DISTRIBUTED STATISTICAL INFERENCE FOR MASSIVE DATA

BY SONG XI CHEN†,* AND LIUHUA PENG†

Peking University* and the University of Melbourne†

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 non-degenerate 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 method, 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

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

MSC 2010 subject classifications: Primary 62G09; secondary 62G20

Keywords and phrases: Distributed bootstrap, Distributed statistics, Massive data, Pseudo-distributed bootstrap
instance the M-estimation by Zhang, Duchi and Wainwright (2013) and the
generalized linear models by Chen and Xie (2014), see also Volgushev, Chao

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, Volgu-
shev and Shao (2015) 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 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 symmet-
ric statistics (Lai and Wang, 1993), which encompass both the linear and
nonlinear statistics, and cover both non-degenerate and degenerate cases.
Although we use the standard SaC formulation, our study reveals new find-
ings. For the degenerate case, the distributed formulation no longer attains
the same efficiency as the full sample statistic, which differs from the non-
degenerate case. The asymptotic distributions of the distributed statistics
under the degeneracy are, respectively, the weighted \( \chi^2 \) and Gaussian, de-
pending 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 re-
samples 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 non-degenerate 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 distrib-
utions 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).

2. Distributed symmetric statistics. Let \( X_N = \{X_1, \ldots, X_N\} \) 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 \( T_N = T(X_N) \) for a parameter \( \theta = \theta(F) \) is invariant under data permutations, that admits a general non-linear expansion

\[
T_N = \theta + N^{-1} \sum_{i=1}^{N} \alpha(X_i; F) + N^{-2} \sum_{1 \leq i < j \leq N} \beta(X_i, X_j; F) + R_N,
\]

where \( \alpha(x; F) \) and \( \beta(x, y; F) \) are known functions, and \( R_N = R(X_N; F) \) is a remainder term. We assume the following conditions on \( \alpha \) and \( \beta \), 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 \( \mathbb{E}\{\alpha(X_1; F)\} = 0 \) and \( \text{Var}\{\alpha(X_1; F)\} = \sigma_\alpha^2 \in [0, \infty) \), and \( \beta(x, y; F) \) being symmetric in \( x \) and \( y \) such that \( \mathbb{E}\{\beta(X_1, X_2; F)|X_1\} = 0 \) and \( \text{Var}\{\beta(X_1, X_2; F)\} = \sigma_\beta^2 \in [0, \infty) \).
(ii) \( \mathbb{E}|\alpha(X_1; F)|^{2+\delta} < \infty \) and \( \mathbb{E}|\beta(X_1, X_2; F)|^{2+\delta'} < \infty \) for some constants \( 0 \leq \delta \leq 1 \) and \( 0 \leq \delta' \leq 1 \). (iii) The distribution of \( \alpha(X_1; F) \) is non-lattice and \( \mathbb{E}|\alpha(X_1; F)|^3 < \infty \).

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

The statistic \( T_N \) encompasses a wide class of statistics, for instance the \( U \)- and \( L \)-statistics, and the M-estimator (Lai and Wang, 1993; Jing and Wang, 2010; Lahiri, 1994). 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 \( \mathbb{E}(R_N) = 0 \) and \( \text{Var}(R_N) = O(N^{-3}) \). 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 \( \alpha \) and \( \beta \), 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 SM). 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 $X_N$ into $K$ data blocks. Let $X_{N,K} = \{X_{k,1}, \ldots, X_{k,n_k}\}$ be the $k$-th data block of size $n_k$, for $k = 1, \ldots, K$. Such division is naturally available when $X_N$ is stored over $K$ storage facilities. Otherwise, the blocks can be attained by random splitting.

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

By averaging the $K$ block-wise statistics, the distributed statistic

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

which can be expressed as

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

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 $\beta$ and the remainders.

