Grouping Around Different Dimensional Affine Subspaces

L.A. García-Escudero, A. Gordaliza, C. Matrán and A. Mayo-Iscar

Abstract Grouping around affine subspaces and other types of manifolds is receiving a lot of attention in the literature due to its interest in several application fields. Allowing for different dimensions is needed in many applications. This work extends the TCLUST methodology to deal with the problem of grouping data around different dimensional linear subspaces in the presence of noise. Two ways of considering error terms in the orthogonal of the linear subspaces are considered.

Key words: Clustering, linear subspaces, dimensions, robustness.

1 Introduction

Many non-hierarchical clustering methods are based on looking for groups around underlying features. For instance, the well-known k-means method creates groups around k point-centers. However, clusters found in a given data set are sometimes due to the existence of certain relationships among the measured variables.

On the other hand, the Principal Components Analysis method serves to find global correlation structures. However, some interesting correlations are non-global since they may be different in different subgroups of the data set even being the distinctive characteristic of groups. This idea has also been proposed with the aim

L.A. García-Escudero
Universidad de Valladolid, Facultad de Ciencias, Prado de la Magdalena S/N, 47011, Valladolid, Spain, e-mail: lagarcia@eio.uva.es

A. Gordaliza
Universidad de Valladolid, e-mail: alfonsog@eio.uva.es

C. Matrán
Universidad de Valladolid, e-mail: matran@eio.uva.es

A. Mayo-Iscar
Universidad de Valladolid, e-mail: agustinm@eio.uva.es

to overcome the "curse of dimensionality" trouble in high-dimensional problems by considering that the data do not uniformly fill the sample space and that data points are indeed concentrated around low dimensional manifolds.

There exist many references about clustering around affine subspaces with equal dimensions within the statistical literature (see, e.g., [10] and [6] and the references therein). We can distinguish between two different approaches: "clusterwise regression" and "orthogonal residuals methods". In clusterwise regression techniques, it is assumed the existence of a privileged response or outcome variable that want to be explained in terms of the explicative ones. Throughout this work, we will be assuming that no privileged outcome variables do exist. Other model-based approaches have been already proposed based on fitting mixtures of multivariate normals assuming that the smallest groups' covariances eigenvalues are small (see, e.g, [3]) but they are not directly aimed at finding clusters around linear subspaces (see [10]).

It is not difficult to find problems where different dimensionalities appear. In fact, this problem has been already addressed by the Machine Learning community. For instance, we can find approaches like "projected clustering" (PROCLUS, ORCLUS, DOC, k-means projective clustering), "correlation connected objects" (4C method), "intrinsic dimensions", "Generalized PCA", "mixture probabilistic PCA", etc.

We will propose in Section 2 suitable statistical models for clustering around affine subspaces with different dimensions. They come from extending the TCLUST modeling in [5]. The possible presence of a fraction α of outlying data is also taken into account. Section 3 provides a feasible algorithm for fitting them. Finally, Section 4 shows some simulations and a real data example.

2 Data models

Clustering around affine subspaces: We assume the existence of k feature affine subspaces in $\mathbb{R}^p$ denoted by H_j with possible different dimensions d_j satisfying $0 \leq d_j \leq p-1$ (a single point if $d_j = 0$). Each subspace H_j is so determined from $d_j + 1$ independent vectors. Namely, a group "center" m_j where the subspace is assumed to pass through and d_j unitary and orthogonal vectors u_j^l, $l = 1, ..., d_j$, spanning the subspace. We can construct a $p \times d_j$ orthogonal matrix U_j from these u_j^l vectors such that each subspace H_j may be finally parameterized as $H_j \equiv \{m_j, U_j\}$.

We assume that an observation x belonging to the j-th group satisfies $x = \mathrm{Pr}_{H_j}(x) + \varepsilon_j^*$, with Pr_{H_j} denoting the orthogonal projection of x onto the subspace H_j given by $\mathrm{Pr}_{H_j}(x) = m_j + U_j U_j'(x - \mu_j)$ and ε_j^* being a random error term chosen in the orthogonal of the linear subspace spanned by the columns of U_j. If ε_j is a random distribution in $\mathbb{R}^{p-d_j}$, we can chose $\varepsilon_j^* = U_j^\perp \varepsilon_j$ with $U_j^\perp$ being a $p \times (p - d_j)$ orthogonal matrix whose columns are orthogonal to the columns of U_j (the Gram-Schmidt procedure may be applied to obtain the matrix $U_j^\perp$). We will further assume that ε_j has a $(p - d_j)$-elliptical distribution with density $|\Sigma_j|^{-1/2} g(x' \Sigma_j^{-1} x)$.

