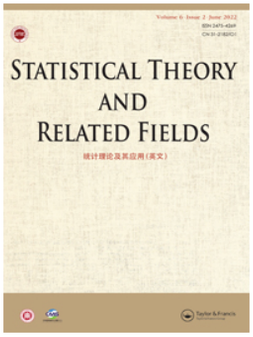

Statistical Theory and Related Fields

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

Discussion on the paper 'A review of distributed statistical inference'

Junlong Zhao

To cite this article: Junlong Zhao (2022) Discussion on the paper 'A review of distributed statistical inference', Statistical Theory and Related Fields, 6:2, 108-110, DOI: 10.1080/24754269.2021.2015861

To link to this article: https://doi.org/10.1080/24754269.2021.2015861

Published online: 16 Dec 2021.

STATISTICAL THEORY AND RELATED FIELDS
2022, VOL. 6, NO. 2, 108–110
https://doi.org/10.1080/24754269.2021.2015861

Taylor & Francis
Taylor & Francis Group

SHORT COMMUNICATION

Discussion on the paper 'A review of distributed statistical inference'

Junlong Zhao

School of Statistics, Beijing Normal University, Beijing, China

ABSTRACT

Distributed statistical inferences have attracted more and more attention in recent years with the emergence of massive data. We are grateful to the authors for the excellent review of the literature in this active area. Besides the progress mentioned by the authors, we would like to discuss some additional development in this interesting area. Specifically, we focus on the balance of communication cost and the statistical efficiency of divide-and-conquer (DC) type estimators in linear discriminant analysis and hypothesis testing. It is seen that the DC approach has different behaviours in these problems, which is different from that in estimation problems. Furthermore, we discuss some issues on the statistical inferences under restricted communication budgets.

ARTICLE HISTORY
Received 11 November 2021
Accepted 20 November 2021

KEYWORDS
Divide and conquer; linear discriminant analysis; hypothesis testing; normal mean

1. Linear discriminant analysis

Linear discriminant analysis (LDA) is a classical classification method (Anderson, 2003). For simplicity, we consider the two-sample problem, assuming that

$$X \sim N_p(\mu_1, \Sigma), \quad Y \sim N_p(\mu_2, \Sigma),$$

where $\mu_i \in \mathbb{R}^p$, $i = 1, 2$ are the mean vectors with $\mu_1 \neq \mu_2$ and $\Sigma \in \mathbb{R}^{p\times p}$ is the covariance matrix. Furthermore, assume that observations come either from X with probability π_1 or from Y with probability π_2 such that $\pi_1 + \pi_2 = 1$. For a new observation Z, Fisher's linear discriminant rule is defined as follows:

$$\psi(Z) = \mathbb{1}\{(Z - \mu_a)^\top \Theta \mu_d > \log(\pi_1/\pi_2)\}, \qquad (1)$$

where $\mu_a = (\mu_1 + \mu_2)/2$, $\mu_d = \mu_1 - \mu_2$, and $\Theta = \Sigma^{-1}$ represents the precision matrix, and $\mathbb{1}\{\cdot\}$ is the indicator function. Suppose that $\{X_i,\ i = 1, \ldots, N_1\}$ and $\{Y_i,\ i = 1, \ldots, N_2\}$ are the independently and identically distributed copies of X and Y, respectively. Let $N = N_1 + N_2$ be the total sample size and suppose that $N > p$. For $i = 1, 2$, denote $\hat{\mu}_i$ as the sample means, and $\hat{\Sigma}_i$ as the sample covariance using observations X_i's and Y_i's, respectively. Then the estimators of μ_a, μ_d and Θ can be defined respectively as follows

$$\hat{\mu}_a = (\hat{\mu}_1 + \hat{\mu}_2)/2, \quad \hat{\mu}_d = \hat{\mu}_1 - \hat{\mu}_2,$$
$$\hat{\Theta} = (\hat{\Sigma}_{\text{pool}})^{-1},$$

where $\hat{\Sigma}_{\text{pool}} = (N_1/N)\hat{\Sigma}_1 + (N_2/N)\hat{\Sigma}_2$ denotes the pooled sample covariance matrix. Then the empirical version of $\psi(Z)$, denoted as $\hat{\psi}(Z)$, can be derived by plugging in the above estimators into (1).

In a distributed setting, one has a central machine (or hub) and many local machines. Suppose that data are split randomly and evenly, and are stored at K local machines. Denote by $\{X_i^{(k)},\ i = 1, \ldots, N_1/K\}$ and $\{Y_i^{(k)},\ i = 1, \ldots, N_2/K\}$ the samples from two classes on the k-th local machine $k = 1, \ldots, K$. Tian and Gu (2017) considered sparse LDA in the high dimensional regime in the case of $\pi_1 = \pi_2 = 1/2$, under the assumption that $\beta = \Theta\mu_d$ is a sparse vector. They proposed a one-shot estimator, which is communication efficient and attains the same convergence rate as the global estimator if $K = O(\sqrt{N/\log p}/\max\{s, s'\})$, where s and s' stand for the sparsity of some parameters.