While the SaC formulation is not new, our analysis on the general symmetric statistics contain fresh results for both non-degenerate ($\sigma^2_\alpha > 0$) and degenerate ($\sigma^2_\alpha = 0$) 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 $c_1 \leq \inf_{k_1, k_2} n_{k_1}/n_{k_2} \leq \sup_{k_1, k_2} n_{k_1}/n_{k_2} \leq c_2$, 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 $E(R_N) = b_1 N^{-\tau_1} + o(N^{-\tau_1})$ and $\text{Var}(R_N) = O(N^{-\tau_2})$ for some $b_1 \neq 0$, $\tau_1 \geq 1$ and $\tau_2 > 1$, then $E(R_{N,K}^{(k)}) = b_{1,k} n_k^{-\tau_1} + o(n_k^{-\tau_1})$ for some $b_{1,k} \neq 0$ and $\text{Var}(R_{N,K}^{(k)}) = O(n_k^{-\tau_2})$, 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 \( R^{(k)}_{N,K} \) 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 \( \tau_2 > 2 \), Bias(\( T_N \)) = \( b_1 N^{-\tau_1} + o(N^{-\tau_1}) \) and Bias(\( T_{N,K} \)) = \( N^{-1} \sum_{k=1}^K b_1 k n_k^{-1} - \tau_1 + o(K)N^{-\tau_1} \). In addition,

\[
\text{Var}(T_N) = \sigma^2 \alpha N^{-1} + 2^{-1} \sigma^2 N^{-2} + N^{-1} \sum_{i=1}^N \text{Cov}\{\alpha(X_i; F), R_N\} + o(N^{-2}),
\]

\[
\text{Var}(T_{N,K}) = \sigma^2 \alpha N^{-1} + 2^{-1} \sigma^2 K N^{-2} + N^{-2} \sum_{k=1}^K n_k \sum_{i=1}^n \text{Cov}\{\alpha(X_{k,i}; F), R^{(k)}_{N,K}\} + o(K N^{-2}).
\]

Theorem 3.1 implies that, if \( T_N \) is unbiased to \( \theta \), namely \( \tau_1 = \infty \), \( T_{N,K} \) is also unbiased. For the case of \( \tau_1 < \infty \), as \( N^{-1} \sum_{k=1}^K b_1 n_k^{-1} - \tau_1 \) is of order \( K^{-1} N^{-\tau_1} \) under Condition C3, the bias is enlarged by a factor of \( K^{\tau_1} \) 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 \( \sigma^2 N^{-2} \). While this increase has no leading order impact in the non-degenerate case (\( \sigma^2 > 0 \)), it becomes significant for the degenerate case (\( \sigma^2 = 0 \)) 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 \( o(N^{-\delta_2/2}) \) for \( T_N \) and \( o(K^{1/2} N^{-\delta_2/2}) \) for \( T_{N,K} \) in the non-degenerate case.

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

\[
\text{MSE}(T_N) = \sigma^2 \alpha N^{-1} + 2^{-1} \sigma^2 N^{-2} + b_1^2 N^{-2\tau_1}
\]

\[+ N^{-1} \sum_{i=1}^N \text{Cov}\{\alpha(X_i; F), R_N\} + o(N^{-2} + N^{-2\tau_1}) \text{ and}
\]

\[
\text{MSE}(T_{N,K}) = \sigma^2 \alpha N^{-1} + 2^{-1} \sigma^2 K N^{-2} + N^{-2} \left( \sum_{k=1}^K b_1 k n_k^{-1} - \tau_1 \right)^2
\]

\[+ N^{-2} \sum_{k=1}^K n_k \sum_{i=1}^n \text{Cov}\{\alpha(X_{k,i}; F), R^{(k)}_{N,K}\} + o(K N^{-2} + K^{2\tau_1} N^{-2\tau_1}).
\]
For the non-degenerate case of $\sigma^2_\alpha > 0$, if
\begin{equation}
K = o(N^{1-1/(2\tau_1)}),
\end{equation}
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 ($\sigma^2_\alpha = 0$ but $\sigma^2_\beta > 0$), 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 non-degenerate case.

