The Annals of Statistics
2016, Vol. 44, No. 6, 2360–2365
DOI: 10.1214/16-AOS1423A
Main article DOI: 10.1214/15-AOS1423

DISCUSSION OF "INFLUENTIAL FEATURES PCA FOR HIGH DIMENSIONAL CLUSTERING"[1]

BY ERY ARIAS-CASTRO AND NICOLAS VERZELEN

University of California and INRA

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, Ke and Wang (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 [Azizyan, Singh and Wasserman (2013), Chan and Hall (2010)]. 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, for example, 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 nonexpert reader into believing that the problem is new, which it is not, and the same reader will not be

Received May 2016.

[1] Supported in part 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).