The Annals of Statistics
2016, Vol. 44, No. 6, 2372–2381
DOI: 10.1214/16-AOS1423C
Main article DOI: 10.1214/15-AOS1423

DISCUSSION OF "INFLUENTIAL FEATURE PCA FOR HIGH DIMENSIONAL CLUSTERING"

BY T. TONY CAI[1] AND LINJUN ZHANG

University of Pennsylvania

We would like to congratulate the authors for an interesting paper and a novel proposal for clustering high-dimensional Gaussian mixtures with a diagonal covariance matrix. The proposed two-stage procedure first selects features based on the Kolmogorov–Smirnov statistics and then applies a spectral clustering method to the post-selected data. A rigorous theoretical analysis for the clustering error is given and the results are supported by a competitive performance in numerical studies.

The following discussion is divided into two parts. We will discuss a clustering method based on the sparse principal component analysis (SPCA) method proposed in [5] under mild conditions and compare it with the proposed IF-PCA method. We then discuss the dependent case where the covariance matrix Σ is not necessarily diagonal. To be consistent, we will follow the same notation used in the present paper.

1. A clustering method based on the SPCA procedure given in [5].

In Section 1.6 of the current paper, the authors showed numerically that the proposed IF-PCA method outperforms a clustering method using the SPCA algorithm introduced in [8]. However, the SPCA method in [8] is not designed for the optimal control of principal subspace estimation error and thus does not perform well in the subsequent clustering. The problem of SPCA has been actively studied recently and several rate-optimal procedures for estimating the principal components and principal subspaces have been proposed. See, for example, [2, 4–6].

In this section, we first introduce a clustering algorithm in the setting considered in the present paper using the SPCA procedure introduced in [5], which was shown to be rate-optimal for estimating the principal subspace under a joint sparsity assumption. We then make a comparison of the performance of this SPCA clustering procedure with that of the proposed IF-PCA method both theoretically and numerically. The results show that this SPCA based clustering procedure yields a comparable bound for clustering error rate with that of IF-PCA under mild assumptions and it also performs well numerically.

Received May 2016.

[1] Supported in part by NSF Grants DMS-12-08982 and DMS-14-03708, and NIH Grant R01 CA127334.