**Theorem 3.2.** Suppose C1(i), C3 hold and $\sigma^2_\alpha > 0$, then
(i) if $R_N = o_p(N^{-1/2}), N^{1/2}\sigma^-1_\alpha(T_N - \theta) \xrightarrow{d} N(0,1)$ as $N \to \infty$;
(ii) if $R_{N,K} = o_p(N^{-1/2}), N^{1/2}\sigma^-1_\alpha(T_{N,K} - \theta) \xrightarrow{d} N(0,1)$ 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 $\sigma^2_\alpha > 0$, 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 $\sigma^2_\alpha > 0$, 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 $\sigma^2_\alpha > 0$, $K = O(N^{\tau'})$ for a positive constant $\tau'$.
(i) Assume C2(i) and C4(i), and $\tau' < 1 - 1/(2\tau_1)$, then as $N \to \infty$,
\begin{equation}
\sup_{x \in \mathbb{R}} \left| \Pr \left\{ N^{1/2}\sigma^-1_\alpha(T_{N,K} - \theta) \leq x \right\} - \Pr \left\{ N^{1/2}\sigma^-1_\alpha(T_N - \theta) \leq x \right\} \right| = O(N^{-\min(\tau_1-\tau_1\tau'-1/2,\tau_2-1)(1-\tau')/3,1/2)}),
\end{equation}
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 \( N \to \infty \),
\[
\sup_{x \in \mathbb{R}} \left| \left( \frac{N^{1/2}}{\sigma_\alpha} \right)^{-1} (T_{N,K} - \theta) \right| \leq \mathbb{P} \left\{ \frac{N^{1/2}}{\sigma_\alpha} (T_N - \theta) \leq x \right\} - \mathbb{P} \left\{ \frac{N^{1/2}}{\sigma_\alpha} (T_N - \theta) \leq x \right\}
\]
(3.3) \( = O(N^{-\min\{\tau_3 - \tau_3\tau' - 1/2, \tau_4 - \tau_4\tau' - \tau'\}}) \).

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 (Lemma 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 \( \tau_1 \leq 1 \), it can be shown that the rate in (3.2) is bounded below by \( K^\tau N^{1/2-\tau_1} \). 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 \( O(N^{-\min\{\tau_3 - \tau_3\tau' - 1/2, \tau_4 - \tau_4\tau' - \tau'\}}) \), is sharp in some cases. For the M-estimator \( \hat{\theta}_N \) which solves \( \sum_{i=1}^N \psi(X_i, \theta) = 0 \) with \( \theta_0 \) being the true value of \( \theta \), it can be expanded in the form of (2.1) with \( \alpha \) and \( \beta \) defined in Proposition S2.1 of the SM under some regularity conditions on \( \psi \). 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 \( \mathbb{P} \{ N^{1/2}/\alpha \leq \theta_N - \theta_0 \leq x \} \) and \( \mathbb{P} \{ N^{1/2}/\alpha \leq \hat{\theta}_{N,K} - \theta_0 \leq x \} \) are
\[
F_{N,E1}(x) = \Phi(x) - N^{-1/2}\alpha \phi(x) \overline{\psi} \left( \hat{\theta}_0 \right)
\]
\[
-6^{-1}N^{-1/2}\alpha \phi(3)(x) \left[ \hat{\mathbb{E}}_1(\theta_0) + 6 \left\{ \phi^{(1)}(\theta_0) \right\}^{-2} \varphi_2(\theta_0) \overline{\psi}(\theta_0) \right]
\]
and
\[
F_{N,E3}(x) = \Phi(x) - KN^{-1/2}\alpha \phi(x) \overline{\psi} \left( \hat{\theta}_0 \right)
\]
\[
-6^{-1}N^{-1/2}\alpha \phi(3)(x) \left[ \hat{\mathbb{E}}_1(\theta_0) + 6 \left\{ \phi^{(1)}(\theta_0) \right\}^{-2} \varphi_2(\theta_0) \overline{\psi}(\theta_0) \right],
\]
respectively. We refer to Proposition S2.1 in the SM for the definitions of \( \overline{\psi} \), \( \hat{\mathbb{E}}_1 \), \( \phi^{(1)}(\theta_0) \), and \( \varphi_2 \). So when \( \overline{\psi}(\theta_0) \neq 0 \), \( \sup_{x \in \mathbb{R}} \left| \mathbb{P} \{ N^{1/2}/\alpha \leq \hat{\theta}_{N,K} - \theta_0 \leq x \} - \mathbb{P} \{ N^{1/2}/\alpha \leq \hat{\theta}_N - \theta_0 \leq x \} \right| \) 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 $\sigma_n^2 = 0$, $T_{N,K}$ can not achieve the same efficiency as $T_N$ according to Theorem 3.1. If $\beta$ 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 $\lim_{L \to \infty} E\{\beta(X_1,X_2;F) - \sum_{\ell=1}^L \lambda_\ell \beta_\ell(X_1;F) \beta_\ell(X_2;F)\}^2 = 0$. The next theorem gives the asymptotic distributions of $T_N$ and $T_{N,K}$ under degeneracy.

