TRIMMED FACTORIAL K-MEANS

Matteo Farnè [1]

ABSTRACT: This paper provides the definition of trimmed factorial k-means (TFKM) algorithm. TFKM is a robust version of factorial k-means, where a robust covariance matrix input is used, and outliers in the identified reduced space are iteratively removed via a trimming procedure. The selected latent rank, number of clusters and outlier proportion are those which maximize Hartigan's statistic.

KEYWORDS: robust clustering, dimension reduction, factorial k-means, trimming

1 Introduction

Clustering high-dimensional data with many objects is a challenging task for several reasons. First, a high dimension and a large sample size make agglomerative hierarchical methods like Ward's one (Ward, 1963) computationally intractable. Second, hierarchical partitioning methods like k-means algorithm (MacQueen, 1967) may become very unstable in high dimensions, due to numerical instability and multicollinearity. Third, any non-robust methodology applied to a large dataset is likely to be affected by outliers, so that there is the need to develop and apply robust versions of traditional methods to prevent the identification of uninformative partitions, like trimmed k-means (TKM) (Cuesta-Albertos *et al.* , 1997).

In order to approach dimension reduction, Vichi & Kiers, 2001 proposed factorial k-means (FKM), a method to identify the latent space most able to maximize the distinctiveness of projected objects. The strong consistency of FKM was proved in Terada, 2015. In this paper, we present a robust version of factorial k-means, named trimmed factorial k-means (TFKM), where outliers are iteratively removed in the reduced space, thus simultaneously identifying radial outliers and designing better shaped clusters. This is obtained by minimising the trimmed least squares criterion in the reduced space. A preliminary version of TFKM was first described in Farnè & Vouldis, 2021. Here, we employ MCD (Minimum Covariance Determinant, see Rousseeuw & Driessen, 1999) or ROBPCA (Hubert *et al.* , 2005) to robustly estimate the input covariance matrix, and we then iteratively apply the trimming procedure to estimated factor scores.

[1] [0] Department of Statistical Sciences, University of Bologna, (e-mail: matteo.farne@unibo.it)

2 Trimmed factorial k-means algorithm

Let us consider a $n \times p$ data matrix $\mathbf{X}$. The trimmed factorial k-means of Vichi & Kiers, 2001 assumes that

$$\mathbf{XAA}' = \mathbf{U}\overline{\mathbf{Y}}\mathbf{A}' + \mathbf{E}, \tag{1}$$

where $\mathbf{A}$ is a $p \times r$ semi-orthogonal *coefficient matrix*, such that $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$; $\mathbf{U}$ is a $n \times c$ *membership matrix* such that $\mathbf{U}_{ij} = 1$, $i = 1,\ldots,n$, $j = 1,\ldots,c$, if observation i belongs to cluster j; $\overline{\mathbf{Y}}$ is a $c \times r$ *centroid matrix*; r is the *latent rank* and c is the *number of clusters*. Model (1) assumes that the variable space is approximately isomorphic to a latent linear space, spanned by the same variables, on which the projected data vectors are maximally apart. It is recovered by minimizing $FKM(\mathbf{A},\mathbf{U},\overline{\mathbf{Y}}) = \|\mathbf{XAA}' - \mathbf{U}\overline{\mathbf{Y}}\mathbf{A}'\|^2 = \|\mathbf{XA} - \mathbf{U}\overline{\mathbf{Y}}\|^2$, which is the deviance within clusters in the reduced space, where by least squares we can obtain $\overline{\mathbf{Y}} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{XA}$.

Denoting the $n \times r$ factor score matrix by $\mathbf{F} = \mathbf{XA}$, in this paper we assume that $(100\alpha)\%$ of the n true factor score vectors, with $\alpha \in [0, 0.5]$, are arbitrarily distant from the bulk of the rest of factor score vectors. Therefore, in this situation it is appropriate to minimize $FKM(\mathbf{A},\mathbf{U},\overline{\mathbf{Y}})$ under the constraint $\sum_{i=1}^{n}\sum_{j=1}^{c}\mathbf{U}_{ij} = [(1-\alpha)n]$, with $\sum_{j=1}^{c}\mathbf{U}_{ij} = \{0,1\}$, for each $i = 1,\ldots,n$. This problem can be numerically solved by adapting the original Alternated Least Squares (ALS) algorithm of Vichi & Kiers, 2001 to the framework of Rousseeuw & Van Driessen, 2000 (see also Farnè & Vouldis, 2021). In particular, once initialized, $\mathbf{A}$, $\mathbf{U}$, and $\overline{\mathbf{Y}}$ are first recovered by the original ALS algorithm, which is the H-step, and a trimming procedure is subsequently applied by excluding the $[\alpha n]$ observations most apart from the respective cluster centroids in the reduced space, which is the C-step.

