})$. *In addition*,

$$\mathrm{Var}(T_N) = \sigma_\alpha^2 N^{-1} + 2^{-1}\sigma_\beta^2 N^{-2} + N^{-1}\sum_{i=1}^N \mathrm{Cov}\{\alpha(X_i; F), R_N\} + o(N^{-2}),$$

$$\begin{aligned}\mathrm{Var}(T_{N,K}) &= \sigma_\alpha^2 N^{-1} + 2^{-1}\sigma_\beta^2 K N^{-2} \\ &\quad + N^{-2}\sum_{k=1}^K n_k \sum_{i=1}^{n_k} \mathrm{Cov}\{\alpha(X_{k,i}; F), R_{N,K}^{(k)}\} + o(KN^{-2}).\end{aligned}$$

Theorem 3.1 implies that, if T_N is unbiased to θ, namely $\tau_1 = \infty$, $T_{N,K}$ is also unbiased. For the case of $\tau_1 < \infty$, as $N^{-1}\sum_{k=1}^K b_{1,k} n_k^{1-\tau_1}$ is of order $K^{\tau_1} N^{-\tau_1}$ under Condition C3, the bias is enlarged by a factor of K^{τ_1} for $T_{N,K}$ relative to that of T_N. For the variance of $T_{N,K}$, there is a factor K increase in the term involving $\sigma_\beta^2 N^{-2}$. While this increase has no leading order impact in the nondegenerate case ($\sigma_\alpha^2 > 0$), it becomes significant for the degenerate case ($\sigma_\alpha^2 = 0$) as the $O(N^{-1})$ terms and the covariance terms all vanish. This is another price paid for the distributed formulation. It is noted that the covariance terms are $o(N^{-3/2})$ for T_N and $o(K^{1/2} N^{-3/2})$ for $T_{N,K}$ in the nondegenerate case.

From Theorem 3.1, the mean square errors (MSEs) of T_N and $T_{N,K}$ are

$$\begin{aligned}\mathrm{MSE}(T_N) &= \sigma_\alpha^2 N^{-1} + 2^{-1}\sigma_\beta^2 N^{-2} + b_1^2 N^{-2\tau_1} \\ &\quad + N^{-1}\sum_{i=1}^N \mathrm{Cov}\{\alpha(X_i; F), R_N\} + o(N^{-2} + N^{-2\tau_1}) \quad \text{and}\end{aligned}$$

$$\begin{aligned}\mathrm{MSE}(T_{N,K}) &= \sigma_\alpha^2 N^{-1} + 2^{-1}\sigma_\beta^2 K N^{-2} + N^{-2}\left(\sum_{k=1}^K b_{1,k} n_k^{1-\tau_1}\right)^2 \\ &\quad + N^{-2}\sum_{k=1}^K n_k \sum_{i=1}^{n_k} \mathrm{Cov}\{\alpha(X_{k,i}; F), R_{N,K}^{(k)}\} + o(KN^{-2} + K^{2\tau_1} N^{-2\tau_1}).\end{aligned}$$

For the nondegenerate case of $\sigma_\alpha^2 > 0$, if

$$K = o(N^{1-1/(2\tau_1)}), \tag{3.1}$$

then $K^{2\tau_1} N^{-2\tau_1} = o(N^{-1})$, which means the increase in the bias is confined in the second order of the MSE as the variance inflation is of the second order as $K = o(N)$. For the degenerate case ($\sigma_\alpha^2 = 0$ but $\sigma_\beta^2 > 0$), the bias increase and the variance inflation of $T_{N,K}$ are the leading order events, which means that the distributed formulation cannot attain the efficiency as T_N, which becomes a bigger price paid for the computational scalability.

The following theorem indicates that T_N and $T_{N,K}$ share the same asymptotic distribution in the nondegenerate case.

THEOREM 3.2. *Suppose* C1(i), C3 *hold and* $\sigma_\alpha^2 > 0$, *then*:

(i) *if* $R_N = o_p(N^{-1/2})$, $N^{1/2}\sigma_\alpha^{-1}(T_N - \theta) \xrightarrow{d} \mathcal{N}(0,1)$ *as* $N \to \infty$;

(ii) *if* $R_{N,K} = o_p(N^{-1/2})$, $N^{1/2}\sigma_\alpha^{-1}(T_{N,K} - \theta) \xrightarrow{d} \mathcal{N}(0,1)$ *as* $N \to \infty$.

A key aspect of the result is in requiring $R_{N,K} = o_p(N^{-1/2})$, which is the case under Conditions C2(i) and C4(i) while $K = o(N^{1-1/(2\tau_1)})$. Alternatively, if Conditions C2(ii) and C4(ii) are satisfied, $R_{N,K} = o_p(N^{-1/2})$ is also guaranteed if $K = o(N^{1-1/(2\tau_3)})$ and $K = O(N^{1-1/(\tau_4+1)})$. These indicate that, when $\sigma_\alpha^2 > 0$, the smaller order R_N is, the higher K can be (and the more efficient with the computation) for $T_{N,K}$ to attain the same asymptotic normality as T_N. We note that $K = o(N^{1-1/(2\tau_1)})$ is just (3.1) for T_N and $T_{N,K}$ having the same leading order MSE.

Although T_N and $T_{N,K}$ share the same asymptotic distribution if $\sigma_\alpha^2 > 0$, a study on their higher order agreement requires Condition C1(iii) for the needed Edgeworth expansions.

THEOREM 3.3. *Suppose* C1(i) *and* (iii), C3 *hold and* $\sigma_\alpha^2 > 0$, $K = O(N^{\tau'})$ *for a positive constant* τ'.

(i) *Assume* C2(i) *and* C4(i), *and* $\tau' < 1 - 1/(2\tau_1)$, *then as* $N \to \infty$,

$$\begin{aligned}&\sup_{x\in\mathbf{R}}\left|\mathrm{P}\left\{N^{1/2}\sigma_\alpha^{-1}(T_{N,K}-\theta)\le x\right\} - \mathrm{P}\left\{N^{1/2}\sigma_\alpha^{-1}(T_N-\theta)\le x\right\}\right| \\ &\quad = O\left(N^{-\min\{\tau_1-\tau_1\tau'-1/2,(\tau_2-1)(1-\tau')/3,1/2\}}\right).\end{aligned} \tag{3.2}$$

In addition, if $\tau_1 > 1$, $\tau_2 > 5/2$ *and* $\tau' < \min\{1 - 1/\tau_1, 1 - 3/(2\tau_2 - 2), 1/2\}$, *the rate in* (3.2) *becomes* $o(N^{-1/2})$.

(ii) *Assume* C2(ii) *and* C4(ii), *and* $\tau' < \min(1-1/(2\tau_3), 1-1/(\tau_4+1))$, *then as* $N \to \infty$,

$$\begin{aligned}&\sup_{x\in\mathbf{R}}\left|\mathrm{P}\left\{N^{1/2}\sigma_\alpha^{-1}(T_{N,K}-\theta)\le x\right\} - \mathrm{P}\left\{N^{1/2}\sigma_\alpha^{-1}(T_N-\theta)\le x\right\}\right| \\ &\quad = O\left(N^{-\min\{\tau_3-\tau_3\tau'-1/2,\tau_4-\tau_4\tau'-\tau',1/2\}}\right).\end{aligned} \tag{3.3}$$

In addition, if $\tau_3 > 1$, $\tau_4 > 1/2$ *and* $\tau' < \min\{1 - 1/\tau_3, 1 - 3/(2\tau_4 + 2), 1/2\}$, *the rate in* (3.3) *becomes* $o(N^{-1/2})$.

Theorem 3.3 quantifies that the higher order difference between the distributions of the standardized T_N and $T_{N,K}$ depends on the orders of K and R_N. When R_N is small enough and K does not diverges too fast, the difference is of $o(N^{-1/2})$, which is assured by the standardized T_N and $T_{N,K}$ sharing the two leading order terms in the Edgeworth expansions (Lemmas S1.3 and S1.4 in the SM), despite $T_{N,K}$ contains less pairs of $\beta(X_i, X_j; F)$ than T_N. However, when T_N contains a bias term of order $N^{-\tau_1}$ with $\tau_1 \le 1$, it can be shown that the rate in (3.2) is bounded below by $K^{\tau_1}N^{1/2-\tau_1}$. This indicates that $\tau_1 > 1$ is necessary for the rate in (3.2) is $o(N^{-1/2})$.

