Submitted to the Annals of Statistics
PREPRINT

DISCUSSION OF INFLUENTIAL FEATURES PCA FOR HIGH DIMENSIONAL CLUSTERING

By Boaz Nadler

Weizmann Institute of Science

We 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 Witten and Tibshirani (2012); Friedman and Meulman (2004); Law, Figueiredo and Jain (2004) 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 $p, n \to \infty$, where even at the influential features, the separation between the unknown classes all tend to zero.

Keywords and phrases: feature selection, sparse PCA, clustering

The approach proposed by the authors to select influential features is based on the Kolmogorov-Smirnov distance of the empirical CDF of each feature from the standard normal distribution. Namely, each feature is processed *separately*, without taking into account the possible correlation structure between different features.

Several issues come to mind. The first is whether processing each feature *separately* is indeed optimal. In this context it is interesting to compare the situation to sparse-PCA. There, Johnstone and Lu (2012) proposed a diagonal thresholding method, which selects features based on their individual variances, and proved the asymptotic consistency of their approach as $p, n \to \infty$. However, as explained in Nadler (2009), and proven in Birnbaum *et al.* (2013), this method is not rate-optimal and to achieve the minimax rate one must also consider the covariance structure between the different features. This raises the following question: Is the situation here different, or could one gain improved feature selection (and subsequently improved clustering) by jointly looking at subsets of candidate features as all being influential? Furthermore, can improved feature selection be done in a computationally efficient way ?

To this end, let me propose the following simple *correlation-based* approach to detect influential features: Given a mean-centered $p \times n$ data matrix X, (i) compute the $p \times p$ correlation matrix $\hat{R} = D^{-1/2} \frac{1}{n} X X^T D^{-1/2}$, where D is diagonal with D_{ii} storing the variance in the i-th feature; (ii) Define the following threshold,

$$t(\alpha) = \sqrt{\frac{2}{n}} \sqrt{\log p - \log(\log p) - \log(4\pi) - 2\log(\alpha)}$$

where $0 < \alpha \ll 1$, and may possibly decrease to zero with p. Declare variable i as influential if there exist at least one $j \neq i$ such that $|\hat{R}_{ij}| > t(\alpha)$; (iv) As in IF-PCA, apply k-means clustering on the first $K-1$ singular vectors of the reduced matrix of the data restricted to the influential features.

In Krauthgamer *et al.* (2015) we proposed a similar approach in the context of sparse-PCA with very weak and sparse signals, and observed that empirically it works better than diagonal thresholding. This was subsequently proven by Deshpande and Montanari (2014).

Let us briefly analyze the capabilities of this method, which I shall call `Corr-IF` (correlated influential features). Consider for simplicity the case of $K = 2$ clusters, equally balanced $\delta_1 = \delta_2 = 0.5$, and $s > 1$ influential features all separated at the same distance $2u$. Namely at each of the s influential features, class 1 follows $N(-u, 1)$ and class 2 follows $N(u, 1)$. Clearly, the correlation between an influential and a non-influential feature

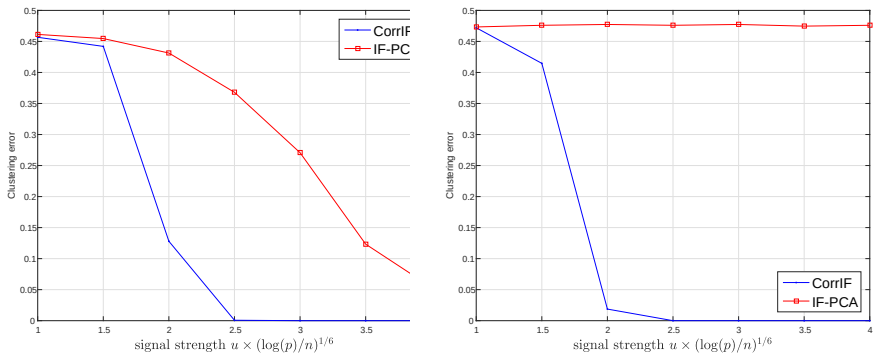

FIG 1. *Comparison of clustering error of* `CorrIF` *and* `IF-PCA`*. (left) Gaussian noise; (right) Non-gaussian noise (exponential and chi-squared) in several non-influential features.*

