*Submitted to the Annals of Statistics*

# DISCUSSION

By Ery Arias-Castro

*University of California, San Diego, USA*

AND

By Nicolas Verzelen

*INRA, UMR 729 MISTEA, F-34060 Montpellier, FRANCE*

We offer below some constructive criticism that, we hope, will shed some light, or least provide a different perspective, on different points touched in the paper as regards to the problem of sparse clustering. We hope this will stimulate a fruitful discussion of the topic.

Before that, we want to congratulate the authors for a *tour de force* in mathematical technique. The authors went for the apparently unreachable goal of obtaining a performance result — sharp to the multiplicative constant — for a sophisticated method addressing a complex problem. This continues an impressive line of papers by Jiashun Jin and his students, postdocs and collaborators. Every time, the goal is extremely ambitious: that of providing constant-sharp phase transition results for central problems in high-dimensional statistics. In fact, despite the fact that the paper under discussion is quite substantial, it is only part of a larger program that aims at precisely describing the phase transitions in the context of sparse clustering — see (Jin et al., 2015, 2016) and also (Jin, 2015).

**1. The review of the literature.** The problem of sparse clustering can be defined as that of clustering possibly high-dimensional (feature) vectors in a setting where only a few features are useful. In their review of the literature, the authors discuss two papers addressing the problem of sparse clustering (Chan and Hall, 2010; Azizyan et al., 2013). They also cite ours (Verzelen and Arias-Castro, 2014) somewhere in the middle of the paper. These papers all appeared in the last few years and this may give the impression that the problem was only considered recently. This is in fact not the case. Although minuscule relative to the literature on sparse regression and classification, the literature on sparse clustering is nontrivial. Friedman and Meulman (2004), in their impactful paper on the topic, cite papers from the 1980's, e.g., (De Soete, 1986). Another important paper is that of Witten and Tibshirani (2010).

Not mentioning this literature, or discussing it properly, weakens the paper in at least two respects. First, it has the potential of misleading the non-expert reader into believing that the problem is new, which it is not, and the same reader will not be able to appreciate the amount of novelty offered here (more on this in the sections that follow). Second, it severely limits the scope of the numerical experiments performed in the paper, as the method proposed by the authors is only compared to methods that are not tailored to sparse clustering, such as K-means and SpectralGEM. It seems more reasonable to use COSA (Friedman and Meulman, 2004) or Sparse K-means (Witten and Tibshirani, 2010) as benchmarks — not only on real data but also in simulated data.

To tackle the problem of sparse clustering, the general strategy followed by the authors is very natural:

1. Select the features that are useful for clustering.
2. Apply a clustering algorithm based only on the selected features.

In this contribution, the authors combine a new feature selection procedure calibrated by higher-criticism thresholding with a spectral clustering method. Although the mathematical analysis is really impressive, it is difficult to disentangle the respective merits of the different ingredients of the procedure in the paper (IF-PCA):

i. Coordinate-wise normality testing by Kolmogorov-Smirnov (KS).
ii. Calibration by the higher criticism (HC).
iii. Spectral graph partitioning using a measure of similarity based on the KS statistics.

For example, would the method perform as well if HC were replaced by the Benjamini-Hochberg method for FDR control?

**2. The selection step.**

*Robustness to non-normality.* The selection step in IF-PCA is based on computing, for each coordinate, the KS statistic with the standard normal distribution as null distribution. This is done after each variable is standardized. This test will find significance when there is substantial departure from normality even in useless features, for example, coordinates where the data are strongly unimodal but not normal. There is no averaging — and therefore no central limit theorem — that can help here so that this issue persists even in the large-sample limit. Although the authors adjust the $p$-values following the empirical null approach proposed by Efron (2004), we do not see how this can correct this issue. It would be interesting to see how

the method behaves when the data are not normal. The authors point to the fact that their method does well on microarray data, but the sample sizes are a bit small, and we believe that trying the method of larger simulated datasets would be more revealing.

As a truly nonparametric method, we find the (coordinate-wise) procedure of Chan and Hall (2010) appealing and possibly promising in some applications. The method is based on coordinate-wise testing for unimodality. It is clear, however, that this method is substantially inferior to the KS normality test in the normal setting considered in the paper under discussion.