REMARK 3.1. It is worth mentioning that the first part of the rate in (3.2) or (3.3), namely $O(N^{-(\tau_1-\tau_1\tau'-1/2)})$ or $O(N^{-(\tau_3-\tau_3\tau'-1/2)})$, is sharp in some cases. For the M-estimator $\hat\theta_N$, which solves $\sum_{i=1}^N \psi(X_i, \theta) = 0$ with θ_0 being the true value of θ, it can be expanded in the form of (2.1) with α and β defined in Proposition S2.1 of the SM under some regularity conditions on ψ. The corresponding reminder term satisfies C2(i) with $\tau_1 = 1$ and $\tau_2 = 3$, or C2(ii) with $\tau_3 = 1$ and $\tau_4 = 1/2$. Let $\hat\theta_{N,K}$ be the distributed version of $\hat\theta_N$. Then the leading order terms in the Edgeworth expansions of $\mathrm{P}\{N^{1/2}\sigma_\alpha^{-1}(\hat\theta_N - \theta_0) \le x\}$

and $\mathrm{P}\{N^{1/2}\sigma_\alpha^{-1}(\hat{\theta}_{N,K}-\theta_0)\le x\}$ are

$$F_{N,E1}(x)=\Phi(x)-N^{-1/2}\sigma_\alpha^{-1}\phi(x)\varpi(\theta_0)$$
$$-6^{-1}N^{-1/2}\sigma_\alpha^{-3}\Phi^{(3)}(x)\big[\mathfrak{K}_1(\theta_0)+6\{\varphi^{(1)}(\theta_0)\}^{-2}\varphi_2(\theta_0)\varpi(\theta_0)\big] \quad \text{and}$$
$$F_{N,E3}(x)=\Phi(x)-KN^{-1/2}\sigma_\alpha^{-1}\phi(x)\varpi(\theta_0)$$
$$-6^{-1}N^{-1/2}\sigma_\alpha^{-3}\Phi^{(3)}(x)\big[\mathfrak{K}_1(\theta_0)+6\{\varphi^{(1)}(\theta_0)\}^{-2}\varphi_2(\theta_0)\varpi(\theta_0)\big],$$

respectively. We refer to Proposition S2.1 in the SM for the definitions of ϖ, $\mathfrak{K}_1$, $\varphi^{(1)}$, and φ_2. So when $\varpi(\theta_0)\neq 0$, $\sup_{x\in\mathbf{R}}|\mathrm{P}\{N^{1/2}\sigma_\alpha^{-1}(\hat{\theta}_{N,K}-\theta_0)\le x\}-\mathrm{P}\{N^{1/2}\sigma_\alpha^{-1}(\hat{\theta}_N-\theta_0)\le x\}|$ is bounded below by a constant term of order $O(KN^{-1/2})$, which is of order $O(N^{-(1/2-\tau')})$ if $K=O(\tau')$ for $\tau'<1/2$. This lower bound is the same as $O(N^{-(\tau_1-\tau_1\tau'-1/2)})$ or $O(N^{-(\tau_3-\tau_3\tau'-1/2)})$ when $\tau_1=1$ or $\tau_3=1$.

For the degenerate case of $\sigma_\alpha^2=0$, $T_{N,K}$ cannot achieve the same efficiency as T_N according to Theorem 3.1. If β has a finite second moment, there exist sequences of eigenvalues $\{\lambda_\ell\}_{\ell=1}^\infty$ and eigenfunctions $\{\beta_\ell\}_{\ell=1}^\infty$ (Serfling (1980)) such that $\beta(x,y;F)=\sum_{\ell=1}^\infty\lambda_\ell\beta_\ell(x;F)\beta_\ell(y;F)$ in the sense that $\lim_{L\to\infty}\mathrm{E}\{|\beta(X_1,X_2;F)-\sum_{\ell=1}^L\lambda_\ell\beta_\ell(X_1;F)\beta_\ell(X_2;F)|^2\}=0$. The next theorem gives the asymptotic distributions of T_N and $T_{N,K}$ under degeneracy.

THEOREM 3.4. *Under* C1(i) *and* C3, $\sigma_\alpha^2=0$ *and* $\sigma_\beta^2>0$:

(i) *if* $R_N=o_p(N^{-1})$, *then as* $N\to\infty$, $2N(T_N-\theta)\xrightarrow{d}\sum_{\ell=1}^\infty\lambda_\ell(\chi_{1\ell}^2-1)$, *where* $\{\chi_{1\ell}^2\}_{\ell=1}^\infty$ *are independent* χ_1^2 *random variables*;

(ii) *if* K *is finite and* $R_{N,K}=o_p(N^{-1})$, $2N(T_{N,K}-\theta)\xrightarrow{d}\sum_{\ell=1}^\infty\lambda_\ell(\chi_{K\ell}^2-K)$ *as* $N\to\infty$, *where* $\{\chi_{K\ell}^2\}_{\ell=1}^\infty$ *are independent* χ_K^2 *random variables*;

(iii) *if* $K\to\infty$, C1(ii) *holds with* $0<\delta'<1$, *and* $R_{N,K}=o_p(K^{1/2}N^{-1})$, *then* $2^{1/2}\times K^{-1/2}N\sigma_\beta^{-1}(T_{N,K}-\theta)\xrightarrow{d}\mathcal{N}(0,1)$ *as* $N\to\infty$.

There are two limiting distributions for $T_{N,K}$ depending on whether K is finite or diverging. If K is finite, the limiting distribution of $T_{N,K}$ is a summation of K independent mixture of weighted chi-squares. If $K\to\infty$, $T_{N,K}$ is asymptotically normal as the chi-squares may be asymptotically normal when the degree of freedom diverges. Regarding the condition for $R_{N,K}$, if K is finite, that $R_{N,K}^{(k)}=o_p(n_k^{-1})$ for $k=1,\dots,K$ ensure $R_{N,K}=o_p(N^{-1})$. When $K\to\infty$, if $\mathrm{E}(R_N)=b_1N^{-\tau_1}+o(N^{-\tau_1})$ and $\mathrm{Var}(R_N)=o(N^{-2})$, then under Condition C4(i), it requires $K^{\tau_1}N^{-\tau_1}=o(K^{1/2}N^{-1})$, that is, $K=o(N^{1-1/(2\tau_1-1)})$ in order for $R_{N,K}=o_p(K^{1/2}N^{-1})$. This is a slower growth rate for K comparing to the nondegenerate case in (3.1). The distributed inference for the degenerate U-statistics is studied in Atta-Asiamah and Yuan (2019) with an emphasis on hypothesis testing. The asymptotic distribution of $T_{N,K}$ in the case of U-statistics that they derived when $K\to\infty$ is similar to Theorem 3.4(iii).

4. Distributed bootstrap. An important issue is how to approximate the distribution of $T_{N,K}$. This motivates our study of the bootstrap, which has been a powerful tool of statistical inference, especially in approximating distributions of statistics (Efron (1979), Hall (1992)).

A naive bootstrap proposal would randomly select K data subsets with replacement from the full sample to get the Monte Carlo versions of $T_{N,K}$. However, it is not computationally feasible for massive data in addition to require full sample communication. The bag of

little bootstrap (BLB) (Kleiner et al. (2014)) and the subsampled double bootstrap (SDB) (Sengupta, Volgushev and Shao (2016)) are two more variates of the bootstrap for distributed statistics. The BLB first generates data subsets of smaller sizes, say S subsets of size n by sampling from the original dataset $\mathfrak{X}_N$ without replacement. Then for each subset of size n, the BLB constructs B inflated resamples of full size N by repeated sampling with replacement from the subset. Finally, the BLB estimator is obtained by averaging over each small subset. Sengupta, Volgushev and Shao (2016) proposed the SDB that combines the idea of the BLB and a fast double bootstrap (Chang and Hall (2015), Davidson and MacKinnon (2002)). SDB first generates a large number (S) of random subsets of size n from the original full sample, which is similar to BLB's first step. For each small subset, SDB generates only one inflated resample of size N.

BLB and SDB carry out resampling from smaller size subsamples via sampling weights from the multinomial distributions. They have computational advantages when the underlying estimator admits a weighted empirical function representation. However, for nonlinear estimators like the symmetric statistics considered in this paper, the computational advantages via a weighted empirical function representation may not be available. As noted in Sengupta, Volgushev and Shao (2016), BLB uses a small number of subsets but a large number of resamples for each subset, which may lead to only a small portion of the full dataset being covered. The SDB, in its current form, cannot be implemented distributively as it conducts the first level resampling from the entire sample. The first level resampling can be made distributively, such that one can generate subsets from each data block. Simulations on this modification of the SDB can be found in the SM.

We propose two versions of distributed bootstrap, which overcome these issues in this and the next sections.

Given the K subsets $\mathfrak{X}_{N,K}^{(1)}, \ldots, \mathfrak{X}_{N,K}^{(K)}$ in the formulation of $T_{N,K}$, let $F_{N,K}^{(k)}$ be the empirical distribution of $\mathfrak{X}_{N,K}^{(k)}$, $\hat{\theta}_{N,K}^{(k)} = \theta(F_{N,K}^{(k)})$ be the analogy of θ under $\mathfrak{X}_{N,K}^{(k)}$, and $\hat{\theta}_{N,K} = N^{-1}\sum_{k=1}^{K} n_k \hat{\theta}_{N,K}^{(k)}$. Our aim is to avoid resampling from the entire dataset when estimating the distribution of $N^{1/2}(T_{N,K} - \theta)$.

To respect the distributive nature of $T_{N,K}$, we propose resampling within each data subset. Let $\mathfrak{X}_{N,K}^{*(k)} = \{X_{k,1}^*, \ldots, X_{k,n_k}^*\}$ be an independent and identically distributed (i.i.d.) sample from $F_{N,K}^{(k)}$. Repeating it B times, one obtains B resampled subsets $\mathfrak{X}_{N,K}^{*1(k)}, \ldots, \mathfrak{X}_{N,K}^{*B(k)}$. Compute the corresponding statistic $T_{N,K}^{*b(k)} = T(\mathfrak{X}_{N,K}^{*b(k)})$, and average them over the K subsets to get the bth copy of the bootstrap distributed statistics

$$T_{N,K}^{*b} = N^{-1}\sum_{k=1}^{K} n_k T_{N,K}^{*b(k)} \quad \text{for } b = 1, \ldots, B.$$

Then the empirical distribution of $\{N^{1/2}(T_{N,K}^{*b} - \hat{\theta}_{N,K})\}_{b=1}^{B}$ is used to approximate the distribution of $N^{1/2}(T_{N,K} - \theta)$.