is zero, whereas if both features i, j are influential, then

$$R_{ij} = \mathbb{E}[\hat{R}_{ij}] = u^2.$$

Let us analyze at what separation values u could a pair of influential features be detected. Consider thus a $p \times p$ sample correlation matrix of multivariate Gaussian observations with diagonal covariance. In the limit as $p, n \to \infty$, each correlation coefficient is approximately distributed as $N(0, 1/n)$, and thus at each row the maximal null sample correlation is sharply concentrated around $\sqrt{(2 \log p)/n}$. Since $\hat{R}_{ij} - R_{ij} = O_P(1/\sqrt{n})$, intuitively, we need $u^2 \gg \sqrt{(\log p)/n}$. Namely, even pairs of influential variables separated at $u \gg (\log p/n)^{1/4}$ will be detected by our proposed `Corr-IF` approach. Since $\log p/n \ll 1$, these are significantly smaller separations than those needed for the IF-PCA method, namely $u = (r \log p/n)^{1/6}$. In fact, if $u > C(\log p/n)^{1/4}$ for a sufficiently large constant C, and the number of influential features $s \ll n$, and $\alpha \to 0$ sufficiently fast (say $\alpha = 1/(p \log p)$), then asymptotically all s influential features will be detected by this approach and at most $O(1)$ non-influential features will be detected, which would lead to perfect clustering. Hence, for this specific setting, this simple method, with computational complexity $O(np^2)$, seems to achieve the minimax rates established in Verzelen and Arias-Castro (2015).

We studied the performance of our method in several simulations. In the first $p = 3000$, $n = 150$, $s = 20$, and the separation in each influential feature is scaled as $u \times (\log p/n)^{1/6}$. Fig. 1 (left) compares the clustering errors of `CorrIF` with parameter $\alpha = 1/p$ and `IF-PCA` as a function of the separation parameter u. As seen in the figure, our method detects the influential features and thus clusters correctly at lower separations where

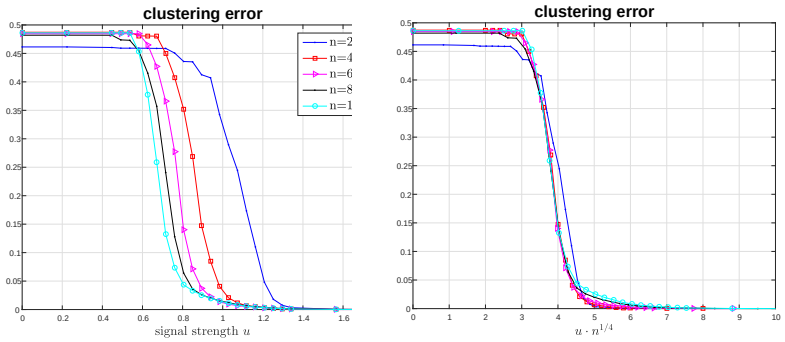

FIG 2. *Clustering error of* `CorrIF` *method vs. signal strength (left) and vs. scaled signal strength (right), for different samples sizes* n.

`IF-PCA` still fails.

In the second simulation we kept $p = 3000$ fixed, considered different sample sizes $n = 200, 400, 600, 800, 1000$, a false alarm parameter $\alpha = 1/p$, a fixed number of influential features $s = 20$, and studied the clustering error of `CorrIF` as a function of signal strength u. The resulting error as a function of u is shown in Fig. 2 (left). Clearly, with a larger sample size, a smaller separation u suffices to detect the influential features and obtain accurate clusters. In Fig. 2 (right) we scaled the signal strength by $n^{1/4}$ as suggested by our analysis. Now all the curves nicely align with each other.

Practical Aspects. Whenever a new procedure is proposed, and in particular when its properties are analyzed under a particular model, a natural question is how robust is the method to deviations from the assumed model. The authors present some very convincing results on several gene expression data. They also perform several simulation studies where the noise is not Gaussian. However, one key assumption in their model as well as in their simulations is that all p features have the *same* noise characteristics. This raises the following question: What happens if the noise in some influential features is of some type, say Gaussian, but in some other non-influential features, it has different characteristics, such as a much slower tail decay ?