**Theorem 3.4.** Under $C1$ (i) and $C3$, $\sigma_n^2 = 0$ and $\sigma_n^2 > 0$,

(i) if $R_N = o_p(N^{-1})$, then as $N \to \infty$, $2N(T_N - \theta) \xrightarrow{d} \sum_{\ell=1}^\infty \lambda_\ell (\chi_\ell^2 - 1)$, where $\{\chi_\ell^2\}_{\ell=1}^\infty$ are independent $\chi_1^2$ 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 $\chi_K^2$ 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 $2^{-1/2}K^{-1/2}N\sigma_n^{-1}(T_{N,K} - \theta) \xrightarrow{d} N(0,1)$ 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 $k = 1, \ldots, K$ ensure $R_{N,K} = o_p(N^{-1})$. When $K \to \infty$, if $E(R_N) = b_1 N^{-\tau_1} + o(N^{-\tau_1})$ and $\text{Var}(R_N) = o(N^{-2})$, then under Condition $C4$ (i), it requires $K\tau_1 N^{-\tau_1} = o(K^{1/2}N^{-1})$, that is $K = o(N^{-1/(2(\tau_1-1))})$ in order for $R_{N,K} = o_p(K^{1/2}N^{-1})$. This is a slower growth rate for $K$ comparing to the non-degenerate case in (3.1). The distributed inference for the degenerate $U$-statistics is studied in Atta-Asiamah and Yuan (2019) with an emphasize 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, 2015) 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 \( X \) 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 (2015) proposed the SDB that combines the idea of the BLB and a fast double bootstrap (Davidson and MacKinnon, 2002; Chang and Hall, 2015). 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 non-linear 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 (2015), 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, can not 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 \( \mathcal{X}_{N,K}^{(1)}, \ldots, \mathcal{X}_{N,K}^{(K)} \) in the formulation of \( T_{N,K} \), let \( F_{N,K}^{(k)} \) be the empirical distribution of \( \mathcal{X}_{N,K}^{(k)} \), \( \hat{\theta}_{N,K}^{(k)} = \theta\left(F_{N,K}^{(k)}\right) \) be the analogy of \( \theta \) under \( \mathcal{X}_{N,K}^{(k)} \), 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 \( \mathcal{X}_{N,K}^{*,(k)} = \{ X_{k,1}^*, \ldots, X_{k,n_k}^* \} \) be an independent and identically distributed (IID) sample from \( F_{N,K}^{(k)} \). Repeating it \( B \) times, one obtains \( B \) resampled subsets \( \mathcal{X}_{N,K}^{*1,(k)}, \ldots, \mathcal{X}_{N,K}^{*B,(k)} \). Compute the corresponding
statistic $T_{N,K}^{*b(k)} = T(x_{N,K}^{*b(k)})$, and average them over the $K$ subsets to get the $b$-th 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 notations and assumptions. Specifically, $T_{N,K}^{*b} = T(x_{N,K}^{*b})$ is assumed to admit

\begin{equation}
(4.1) \quad T_{N,K}^{*b} = \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 \leq i < j \leq n_k} \hat{\beta}(X_{k,i}^{*}, X_{k,j}^{*}; F_{N,K}^{(k)}) + R_{N,K}^{(k)},
\end{equation}

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

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

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 \( F_{N,K}^{(k)} \), \( k = 1, \ldots, K \), \( E\{\hat{\alpha}(X_{k,i}^*; F_{N,K}^{(k)}) | F_{N,K}^{(k)}\} = 0 \) and \( E\{\hat{\beta}(X_{k,i}^*, y; F_{N,K}^{(k)}) | F_{N,K}^{(k)}\} = 0 \) 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 uniformly 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 \( E\{\alpha(X_1; F)\} = 0 \), 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 \( \sum n_k^{-1} \alpha(X_{k,i}; F_{N,K}^{(k)}) = 0 \) satisfied exactly, we can centralize \( \alpha(x; F_{N,K}^{(k)}) \) by the empirical sub-sample 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 \( E\{\hat{\alpha}(X_{k,i}^*; F_{N,K}^{(k)}) | F_{N,K}^{(k)}\} = 0 \), which is sufficient for Theorem 4.1 presented below. Similar arguments can be made on the assumption that \( \sum n_k^{-1} \hat{\beta}(X_{k,i}, y; F_{N,K}^{(k)}) = 0 \) for any \( y \in S(F) \).

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

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 non-lattice 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 non-degenerated case of \( \sigma^2_\alpha > 0 \).