The algorithm input is the Minimum Covariance Determinant (MCD) covariance matrix estimate, if $n \geq 2p$, or the ROBPCA-based reduced covariance matrix with fixed rank $p/10$, otherwise. We call the algorithm input $\mathbf{C}$. We then fix the latent rank r, the number of clusters c, and the outlier proportion α, and we apply the following procedure.

- **Step 0**. We derive the best r-ranked approximation of $\mathbf{C}$ as $\mathbf{C}_r = \mathbf{V}_r\mathbf{D}_r\mathbf{V}_r'$ by extracting the top r principal components of $\mathbf{C}$. We generate a permutation square matrix of size p, $\mathbf{P}$, we orthogonalize it by Gram-Schmidt algorithm, getting $\tilde{\mathbf{P}}$, and we obtain the initial coefficient matrix as $\mathbf{A}_0 = \tilde{\mathbf{P}}\mathbf{V}_r$. Then, we calculate $\mathbf{F}_0 = \mathbf{XA}_0$, the mean factor score $\overline{\mathbf{F}}_0$, and the distances $\mathbf{d}_{i,0} = \mathbf{F}_{i,0} - \overline{\mathbf{F}}_0$, for $i = 1,\ldots,n$. We derive for each i a T-score

as follows: $T_{i,0} = n\mathbf{d}'_{i,0}\mathbf{C}^{-1}_{F,0}\mathbf{d}_{i,0}$, where $\mathbf{C}_{F,0}$ is the $r \times r$ covariance matrix of $\mathbf{F}_0$. Then, we calculate the $2c$ quantiles of $T_{i,0}$, and we allocate each object to the closest quantile among the first, the third, ..., the $(c-1)$-th. We thus obtain the initial membership matrix $\mathbf{U}_0$, and the initial centroid matrix $\overline{\mathbf{Y}}_0 = (\mathbf{U}'_0\mathbf{U}_0)^{-1}\mathbf{U}'_0\mathbf{X}\mathbf{A}_0$. We set $k = 1$, and we proceed as follows.

- **Step 1**. We minimize $FKM(\mathbf{A}_{k-1}, \mathbf{U}_k, \overline{\mathbf{Y}}_{k-1})$ with respect to $\mathbf{U}_k$ given the values of $\mathbf{A}_{k-1}$ and $\overline{\mathbf{Y}}_{k-1}$. For each row i of $\mathbf{U}_k$, we first impose for each $\nu = 1, \ldots, c$ that $\mathbf{U}_{i\nu,k} = 1$, and we then set $\mathbf{U}_{ij,k} = 1$ if and only if

$$\arg \min_{\nu=1,\ldots,c} FKM(\mathbf{A}_{k-1}, \mathbf{U}_{i\nu,k}, \overline{\mathbf{Y}}_{k-1}) = j.$$

- **Step 2**. We calculate $\mathbf{F}_{k-1} = \mathbf{X}\mathbf{A}_{k-1}$, and the distances $\mathbf{d}_{i,k} = \mathbf{F}_{i,k-1} - \overline{\mathbf{Y}}_{l_i,k-1}$, where l_i is s.t. $\mathbf{U}_{il_i,k} = 1$. Then, we derive for each object a T-score as follows: $T_{i,k} = n\mathbf{d}'_{i,k}\mathbf{C}^{-1}_{F,k-1}\mathbf{d}_{i,k}$, $i = 1, \ldots, n$, where $\mathbf{C}_{F,k-1}$ is the $r \times r$ covariance matrix of $\mathbf{F}_{k-1}$. At this stage, we derive the $(1-\alpha)$-quantile of $\mathbf{T}_k$, $T_{1-\alpha,k}$, and we set $\mathbf{U}_{il_i,k} = 0$ if $T_{i,k} > T_{1-\alpha,k}$.
- **Step 3**. $FKM(\mathbf{A}_k, \mathbf{U}_k, \overline{\mathbf{Y}}_k)$ is minimized keeping fixed $\mathbf{U}_k$, to jointly update $\mathbf{A}_k$ and $\overline{\mathbf{Y}}_k$. Among all the linear combinations of $\mathbf{X}$, the ones closer to the centroids (in the transformed space) are derived by taking the first r eigenvectors of $\mathbf{X}'(\mathbf{U}_k(\mathbf{U}'_k\mathbf{U}_k)^{-1}\mathbf{U}'_k - \mathbf{I}_n)\mathbf{X}$ (see Ten Berge, 1993). Based on the optimal $\mathbf{A}_k$, we can then update $\overline{\mathbf{Y}}_k$ using the expression $(\mathbf{U}'_k\mathbf{U}_k)^{-1}\mathbf{U}'_k\mathbf{X}\mathbf{A}_k$.
- **Step 4** $FKM(\mathbf{A}_k, \mathbf{U}_k, \overline{\mathbf{Y}}_k)$ is computed for the current values of $\mathbf{U}_k$, $\mathbf{A}_k$, and $\overline{\mathbf{Y}}_k$. If $FKM(\mathbf{A}_k, \mathbf{U}_k, \overline{\mathbf{Y}}_k) < FKM(\mathbf{A}_{k-1}, \mathbf{U}_{k-1}, \overline{\mathbf{Y}}_{k-1})$, we increase k by 1 and we go again with Steps 1, 2 and 3. Otherwise, the process has converged, we set $k^* = k-1$ and we retain as solutions $\mathbf{U}_{k^*}$, $\mathbf{A}_{k^*}$, and $\overline{\mathbf{Y}}_{k^*}$.