Li and Zhao (2021) considered the distributed LDA without sparsity assumption under the settings where $p/N \to 0$ and $Kp/N \to r \in [0, 1)$. Note that to compute $\hat{\Sigma}^{-1}$, one needs to transfer p by p matrices to the central machine, of which the communication costs can be expensive. Li and Zhao (2021) proposed a two-round estimator and a one-shot estimator, defined as follows.

Denote by $\hat{\mu}_i^{(k)}$ the estimator of μ_i with data at the kth machine, for $i = 1, 2$, and $k = 1, \ldots, K$. The one-shot estimator considers the following decision rule,

$$\psi_{\text{one}}(Z) = \mathbb{1}\left\{Z^\top \left(K^{-1}\sum_{k=1}^{K} \hat{\Theta}^{(k)}\hat{\mu}_d^{(k)}\right)\right.$$

CONTACT Junlong Zhao zhaojunlong928@126.com

$$-K^{-1}\sum_{k=1}^{K}(\hat{\mu}_a^{(k)})^\top\hat{\Theta}^{(k)}\hat{\mu}_d^{(k)} > \log(N_1/N_2)\Bigg\}, \quad (2)$$

where $\hat{\Theta}^{(k)} = (\hat{\Sigma}^{(k)}_{\text{pool}})^{-1}$ is the pooled sample covariance matrix using the data at the kth machine, $\hat{\mu}_a^{(k)} = (\hat{\mu}_1^{(k)} + \hat{\mu}_2^{(k)})/2$ and $\hat{\mu}_d^{(k)} = \hat{\mu}_1^{(k)} - \hat{\mu}_2^{(k)}$. Note that $\hat{\Theta}^{(k)}$ and $\mu_i^{(k)}$ can be computed with the data only at the kth machine and that it is sufficient to transmit the vectors $\hat{\Theta}^{(k)}\hat{\mu}_d^{(k)} \in \mathbb{R}^p$ and the scalars $(\hat{\mu}_a^{(k)})^\top\hat{\Theta}^{(k)}\hat{\mu}_d^{(k)}$ for all k to the hub. The two-round estimator is an improved version of $\psi_{one}(Z)$, just replacing the local estimators $\hat{\mu}_a^{(k)}, \hat{\mu}_d^{(k)}$ in (2) by the global ones $\hat{\mu}_a, \hat{\mu}_d$ with an additional round of communication. In fact, by transferring $\hat{\mu}_i^{(k)}$'s to the central hub, we can obtain $\hat{\mu}_i = K^{-1}\sum_{k=1}^{K}\hat{\mu}_i^{(k)}$ and consequently $\hat{\mu}_a = (\hat{\mu}_1 + \hat{\mu}_2)/2$ and $\hat{\mu}_d = \hat{\mu}_1 - \hat{\mu}_2$.

Li and Zhao (2021) compared the classification accuracy of the global estimator with those of distributed ones. They showed that when $K = o(N/p)$, both the two-round estimator and the one-shot estimator can be as good as the global one under mild conditions. Moreover, they found if $Kp/N \to r \in [0, 1)$ and $\pi_1 = \pi_2$, the two-round estimator can be as good as the global one, but the one-shot estimator is inferior to the global one. This is an interesting result, since when $Kp/N \to r > 0$, $\hat{\Sigma}^{(k)}_{\text{pool}}$ is not a consistent estimator of Σ by the random matrix theory. Therefore, at the price of more communication cost, the two-round estimator achieves better statistical efficiency.

2. Hypothesis testing of the mean vectors

In this section, we discuss the DC approach in the one-sample testing problem in the distributed system. We observe that DC type test statistics always lead to the loss of power, which is different from that of point estimation where the DC type estimator can be as good as the global one.

Suppose that $X \in \mathbb{R}^p$ is a random vector with $E(X) = \mu$. For a given vector μ_0, consider the hypothesis testing problem

$$H_0 : \mu = \mu_0 \quad \text{v.s.} \quad H_1 : \mu \neq \mu_0.$$

Suppose that X follows the normal distribution $N(\mu, \Sigma)$ with unknown covariance matrix Σ. Let $\{X_i, i = 1, \ldots, n\}$ are independent and identically distributed copies of X. In the setting of $p < n$, the classical test statistic is Hotelling T^2 (Anderson, 2003), defined as follows,

$$T^2 = (n-1)(\bar{X} - \mu_0)^\top\hat{\Theta}(\bar{X} - \mu_0),$$

where $\bar{X}$ denotes the sample mean and $\hat{\Theta} = (\hat{\Sigma})^{-1}$ with $\hat{\Sigma}$ being the sample covariance matrix. In high dimensional cases with $p > n$, the sample covariance matrix is singular and the Hotelling T^2 test statistic is not well defined. Many works are developed to extend the Hotelling T^2 to large or high dimensional regimes (Bai & Saranadasa, 1996; Srivastava & Du, 2008; Wang et al., 2015, etc.).