**Theorem 4.1.** Assume \( \sigma^2_\alpha > 0 \), under Condition C1 (i) and (ii), C3 and C5 with \( \delta > 0 \) and \( \delta' = 0 \), \( K = O(N^{\tau'}) \) for a positive constant \( \tau' \), and \( R_{N,K}^* \) in (4.2) satisfies
\[
P\{|R_{N,K}^*| \geq N^{-1/2}(\ln N)^{-1} | F_{N,K}^{(1)}, \ldots, F_{N,K}^{(K)}\} = o_p(1).
\]
In addition, assume C1 (iii) and C6, \( \delta \), \( (4.3) \)

\[ E \) and \( |E| \) then \( (4.3) \)

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

\[ \sup_{x \in \mathbb{R}} \left| P \left\{ N^{1/2}(T^*_{N,K} - \hat{\theta}_{N,K}) \leq x \mid F_{N,K}^{(1)}, \ldots, F_{N,K}^{(K)} \right\} \right| = o_p(1). \] (4.3)

\[ -P \left\{ N^{1/2}(T_{N,K} - \theta) \leq x \right\} = o_p(1). \]

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 \( P\{ |R^*_{N,K}| \geq K^{1/2}N^{-1}F_{N,K}^{(1)}, \ldots, F_{N,K}^{(K)} \} = O_p(K^{1/2}N^{-1/2}) \), 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 \( \tau' < \min\{1 - 1/(2\tau_3 - 1), 1 - 2/(\tau_4 + 1), 1/2\} \), \( R^*_{N,K} \) satisfies \( P\{ |R^*_{N,K}| \geq K^{1/2}N^{-1}F_{N,K}^{(1)}, \ldots, F_{N,K}^{(K)} \} = O_p(K^{1/2}N^{-1/2}) \), 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 non-degenerate 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 \( u^*_{\tau} \) as the sample \( \tau \)-th quantile of \( \left\{ N^{1/2}(T^*_{N,K} - \hat{\theta}_{N,K}) \right\}_{\ell=1}^{B} \), then an equal-tail two-sided CI for \( \theta \) with confidence level \( 1 - \tau \) can be constructed as

\[ \left( T_{N,K} - N^{-1/2}u^*_{1-\tau/2}, T_{N,K} - N^{-1/2}u^*_\tau \right). \] (4.4)

The bootstrap resample from the DB algorithm can be also used to estimate \( \text{Var}(T_{N,K}) \). 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 \( \text{Var}(T_{N,K}) \).

**Theorem 4.2.** Under the conditions of Theorem 4.1 (i), assume that \( \text{E}\{\beta(X_1, X_i; F)\}^2 < \infty \) with \( 0 < \delta' < 1 \) for \( X_i = X_1 \) and \( X_2 \), respectively, and \( \text{E}\{N^{1/2}R^*_{N,K}\}^2 < \infty \). Then, \( \hat{\sigma}_{DB}^2/\text{Var}(T_{N,K}) \to 1 \) 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, \( E\{\hat{\beta}(X_{k,i}^*, y; F_{N,K}^{(k)})|F_{N,K}^{(k)}\} = 0 \) 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 \( \tau' \).

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

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

(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 \( 2^{1/2}K^{-1/2} \sum_{k=1}^{K} n_k^{-1} \sum_{1 \leq i < j \leq n_k} \hat{\beta}(X_{k,i}^*, X_{k,j}^*; F_{N,K}^{(k)}) \) is consistent to that of \( 2^{1/2}K^{-1/2}N(T_{N,K} - \theta) \). When \( K \) is fixed, the conditional distribution of \( 2 \sum_{k=1}^{K} n_k^{-1} \sum_{1 \leq i < j \leq n_k} \hat{\beta}(X_{k,i}^*, X_{k,j}^*; F_{N,K}^{(k)}) \) may be used to estimate that of \( 2N(T_{N,K} - \theta) \). The corresponding theoretical analysis requires the eigen-decomposition of \( \beta \), 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 re-calculated 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 non-studentized distributed statistics.** We first consider the non-degenerate case. Due to different subset sizes, we need to scale $\{T_{N,K}^{(k)}\}_{k=1}^K$ before the resampling. Let $T_{N,K}^{(k)} = N^{-1/2}K^{1/2}n_kT_{N,K}^{(k)}$ for $k = 1, \ldots, K$, and $F_{K,T}$ be the empirical distribution of $\{T_{N,K}^{(k)}\}_{k=1}^K$. Suppose $\{T_{N,K}^{s(k)}\}_{k=1}^K$ is an IID sample from $F_{K,T}$ and $T_{N,K}^{s} = K^{-1}\sum_{k=1}^K T_{N,K}^{s(k)}$. Then, the distribution of $K^{1/2}(T_{N,K}^{s} - N^{1/2}K^{-1/2}T_{N,K}^{(1)})$ conditional on $F_{K,T}$ 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 $\{T_{N,K}^{(k)}\}_{k=1}^K$, which are independent but not necessarily identically distributed. The bootstrap under non-IID 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 non-degenerate case.

