Submitted to the Annals of Statistics

DISCUSSION OF "INFLUENTIAL FEATURES PCA FOR HIGH DIMENSIONAL CLUSTERING", BY J. JIN AND W. WANG

By Natalia A. Stepanova* and Alexandre B. Tsybakov †

School of Mathematics and Statistics, Carleton University, Ottawa, CREST- ENSAE †*

1. Overview of the proposed clustering method. The paper by Jiashun Jin and Wanjie Wang (further referred to as [JW]) addresses an important issue of clustering in Gaussian mixture models. To establish a conceptual framework, one may consider a model which is slightly simpler than in [JW], but presents the same difficulties. Namely, assume that we observe an $n \times p$ matrix $\mathbb{X}$ with rows

$$X_i = \mu_{z(i)} + Z_i, \quad i = 1, \dots, n, \tag{1}$$

where $z : \{1, \dots, n\} \to \{1, \dots, K\}$ is an unknown assignment of the observations to K classes, $\mu_1, \dots, \mu_K$ are unknown vectors in $\mathbb{R}^p$, and $Z_i \in \mathbb{R}^p$ are i.i.d. normal vectors with mean 0 and covariance matrix $\sigma^2 I_p$. Here I_p is the $p \times p$ identity matrix. In [JW], the covariance matrix of Z_1 is diagonal, with the diagonal elements bounded from below and from above by constants red that are independent of p, and there is an additional common mean vector $\bar{\mu}$ in the model. This adds some technicalities but does not change the essence of the problem. Jiashun Jin and Wanjie Wang consider an asymptotic setting where $p \to \infty$ and $n = o(p)$, but K is fixed. It is assumed that the classes $z^{-1}(k), k = 1, \dots, K$, are "balanced" in the sense that their cardinalities $\delta_k = |z^{-1}(k)|$ are greater than $C_0 n$ for some constant $C_0 > 0$ independent of n and p. It is also assumed that the vectors $\mu_1, \dots, \mu_K$ are linearly independent and s-sparse in a group sense, that is, all their non-zero components belong to the same set of indices of size s (called here the sparsity pattern), where $s = p^{1-\vartheta}$ for some $0 < \vartheta < 1$. Note that we may write the model (1) in the form (which is a simplified version of (2.6) in [JW]):

$$\mathbb{X} = LM + Z,$$

where M is an $K \times p$ matrix with rows $\mu_1, \dots, \mu_K$, and L is an $n \times K$ binary matrix with the ith row is equal to the $z(i)$th canonical basis vector.

The clustering problem that is addressed in the paper is to find an estimator $\hat{z}(\cdot)$ of the class assignment $z(\cdot)$ such that the normalized Hamming loss

$$\min_{\phi} \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}\Big(\hat{z}(i) \neq \phi(z(i))\Big)$$

converges to 0 as $p \to \infty$. Here, $\min_\phi$ denotes the minimum over all permutations $(\phi(1), \dots, \phi(n))$ of $(1, \dots, n)$.

Clearly, it is natural to take advantage of sparsity. The IF-PCA procedure of [JW] first selects the sparsity pattern of the vectors $\mu_1, \dots, \mu_K$ based on the Kolmogorov-Smirnov (KS) statistics. The KS statistics are computed independently for each $j \in \{1, \dots, p\}$ based on n-samples $(X_1(j), \dots, X_n(j))$ corresponding to the columns of $\mathbb{X}$. Then, they are compared to a suitably chosen threshold t to perform selection. This is called the feature selection step. Two main definitions of the threshold are considered. One of them is $t = A\sqrt{\log p}$ with a carefully chosen $A > 0$, for which the theoretical results are proved. The second one is a data-driven choice of t based on the version of the Higher Criticism (HC) statistic, for which simulations are performed. We discuss this second choice in

detail below. Assuming that the sparsity pattern is correctly selected with high probability, the columns of $\mathbb{X}$ marked as non-selected are dropped. This reduces the dimension from p to s, leading to a new matrix $\mathbb{X}'$. Finally, the first K unit-norm left singular vectors of $\mathbb{X}'$ are computed and a k-means clustering procedure is applied to obtain the estimated assignment $\hat{z}(\cdot)$.