To this end we made a simulation at parameters identical to the ones described in the previous section, only that 100 of the non-influential features had exponential distribution with parameter $\lambda = 1$, and another 100 non-influential features had a χ^2 distribution with 5 degrees of freedom. In this case, these 200 features have a distribution which is far Gaussian, and thus have large KS deviations. The `IF-PCA` method selects the 100 exponential variables as influential and a few additional ones from the other features, and

thus fails to cluster correctly at the considered signal separations. In contrast, since sample correlation coefficients are quite robust to the underlying distribution, our proposed method is still able to detect the true influential features and cluster correctly. As shown in Fig. 1(right), its performance is hardly affected by different variables having different noise characteristics.

Due to time and space limitations, I have tried to apply the proposed `Corr-IF` approach to only one real dataset - the Lung1 data. Here, my proposed method did not work so well, simply because this data deviates significantly from the Gaussian mixture assumption. A closer examination reveals that quite a few out of the 12,533 features exhibit "outliers" with extremely large values, hence easily detected by the KS statistic. It turns out these outliers predominantly belong to the class-1 samples, but each sample of class 1 is an outlier on different subsets of features (and not as in the mixture of Gaussian model). Perhaps this calls for a more complex hierarchical Dirichlet type model where at feature j, $X_j = Z_j\mu_j(Y) + noise$, where Z_j is a 0/1 latent variable that decides on whether there is activation at this feature for this sample.

Future Research. The approach presented in this paper raises several questions for future research. Let me just mention one of them: how to adapt the proposed approach to a *semi-supervised setting*, whereby for few out of the n available samples, we do have their class labels. It is then also also natural to ask – what benefit do these few labeled instances provide? In this context, in their seminal work, Castelli and Cover (1995) showed that labeled data has exponentially larger value over unlabeled data, in the sense that each additional labeled sample reduces the probability of error exponentially fast to the Bayes risk. Their analysis assumed fixed p and fixed distributions of the two classes. The situation here is different, as the model assumes that both $p, n \to \infty$, with the separation between different classes at the level of individual coordinates tending to zero. It is thus of interest to better understand the value of few labeled samples under this model.

To conclude, I again wish to comment the authors on an interesting and timely piece of work. In my opinion, despite decades of research, the problem of clustering is still far from being resolved. This work not only suggests a new method in the context of high dimensions, but also raises many new interesting questions for future research.

References.

Birnbaum, A., Johnstone, I. M., Nadler, B. and Paul, D. (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *Annals of statistics* **41** 1055.

Castelli, V. and Cover, T. M. (1995). On the exponential value of labeled samples. *Pattern Recognition Letters* **16** 105–111.

Dash, M., Choi, K., Scheuermann, P. and Liu, H. (2002). Feature selection for clustering-a filter solution. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on* 115–122. IEEE.

Deshpande, Y. and Montanari, A. (2014). Sparse pca via covariance thresholding. In *Advances in Neural Information Processing Systems* 334–342.

Friedman, J. and Meulman, J. (2004). Clustering objects on subsets of attributes. *Journal of the Royal Statistical Society B* **66** 815-849.

He, X., Cai, D. and Niyogi, P. (2005). Laplacian score for feature selection. In *Advances in neural information processing systems* 507–514.

Johnstone, I. M. and Lu, A. Y. (2012). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association* **104** 682-693.

Krauthgamer, R., Nadler, B., Vilenchik, D. et al. (2015). Do semidefinite relaxations solve sparse PCA up to the information limit? *The Annals of Statistics* **43** 1300–1322.

Law, M., Figueiredo, M. and Jain, A. K. (2004). Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26** 1154-1166.

Nadler, B. (2009). Discussion. *Journal of the American Statistical Association* **104** 694–697.

Verzelen, N. and Arias-Castro, E. (2015). Detection and Feature Selection in Sparse Mixture Models. *arxiv.org/1405.1478.v2*.

Witten, D. M. and Tibshirani, R. (2012). A framework for feature selection in clustering. *Journal of the American Statistical Association* **105** 713-726.

Department of Computer Science
Weizmann Institute of Science
Rehovot, Israel