Du and Zhao (2021) considered the distributed version of these test statistics. Specifically, based on the DC approach, they extended the Hotelling T^2 statistics under the setting $Kp/n \to r \in [0, 1)$ and the nonparametric test statistics of Wang et al. (2015) for high dimensional settings. The ratio of the communication cost of deriving the global test statistics over that of the distributed test statistics is of order $O(p^2)$ in the case of $Kp/n \to r \in [0, 1)$, and $O(p)$ in high dimensional regimes.

They compared the power of distributed statistics with those of global ones, showing that the distributed test statistics are less efficient than those of the global ones whenever $K > 1$. Denote by $\beta_d(n)$ and $\beta_g(n)$ the powers of the distributed and global test statistics as the function of sample size n, respectively, and define n_g/n_d such that $\beta_d(n_d) = \beta_g(n_g)$ as the relative efficiency. The asymptotic relative efficiencies of distributed test statistics have the order $1/\sqrt{K}$.

Hence, the story of the DC approach in the hypothesis problem above is quite different from that of the point estimation, where the mean square error (MSE) of the DC estimators can be as good as that of global ones (Lee et al., 2017; Volgushev et al., 2019; Zhang et al., 2013, etc.). On the other hand, Shi et al. (2018) and Banerjee et al. (2019) showed that, in some non-standard problems, the DC estimators converge at a rate much faster than the global ones. These results show the different behaviours of the DC approach in statistical inferences.

3. Statistical inferences under a restricted communication budget

As discussed before, it is seen that the DC method is communication efficient compared with the global one. But the statistical efficiencies of DC estimators are inferior to the global ones in many cases. To improve the efficiency of the DC estimators, some iterative methods are proposed in the literature at the price of more communication costs. This leads to an interesting problem of how to implement statistical inferences with the given communication budgets.

For the distributed mean estimation, Garg et al. (2014) proved the bounds of the bits in communication required to achieve the minimax square loss. Zhang et al. (2013) and Braverman et al. (2016) found the minimax rate when estimating the mean vector with a restricted communication cost. Cai and Wei (2020) discussed the estimation of the mean vector of a Gaussian distribution with the restriction on communication budget.

However, how to handle the statistical problems with the restricted budget in other settings is an interesting problem for future work. For example, for the hypothesis testing problem discussed in Section 2, how to design test statistics that can achieve good statistical efficiency under a given communication budget needs further investigation.

Disclosure statement

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

ORCID

Junlong Zhao http://orcid.org/0000-0002-1606-7723

References

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis* (3rd ed.). John Wiley & Sons.

Bai, Z., & Saranadasa, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statistica Sinica*, *6*(2), 311–329. https://doi.org/10.1007/s004400050035

Banerjee, M., Durot, C., & Sen, B. (2019). Divide and conquer in nonstandard problems and the super-efficiency phenomenon. *Annals of Statistics*, *47*(2), 720–757. https://doi.org/10.1214/17-AOS1633

Braverman, M., Garg, A., Ma, T., Nguyen, H. L., & Woodruff, D. P. (2016). Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing* (pp. 1011–1020).

Cai, T. T., & Wei, H. (2020). *Distributed Gaussian mean estimation under communication constraints: Optimal rates and communication-efficient algorithms*. arXiv:2001.08877.

Du, B., & Zhao, J. (2021). *Hypothesis testing of one sample mean vector in distributed frameworks*. arXiv:2110.02588.

Garg, A., Ma, T., & Nguyen, H. (2014). On communication cost of distributed statistical estimation and dimensionality. *Advances in Neural Information Processing Systems*, *27*, 2726–2734. https://doi.org/10.1109/hipc.1997.634533

Lee, J. D., Liu, Q., Sun, Y., & Taylor, J. E. (2017). Communication-efficient sparse regression. *Journal of Machine Learning Research*, *18*(1), 115–144. https://doi.org/10.17077/etd.005893

Li, M., & Zhao, J. (2021). Communication-efficient distributed linear discriminant analysis for binary classification. *Statistica Sinica*. https://doi.org/10.5705/ss.202020.0374

Shi, C., Lu, W., & Song, R. (2018). A massive data framework for M-estimators with cubic-rate. *Journal of the American Statistical Association*, *113*(524), 1698–1709. https://doi.org/10.1080/01621459.2017.1360779

Srivastava, M. S., & Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, *99*(3), 386–402. https://doi.org/10.1016/j.jmva.2006.11.002

Tian, L., & Gu, Q. (2017). Communication-efficient distributed sparse linear discriminant analysis. In *Artificial Intelligence and Statistics* (pp. 1178–1187).

Volgushev, S., Chao, S. K., & Cheng, G. (2019). Distributed inference for quantile regression processes. *Annals of Statistics*, *47*(3), 1634–1662. https://doi.org/10.1214/18-AOS1730

Wang, L., Peng, B., & Li, R. (2015). A high-dimensional nonparametric multivariate test for mean vector. *Journal of the American Statistical Association*, *110*(512), 1658–1669. https://doi.org/10.1080/01621459.2014.988215

Zhang, Y., Duchi, J. C., Jordan, M. I., & Wainwright, M. J. (2013). Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Neural Information Processing Systems* (pp. 2328–2336).