Journal of Multivariate Analysis 117 (2013) 14–31

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

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

Strong consistency of k-parameters clustering

María Teresa Gallegos [a], Gunter Ritter [a,b,*]

[a] Institute for Data Analysis, D-94121 Salzweg, Germany
[b] Faculty of Informatics and Mathematics, University of Passau, D-94030 Passau, Germany

ARTICLE INFO

Article history:
Received 2 June 2012
Available online 31 January 2013

AMS subject classifications:
primary 62H30
secondary 62F12

Keywords:
Cluster analysis
Classification models
Elliptical models
Maximum likelihood estimation
Strong consistency

ABSTRACT

Pollard showed for k-means clustering and a very broad class of sampling distributions that the optimal cluster means converge to the solution of the related population criterion as the size of the data set increases. We extend this consistency result to k-parameters clustering, a method derived from the *heteroscedastic, elliptical* classification model. It allows a more sensitive data analysis and has the advantage of being affine equivariant. Moreover, the present theory yields a consistent criterion for selecting the number of clusters in such models.

1. Introduction

The k-means algorithm, Steinhaus [32], enjoys great popularity in data analysis, knowledge discovery, and vector quantization in order to partition a data set $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^d$ in a given number $g \geq 2$ of clusters. Let $x_1, \dots, x_n$ be the data set to be clustered, let ℓ_i be the label of the cluster of data point x_i, $1 \leq i \leq n$, and write $\boldsymbol{\ell} = (\ell_1, \dots, \ell_n)$. If $\overline{x}_j(\boldsymbol{\ell})$ stands for the cluster mean and $W_j(\boldsymbol{\ell}) = \sum_{i:\ell_i=j}(x_i - \overline{x}_j(\boldsymbol{\ell}))(x_i - \overline{x}_j(\boldsymbol{\ell}))^\top$ for the SSP matrix ("sum of squares and products") of cluster j w.r.t. $\boldsymbol{\ell}$, this algorithm computes the *Pooled Trace* (or *Ward's*) *criterion*

$$\underset{\boldsymbol{\ell}}{\operatorname{argmin}}\ \operatorname{tr} \sum_{j=1}^{g} W_j(\boldsymbol{\ell}) = \underset{\boldsymbol{\ell}}{\operatorname{argmin}} \sum_{j=1}^{g} \sum_{\ell_i=j} \|x_i - \overline{x}_j(\boldsymbol{\ell})\|^2.$$

It tends to produce spherical clusters of about equal size and about equal scatter, if the data set allows this. In fact, Bock [5] revealed the criterion as the ML estimator of a homoscedastic, isobaric, normal clustering model with spherical covariance matrices; see also Bock [6]. Properties of the estimator and the algorithm are well known. In particular, MacQueen [25] showed that the k-means algorithm reduces the criterion (and coined its name). Bryant and Williamson [8] studied the asymptotic behavior of a class of classification ML estimators and applied their result to a univariate, homoscedastic mixture of normal populations. Pollard [29,30] proved for a very broad class of sampling distributions and *homoscedastic, isobaric, spherical* statistical models that the optimal means converge as the size of the data set increases. He also identified the limit as the solution to the related population criterion. This means that the *global* maximum is the favorite solution if the data set is large. His result is remarkable inasmuch as the sampling distribution may be very general and very different from

[*] Corresponding author at: Institute for Data Analysis, D-94121 Salzweg, Germany.
E-mail address: ritter@fim.uni-passau.de (G. Ritter).

0047-259X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2013.01.013

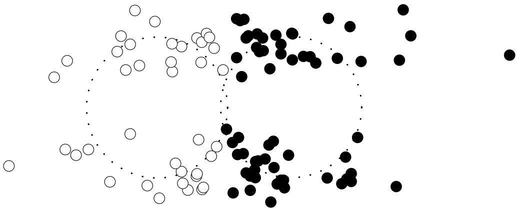

Fig. 1. Example of a partition obtained from improper use of Ward's criterion. The estimated means and scales are indicated by the two circles.

the model. This property is, however, not specific to the classification model. In fact, White [35] proved consistency of ML estimators for independent observations coming from an unspecified parent distribution.

