Journal of the Korean Statistical Society 39 (2010) 135–136

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Discussion

Discussion: The forward search: Theory and data analysis☆

L.A. García-Escudero, A. Gordaliza*, A. Mayo-Iscar

Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Valladolid, Spain[1]

ARTICLE INFO

Article history:
Received 26 January 2010
Available online 18 February 2010

First of all, we would like to congratulate the authors for the interesting and exhaustive research work they have developed over the last years to explore the "forward search" methodology and its useful applications. This comprehensive survey nicely summarizes all the stimulating work done by the research group on this topic.

We completely agree with the authors regarding the claim that a *dynamic* view of data sets is a more interesting approach than a *static* one. This opinion is clearly confirmed through the close connections existing between some works developed by our research team (Cuesta-Albertos, Matran, & Mayo-Iscar, 2008a,b; García-Escudero & Gordaliza, 2005, 2007; García-Escudero, Gordaliza, & Matrán, 2003; García-Escudero, Gordaliza, Matrán, & Mayo-Iscar, submitted for publication) and the surveyed "forward search" techniques.

A dynamic view of data sets throughout a moving range of trimming levels is especially appealing in Robustness and Clustering settings. Let us just briefly comment on two problems tackled by our research team for which the "forward search" approach makes perfect sense:

1. *"Automatic" determination of the contamination level:* Several trimming approaches existing in the statistical literature are designed for removing a proportion of "most outlying" data which is fixed in advance. However, it is not at all obvious which is the "right" trimming level α that should be chosen for a given data set. Notice that the proper determination of α actually serves to declare which are the data points that will finally be labeled as outliers. Riani, Atkinson, and Cerioli (2009) presents a "forward search"-based methodology for making this automatic choice through a graphical tool based on "envelopes". A closely related approach was followed in García-Escudero and Gordaliza (2007) by using the so-called radius process and the associated confidence bands. Theoretical properties of this radius process (viewed as a stochastic process indexed on the trimming level α) were also analyzed there. A new look of the "forward search" methodology from the stochastic processes point of view, may be interesting, for instance, for studying the optimal envelopes that accompany the "forward search" plots. An iterative procedure was also presented in García-Escudero and Gordaliza (2007) in order to obtain an automatic determination of the contamination level.

   We claim that the proper estimation of the scale parameter ($\sigma = |\Sigma|^{1/p}$, with Σ being the covariance matrix), by knowing only a "central" part of the random sample, is the key point for the proper determination of the contamination level. This idea has been explored in Cuesta-Albertos et al. (2008b), obtaining theoretical results of interest.
2. *Robust clustering:* As the authors also point out, there exist clear connections between Cluster Analysis and (multivariate) outlier identification techniques. Dynamic trimming techniques can provide appropriate tools to unify the simultaneous exploration of the presence of clusters and outliers. Bearing this in mind, it could be said that the SSC methodology

☆ Research partially supported by the Spanish Ministerio de Ciencia e Innovación, grant MTM2008-06067-C02-01, and 02 and by Consejería de Educación y Cultura de la Junta de Castilla y León, GR150.

* Corresponding author.
E-mail address: alfonsog@eio.uva.es (A. Gordaliza).

[1] Departamento de Estadística e Investigación Operativa, Facultad de Ciencias, Universidad de Valladolid, 47002, Valladolid, Spain.

doi:10.1016/j.jkss.2010.01.005

presented in García-Escudero and Gordaliza (2007) is a particular type of "forward search" methodology. The main difference with the approach followed in Atkinson, Riani, and Cerioli (2006) and Atkinson and Riani (2007) is that the SSC methodology starts from a "complete" robust clustering solution (k initial groups are obtained by using a very high trimming level) while Atkinson and co-authors' proposals start from several random initializations made of just one group ($k = 1$). The proper estimation of the scales of the groups' covariance matrices again plays a key role in the SSC approach. This proper estimation of the scale parameters is not an easy task since it is strongly influenced by the presence of possible different sizes for the groups.

Another important task where the "forward search" methodology may be extremely useful, is the determination of the number of groups k in a Cluster Analysis problem. This problem is hard to solve without taking into account simultaneously the one of estimating the contamination level α. In fact, they are two closely connected problems. The dynamic view of data in terms of the trimming fraction α and the number of groups k is an ingenious way to perform this simultaneous analysis. This was the methodology followed in García-Escudero et al. (2003) for trimmed k-means, which work well under the implicit constraint of spherical covariance structures for the groups. Other more general constraints on the groups' covariance matrices have been taken into account in García-Escudero et al. (submitted for publication).

References

Atkinson, A. C., Riani, M., & Cerioli, A. (2006). Random start forward searches with envelopes for detecting clusters in multivariate data. In S. Zani, A. Cerioli, M. Riani, & M. Vichi (Eds.), *Data analysis, classification and the forward search* (pp. 163–171). Berlin: Springer-Verlag.

Atkinson, A. C., & Riani, M. (2007). Exploratory tools for clustering multivariate data. *Computational Statistics & Data Analysis*, *52*, 272–285.

Cuesta-Albertos, J. A., Matran, C., & Mayo-Iscar, A. (2008a). Robust estimation in the normal mixture model based on robust clustering. *Journal of the Royal Statistical Society. Series B*, *70*, 779–802.

Cuesta-Albertos, J. A., Matran, C., & Mayo-Iscar, A. (2008b). Trimming and likelihood: Robust location and dispersion estimation in the elliptical model. *The Annals of Statistics*, *36*, 2284–2318.

García-Escudero, L. A., & Gordaliza, A. (2005). Generalized radius processes for elliptically contoured distributions. *Journal of the American Statistical Association*, *471*, 1036–1045.

García-Escudero, L. A., & Gordaliza, A. (2007). The importance of the scales in heterogeneous robust clustering. *Computational Statistics & Data Analysis*, *51*, 4403–4412.

García-Escudero, L. A., Gordaliza, A., & Matrán, C. (2003). Trimming tools in exploratory data analysis. *Journal of Computational and Graphical Statistics*, *12*, 434–449.

García-Escudero, L. A., Gordaliza, A., Matrán, C., & Mayo-Iscar, A. (2010). Exploring the number of groups in robust model-based clustering (Submitted for publication). Preprint available at: http://www.eio.uva.es/infor/personas/langel.html.

Riani, M., Atkinson, A. C., & Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society. Series B*, *71*, 447–466.