We call the procedure the distributed bootstrap (DB) as the resampling is conducted within each data subset without the sample inflation as BLB. For each data subset in each iteration, we can calculate $T_{N,K}^{*b(k)}$ and $\hat{\theta}_{N,K}^{(k)}$ locally avoiding data communication among data subsets.

To explore the theoretical properties of the DB, we need some notation and assumptions. Specifically, $T_{N,K}^{*(k)} = T(\mathfrak{X}_{N,K}^{*(k)})$ is assumed to admit

$$\begin{aligned} T_{N,K}^{*(k)} = \hat{\theta}_{N,K}^{(k)} &+ n_k^{-1}\sum_{i=1}^{n_k} \hat{\alpha}(X_{k,i}^*; F_{N,K}^{(k)}) \\ &+ n_k^{-2} \sum_{1 \le i < j \le n_k} \hat{\beta}(X_{k,i}^*, X_{k,j}^*; F_{N,K}^{(k)}) + R_{N,K}^{*(k)}, \end{aligned} \tag{4.1}$$

where $\hat{\theta}_{N,K}^{(k)} = \theta(F_{N,K}^{(k)})$ and $R_{N,K}^{*(k)} = R(\mathfrak{X}_{N,K}^{*(k)}; F_{N,K}^{(k)})$ are analogues of θ and $R_{N,K}^{(k)}$ under $F_{N,K}^{(k)}$. Then $T_{N,K}^{*} = N^{-1}\sum_{k=1}^{K} n_k T_{N,K}^{*(k)}$ can be written as

$$
\begin{aligned}
T_{N,K}^{*} &= \hat{\theta}_{N,K} + N^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\hat{\alpha}\big(X_{k,i}^{*}; F_{N,K}^{(k)}\big) \\
&\quad + N^{-1}\sum_{k=1}^{K} n_k^{-1}\sum_{1\le i<j\le n_k}\hat{\beta}\big(X_{k,i}^{*}, X_{k,j}^{*}; F_{N,K}^{(k)}\big) + R_{N,K}^{*},
\end{aligned}
\tag{4.2}
$$

where $\hat{\theta}_{N,K} = N^{-1}\sum_{k=1}^{K} n_k\hat{\theta}_{N,K}^{(k)}$ and $R_{N,K}^{*} = N^{-1}\sum_{k=1}^{K} n_k R_{N,K}^{*(k)}$. Here, $\{\hat{\alpha}(x; F_{N,K}^{(k)})\}_{k=1}^{K}$ and $\{\hat{\beta}(x, y; F_{N,K}^{(k)})\}_{k=1}^{K}$ are the empirical versions of $\alpha(x; F)$ and $\beta(x, y; F)$, which we regulate in the following condition.

CONDITION C5. *For $k = 1, \ldots, K$, $\sum_{i=1}^{n_k}\hat{\alpha}(X_{k,i}; F_{N,K}^{(k)}) = 0$, $\hat{\beta}(x, y; F_{N,K}^{(k)})$ is symmetric in x and y, $\sum_{i=1}^{n_k}\hat{\beta}(X_{k,i}, y; F_{N,K}^{(k)}) = 0$ for any $y \in S(F)$, the support of F. In addition, as $n_k \to \infty$, $\sup_{x\in S(F)}|\hat{\alpha}(x; F_{N,K}^{(k)}) - \alpha(x; F)| = o_p(1)$ and $\sup_{x,y\in S(F)}|\hat{\beta}(x, y; F_{N,K}^{(k)}) - \beta(x, y; F)| = o_p(1)$.*

Condition C5 indicates that for $X_{k,i}^{*}$ with distribution $F_{N,K}^{(k)}$, $k = 1, \ldots, K$, $\mathrm{E}\{\hat{\alpha}(X_{k,i}^{*}; F_{N,K}^{(k)})|F_{N,K}^{(k)}\} = 0$ and $\mathrm{E}\{\hat{\beta}(X_{k,i}^{*}, y; F_{N,K}^{(k)})|F_{N,K}^{(k)}\} = 0$ for any $y \in S(F)$. Moreover, it requires that $\{\hat{\alpha}(x; F_{N,K}^{(k)})\}_{k=1}^{K}$ and $\{\hat{\beta}(x, y; F_{N,K}^{(k)})\}_{k=1}^{K}$ are uniformally consistent to $\alpha(x; F)$ and $\beta(x, y; F)$. Similar conditions are assumed in Lai and Wang (1993).

REMARK 4.1. Under C1(i) that $\mathrm{E}\{\alpha(X_1; F)\} = 0$, the first part of C5 with $\hat{\alpha}(x; F_{N,K}^{(k)}) = \alpha(x; F_{N,K}^{(k)})$ is satisfied approximately in large samples via the conditional law of large numbers. To make $\sum_{i=1}^{n_k}\hat{\alpha}(X_{k,i}; F_{N,K}^{(k)}) = 0$ satisfied exactly, we can centralize $\alpha(x; F_{N,K}^{(k)})$ by the empirical subsample mean $n_k^{-1}\sum_{i=1}^{n_k}\alpha(X_{k,i}; F_{N,K}^{(k)})$, namely to choose $\hat{\alpha}(x; F_{N,K}^{(k)}) = \alpha(x; F_{N,K}^{(k)}) - n_k^{-1}\sum_{i=1}^{n_k}\alpha(X_{k,i}; F_{N,K}^{(k)})$. It follows that $\mathrm{E}\{\hat{\alpha}(X_{k,i}^{*}; F_{N,K}^{(k)})|F_{N,K}^{(k)}m\} = 0$, which is sufficient for Theorem 4.1 presented below. Similar arguments can be made on the assumption that $\sum_{i=1}^{n_k}\hat{\beta}(X_{k,i}, y; F_{N,K}^{(k)}) = 0$ for any $y \in S(F)$.

REMARK 4.2. According to the proof of Theorem 4.1 in the SM, that $\sup_{x\in S(F)}|\hat{\alpha}(x; F_{N,K}^{(k)}) - \alpha(x; F)| = o_p(1)$ and $\sup_{x,y\in S(F)}|\hat{\beta}(x, y; F_{N,K}^{(k)}) - \beta(x, y; F)| = o_p(1)$ in Condition C5 can be replaced by the following assumptions: there exists a constant positive δ such that $\mathrm{E}\{|\hat{\alpha}(X_{k,i}^{*}; F_{N,K}^{(k)})|^{2+\delta}|F_{N,K}^{(k)}\} < \infty$ and $\mathrm{E}\{|\hat{\beta}(X_{k,i}^{*}, X_{k,j}^{*}; F_{N,K}^{(k)})|^{2}|F_{N,K}^{(k)}\} < \infty$ in probability for $k = 1, \ldots, K$. In addition, $N^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_k}[\{\hat{\alpha}(X_{k,i}; F_{N,K}^{(k)})\}^2 - \{\alpha(X_{k,i}; F)\}^2] = o_p(1)$.

Additional conditions are needed when establishing more accurate approximation result on the DB with Condition C6 extending that in C1(iii).

CONDITION C6. *Conditional on $F_{N,K}^{(k)}$, the distribution of $\hat{\alpha}(X_{N,K}^{*}; F_{N,K}^{(k)})$ is nonlattice almost surely and $\sup_{x\in S(F)}|\hat{\alpha}(x; F_{N,K}^{(k)}) - \alpha(x; F)| = O_p(n_k^{-1/2})$ for $k = 1, \ldots, K$.*

The next theorem establishes the consistency and approximation accuracy of the DB for the nondegenerated case of $\sigma_a^2 > 0$.

THEOREM 4.1. *Assume* $\sigma_\alpha^2 > 0$, *under Condition* C1(i) *and* (ii), C3 *and* C5 *with* $\delta > 0$ *and* $\delta' = 0$, $K = O(N^{\tau'})$ *for a positive constant* τ', *and* $R_{N,K}^*$ *in* (4.2) *satisfies* $\mathrm{P}\{|R_{N,K}^*| \geq N^{-1/2}(\ln N)^{-1}|F_{N,K}^{(1)}, \ldots, F_{N,K}^{(K)}\} = o_p(1)$.

(i) *Suppose* C2(i), C4(i), *and* $\tau' < 1 - 1/(2\tau_1)$, *then as* $N \to \infty$,

$$\sup_{x\in\mathbf{R}}\big|\mathrm{P}\big\{N^{1/2}\big(T_{N,K}^* - \hat{\theta}_{N,K}\big) \leq x|F_{N,K}^{(1)}, \ldots, F_{N,K}^{(K)}\big\} - \mathrm{P}\big\{N^{1/2}(T_{N,K} - \theta) \leq x\big\}\big| = o_p(1). \tag{4.3}$$

In addition, assume C1(iii) *and* C6, $\delta = 1$, $\tau_2 > 5/2$ *and* $\tau' < \min\{1 - 1/(2\tau_1 - 1), 1/2\}$, $R_{N,K}^*$ *satisfies* $\mathrm{P}\{|R_{N,K}^*| \geq K^{1/2}N^{-1}|F_{N,K}^{(1)}, \ldots, F_{N,K}^{(K)}\} = O_p(K^{1/2}N^{-1/2})$, *then as* $N \to \infty$, $o_p(1)$ *in* (4.3) *becomes* $O_p(K^{1/2}N^{-1/2})$.