The reported algorithm is repeated $N = 1000$ times, and the final solution is chosen as the one with minimum objective $FKM(\mathbf{A}_{k^*}, \mathbf{U}_{k^*}, \overline{\mathbf{Y}}_{k^*})$ across the N trials.

A grid of possible values for the latent rank r, the number of clusters c and the outlier proportion α is specified. Given that $\overline{\mathbf{Y}} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{X}\mathbf{A}$, and $\text{rk}((\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{X}\mathbf{A}) = \min(c-1, r)$, we cannot explore any combination violating the condition $r \leq c-1$, to avoid singularity in the reduced space. We denote the solutions for each triple (r, c, α) as $\mathbf{U}(r, c, \alpha)$, $\mathbf{A}(r, c, \alpha)$, $\overline{\mathbf{Y}}(r, c, \alpha)$, obtained under the constraint $\sum_{i=1}^{n}\sum_{j=1}^{c}\mathbf{U}_{ij} = [(1-\alpha)n]$, with $\sum_{j=1}^{c}\mathbf{U}_{ij} = \{0, 1\}$, for each $i = 1, \ldots, n$. The optimal values of r, c, and α are then identified by employing Hartigan's statistic (1975), which can be obtained as follows.

First, within clusters deviance is computed for each triple (r, c, α) as $W(r, c, \alpha) = \sum_{i=1}^{n} \|\mathbf{d}_i\|$, where $\mathbf{d}_i = \mathbf{F}_i(r, c, \alpha) - \overline{\mathbf{Y}}_{l_i}(r, c, \alpha)$, l_i is such that

$\mathbf{U}_{il_i}(r,c,\alpha) = 1$, $\mathbf{F}(r,c,\alpha) = \mathbf{XA}(r,c,\alpha)$,
$\overline{\mathbf{Y}}(r,c,\alpha) = (\mathbf{U}(r,c,\alpha)'\mathbf{U}(r,c,\alpha))^{-1}\mathbf{U}(r,c,\alpha)'\mathbf{XA}(r,c,\alpha)$. Second, Hartigan's statistic $H(r,c,\alpha)$ is obtained as

$$H(r,c,\alpha) = (p-c-1)\left(\frac{W(r,c,\alpha)}{W(r,c-1,\alpha)} - 1\right).$$

Finally, we select the triple (r^*,c^*,α^*) returning the maximum $H(r,c,\alpha)$ across selected grid values.

References

CUESTA-ALBERTOS, J., GORDALIZA, A., & MATRÁN, C. 1997. Trimmed k -means: an attempt to robustify quantizers. *The Annals Of Statistics*, **25**, 553–576.

FARNÈ, M., & VOULDIS, A. 2021. Banks' business models in the euro area: a cluster analysis in high dimensions. *Annals Of Operations Research*, **305**, 23–57.

HARTIGAN, J. 1975. Clustering algorithms. *John Wiley & Sons*.

HUBERT, M., ROUSSEEUW, P., & VANDEN BRANDEN, K. 2005. ROBPCA: a new approach to robust principal component analysis. *Technometrics*, **47**, 64–79.

MACQUEEN, J. 1967. Classification and analysis of multivariate observations. *5th Berkeley Symp. Math. Statist. Probability*, 281–297.

ROUSSEEUW, P., & DRIESSEN, K. 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212–223.

ROUSSEEUW, P., & VAN DRIESSEN, K. 2000. An algorithm for positive-breakdown regression based on concentration steps. *Data Analysis*, 335–346.

TEN BERGE, J. 1993. Least squares optimization in multivariate analysis. *Leiden University Leiden*.

TERADA, Y. 2015. Strong consistency of factorial k-means clustering. *Annals Of The Institute Of Statistical Mathematics*, **67**, 335–357.

VICHI, M., & KIERS, H. 2001. Factorial k-means analysis for two-way data. *Computational Statistics & Data Analysis*, **37**, 49–64.

WARD, J. 1963. Hierarchical grouping to optimize an objective function. *Journal Of The American Statistical Association*, **58**, 236–244.