**Theorem 5.1.** (Non-degenerate case) Under Condition C1 (i), C2 (i), C3, C4 (i) and $\sigma_\alpha^2 > 0$; $K = O(N^{\tau'})$ for the $\tau'$ 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 $\sup_k E|n_k^{1/2}R_{N,K}^{(k)}|^{2+\delta} < \infty$. Then, as $K \to \infty$,

\[
\sup_{x \in \mathbb{R}} \left| P \left\{ K^{1/2}(T_{N,K}^{s} - N^{1/2}K^{-1/2}T_{N,K}^{(1)}) \leq x \mid F_{K,T} \right\} - P \left\{ N^{1/2}(T_{N,K} - \theta) \leq x \right\} \right| = o_p(1).
\]

(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 $\sup_k E|n_k^{1/2}R_{N,K}^{(k)}|^{3} < \infty$; the distribution of $T_{N,K}^{(1)}$ is non-lattice. 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 \( \{ T_{N,K}^{(k)} \}_{k=1}^K \), 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 rescaled \( \{ T_{N,K}^{(k)} \}_{k=1}^K \) with a different scaling factor such that \( T_{N,K}^{(k)} = n_k T_{N,K}^{(k)} \) followed by the same resampling procedures for the non-degenerate case. Suppose that \( T_{N,K}^{(1)}, \ldots, T_{N,K}^{(K)} \) is an IID sample from \( F_{K,T} \) and denote \( T_{N,K}^* = K^{-1} \sum_{k=1}^K T_{N,K}^{(k)} \).

**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 \( K = O(N^{\tau_2}) \) for a positive constant \( \tau_2 \) specified as below.

(i) Assume \( \tau' < 1 - 1/(2\tau_1 - 1) \) and C1 (ii) holds with \( 0 < \delta, \delta' < 1 \) and
\[
\sup_k E |n_k R_{N,K}^{(k)}|^{2+\delta} < \infty, \text{ then as } K \to \infty,
\]
\[
(5.2) \quad \sup_{x \in \mathbb{R}} \left| P \left\{ K^{1/2} \left( T_{N,K}^* - NK^{-1} T_{N,K} \right) \leq x \right| F_{K,T} \right| - P \left( NK^{-1/2} (T_{N,K} - \theta) \leq x \right) \right| = o_p(1).
\]

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

Compared to the case of non-degenerate case, stronger conditions are needed for the degenerate case to control \( R_{N,K}^{(k)} \). This is reflected firstly in that \( \tau_1 \) needs to be strictly larger than 1 such that \( R_{N,K} \) can be dominated by the quadratic term involving \( \beta \). Secondly, \( n_k \) 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 non-generate case. Despite these, PDB for both the non-degenerate 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 \( \text{Var}(T_{N,K}) \) 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

\[
S_{K}^{2} = (K - 1)^{-1} \sum_{k=1}^{K} \left( \frac{N^{-1}Kn_kT_{N,K}^{(k)} - T_{N,K}}{K} \right)^{2}.
\]

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, \( K^{-2}N^{2}S_{K}^{2} \to 2^{-1}\sigma_{\beta}^{2} \) almost surely as \( K \to \infty \).

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

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 \( \text{Var}(T_{N,K}) \), it may not be competitive to the variance estimator obtained by the DB. Furthermore, using a very large \( K \) can lead to a non-ignorable 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), \( K^{1/2}S_{K}^{-1}(T_{N,K} - \theta) \overset{d}{\to} \mathcal{N}(0,1) \) as \( K \to \infty \).

Based on Theorem 5.3, a Wald-type confidence interval of \( \theta \) based on \( T_{N,K} \) and \( S_{K}^{2} \) can be established as

\[
\left( T_{N,k} - z_{1-\tau/2}K^{-1/2}S_{K}, T_{N,k} + z_{1-\tau/2}K^{-1/2}S_{K} \right),
\]

where \( z_{1-\tau/2} \) is the \((1 - \tau/2)\)-th upper quantile of \( \mathcal{N}(0,1) \). 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 non-studentized distributed statistic in Section 5.1, the only adjustment is in the studentization for each bootstrap pseudo sample. Denote
\[
T^{(k)}_{N,K} = N^{-1} K n_k T^{(k)}_{N,K}
\]
for \( k = 1, \ldots, K \).

