A Generalized Mean Approach for Distributed-PCA

Zhi-Yu Jou, Su-Yun Huang,
Institute of Statistical Science, Academia Sinica, Taiwan
Hung Hung*,
Institute of Health Data Analytics and Statistics, National Taiwan University, Taiwan
and
Shinto Eguchi
Institute of Statistical Mathematics, Japan

Abstract

Principal component analysis (PCA) is a widely used technique for dimension reduction. As datasets continue to grow in size, distributed-PCA (DPCA) has become an active research area. A key challenge in DPCA lies in efficiently aggregating results across multiple machines or computing nodes due to computational overhead. Fan et al. (2019) introduced a pioneering DPCA method to estimate the leading rank-r eigenspace, aggregating local rank-r projection matrices by averaging. However, their method does not utilize eigenvalue information. In this article, we propose a novel DPCA method that incorporates eigenvalue information to aggregate local results via the matrix β-mean, which we call β-DPCA. The matrix β-mean offers a flexible and robust aggregation method through the adjustable choice of β values. Notably, for $\beta = 1$, it corresponds to the arithmetic mean; for $\beta = -1$, the harmonic mean; and as $\beta \to 0$, the geometric mean. Moreover, the matrix β-mean is shown to associate with the matrix β-divergence, a subclass of the Bregman matrix divergence, to support the robustness of β-DPCA. We also study the stability of eigenvector ordering under eigenvalue perturbation for β-DPCA. The performance of our proposal is evaluated through numerical studies.

Keywords: distributed computing; eigenvalue perturbation; generalized matrix mean; matrix divergence; PCA

*Corresponding author. *Email:* hhung@ntu.edu.tw