(ii) *Suppose* C2(ii), C4(ii), *and* $\tau' < \min(1 - 1/(2\tau_3), 1 - 1/(\tau_4 + 1))$, *then* (4.3) *holds as* $N \to \infty$. *In addition, suppose* C1(iii) *and* C6, $\delta = 1$, $\tau_3 > 1$ *and* $\tau_4 > 1/2$ *and* $\tau' < \min\{1 - 1/(2\tau_3 - 1), 1 - 2/(2\tau_4 + 1), 1/2\}$, $R_{N,K}^*$ *satisfies* $\mathrm{P}\{|R_{N,K}^*| \geq K^{1/2}N^{-1}|F_{N,K}^{(1)}, \ldots, F_{N,K}^{(K)}\} = O_p(K^{1/2}N^{-1/2})$, *then as* $N \to \infty$, $o_p(1)$ *in* (4.3) *becomes* $O_p(K^{1/2}N^{-1/2})$.

Theorem 4.1 provides the accuracy of the DB approximation to the distribution of the distributed statistics for the nondegenerate case. The reason for directly imposing condition on the resampled quantities $R_{N,K}^*$ is due to the implicit nature of the $R_{N,K}$ in the symmetric statistic formulation, which is not unusual as it is also conducted in Lai and Wang (1993). For a specific statistic, we need to check if $T_{N,K}^*$ satisfies the conditions in Theorem 4.1. Section S2.1 and S2.2 in the SM provide details on the conditions required for the U-statistics and M-estimators. Theorem 4.1 also indicates that by adding extra conditions, the DB's approximation accuracy is improved from $o_p(1)$ to $O_p(K^{1/2}N^{-1/2})$; see Section S2.1 in the SM for the U-statistics.

Theorem 4.1 ensures that the DB can be used by combining with the continuous mapping theorem and the delta method for inference purposes, for instance in constructing confidence intervals (CIs). Specifically, denote u_τ^* as the sample τth quantile of $\{N^{1/2}(T_{N,K}^{*b} - \hat{\theta}_{N,K})\}_{b=1}^B$, then an equal-tail two-sided CI for θ with confidence level $1 - \tau$ can be constructed as

$$\big(T_{N,K} - N^{-1/2}u_{1-\tau/2}^*,\ T_{N,K} - N^{-1/2}u_{\tau/2}^*\big). \tag{4.4}$$

The bootstrap resample from the DB algorithm can be also used to estimate $\mathrm{Var}(T_{N,K})$. Let $\hat{\sigma}_{DB}^2 = B^{-1}\sum_{b=1}^B (T_{N,K}^{*b} - B^{-1}\sum_{\ell=1}^B T_{N,K}^{*\ell})^2$. The following theorem shows the consistency of $\hat{\sigma}_{DB}^2$ to $\mathrm{Var}(T_{N,K})$.

THEOREM 4.2. *Under the conditions of Theorem* 4.1(i), *assume that* $\mathrm{E}|\beta(X_1, X_i; F)|^{2+\delta'} < \infty$ *with* $0 < \delta' < 1$ *for* $X_i = X_1$ *and* X_2, *respectively, and* $\mathrm{E}\{|N^{1/2}R_{N,K}^*|^{2+\delta} \times |F_{N,K}^{(1)}, \ldots, F_{N,K}^{(K)}\} < \infty$. *Then* $\hat{\sigma}_{DB}^2/\mathrm{Var}(T_{N,K}) \to 1$ *in probability as* $N \to \infty$.

For the degenerate case, a key in the bootstrap formulation is to preserve the degeneracy in the bootstrap resamples (Arcones and Giné (1992)). Under C5, $\mathrm{E}\{\hat{\beta}(X_{k,i}^*, y; F_{N,K}^{(k)})|F_{N,K}^{(k)}\} = 0$ for any $y \in S(F)$, which indicates that $\hat{\beta}(x, y; F_{N,K}^{(k)})$ is degenerate conditional on $F_{N,K}^{(k)}$. Motivated by the bootstrap for degenerate U-statistics in Arcones and Giné (1992), the DB can be adapted by replacing $(T_{N,K}^{*(k)} - \hat{\theta}_{N,K}^{(k)})$ with $n_k^{-2}\sum_{1\leq i<j\leq n_k}\hat{\beta}(X_{k,i}^*, X_{k,j}^*; F_{N,K}^{(k)})$ for each data block. The following theorem establishes the consistency of the adapted DB for the degenerate case when $K \to \infty$.

THEOREM 4.3. *Assume* $\sigma_\alpha^2 = 0$ *and* $\sigma_\beta^2 > 0$, *under* C1(i) *and* (ii), C3 *and* C5 *with* $\delta' > 0$, $K \to \infty$ *and* $K = O(N^{\tau'})$ *for a positive constant* τ'.

(i) *Suppose* C2(i), C4(i) *with* $\tau_1 > 1$ *and* $\tau_2 > 2$, *and* $\tau' < 1 - 1/(2\tau_1 - 1)$, *then as* $K \to \infty$,

$$\sup_{x\in\mathbf{R}}\left|\mathrm{P}\left\{2^{1/2}K^{-1/2}\sum_{k=1}^{K} n_k^{-1}\sum_{1\le i<j\le n_k}\hat{\beta}(X_{k,i}^*, X_{k,j}^*; F_{N,K}^{(k)}) \le x \mid F_{N,K}^{(1)}, \ldots, F_{N,K}^{(K)}\right\}\right. \tag{4.5}$$
$$\left. - \mathrm{P}\{2^{1/2}K^{-1/2}N(T_{N,K} - \theta) \le x\}\right| = o_p(1).$$

(ii) *Suppose* C2(ii), C4(ii) *with* $\tau_3 > 1$ *and* $\tau_4 > 0$, *and* $\tau' < \min(1 - 1/(2\tau_3 - 1), 1 - 1/(\tau_4 + 1))$, *then* (4.5) *also holds as* $K \to \infty$.

Theorem 4.3 ensures that, when $K \to \infty$, the conditional distribution of $2^{1/2}K^{-1/2} \times \sum_{k=1}^{K} n_k^{-1}\sum_{1\le i<j\le n_k}\hat{\beta}(X_{k,i}^*, X_{k,j}^*; F_{N,K}^{(k)})$ is consistent to that of $2^{1/2}K^{-1/2}N(T_{N,K} - \theta)$. When K is fixed, the conditional distribution of $2\sum_{k=1}^{K} n_k^{-1}\sum_{1\le i<j\le n_k}\hat{\beta}(X_{k,i}^*, X_{k,j}^*; F_{N,K}^{(k)})$ may be used to estimate that of $2N(T_{N,K} - \theta)$. The corresponding theoretical analysis requires the eigen-decomposition of $\hat{\beta}$, which does not have a general form for the symmetric statistic. Thus, for the case of K being finite, we only provide the consistency of the adapted DB for the U-statistics in Theorem S2.4(ii) of the SM.

5. Pseudo-distributed bootstrap. Although the DB leads to substantial computational saving, it is still computationally involved as the distributed statistics need to be recalculated for each bootstrap replication. To further reduce the computational burden, we consider another way to approximate the distribution of $T_{N,K}$ for the case of diverging K.

The idea comes from the expression

$$T_{N,K} = N^{-1}\sum_{k=1}^{K} n_k T_{N,K}^{(k)} = K^{-1}\sum_{k=1}^{K}(Kn_k/N)T_{N,K}^{(k)}.$$

Hence, when K is large, approximating the distribution of $T_{N,K}$ is similar to that of the sample mean of independent but not necessary identically distributed data. This leads us to propose directly resampling $\{T_{N,K}^{(k)}\}_{k=1}^{K}$.

5.1. *PDB for nonstudentized distributed statistics.* We first consider the nondegenerate case. Due to different subset sizes, we need to scale $\{T_{N,K}^{(k)}\}_{k=1}^{K}$ before the resampling. Let $\mathcal{T}_{N,K}^{(k)} = N^{-1/2}K^{1/2}n_k T_{N,K}^{(k)}$ for $k = 1, \ldots, K$, and $F_{K,\mathcal{T}}$ be the empirical distribution of $\{\mathcal{T}_{N,K}^{(k)}\}_{k=1}^{K}$. Suppose $\{\mathcal{T}_{N,K}^{*(k)}\}_{k=1}^{K}$ is an i.i.d. sample from $F_{K,\mathcal{T}}$ and $\mathcal{T}_{N,K}^{*} = K^{-1}\sum_{k=1}^{K}\mathcal{T}_{N,K}^{*(k)}$. Then the distribution of $K^{1/2}(\mathcal{T}_{N,K}^{*} - N^{1/2}K^{-1/2}T_{N,K})$ conditional on $F_{K,\mathcal{T}}$ is used to estimate that of $N^{1/2}(T_{N,K} - \theta)$. We call this the pseudo-distributed bootstrap (PDB).

The PDB is the bootstrap on the scaled $\{\mathcal{T}_{N,K}^{(k)}\}_{k=1}^{K}$, which are independent but not necessarily identically distributed. The bootstrap under non-i.i.d. models has been studied in Liu (1988). The PDB is similar to Volgushev, Chao and Cheng (2019)'s proposal of a weighted bootstrap algorithm that resamples the subsample estimators for quantile regression. The following theorem establishes the asymptotic properties of the PDB for the nondegenerate case.