Let \( F_{K,T} \) be the empirical distribution of \( \{T^{(k)}_{N,K}\}^K_{k=1} \). Suppose \( \{T^{*,(k)}_{N,K}\}^K_{k=1} \) is an IID sample from \( F_{K,T} \). Let \( T^*_{N,K} = K^{-1} \sum_{k=1}^K T^{*,(k)}_{N,K} \) and \( S^*_K = \left\{ \left( K - 1 \right)^{-1} \sum_{k=1}^K \left( T^{*(k)}_{N,K} - T^{*,(k)}_{N,K} \right)^2 \right\}^{1/2} \), then the distribution of \( K^{1/2} \{S^*_K\}^{-1} (T^*_{N,K} - T_{N,K}) \) conditional on \( F_{K,T} \) is used to estimate that of \( K^{1/2} S^{-1}_K (T_{N,K} - \theta) \).

**Theorem 5.4.** Under the conditions of Theorem 5.1 (i) or Theorem 5.2 (i), as \( K \to \infty \),
\[
\sup_{x \in \mathbb{R}} \left| \Pr \left\{ K^{1/2} \{S^*_K\}^{-1} (T^*_{N,K} - T_{N,K}) \leq x \mid F_{K,T} \right\} - \Pr \left\{ K^{1/2} S^{-1}_K (T_{N,K} - \theta) \leq x \right\} \right| = o_p(1).
\]

Theorem 5.4 indicates that the PDB works for the studentized distributed statistics for both non-degenerate 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 \( T^{(k)}_{N,K} \) is non-lattice. In addition, we assume

(a) if \( \sigma_\alpha^2 > 0 \), \( E|n_k^{1/2} T^{(k)}_{N,K}|^3 < \infty \); (b) if \( \sigma_\beta^2 = 0 \) but \( \sigma_\alpha^2 > 0 \), \( E|n_k T^{(k)}_{N,K}|^3 < \infty \). Then, as \( K \to \infty \),
\[
\sup_{x \in \mathbb{R}} \left| \Pr \left\{ K^{1/2} S^{-1}_K (T_{N,K} - \theta) \leq x \right\} - \Phi(x) \right| = O_p(K^{-1/2}) \quad \text{and}
\]
\[
\sup_{x \in \mathbb{R}} \left| \Pr \left\{ K^{1/2} \{S^*_K\}^{-1} (T^*_{N,K} - T_{N,K}) \leq x \mid F_{K,T} \right\} - \Pr \left\{ K^{1/2} S^{-1}_K (T_{N,K} - \theta) \leq x \right\} \right| = o_p(K^{-1/2}).
\]

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 non-studentized PDB. The approximation error \( o_p(K^{-1/2}) \) can be made to \( O_p(K^{-1}) \) if we impose stronger moment conditions on \( T^{(k)}_{N,K} \).
6. Simulation studies. In this section, we use Gini’s mean difference as a non-degenerate 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. Non-degenerate 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
\[ U_N = 2 \sum_{1 \leq i < j \leq N} |X_i - X_j|. \]
It is an unbiased estimator of the dispersion parameter \( \theta_U = \mathbb{E}|X_i - X_j| \), and is a U-statistic of degree two. Suppose the full data are divided into \( K \) blocks with the \( k \)-th of size \( n_k \), and let \( U_{N,K}^{(k)} \) be the Gini’s difference from the \( k \)-th 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 \( X \) were experimented: (I) \( \mathcal{N}(1, 1) \); (II) \( \text{Gamma}(3, 1) \) and (III) \( \text{Poisson}(4) \). 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 \( \theta \) 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 (2015), and the \( s \)-th subset had size \( N/K \). For the SDB, the size of the random subset was also \( N/K \). Table 1 summarizes the number of 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 \( \theta \) 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 under-coverage for the other four methods for relatively small \( K \leq 20 \). The reason for the DB, BLB and SDB having under-coverage when the block-size \( K \) 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 under-coverage. If we are not concerned with the time of data communication among multiple data storage locations, for large enough $K$, DB and 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 Sen-gupta, Volgushev and Shao (2015), these five methods are compared with

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 Sen-gupta, Volgushev and Shao (2015), 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 result 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 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 PBDS quickly decreased to an acceptable rate and became comparable to those of the DB and SDB towards 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 PBDS 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 better estimator than the PDB and PBDS 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 $dcov^2_{N,K}(\mathbf{Y}, \mathbf{Z})$ 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 IID 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 IID from student-$t$ distribution...
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 non-studentized 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).
with 5 degrees of freedom; (III) For both $G_1$ and $G_2$, their components are IID 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 $\hat{\sigma}_{N,K}^2$ (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.