2. Optimality issue. One of the messages of the paper stated in Section 2.4 is that "clustering and feature selection are possible and non-trivial" only when the non-zero components of the vectors μ_k are greater than the critical values of order $((\log p)/n)^{-1/6}$. Section 2.4 also discusses some related "phase transitions". However, the results deal with one particular method suggested in the paper. There is no guarantee that the method is optimal. Moreover, only some upper bounds on the rates are obtained, and there is no guarantee that the bounds are tight even for this particular method. So, the phase transitions are only related to these upper bounds for the proposed procedure and are not shown to represent a general phenomenon. The $((\log p)/n)^{-1/6}$ critical value and the corresponding rates for the Hamming loss appear to be too pessimistic[1] - the critical features are required to be rather strong.

An interesting question is to get more insight into the problem and to investigate critical thresholds that cannot be improved with any method. This task needs developing minimax lower bounds on the Hamming loss for suitably defined sparsity classes of vectors $\mu_1, \ldots, \mu_K$. Since there is no minimax setting in [JW], it is difficult to say what would be the best rate or the smallest critical value of useful features. However, using other methods than in [JW], one can achieve better critical thresholds in a natural minimax setting for this model. The argument is as follows. The fact that the jth column of the matrix LM contains at least one non-zero element implies that it contains at least $C_0 n$ non-zero elements, since the same row appears in LM at least $C_0 n$ times (recall that $C_0 n$ is a lower bound on the size of any of K classes). Thus, for each j, we deal with a detection problem for a normal means model in $\mathbb{R}^n$ where the mean vector is $C_0 n$-sparse. This is a "dense" case of testing problem, and it is well-known that the minimal absolute value of the non-zero components, for which successful detection is possible, is of order $n^{-1/4}$ (see, for example, [2], Corollary 2, taking there $d = n$, $s = C_0 n$). The optimal test is based on a chi-square type statistic and the testing errors decrease exponentially in n. Thus, taking the maximum over $j = 1, \ldots, p$ (that can only have a logarithmic influence on rates), we find that the sparsity pattern is correctly selected with high probability when the non-zero components of vectors μ_k are of order $n^{-1/4}$ (maybe up to logs), which is much better than $((\log p)/n)^{-1/6}$. One should note that this argument is valid under the assumption that σ is known. There remains a question of whether the knowledge of σ is so crucial that the rate of testing changes dramatically when σ is unknown. There are some problems where it is indeed the case. However, the analogy with Verzelen and Arias-Castro [6, Section 3.3], who considered a very similar problem, suggests that the case of diagonal covariance matrix does not present this anomaly. Usually, it is enough to get a rough over-estimate of σ for the chi-square statistic to work.

3. The form of the HC statistic for feature selection. A key step of the procedure in [JW] consists in selecting the sparsity pattern by means of thresholding at some level t. Section 1.3 in [JW] introduces a data-driven threshold selection method based on the HC statistic (the thresholding step of the IF-PCA procedure), which is only explored in [JW] numerically. We have some questions concerning this method of threshold selection.

[1] The rate $((\log p)/n)^{-1/6}$ is a logarithmic rate for all reasonable sample sizes since $n^{-1/6} < \log n$ if $n < 10^8$.

The HC statistic defined in (1.11) of [JW] has the form

$$\mathrm{HC}_{p,j} = \frac{\sqrt{p}(j/p - \pi_{(j)})}{\sqrt{\max\left\{\sqrt{n}(j/p - \pi_{(j)}), 0\right\} + j/p}}, \quad 1 \le j \le p/2. \tag{2}$$

Based on it, the index

$$\hat{j} = \mathrm{argmax}_{1 \le j \le p/2, \pi_{(j)} > (\log p)/p} \mathrm{HC}_{p,j},$$

is selected, and the threshold $t = t_p^{\mathrm{HC}}$ is defined as the $\hat{j}$th largest KS score given by formula (1.6) in [JW].

Observe that it is a one-sided statistic, which makes sense only if the non-negative scores $\psi_{n,j}$ are considered. However, the summary of the IF-HCT-PCA method in Table 3 of [JW] employs the centered scores $\psi^*_{n,j}$. For such scores, some significant features may correspond to highly negative values of $j/p - \pi_{(j)}$. Thus, if most of the scores are negative, the method nevertheless will stop at $\hat{j}$, corresponding to a positive score, no matter how small it is in absolute value. Therefore, we wonder whether the authors mean using $|\mathrm{HC}_{p,j}|$ rather than $\mathrm{HC}_{p,j}$.