THEOREM 5.1. (*Nondegenerate case*) *Under Condition* C1(i), C2(i), C3, C4(i) *and* $\sigma_\alpha^2 > 0$; $K = O(N^{\tau'})$ *for the* τ' *specified below.*

(i) *Assume* $\sup_k N^{-1/2}K^{1/2}|n_k - NK^{-1}| \to 0$ *and* $\tau' < 1 - 1/(2\tau_1)$; C1(ii) *holds with* $0 < \delta, \delta' < 1$ *and* $\sup_k \mathrm{E}|n_k^{1/2}R_{N,K}^{(k)}|^{2+\delta} < \infty$. *Then, as* $K \to \infty$,

$$\sup_{x\in\mathbf{R}}\big|\mathrm{P}\big\{K^{1/2}\big(\mathcal{T}_{N,K}^{*} - N^{1/2}K^{-1/2}T_{N,K}\big) \le x|F_{K,\mathcal{T}}\big\} - \mathrm{P}\big\{N^{1/2}(T_{N,K} - \theta) \le x\big\}\big| = o_p(1). \tag{5.1}$$

(ii) *Assume* $n_k = NK^{-1}$ *for* $k = 1, \ldots, K$ *and* $\tau' < 1 - 2/(2\tau_1 + 1)$; C1(ii) *holds with* $\delta = \delta' = 1$ *and* $\sup_k \mathrm{E}|n_k^{1/2}R_{N,K}^{(k)}|^3 < \infty$; *the distribution of* $T_{N,K}^{(1)}$ *is nonlattice. Then as* $K \to \infty$, *the* $o_p(1)$ *in* (5.1) *becomes* $O_p(K^{-1/2})$.

Theorem 5.1(i) shows that under moderate conditions on n_k, K and the moments of $T_{N,K}^{(k)}$, the PDB offers consistent approximation to the distribution of $N^{1/2}(T_{N,K} - \theta)$. By imposing stronger conditions on n_k, K and $T_{N,K}^{(k)}$, Theorem 5.1(ii) indicates that the approximation accuracy of the PDB can be improved to $O_p(K^{-1/2})$. As the resampling of PDB is on the pseudo sample $\{\mathcal{T}_{N,K}^{(k)}\}_{k=1}^{K}$, which is of size K, the accuracy of approximation of the PDB is of an order that depend on K, rather than N. Compared to the DB, besides a large computational saving, an appealing property of the PDB is its avoiding (4.2) required for the DB in Theorem 4.1. This makes the PDB more versatile and easier to implement despite requiring $K \to \infty$.

The PDB is also workable for the degenerate case. Indeed, when $\sigma_\alpha^2 = 0$ but $\sigma_\beta^2 > 0$, $T_{N,K}$ is still an average of K independent random variables. One only needs to rescale $\{T_{N,K}^{(k)}\}_{k=1}^{K}$ with a different scaling factor such that $\mathcal{T}_{N,K}^{(k)} = n_k T_{N,K}^{(k)}$ followed by the same resampling procedures for the nondegenerate case. Suppose that $\mathcal{T}_{N,K}^{*(1)}, \ldots, \mathcal{T}_{N,K}^{*(K)}$ is an i.i.d. sample from $F_{K,\mathcal{T}}$ and denote $\mathcal{T}_{N,K}^{*} = K^{-1}\sum_{k=1}^{K}\mathcal{T}_{N,K}^{*(k)}$.

THEOREM 5.2. (*Degenerate case*) *If* $\sigma_\alpha^2 = 0$ *but* $\sigma_\beta^2 > 0$, *under Condition* C1(i), C2(i) *and* C4(i) *with* $\tau_1 > 1$, $\tau_2 > 2$; *assume* $n_k = NK^{-1}$ *for* $k = 1, \ldots, K$ *and* $K = O(N^{\tau'})$ *for a positive constant* τ' *specified as below.*

(i) *Assume* $\tau' < 1 - 1/(2\tau_1 - 1)$ *and* C1(ii) *holds with* $0 < \delta, \delta' < 1$ *and* $\sup_k \mathrm{E}|n_k R_{N,K}^{(k)}|^{2+\delta} < \infty$, *then as* $K \to \infty$,

$$\sup_{x\in\mathbf{R}}\big|\mathrm{P}\big\{K^{1/2}\big(\mathcal{T}_{N,K}^{*} - NK^{-1}T_{N,K}\big) \le x|F_{K,\mathcal{T}}\big\} - \mathrm{P}\big\{NK^{-1/2}(T_{N,K} - \theta) \le x\big\}\big| = o_p(1). \tag{5.2}$$

(ii) *Assume* $\tau' < 1 - 1/\tau_1$, *the distribution of* $T_{N,K}^{(1)}$ *is nonlattice*, C1(ii) *holds with* $\delta = \delta' = 1$ *and* $\sup_k \mathrm{E}|n_k R_{N,K}^{(k)}|^3 < \infty$, *then as* $K \to \infty$, *the* $o_p(1)$ *in* (5.2) *becomes* $O_p(K^{-1/2})$.

Comparing the degenerate case to the nondegenerate case, stronger conditions are needed for the degenerate case to control $R_{N,K}^{(k)}$. This is reflected first in that τ_1 needs to be strictly larger than 1 such that $R_{N,K}$ can be dominated by the quadratic term involving β. Second, n_k is assumed to be the same for all subsets and K has a slower growth rate to N. Finally, a stronger moment condition is needed for $\{R_{N,K}^{(k)}\}_{k=1}^{K}$. In addition, under the stronger conditions in (ii), approximation accuracy to the order of $O_p(K^{-1/2})$ is attained as in the nondegenerate case. Despite these, PDB for both the nondegenerate and degenerate cases offers easier implementation than the DB as it avoids (4.2) and offers faster computation with reasonable theoretical properties.

5.2. *PDB for studentized distributed statistics.* The PDB introduced in Section 5.1 is conducted on the statistics $N^{1/2}(T_{N,K}-\theta)$ or $NK^{-1/2}(T_{N,K}-\theta)$, which are not studentized. Given the more accurate approximation offered by the percentile-t bootstrap (Hall (1992)), we consider implementing the PDB on studentized statistics in this subsection.

For $K\to\infty$, a straightforward estimator of $\mathrm{Var}(T_{N,K})$ is the sample variance of the pseudo sample $\{N^{-1}Kn_kT_{N,K}^{(k)}\}_{k=1}^K$ as considered in Volgushev, Chao and Cheng (2019). Define a pseudo-sample variance estimator

$$S_K^2=(K-1)^{-1}\sum_{k=1}^K\big(N^{-1}Kn_kT_{N,K}^{(k)}-T_{N,K}\big)^2. \tag{5.3}$$

PROPOSITION 5.1. (i) *For the case of $\sigma_\alpha^2>0$, under the conditions assumed in Theorem* 5.1(i), $K^{-1}NS_K^2\to\sigma_\alpha^2$ *almost surely as* $K\to\infty$.

(ii) *When $\sigma_\alpha^2=0$ but $\sigma_\beta^2>0$, assume the conditions in Theorem* 5.2(i) *hold,* $K^{-2}N^2S_K^2\to 2^{-1}\sigma_\beta^2$ *almost surely as* $K\to\infty$.

As Proposition 5.1 indicates that $K^{-1}S_K^2$ is a consistent estimator of $\mathrm{Var}(T_{N,K})$ for both the nondegenerate and degenerate case, we can attain the studentized statistic $K^{1/2}S_K^{-1}\times(T_{N,K}-\theta)$.

REMARK 5.1. As S_K^2 is the sample variance of the pseudo sample, its estimation accuracy depends on K. So when K is not sufficient large, despite $K^{-1}S_K^2$ is consistent for $\mathrm{Var}(T_{N,K})$, it may not be competitive to the variance estimator obtained by the DB. Furthermore, using a very large K can lead to a nonignorable loss in the statistical efficiency of $T_{N,K}$ when compared to T_N. Numerical results illustrate this aspect are available in the SM.

Incorporating Proposition 5.1 with Theorems 3.2 and 3.4, we have the following results on the asymptotic distributions of studentized $T_{N,K}$.

THEOREM 5.3. *Under the conditions of Theorem* 5.1(i) *or Theorem* 5.2(i), $K^{1/2}S_K^{-1}\times(T_{N,K}-\theta)\xrightarrow{d}\mathcal{N}(0,1)$ *as* $K\to\infty$.

Based on Theorem 5.3, a Wald-type confidence interval of θ based on $T_{N,K}$ and S_K^2 can be established as

$$\big(T_{N,k}-z_{1-\tau/2}K^{-1/2}S_K,\,T_{N,k}+z_{1-\tau/2}K^{-1/2}S_K\big), \tag{5.4}$$

where $z_{1-\tau/2}$ is the $(1-\tau/2)$th upper quantile of $N(0,1)$. See Section 4.1 of Volgushev, Chao and Cheng (2019) for an implementation under the quantile regression scenario.

Next, we propose a PDB algorithm for the studentized distributed statistics. The procedure is similar to the PDB for the nonstudentized distributed statistic in Section 5.1, the only adjustment is in the studentization for each bootstrap pseudo sample. Denote

$$\mathbb{T}_{N,K}^{(k)}=N^{-1}Kn_kT_{N,K}^{(k)}\quad\text{for } k=1,\dots,K.$$

Let $F_{K,\mathbb{T}}$ be the empirical distribution of $\{\mathbb{T}_{N,K}^{(k)}\}_{k=1}^K$. Suppose $\{\mathbb{T}_{N,K}^{*(k)}\}_{k=1}^K$ is an i.i.d, sample from $F_{K,\mathbb{T}}$. Let $\mathbb{T}_{N,K}^*=K^{-1}\sum_{k=1}^K\mathbb{T}_{N,K}^{*(k)}$ and $S_K^*=\{(K-1)^{-1}\sum_{k=1}^K(\mathbb{T}_{N,K}^{*(k)}-\mathbb{T}_{N,K}^*)^2\}^{1/2}$, then the distribution of $K^{1/2}\{S_K^*\}^{-1}(\mathbb{T}_{N,K}^*-T_{N,K})$ conditional on $F_{K,\mathbb{T}}$ is used to estimate that of $K^{1/2}S_K^{-1}(T_{N,K}-\theta)$.