Table 3  
Sizes of Independence tests based on distributed distance covariance $d_{cov}^2_{N,K}(Y, Z)$. $T_{Var}$: test using variance estimation in (A.17); $T_{PDBS}$: test using the pseudo-distributed bootstrap for studentized distributed distance covariance (A.19).

<table>
<thead>
<tr>
<th>$K$</th>
<th>$p = 5$</th>
<th>$p = 10$</th>
<th>$p = 20$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T_{Var}$</td>
<td>$T_{PDBS}$</td>
<td>$T_{Var}$</td>
</tr>
<tr>
<td>20</td>
<td>0.053</td>
<td>0.053</td>
<td>0.055</td>
</tr>
<tr>
<td>50</td>
<td>0.056</td>
<td>0.049</td>
<td>0.047</td>
</tr>
<tr>
<td>100</td>
<td>0.046</td>
<td>0.044</td>
<td>0.047</td>
</tr>
<tr>
<td>200</td>
<td>0.043</td>
<td>0.042</td>
<td>0.049</td>
</tr>
<tr>
<td>20</td>
<td>0.058</td>
<td>0.043</td>
<td>0.048</td>
</tr>
<tr>
<td>50</td>
<td>0.057</td>
<td>0.056</td>
<td>0.056</td>
</tr>
<tr>
<td>100</td>
<td>0.050</td>
<td>0.050</td>
<td>0.051</td>
</tr>
<tr>
<td>200</td>
<td>0.046</td>
<td>0.043</td>
<td>0.054</td>
</tr>
<tr>
<td>20</td>
<td>0.059</td>
<td>0.048</td>
<td>0.045</td>
</tr>
<tr>
<td>50</td>
<td>0.056</td>
<td>0.059</td>
<td>0.051</td>
</tr>
<tr>
<td>100</td>
<td>0.050</td>
<td>0.049</td>
<td>0.044</td>
</tr>
<tr>
<td>200</td>
<td>0.049</td>
<td>0.049</td>
<td>0.050</td>
</tr>
</tbody>
</table>

To investigate the powers of these two tests, we generated IID samples $\{Y_1, \ldots, Y_N\}$ from $G_1$ and $\{Z_1, \ldots, Z_N\}$ from $G_2$, and the same three combinations of $G_1$ and $G_2$ were considered as before. For $Y_i = (Y_{i1}, \ldots, Y_{ip})^T$ and $Z_i = (Z_{i1}, \ldots, Z_{ip})^T$, we simulate $\text{cor}(Y_{ij}, Z_{ik}) = \varrho^{\left|j-k-p\right|}$ for $j, k = 1, \ldots, p$, and $\varrho = 0.05, 0.1$ were considered. Under these setups, $Y_i$ and $Z_i$ are dependent. Table 4 gives the empirical powers of $T_{Var}$ and $T_{PDBS}$ 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 tests also decrease. This is due to the increase in
the variance of $dcov^2_{N,K}(Y, Z)$ 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 non-degenerate 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, e.g., 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 IID data. In practice, data with heterogeneity is a realistic situation. It can be modeled by having a common parameter $\theta$ 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 non-smoothed 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 non-smoothed 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 non-smoothed 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 non-smooth case would require more works and we leave it to further study.

SUPPLEMENTARY MATERIAL

Supplement to “Distributed Statistical Inference for Massive Data”

(doi: COMPLETED BY THE TYPESETTER; .pdf). In the SM, we present technical details, proofs of main theorems and additional numerical results.

References.


DISTRIBUTED INFERENCE FOR MASSIVE DATA


guanghua School of Management and Center for Statistical Science
Peking University
Beijing 100651, China
E-mail: csx@gsm.pku.edu.cn

School of Mathematics and Statistics
the University of Melbourne
Victoria 3010
Australia
E-mail: liuhua.peng@unimelb.edu.au