

## DISCUSSION OF “INFLUENTIAL FEATURES PCA FOR HIGH DIMENSIONAL CLUSTERING”

BY BOAZ NADLER

*Weizmann Institute of Science*

I commend Jin and Wang on a very interesting paper introducing a novel approach to feature selection within clustering and a detailed analysis of its clustering performance under a Gaussian mixture model.

I shall divide my discussion into several parts: (i) prior work on feature selection and clustering; (ii) theoretical aspects; (iii) practical aspects; and finally (iv) some questions and directions for future research.

**On feature selection and clustering.** Jin and Wang write that the idea of two-stage clustering, consisting of feature selection followed by clustering is not completely new, and cite a paper by Chan and Hall from 2010. I would like to point out that the fact that in high dimensional settings, feature selection is a critical component in successful clustering has been long recognized in the clustering community. Indeed, several methods that select variables on which to cluster have been proposed, see for example [Friedman and Meulman \(2004\)](#), [Law, Figueiredo and Jain \(2004\)](#), [Witten and Tibshirani \(2010\)](#) and references therein for earlier works. These methods are different from IF-PCA as they propose joint feature selection and clustering, solving the resulting non-convex problem by an expectation-minimization approach. However, similar in spirit to the approach presented here, there are also filter methods that choose variables irrespective of the clustering method that will follow, for example, [Dash et al. \(2002\)](#), [He, Cai and Niyogi \(2005\)](#).

In this sense, and also from a practical perspective, it would be interesting to see a comparison of the proposed IF-PCA method to some of these methods and a discussion of their similarities and differences.

**Theoretical aspects.** The model assumed by Jin and Wang is that of a Gaussian mixture model with diagonal covariance, where the mean vectors of the different classes are all sparse, so most features are distributed as a single Gaussian and only a few of them as a mixture of say  $K$  Gaussians with different means. The theoretical analysis considers the limit as both  $p, n \rightarrow \infty$ , where even at the influential features, the separation between the unknown classes all tend to zero.

---

Received May 2016.

*Key words and phrases.* Feature selection, sparse PCA, clustering.