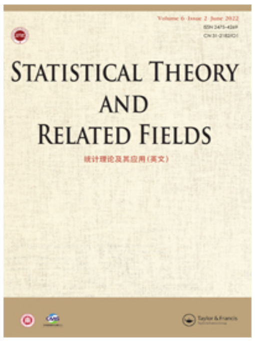

STATISTICAL THEORY AND RELATED FIELDS

Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Discussion on ‘A review of distributed statistical inference’

Yang Yu & Guang Cheng

To cite this article: Yang Yu & Guang Cheng (2022) Discussion on ‘A review of distributed statistical inference’, Statistical Theory and Related Fields, 6:2, 102-103, DOI: 10.1080/24754269.2022.2030107

To link to this article: https://doi.org/10.1080/24754269.2022.2030107

Published online: 04 Feb 2022.

STATISTICAL THEORY AND RELATED FIELDS
2022, VOL. 6, NO. 2, 102–103
https://doi.org/10.1080/24754269.2022.2030107

Taylor & Francis
Taylor & Francis Group

SHORT COMMUNICATION

OPEN ACCESS Check for updates

Discussion on 'A review of distributed statistical inference'

Yang Yu and Guang Cheng

Department of Statistics, Purdue University, West Lafayette, IN, USA

ARTICLE HISTORY Received 22 November 2021; Accepted 8 December 2021

We congratulate the authors on an impressive team effort to comprehensively review various statistical estimation and inference methods in distributed frameworks. This paper is an excellent resource for anyone wishing to understand why distributed inference is important in the era of big data, what the challenges of conducting distributed inference instead of centralized inference are, and how statisticians propose solutions to overcome these challenges.

First, we notice that this paper focuses mainly on distributed estimation, and we would like to point out several other works on distributed inference. For smooth loss functions, Jordan et al. (2018) established asymptotic normality for their multi-round distributed estimator, which yields two communication-efficient approaches to constructing confidence regions using a sandwiched covariance matrix. For non-smooth loss functions, Chen et al. (2021) similarly proposed a sandwich-type confidence interval based on the asymptotic normality of their distributed estimator. More generic inference approaches, such as bootstrap, have also been studied in the massive data setting including the distributed framework. The authors reviewed the Bag of Little Bootstraps (BLB) method proposed by Kleiner et al. (2014), which is to repeatedly resample and refit the model at each local machine and finally aggregate the bootstrap statistics. Considering the huge computational cost of BLB, Sengupta et al. (2016) proposed the Subsampled Double Bootstrap (SDB) method, which has higher computational efficiency but requires a large number of local machines to maintain statistical accuracy.

In addition to distributed samples, the dimensionality can also become large in the big data era, and in this case researchers may be more interested in simultaneous inference on multiple parameters. In the centralized setting, bootstrap is one of the solutions to the simultaneous inference problems (Zhang & Cheng, 2017). In a distributed framework where the dimensionality grows, Yu et al. (2020) proposed distributed bootstrap methods for simultaneous inference, which not only are efficient in terms of both communication and computation, but also allow a flexible number of local machines. The idea of their first method `k-grad` is to gather gradient vectors from all of the K local machines and conduct a multiplier bootstrap, which requires a large K. Based on `k-grad`, they developed the second method `n+k-1-grad` by using n gradient vectors from the central machine so that it also allows a small K. The trade-off between the communication efficiency and the statistical accuracy, as mentioned by the authors, was shown through theoretical and numerical studies. Moreover, their theory characterizes a sufficient number of communication rounds that guarantee the optimal statistical accuracy and efficiency, which also provides a practical guide on how to determine the number of communication rounds. Interestingly, this sufficient number of communication rounds is only logarithmically increasing in the number of local machines.

When the dimensionality is higher than the local sample size, or even higher than the total sample size, simultaneous inference becomes of more interest. Yu et al. (2021) extended `k-grad` and `n+k-1-grad` to the high-dimensional domain using de-biased Lasso for high-dimensional estimation and nodewise Lasso for approximating inverse Hessian matrix (Van de Geer et al., 2014). Under the sparsity assumption, they similarly established a sufficient number of communication rounds for guaranteeing the optimal statistical accuracy and efficiency, which is logarithmically increasing in both the number of local machines and the sparsity level in the true parameter and the inverse population Hessian matrix. Given that these methods depend on hyper-parameters for regularization, they also proposed a communication-efficient cross-validation approach to tuning the hyper-parameters.

Second, we want to point out a direction on non-parametric inferences in the distributed frameworks. The existing non-parametric works are all one-shot methods, which are expected to have a limitation of an upper bound on the number of local machines. It would be interesting to see how to bypass this limitation by developing a distributed estimator that allows

CONTACT Guang Cheng chengg@purdue.edu

multiple rounds of communication as in those parametric works.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by NSF [Grant Number 2134209].

References

Chen, X., Liu, W., & Zhang, Y. (2021). First-order newton-type estimator for distributed estimation and inference. *Journal of the American Statistical Association*, 1–17. https://doi.org/10.1080/01621459.2021.1891925

Jordan, M. I., Lee, J. D., & Yang, Y. (2018). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, *114*(526), 668–681. https://doi.org/10.1080/01621459.2018.1429274

Kleiner, A., Talwalkar, A., Sarkar, P., & Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(4), 795–816. https://doi.org/10.1111/rssb.2014.76.issue-4

Sengupta, S., Volgushev, S., & Shao, X. (2016). A subsampled double bootstrap for massive data. *Journal of the American Statistical Association*, *111*(515), 1222–1232. https://doi.org/10.1080/01621459.2015.1080709

S. Van de Geer, Bühlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, *42*(3), 1166–1202. https://doi.org/10.1214/14-AOS1221

Yu, Y., Chao, S.-K., & Cheng, G. (2020). *Simultaneous inference for massive data: Distributed bootstrap*. International Conference on Machine Learning, PMLR (pp. 10892–10901).

Yu, Y., Chao, S.-K., & Cheng, G. (2021). *Distributed bootstrap for simultaneous inference under high dimensionality*. Preprint. arXiv:2102.10080

Zhang, X., & Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, *112*(518), 757–768. https://doi.org/10.1080/01621459.2016.1166114