THEOREM 5.4. *Under the conditions of Theorem* 5.1(i) *or Theorem* 5.2(i), *as* $K \to \infty$,

$$\sup_{x\in\mathbf{R}}\big|\mathrm{P}\big\{K^{1/2}\big\{S_K^*\big\}^{-1}\big(\mathbb{T}_{N,K}^* - T_{N,K}\big) \le x | F_{K,\mathbb{T}}\big\} - \mathrm{P}\big\{K^{1/2}S_K^{-1}(T_{N,K}-\theta)\le x\big\}\big| = o_p(1).$$

Theorem 5.4 indicates that the PDB works for the studentized distributed statistics for both nondegenerate and degenerate cases.

Compared to the normal approximation, studentization in the conventional bootstrap can correct the first term in the Edgeworth expansion (Hall (1992)). We have similar results for the PBD.

THEOREM 5.5. *Assume* $n_k = NK^{-1}$ *for* $k = 1, \ldots, K$ *and the distribution of* $\mathbb{T}_{N,K}^{(k)}$ *is nonlattice. In addition, we assume*:

(a) *if* $\sigma_\alpha^2 > 0$, $\mathrm{E}|n_k^{1/2}\mathbb{T}_{N,K}^{(k)}|^3 < \infty$; (b) *if* $\sigma_\alpha^2 = 0$ *but* $\sigma_\beta^2 > 0$, $\mathrm{E}|n_k\mathbb{T}_{N,K}^{(k)}|^3 < \infty$. *Then, as* $K \to \infty$,

$$\sup_{x\in\mathbf{R}}\big|\mathrm{P}\big\{K^{1/2}S_K^{-1}(T_{N,K}-\theta)\le x\big\} - \Phi(x)\big| = O_p\big(K^{-1/2}\big) \quad \textit{and}$$

$$\sup_{x\in\mathbf{R}}\big|\mathrm{P}\big\{K^{1/2}\big\{S_K^*\big\}^{-1}\big(\mathbb{T}_{N,K}^* - T_{N,K}\big) \le x | F_{K,\mathcal{T}}\big\} - \mathrm{P}\big\{K^{1/2}S_K^{-1}(T_{N,K}-\theta)\le x\big\}\big| = o_p\big(K^{-1/2}\big).$$

Theorem 5.5 maintains that the PDB for the studentized distributed statistics offers more accurate ($o_p(K^{-1/2})$) distributional approximation than that of the nonstudentized PDB. The approximation error $o_p(K^{-1/2})$ can be made to $O_p(K^{-1})$ if we impose stronger moment conditions on $\mathbb{T}_{N,K}^{(k)}$.

6. Simulation studies. In this section, we use Gini's mean difference as a nondegenerate case and the distributed version of the distance covariance (Székely, Rizzo and Bakirov (2007)) as a degenerate example to demonstrate the empirical performance. All the simulations were conducted in R with a single Intel(R) Core(TM) i7 4790K @4.0 GHz processor.

6.1. *Nondegenerate case.* First, we use Gini's mean difference to compare the proposed bootstrap methods to the BLB and SDB. More simulations results on the proposed distributed approaches can be found in Section S4.1 in the SM. All the simulation results were based on 2000 replications.

The Gini's mean difference is $U_N = 2\{N(N-1)\}^{-1}\sum_{1\le i<j\le N}|X_i - X_j|$. It is an unbiased estimator of the dispersion parameter $\theta_U = \mathrm{E}|X_i - X_j|$, and is a U-statistic of degree two. Suppose the full data are divided into K blocks with the kth of size n_k, and let $U_{N,K}^{(k)}$ be the Gini's difference from the kth data block. Then the distributed estimator is $U_{N,K} = N^{-1}\sum_{k=1}^K n_k U_{N,K}^{(k)}$.

Three distributions of X_i were experimented: (I) $\mathcal{N}(1,1)$, (II) Gamma(3, 1) and (III) Poisson(4). The sample size $N = 100{,}000$. For convenience, we randomly divided the dataset into blocks of equal sizes. In the simulations, we constructed the 95% equal-tailed confidence intervals for θ based on $U_{N,K}$ with $K \in \mathcal{K} = \{5, 10, 20, 50, 100, 200, 500, 1000\}$. For each simulated dataset and $K \in \mathcal{K}$, each method was allowed to run for 30 seconds in order to mimic the fixed time budget scenario. For the BLB, we fixed $B = 100$ as in Kleiner et al. (2014) and Sengupta, Volgushev and Shao (2016), and the sth subset had size N/K. For the SDB, the size of the random subset was also N/K. Table 1 summarizes the number of

TABLE 1

Number of completed iterations with respect to K in 30 seconds for five bootstrap methods: the distributed bootstrap (DB), the pseudo-distributed bootstrap on nonstudentized statistics (PDB), the pseudo-distributed bootstrap on studentized statistics (PDBS), the bag of little bootstrap (BLB) and the subsampled double bootstrap (SDB)

	K							
	5	10	20	50	100	200	500	1000
DB	40	81	162	405	810	1578	3488	5000+
PDB	5000+	5000+	5000+	5000+	5000+	5000+	5000+	5000+
PDBS	5000+	5000+	5000+	5000+	5000+	5000+	5000+	5000+
BLB	0	1	1	4	8	15	33	53
SDB	37	73	154	385	750	1428	3125	4545

iterations completed for each method within the 30 second budget for different K. The table shows that the PDB and PDBS were the fastest that had the most completed iterations among the five methods, while BLB was the slowest. For $K = 5$, when the computation was the most involved, BLB could not finish one iteration within the 30 seconds. The DB and SDB had similar performance. However, it is worth mentioning that these results did not account for the potential time expenditure in data communication among different data blocks.

Table 2 reports the coverage probabilities and widths of the nominal 95% confidence intervals for θ under the fixed 30 second budget for the Gaussian scenario, while the results for Gamma and Poisson distributions can be found in Tables S6 and S7 of the SM. It shows that except the PDBS, there was undercoverage for the other four methods for relatively small $K \leq 20$. The reason for the DB, BLB and SDB having undercoverage when the blocksize K was small was due to their having fewer completed bootstrap iterations as shown in Table 1. It was quite remarkable to observe the PDBS had the best coverage probabilities among the five methods when $K \leq 20$ by adjusting its width. As the number of subsets K increased, and the computational burden was alleviated, the performances of DB, PDB, PDBS and SDB all improved with largely comparable coverage and the best coverage appeared at $K = 100$, while BLB still encountered some undercoverage. If we are not concerned with the time of data communication among multiple data storage locations, for large enough K, DB and

TABLE 2

Coverage probabilities and widths ($\times 1000$ in parentheses) of the 95% confidence intervals of the five bootstrap methods with 30 seconds time budget for the Gaussian data

	K							
	5	10	20	50	100	200	500	1000
DB	0.912	0.930	0.937	0.951	0.953	0.954	0.951	0.952
	(9.030)	(9.450)	(9.730)	(9.892)	(9.941)	(9.966)	(9.948)	(9.926)
PDB	0.834	0.904	0.925	0.948	0.949	0.951	0.956	0.956
	(8.350)	(9.170)	(9.655)	(9.853)	(9.910)	(9.941)	(10.00)	(10.04)
PDBS	0.950	0.944	0.945	0.953	0.950	0.956	0.956	0.956
	(21.95)	(12.10)	(10.76)	(10.24)	(10.09)	(10.03)	(10.03)	(10.06)
BLB	NA	0.938	0.937	0.946	0.944	0.942	0.942	0.937
	(NA)	(9.571)	(9.577)	(9.575)	(9.560)	(9.538)	(9.518)	(9.464)
SDB	0.917	0.930	0.940	0.951	0.951	0.954	0.955	0.952
	(8.977)	(9.418)	(9.695)	(9.888)	(9.947)	(9.961)	(9.981)	(9.977)

SDB offered the best performance (coverage and width). At the same time, the computational more efficient PDB and PDBS had quite comparable performance to DB and SDB.

Next, we evaluate the relative errors in the width of the confidence intervals. Let d be the true width, $\hat{d}$ be the width of the 95% confidence interval by one of the five methods, and the relative error is $|\hat{d} - d|/d$. The exact width d was obtained by 5000 simulations for each distribution. The relative errors were averaged over 1000 replications. Following the strategy in Sengupta, Volgushev and Shao (2016), these five methods are compared with respect to the time evolution of the relative errors under the fixed budget of 30 seconds. This was to explore which method can produce more precise results at a fixed time budget. For the BLB and SDB, one iteration means the completion of estimation procedure for one subset. The relative error was assigned to be 1 before the first iteration was completed.