*Dependency of the rates with respect to the number of clusters $K$.* (Here we use the notation of Section 2 in the paper.) In (2.12), the authors assume that for any relevant feature $j$

$$\tau(j) := \sqrt{n}\Big|\sum_{k=1}^{K} \delta_k m_k^3(j)\Big| \gtrsim \sqrt{\log(p)}\ . \tag{1}$$

This assumption drives the SNR required for successful detection

$$m_k(j) \gtrsim (\log(p)/n)^{1/6}\ , \tag{2}$$

at least in some of the regimes — see Section 2.4. Deriving the moment of $W(j)$, we can provide another interpretation of this rate. Indeed, we have

$$\mathbb{E}[W(j)] \approx 0\ , \quad \mathbb{E}[W^2(j)] \approx 1\ , \quad \mathbb{E}[W^3(j)] \approx \frac{\sum_{k=1}^{K} \delta_k m_k^3(j)}{[1+\sum_{k=1}^{K} \delta_k m_k^2(j)]^{3/2}}\ ,$$

where $\approx$ comes from the fact that $W(j)$ is (only) empirically centered and normalized. The feature detection procedure amounts to testing whether $W(j)$ follows a centered normal distribution. The detection rate is driven by the skewness $\mathbb{E}[W^3(j)]/\,\mathbb{E}[W^2(j)]^{3/2}$ of the distribution — and this is implicitly what $\tau(j)$ is quantifying.

In Section 2.4, the authors observe that Assumption (2.12) may fail in very simple settings such as symmetric mixtures ($K = 2$ and $\delta_1 = \delta_2$), in which case all $\tau(j)$ are equal to zero. In that symmetric case, they argue that the detection rates will be driven by $\tau_4(j) := \sum_{k=1}^{K} \delta_k m_k^4(j)$ — the fourth power instead of the third power as in (1) — and this will result in the following condition on the SNR

$$m_k \gtrsim (\log(p)/n)^{1/8}\ ,$$

which is obviously stronger than (2). Just like $\tau(j)$ appears in the third power of $W(j)$, $\tau_4(j)$ occurs in the fourth moment $\mathbb{E}[W^4(j)]$ so that, by the same

reasoning, the detection rate is driven by the kurtosis $W(j)$ compared to the one of a normal distribution. (This phenomenon is carefully analyzed in our own work (Verzelen and Arias-Castro, 2014).) Interestingly, the Kolmogorov-Smirnov statistic seems to adapt to these moment conditions!

Going one step further, for some choices of parameters, the moments of $W(j)$ coincide with those of a normal distribution up to order $2K-1$. Indeed, this construction is equivalent to a partial moment problem — see (Karlin and Studden, 1966). We can speculate that the detection rate will then be driven by $\tau_{2K}(j) := \sum_{k=1}^{K} \delta_k m_k^{2K}(j)$ which we believe will results in the condition

$$m_k \gtrsim (\log(p)/n)^{1/4K} \ .$$

This exponential dependency of the rates with respect to the number of clusters $K$ is in line with recent results of Moitra and Valiant (2010) (among others) on parameter estimation for Gaussian mixtures.

Although the authors did not conduct the analysis of their procedure in this regime, it is likely that their IF-PCA adapts to this situation. The two main points we want to make are:

- The detection (and clustering rates) are much worse than $(\log(p)/n)^{1/6}$ when Assumption (2.12) is not satisfied.
- IF-PCA may be able to adapt to the minimax rate regardless.

## 3. The clustering step.

*Spectral algorithms.* The authors call their clustering routine 'PCA', which they equate with the SpectralGEM algorithm of Lee et al. (2010). Our understanding is that this may not true to the letter. In a nutshell, their clustering routine (after feature selection) is as follows:

i. Project the standardized observations onto their top $K-1$ principal components.
ii. Apply Lloyd's algorithm[1] for K-means.

This amounts to forming the affinity matrix $\mathbf{A} = \mathbf{W}\mathbf{W}^\top = (A_{ii'} : i, i' \in [n])$, where[2] $A_{ii'} = \langle W_i, W_{i'} \rangle$, and performing spectral clustering directly on $\mathbf{A}$, for example, as in (Lei and Rinaldo, 2015). This has been standard for a while, although working with some form of graph Laplacian seems to be more popular (Von Luxburg, 2007).

[1] This is what the `kmeans` function of Matlab does.
[2] $\langle \cdot, \cdot \rangle$ denotes the inner product.

The latter is essentially what SpectralGEM does. SpectralGEM is indeed very similar to, but not quite the same as applying the spectral graph partitioning of Ng et al. (2002) to the affinity matrix $\mathbf{B} = (B_{ii'} : i, i' \in [n])$, where $B_{ii'} = \sqrt{\max(\langle W_i, W_{i'} \rangle, 0)}$. (A minor detail: SpectralGEM uses Ward's algorithm for K-means. It also includes a step for estimating the number of clusters based on the eigengap.)