The role of the term $Q \triangleq \max\left\{\sqrt{n}(j/p - \pi_{(j)}), 0\right\}$ in (2) remains unclear. The discussion at the end of Section 1.3 only arrives to the conclusion that the function $\mathrm{HC}_p(t)$ is monotone between the adjacent discontinuities. But this property is valid with many other choices of additive terms in the denominator, not necessarily with the term Q. Furthermore, it is not clear why this property is so important. Overall, we were not able to follow the discussion after the word 'Remark' at the end of Section 1.3.

Finally, we wonder where does the exact form of the constraint $\pi_{(j)} > (\log p)/p$ come from. The authors write that it prevents an ill behavior of $\mathrm{HC}_{p,j}$ for small j, by analogy with the HC statistic of Donoho and Jin [4]. However, in [4] we find the truncation at $1/p$ rather than at $(\log p)/p$. Moreover, the $1/p$ truncated and non-truncated test statistics have the same asymptotic distribution under the null hypothesis. This is conjectured in [4, page 974], and proved in [5]. One can also prove that the $(\log p)/p$ truncation leads to the same distribution. Therefore, all these truncations change nothing from the theoretical point of view – the asymptotic distribution is not affected. However, they may turn out to be extremely important for the output of the HC type procedures in practice. In view of this, choosing one of the many, asymptotically equivalent, possible levels, such as $(\log p)/p$, looks more like a rough guide rather than something precisely recommended. Why not, for example, $5(\log p)/p$ or $(\log p)^2/p$? Then, although nothing changes in the theory, the behavior of the procedure in practice may become dramatically different. We suspect that the literal application of the constraint $\pi_{(j)} > (\log p)/p$ may actually result in a poor behavior in reality, and all the story is rather a way to say that some truncation may be needed, though ultimately chosen by hand. An open problem is to choose the truncation via a self-tuning adaptive procedure. An alternative approach is to avoid any truncation and use a penalization in the denominator of the test statistic (see [3, 5]). For example, a possible modification that is based on Theorem 4.2.3 of [3] is to replace $\mathrm{HC}_{p,j}$, $1 \le j \le p/2$, by the statistics

$$\widehat{\mathrm{HC}}_{p,j} \triangleq \frac{\sqrt{p}(j/p - \pi_{(j)})}{\sqrt{\max\left\{\sqrt{n}(j/p - \pi_{(j)}), 0\right\} + (j/p)\log\log(p/j)}}, \quad 1 \le j \le p/2,$$

and set $\hat{j} = \mathrm{argmax}_{1 \le j \le p/2} \widehat{\mathrm{HC}}_{p,j}$.

4. Assumptions on the model and applications. The normality of the errors and the diagonal structure of the covariance matrix are assumed throughout the paper. These seem to be very strong assumptions in view of applications to the analysis of the considered gene microarray data sets. In genomics, where a degree of correlation is high within a group of genes sharing the same biological pathway, the data are of different kind (see, for example, [1]). The assumption that $n = p^\theta$ and $s = p^{1-\vartheta}$ is also very specific and could be made more general. The main result of the paper, Theorem 2.2, that ensures consistency of the proposed estimation procedure is stated under these and some other quite restrictive assumptions. For example, the threshold is not data-driven but should depend on the unknown θ or ϑ via q to get the rates in Corollary 2.2. At the same time, for several data sets used in Section 1 of [JW], the empirical method, for which no theory has been provided, gives reasonably good numerical results. This method, as defined, is completely data-driven. In view of the above comments concerning the constraint $\pi_{(j)} > (\log p)/p$, we wonder whether in the numerical experiments the truncation is done exactly in this form, or it is chosen by hand.

References.

[1] Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, **7**, 781–791.

[2] Comminges, L., Collier, O., and Tsybakov, A.B. (2015). Minimax estimation of linear and quadratic functionals on sparsity classes. To appear in *Ann. Statist. http://arxiv.org/abs/1502.00665*

[3] Csörgő, M., Csörgő, S., Horváth, L., and Mason, D. (1986). Weighted empirical and quantile processes. *Ann. Probab.*, **14**, 31–85.

[4] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, **32**, 962–994.

[5] Stepanova N. and Pavlenko T. (2016). Goodness-of-fit tests based on sup-functionals of weighted empirical processes. *http://arxiv.org/abs/1406.0526*

[6] Verzelen, N. and Arias-Castro, E. (2014). Detection and feature selection in sparse mixture models. *http://arxiv.org/abs/1405.1478*