Given a data set $\{x_1, ..., x_n\}$, we define the clustering problem through the maximization of the "classification log-likelihood":

$$\sum_{j=1}^{k}\sum_{i\in R_j}\log(p_j f(x_i;H_j,\Sigma_j)), \tag{1}$$

with $\cup_{j=1}^{k}R_j=\{1,...,n\}$, $R_j\cap R_l=\emptyset$ for $j\neq l$ and

$$f(x_i;H_j,\Sigma_j)=|\Sigma_j|^{-1/2}g\big((x_i-\mathrm{Pr}_{H_j}(x_i))'U_j^{\perp}\Sigma_j^{-1}(U_j^{\perp})'(x_i-\mathrm{Pr}_{H_j}(x_i))\big). \tag{2}$$

Furthermore, we are assuming the existence of some underlying unknown weights p_j's which satisfy $\sum_{j=1}^{k}p_j=1$ in (1). These weights help to do more logical assignments to groups when they overlap.

Robustness: The term "robustness" may be used in a twofold sense. First, in Machine Learning, the term "robustness" is often employed to refer to procedures which are able to handle certain degree of internal within-cluster variability due, for instance, to measurement errors. This meaning obviously has to do with the consideration of data models as those previously presented. Another meaning for the term "robustness" (more common in the statistical literature) has to do with the ability of the procedure to resist to the effect of certain fraction of "gross errors". The presence of gross errors is unfortunately the rule in many real data sets.

To take into account gross errors, we can modify the "spurious-outliers model" in [4] to define a unified suitable framework when considering these two possible meanings for the term "robustness". Starting from this "spurious-outliers model", it makes sense to search for linear affine subspaces H_j, group scatter matrices Σ_j and a partition of the sample $\cup_{j=0}^{k}R_j=\{1,2,...,n\}$ with $R_j\cap R_l=\emptyset$ for $j\neq l$ and $\#R_0=n-[n\alpha]$ maximizing the "trimmed classification log-likelihood":

$$\sum_{j=1}^{k}\sum_{i\in R_j}\log(p_j f(x_i;H_j,\Sigma_j)). \tag{3}$$

Notice that the fraction α of observations in R_0 is not longer taken into account in (3).

Visual and normal errors: Although several error terms may be chosen under the previous general framework, we focus on two reasonable and parsimonious distributions. They follow from considering $\Sigma_j=\sigma_j I_{p-d_j}$ and the following g functions in (2):

a) *Visual errors model* (VE-model): We assume that the mechanism generating the errors follows two steps. First, we randomly choose a vector v in the sphere $S_{p-d_j}=\{x\in\mathbb{R}^{p-d_j}:\|x\|=1\}$. Afterward, we obtain the error term ε_j as $\varepsilon_j=v\cdot|z|$ with z following a $N_1(0,\sigma_j^2)$ distribution. We call them "visual" errors because we "see" (when $p\leq 3$) the groups equally scattered when the σ_j's are equal independently of the dimensions. The VE-model leads to use:

$$f(x;H_j,\sigma_j)= \tag{4}$$
$$=\frac{\Gamma\big((p-d_j)/2\big)}{\pi^{(p-d_j)/2}\sqrt{2\pi\sigma_j^2}}\|x-\mathrm{Pr}_{H_j}(x)\|^{-(p-d_j-1)/2}\exp\big(-\|x-\mathrm{Pr}_{H_j}(x)\|^2/2\sigma_j^2\big).$$

To derive this expression, consider the stochastic decomposition of a spherical distribution X in $\mathbb{R}^{p-d_j}$ as $X = RU$ with R a "radius" variable and U an uniform distribution on S_{p-d_j}. If h denotes the p.d.f. of R and g the density generator of the spherical family then $h(r) = \frac{2\pi^{(p-d_j)/2}}{\Gamma((p-d_j)/2)} r^{(p-d_j)-1} g(r^2)$. Thus, if $R = |Z|$ with Z being a $N(0,1)$ random variable, we get $h(r) = 2/\sqrt{2\pi} \cdot \exp(-x^2/2)$. Expression (4) just follows from (2). Notice that $g(x) = C_{p-d_j} x^{N-1} \exp(-rx^s)$ with $N-1 = -(p-d_j-1)/2$, $r = 1/2 > 0$, $s = 1 > 0$ (and satisfying the condition $2N+p > 2$). Therefore, this density reduces to the univariate normal distribution whenever $p-d_j = 1$ and, in general, belongs to the symmetric Kotz type family [8].