*Beyond common covariance matrices.* Throughout this paper, it is assumed that all the components share the same diagonal covariance matrix $\Sigma$. It is perhaps possible to modify this procedure to allow different covariance matrices $\Sigma_k = \mathrm{Diag}(\sigma_k^2(1), \ldots, \sigma_k^2(p))$, $k = 1, \ldots, K$. The coordinate-wise KS test will still be able to distinguish coordinates $j$ whose corresponding marginal distribution of $X(j)$ is normal, that is $\sigma_1^2(j) = \ldots = \sigma_K^2(j)$ and $\mu_1(j) = \ldots = \mu_K(j)$, from relevant coordinates $j$ whose corresponding marginal distribution of $X_j$ is a non-trivial Gaussian mixture. It is not clear to us whether constant-sharp bounds can be derived in this setting, but (up to multiplicative constants) rates seem within reach. As for the clustering step, simple spectral methods (such as 'PCA') will fail to recover the clusters when the covariances are different. Nevertheless, there exists an important body of work on provably consistant learning methods for this problem (see e.g., Vempala and Wang, 2004; Achlioptas and McSherry, 2005; Kannan et al., 2005) that one could plug into the feature selection step.

*A tribute to Peter Hall.* Peter Hall passed away very recently (early January, 2016). His legendary prolific contribution to mathematical and methodological statistics, as well as probability theory, includes some work related to the paper under discussion, in particular (Chan and Hall, 2010), and also extensive work on the higher criticism (Hall et al., 2008, 2010; Delaigle et al., 2011; Delaigle and Hall, 2009; Hall et al., 2014), the bulk of it with Jiashun Jin.

*Acknowledgements.* This work was partially supported by the US Office of Naval Research (N00014-13-1-0257) and the French Agence Nationale de la Recherche (ANR 2011 BS01 010 01 projet Calibration).

**References.**

Achlioptas, D. and F. McSherry (2005). On spectral learning of mixtures of distributions. In *Learning Theory*, pp. 458–469. Springer.

Azizyan, M., A. Singh, and L. Wasserman (2013). Minimax theory for high-dimensional gaussian mixtures with sparse mean separation. *Neural Information Processing Systems (NIPS)*.

Chan, Y. and P. Hall (2010). Using evidence of mixed populations to select variables for clustering very high-dimensional data. *J. Amer. Statist. Assoc. 105*(490), 798–809.

De Soete, G. (1986). Optimal variable weighting for ultrametric and additive tree clustering. *Quality and Quantity 20*(2-3), 169–180.

Delaigle, A. and P. Hall (2009). Higher criticism in the context of unknown distribution, non-independence and classification. *Perspectives in Math-ematical Sciences I: Probability and Statistics*, 109–138.

Delaigle, A., P. Hall, and J. Jin (2011). Robustness and accuracy of methods for high dimensional data analysis based on student's t-statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(3), 283–301.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc. 99*(465), 96–104.

Friedman, J. and J. Meulman (2004). Clustering objects on subsets of attributes. *J. Roy. Statist. Soc. Ser. B 66*, 815—849.

Hall, P., J. Jin, et al. (2008). Properties of higher criticism under strong dependence. *The Annals of Statistics 36*(1), 381–402.

Hall, P., J. Jin, et al. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics 38*(3), 1686–1732.

Hall, P., J. Jin, H. Miller, et al. (2014). Feature selection when there are many influential features. *Bernoulli 20*(3), 1647–1671.

Jin, J. (2015). Fast community detection by score. *The Annals of Statistics 43*(1), 57–89.

Jin, J., Z. T. Ke, and W. Wang (2015). Phase transitions for high dimensional clustering and related problems. *arXiv preprint arXiv:1502.06952*.

Jin, J., Z. T. Ke, and W. Wang (2016+). Optimal spectral clustering by higher criticism thresholding. Manuscript.

Kannan, R., H. Salmasian, and S. Vempala (2005). The spectral method for general mixture models. In *Learning Theory*, pp. 444–457. Springer.

Karlin, S. and W. J. Studden (1966). *Tchebycheff systems: With applications in analysis and statistics*. Pure and Applied Mathematics, Vol. XV. Interscience Publishers John Wiley & Sons, New York-London-Sydney.

Lee, A. B., D. Luca, K. Roeder, et al. (2010). A spectral graph approach to discovering genetic ancestry. *The Annals of Applied Statistics 4*(1), 179–202.

Lei, J. and A. Rinaldo (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist. 43*(1), 215–237.

Moitra, A. and G. Valiant (2010). Settling the polynomial learnability of mixtures of gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pp. 93–102. IEEE.

Ng, A., M. Jordan, and Y. Weiss (2002). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems 2*, 849–856.

Vempala, S. and G. Wang (2004). A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences 68*(4), 841–860.

Verzelen, N. and E. Arias-Castro (2014). Detection and feature selection in sparse mixture models. *arXiv preprint arXiv:1405.1478*.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing 17*(4), 395–416.

Witten, D. M. and R. Tibshirani (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association 105*(490), 713–726.