In vector quantization, application of Ward's criterion and the k-means algorithm, here ascribed to Lloyd [24], is justified and standard. The engineer using optimal quantization takes a geometric standpoint and decomposes a data set in subsets that unite nearby points and separate distant ones w.r.t. some *given* metric. Their application is less justified in cluster analysis where we search for causes that generate the data. To this end, we assume that the causes manifest themselves in different populations that induce in the data set *cohesive* (compact) clusters of possibly different sizes and extents and *separated* by location or sometimes by scale. The engineer even decomposes a data set uniform on a square, the cluster analyst finds that this data set bears no cluster structure, it originates from a single source. It is a matter between quantity and quality. Generally accepted, logical, mathematical definitions of the concepts of "cohesion" and "separation" based on the data set do not exist although there are validation methods and tests that are useful in this respect. This is contrary to mathematical topology where the analogous notion of a "connected component" is clearly and logically defined. Both concepts appeal also to intuition.

In the event of elliptical, non-spherical clusters, Ward's criterion (and, hence, the k-means algorithm) may lead to a result unacceptable in cluster analysis. A typical example is presented in Fig. 1. The two-dimensional data set was sampled from two normal populations of equal scales elongated in horizontal direction and lying side by side. Up to small probability, the populations are separated by a horizontal line between them, but Ward's criterion traces a separator perpendicular to it as shown in Fig. 1. The clusters obtained are neither isolated nor cohesive as visual inspection shows. Moreover, the solution created by Ward's criterion is not only inappropriate in the above sense, it also does not reflect the partition induced by the two original populations. The reason for the failure of Ward's criterion in the sense of cluster analysis is that the underlying populations are not spherical. Generally, an inappropriate, narrow model may "create" a wrong structure in the data set. It is therefore important to base cluster criteria on more general statistical models.

Such a model was proposed by Scott and Symons [31]. They used the likelihood paradigm to derive a criterion for the heteroscedastic, isobaric, normal clustering model with arbitrary covariance matrices applicable to such a more general situation, the *heteroscedastic ML Determinant criterion*

$$\sum_{j=1}^{g} n_j(\boldsymbol{\ell}) \log\det S_j(\boldsymbol{\ell}) \tag{1}$$

to be minimized w.r.t. $\boldsymbol{\ell}$. Here, $n_j(\boldsymbol{\ell})$ denotes the size of cluster j w.r.t. $\boldsymbol{\ell}$ and $S_j(\boldsymbol{\ell}) = W_j(\boldsymbol{\ell})/n_j$ is its scatter matrix.

This criterion works nicely in the case of elliptical clusters of about equal sizes but may otherwise run into trouble. Symons [33] corrected this shortcoming in considering the labeling $\boldsymbol{\ell}$ not as a parameter (which it is not since its length increases with the data set) but as a *hidden variable* drawn from the n-fold product $\boldsymbol{\pi} \otimes \cdots \otimes \boldsymbol{\pi}$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$, on $(1..g)^{(1..n)}$. It acts as a prior probability and, since the number g is small, can be estimated by an empirical Bayesian procedure. Symons arrives at the *heteroscedastic MAP Determinant criterion*

$$n\mathrm{H}\left(\frac{n_1(\boldsymbol{\ell})}{n}, \ldots, \frac{n_g(\boldsymbol{\ell})}{n}\right) + \frac{1}{2}\sum_{j=1}^{g} n_j(\boldsymbol{\ell}) \log\det S_j(\boldsymbol{\ell}), \tag{2}$$

again to be minimized w.r.t. $\boldsymbol{\ell}$. The criterion differs from the heteroscedastic ML Determinant criterion (1) in the entropy $\mathrm{H}(p_1, \ldots, p_g) = -\sum_j p_j \log p_j$ of the cluster proportions $p_j = n_j(\boldsymbol{\ell})/n$ which counteracts the tendency of the ML criterion to create clusters of about equal sizes. This is the state of the art concerning the normal classification model of clustering. The related iterative relocation algorithm alternates between clustering and parameter estimation. We call it the *k-parameters* algorithm. This name reminds of the k-means algorithm but expresses the fact that it is not only means that are estimated but also other parameters such as scale matrices and weights. The criterion can be extended to elliptical basic models $E_{\phi,m,V}(x) = \sqrt{\det V^{-1}} e^{-\phi((x-m)^{\mathrm{T}} V^{-1}(x-m))}$ with mean m, scale matrix V, and a fixed *radial function* ϕ. The conditional density becomes

$$f(\boldsymbol{\ell}, \mathbf{x} \mid \boldsymbol{\pi}, \mathbf{m}, \mathbf{V}) = \prod_{i=1}^{n} \pi_{\ell_i} E_{\phi, m_{\ell_i}, V_{\ell_i}}(x_i) \tag{3}$$