Figure 1 displays the evolution of the relative errors in the widths of the confidence intervals with respect to time (in seconds) for the five methods with different block size K for Gaussian $\mathcal{N}(1, 1)$ data. We note that the PDB and PBDS were the fastest to converge for all K. However, the PDBS had bad performance in in terms of relative errors for relatively small K. BLB's relative error was severely affected by its rather limited completion rates of the bootstrap iteration when K is relatively small. When $K = 5$, the distributed bootstrap (DB) and SDB had smaller relative errors than the PDB after 15 seconds. This is expected as the convergence rate of the PDB relies on a larger K. As K was increased to larger than 200, the relative errors of the PDB and PDBS quickly decreased to an acceptable rate and became comparable to those of the DB and SDB toward the 30 seconds. DB and SDB had similar performance and they could produce stable results with a sufficient time budget. Results for Gamma and Poisson scenarios were similar and are reported in Figures S5 and S6 in the SM.

In conclusion, with a small time budget, the PDB and PDBS had better performance than the other three resampling strategies for relatively large K. When K is relatively large, PDB and PBDS have advantage in computing and can produce reasonable inference. With a sufficient time budget, the DB and SDB can work for a wide range of K and obtain a better estimator than the PDB and PDBS in some situations.

6.2. *Degenerate case.* The distributed version of distance covariance and its usage in measuring and testing dependence between two multivariate random vectors has been studied in Section S2.3 in the SM. In this section, we investigate the performance of the distributed distance covariance $d\,\mathrm{cov}^2_{N,K}(\mathbf{Y}, \mathbf{Z})$ in testing independence by numerical studies. All simulation results in this section are based on 1000 iterations with the nominal significant level at 5%. The sample size is fixed at $N = 100{,}000$ and the number of data blocks K is selected in the set $\{20, 50, 100, 200\}$.

For the null hypothesis, we generated i.i.d. samples $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_N\}$ and $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_N\}$ independently from distributions $\mathbf{G}_1$ and $\mathbf{G}_2$, respectively. Three different combinations of $\mathbf{G}_1$ and $\mathbf{G}_2$ were considered: (I) $\mathbf{G}_1$ and $\mathbf{G}_2$ are both $\mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$, where $\mathbf{I}_p$ is the p-dimensional identity matrix; (II) $\mathbf{G}_1$ is $\mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$, for $\mathbf{G}_2$, its p components are i.i.d. from student-t distribution with 5 degrees of freedom; (III) For both $\mathbf{G}_1$ and $\mathbf{G}_2$, their components are i.i.d. from student-t distribution with 5 degrees of freedom. The dimension p was chosen as 5, 10 and 20 for all scenarios.

Table 3 reports the empirical sizes of T_{Var} and T_{PDBS}, which stand for the testing procedures based on distributed variance estimator $\hat{\sigma}^2_{\beta,N,K}$ (A.17) and pseudo-distributed bootstrap for studentized distributed statistics (A.19), respectively. Table 3 shows that the empirical sizes of both methods are close to the nominal level 5% for all combinations of p and K under all three scenarios. Thus, these two test procedures both have good control of Type-I error for a wide range of K. In addition, the performance of these two methods is very comparable to each other.

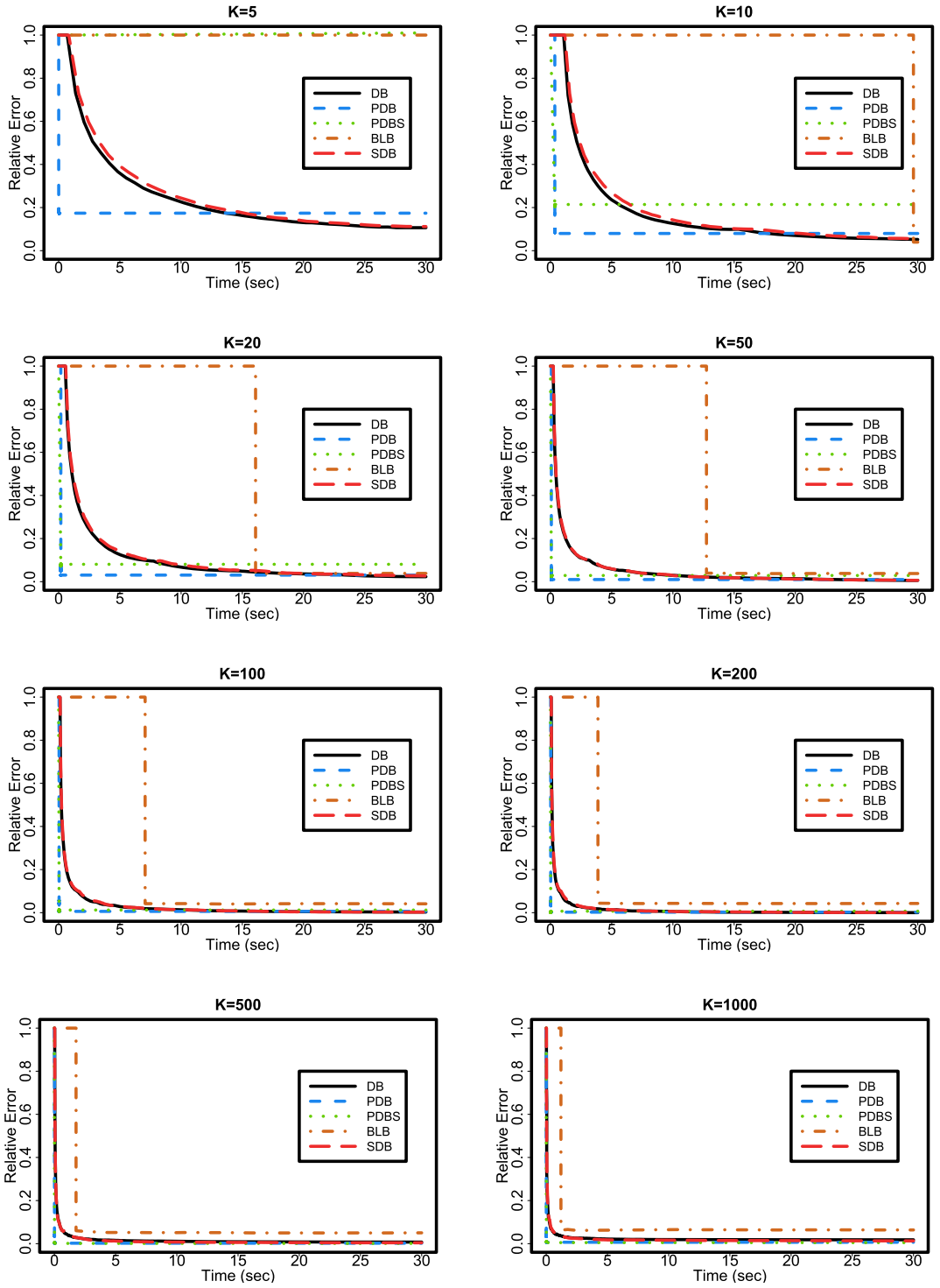

FIG. 1. *Time evolution of the relative errors* $|\hat{d} - d|/d$ *in the width of the confidence interval under the Gaussian scenario with respect to different black size. DB*: *the distributed bootstrap* (*solid lines*); *PDB*: *the pseudo-distributed bootstrap on nonstudentized statistics* (*dashed lines*); *PDBS*: *the pseudo-distributed bootstrap on studentized statistics* (*dotted lines*); *BLB*: *the bag of little bootstrap* (*dot–dashed lines*); *SDB*: *the subsampled double bootstrap* (*long-dashed lines*).

TABLE 3

Sizes of Independence tests based on distributed distance covariance $d\,\mathrm{cov}^2_{N,K}(\mathbf{Y},\mathbf{Z})$. T_{Var}: *test using variance estimation in* (A.17); T_{PDBS}: *test using the pseudo-distributed bootstrap for studentized distributed distance covariance* (A.19)

	$p=5$		$p=10$		$p=20$	
K	T_{Var}	T_{PDBS}	T_{Var}	T_{PDBS}	T_{Var}	T_{PDBS}
			scenario I			
20	0.053	0.053	0.055	0.050	0.045	0.044
50	0.056	0.049	0.047	0.045	0.046	0.044
100	0.046	0.044	0.047	0.047	0.056	0.053
200	0.043	0.042	0.049	0.050	0.054	0.052
			scenario II			
20	0.058	0.043	0.048	0.041	0.053	0.045
50	0.057	0.056	0.056	0.052	0.053	0.049
100	0.050	0.050	0.051	0.047	0.053	0.055
200	0.046	0.043	0.054	0.054	0.052	0.047
			scenario III			
20	0.059	0.048	0.045	0.035	0.066	0.056
50	0.056	0.059	0.051	0.044	0.040	0.042
100	0.050	0.049	0.044	0.045	0.038	0.045
200	0.049	0.049	0.050	0.049	0.047	0.040

To investigate the powers of these two tests, we generated i.i.d. samples $\{\mathbf{Y}_1,\ldots,\mathbf{Y}_N\}$ from $\mathbf{G}_1$ and $\{\mathbf{Z}_1,\ldots,\mathbf{Z}_N\}$ from $\mathbf{G}_2$, and the same three combinations of $\mathbf{G}_1$ and $\mathbf{G}_2$ were considered as before. For $\mathbf{Y}_i=(Y_{i1},\ldots,Y_{ip})^T$ and $\mathbf{Z}_i=(Z_{i1},\ldots,Z_{ip})^T$, we simulate $\mathrm{cor}(Y_{ij},Z_{ik})=\varrho^{|j-k-p|}$ for $j,k=1,\ldots,p$, and $\varrho=0.05,0.1$ were considered. Under these setups, $\mathbf{Y}_i$ and $\mathbf{Z}_i$ are dependent. Table 4 gives the empirical powers of T_{Var} and T_{PDB} for $\varrho=0.05$. The results for $\varrho=0.1$ is provided in Table S8 in the SM.

From Table 4, the empirical powers of these two tests decrease as the dimension p increases. In addition, as the number of data blocks K increases, the empirical powers of the

TABLE 4

Powers of Independence tests based on distributed distance covariance $d\,\mathrm{cov}^2_{N,K}(\mathbf{Y},\mathbf{Z})$ *for* $\varrho=0.05$

	$p=5$		$p=10$		$p=20$	
K	T_{Var}	T_{PDBS}	T_{Var}	T_{PDBS}	T_{Var}	T_{PDBS}
			scenario I			
20	1.000	1.000	0.972	0.959	0.620	0.584
50	0.996	0.991	0.775	0.758	0.332	0.323
100	0.934	0.918	0.533	0.523	0.232	0.230
200	0.735	0.722	0.332	0.326	0.155	0.145
			scenario II			
20	1.000	1.000	0.888	0.866	0.419	0.399
50	0.978	0.971	0.607	0.587	0.256	0.246
100	0.839	0.822	0.392	0.377	0.177	0.176
200	0.600	0.587	0.240	0.230	0.114	0.111
			scenario III			
20	1.000	1.000	0.953	0.928	0.494	0.437
50	0.986	0.985	0.672	0.659	0.278	0.281
100	0.890	0.878	0.443	0.429	0.193	0.185
200	0.674	0.665	0.294	0.294	0.132	0.130

tests also decrease. This is due to the increase in the variance of $d\,\mathrm{cov}^2_{N,K}(\mathbf{Y},\mathbf{Z})$ when K increases. This is the price we need to pay for using the distributed distance covariance. The computing time and memory requirement can be reduced by increasing the number of data blocks; however, this will result in the power loss of the tests.

7. Discussion. The paper investigates distributed inferences on the general symmetric statistics T_N to make the computation scalable. We have analyzed the statistical properties of the distributed statistics as well as their asymptotic distributions when the statistics is nondegenerate or otherwise. Two distributed bootstrap methods are proposed and studied theoretically, which are shown to have advantages over the BLB and SDB.

An important practical issue for the distributed inference is the choice of K. We provide conditions and requirements on K for the theorems that support the distributed approaches. Generally speaking, the requirement on K depends on the stochastic order of the reminder terms. It is also noted that a less specific K means the results are valid for wider situations, which brings flexibility in choosing K in practice as the choice of K is constrained in many aspects, for example, the form of the underlying statistics, the time budget and the available computing resources. As showed in our analysis, an increasing K would decrease the computational cost, but lead to a loss in statistical efficiency. However, it is still an issue on how to select K in practice that balances the computing and statistical efficiency. Instead of considering a fixed time budget, one may minimize computing time subject to certain statistical efficiency, which is similar to the sample size determination problem.

We have focused on i.i.d. data. In practice, data with heterogeneity is a realistic situation. It can be modeled by having a common parameter θ among all data and individual block specific parameters $\{\eta_k\}_{k=1}^K$ for the heterogeneity. The common parameter can be estimated distributively with a modification of the existing algorithm. The parameters $\{\eta_k\}$ for the individual blocks can be improved as the estimation of the common parameter improves while cultivating their dependence to the common parameter estimator.

When T_N is degenerate, the distributed statistic $T_{N,K}$ no longer attains the same efficiency as T_N due lacks of data communications between subsets. It is of interest to find other communication efficient algorithms for degenerate statistics with less efficiency loss. Moreover, when the degeneracy is of a higher order, we would expect more efficiency loss with the distributed formulation. At the same time, as the order of degeneracy increases, the order of the estimation errors gets smaller, which partially compensates for the loss of efficiency.

Furthermore, this paper mainly focuses on the smoothed case for the estimation. Two extensions may be made for the nonsmoothed case. One is to smooth the functions involved in the estimation; for instance, the indication function for quantile estimation can be smoothed via a kernel function (Chen and Hall (1993)). Another is to stay with the nonsmoothed function and employ the asymptotic technique as demonstrated in Huber (1967) and Sherman (1993). The Edgeworth expansion has been established for the distributive statistics. However, the Edgeworth expansion for nonsmoothed statistics is less studied as most of the results are based on the smooth functions of means (Bhattacharya and Ghosh (1978)) and transformation (Skovgaard (1981)). Thus, extensions of the results in this paper to the nonsmooth case would require more works and we leave it to further study.

Funding. Chen's research is partially supported by National Natural Science Foundation of China grants 92046021, 12026607, 12071013 and 71973005 and LMEQF at Peking University.

SUPPLEMENTARY MATERIAL

Supplement to "Distributed statistical inference for massive data" (DOI: 10.1214/21-AOS2062SUPP; .pdf). In the SM, we present technical details, proofs of main theorems and additional numerical results.

REFERENCES

ARCONES, M. A. and GINÉ, E. (1992). On the bootstrap of U and V statistics. *Ann. Statist.* **20** 655–674. MR1165586 https://doi.org/10.1214/aos/1176348650

ATTA-ASIAMAH, E. and YUAN, M. (2019). Distributed inference for degenerate U-statistics. *Stat* **8** e234, 8. MR3978408 https://doi.org/10.1002/sta4.234

BATTEY, H., FAN, J., LIU, H., LU, J. and ZHU, Z. (2018). Distributed testing and estimation under sparse high dimensional models. *Ann. Statist.* **46** 1352–1382. MR3798006 https://doi.org/10.1214/17-AOS1587

BHATTACHARYA, R. N. and GHOSH, J. K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.* **6** 434–451. MR471142 https://doi.org/10.1214/aos/1176344134

CHANG, J. and HALL, P. (2015). Double-bootstrap methods that use a single double-bootstrap simulation. *Biometrika* **102** 203–214. MR3335106 https://doi.org/10.1093/biomet/asu060

CHEN, S. X. and HALL, P. (1993). Smoothed empirical likelihood confidence intervals for quantiles. *Ann. Statist.* **21** 1166–1181. MR1241263 https://doi.org/10.1214/aos/1176349256

CHEN, S. X and PENG, L. (2021). Supplement to "Distributed statistical inference for massive data." https://doi.org/10.1214/21-AOS2062SUPP

CHEN, X. and XIE, M. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statist. Sinica* **24** 1655–1684. MR3308656

DAVIDSON, R. and MACKINNON, J. G. (2002). Fast double bootstrap tests of nonnested linear regression models. *Econometric Rev.* **21** 419–429. MR1951651 https://doi.org/10.1081/ETC-120015384

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR515681 https://doi.org/10.1214/aos/1176344552

HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*. *Springer Series in Statistics*. Springer, New York. MR1145237 https://doi.org/10.1007/978-1-4612-4384-7

HE, X. and SHAO, Q.-M. (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *Ann. Statist.* **24** 2608–2630. MR1425971 https://doi.org/10.1214/aos/1032181172

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (*Berkeley, Calif.*, 1965/66), *Vol. I*: *Statistics* **5.1** 221–233. Univ. California Press, Berkeley, CA. MR0216620

JING, B.-Y. and WANG, Q. (2010). A unified approach to Edgeworth expansions for a general class of statistics. *Statist. Sinica* **20** 613–636. MR2682633

KLEINER, A., TALWALKAR, A., SARKAR, P. and JORDAN, M. I. (2014). A scalable bootstrap for massive data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 795–816. MR3248677 https://doi.org/10.1111/rssb.12050

LAHIRI, S. N. (1994). On two-term Edgeworth expansions and bootstrap approximations for Studentized multivariate M-estimators. *Sankhyā Ser. A* **56** 201–226. MR1664912

LAI, T. L. and WANG, J. Q. (1993). Edgeworth expansions for symmetric statistics with applications to bootstrap methods. *Statist. Sinica* **3** 517–542. MR1243399

LIN, N. and XI, R. (2010). Fast surrogates of U-statistics. *Comput. Statist. Data Anal.* **54** 16–24. MR2558454 https://doi.org/10.1016/j.csda.2009.08.009

LIU, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models. *Ann. Statist.* **16** 1696–1708. MR964947 https://doi.org/10.1214/aos/1176351062

SENGUPTA, S., VOLGUSHEV, S. and SHAO, X. (2016). A subsampled double bootstrap for massive data. *J. Amer. Statist. Assoc.* **111** 1222–1232. MR3561944 https://doi.org/10.1080/01621459.2015.1080709

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York. MR0595165 https://doi.org/0.1002/9780470316481

SHERMAN, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica* **61** 123–137. MR1201705 https://doi.org/10.2307/2951780

SKOVGAARD, I. M. (1981). Transformation of an Edgeworth expansion by a sequence of smooth functions. *Scand. J. Stat.* **8** 207–217. MR0642801

SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35** 2769–2794. MR2382665 https://doi.org/10.1214/009053607000000505

VOLGUSHEV, S., CHAO, S.-K. and CHENG, G. (2019). Distributed inference for quantile regression processes. *Ann. Statist.* **47** 1634–1662. MR3911125 https://doi.org/10.1214/18-AOS1730

ZHANG, Y., DUCHI, J. C. and WAINWRIGHT, M. J. (2013). Communication-efficient algorithms for statistical optimization. *J. Mach. Learn. Res.* **14** 3321–3363. MR3144464