b) *"Normal" errors model* (NE-model): With this approach, the mechanism generating the error terms is based on adding a normal noise in the orthogonal of the feature space H_j. I.e., we take ε_j following a $N_{p-d_j}(0, \sigma_j^2 I_{p-d_j})$ distribution:

$$f(x; H_j, \sigma_j) = (2\pi\sigma_j^2)^{-(p-d_j)/2} \exp\left(-\|x - \mathrm{Pr}_{H_j}(x)\|^2 / 2\sigma_j^2\right) \tag{5}$$

The use of "normal" errors has been already considered in [1] and "visual" errors in [9] when working with 2-dimensional data sets and grouping around (1-dimensional) smooth curves.

Fig. 1 shows two generated data sets with VE- and NE-models. It also shows the boundaries of sets $\{x : d(x, H_j) \leq z_{0.025}/2\}$ with $z_{0.025}$ being the 97.5% percentile of the $N_1(0,1)$ and $d(x,H) = \inf_{y \in H} \|x-y\|$ when H_1 is a point (a ball) and when H_2 is a line (a "strip"). Note the great amount of observations that fall outside the ball in the normal errors case although we had considered the same scatters in both groups.

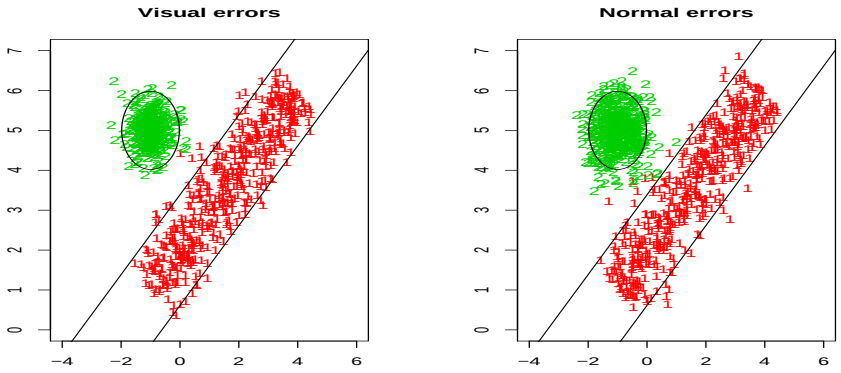

Fig. 1 Simulated data set from the VE- and NE- models.

Constraints on the scatter parameters: Let us consider $d_j + 1$ observations and H_j the affine subspace determined by them. We can easily see that (3) (and (1) too) become unbounded when $|\Sigma_j| \to 0$. Thus, the proposed maximization problems would not be mathematically well-defined without posing any constraint on the Σ_j's.

When $\Sigma_j = \sigma_j I_{p-d_j}$, the constraints introduced in [5] are translated into

$$\max_j \sigma_j^2 \Big/ \min_j \sigma_j^2 \leq c \text{ for a given constant } c \geq 1. \tag{6}$$

Constant c avoids non interesting clustering solutions with clusters containing very few almost collinear observations. This type of restrictions goes back to [7].

3 Algorithm

The maximization of (3) under the restriction (6) has high computational complexity. We propose here an algorithm inspired in the TCLUST one. Some ideas behind the classification EM algorithm [2] and from the RLGA [6] also underly.

1. *Initialize the iterative procedure:* Set initial weights values $p_1^0 = ... = p_k^0 = 1/k$ and initial scatter values $\sigma_1^0 = ... = \sigma_k^0 = 1$. As starting k linear subspaces, randomly select k sets of $d_j + 1$ data points to obtain k initial centers m_j^0 and k initial matrices U_j^0 made up of orthogonal unitary vectors.
2. *Update the parameters in the l-th iteration as*:

2.1. Obtain

$$D_i = \max_{j=1,...,k} \{p_j^l f(x_i; m_j^l, U_j^l, \sigma_j^l)\} \tag{7}$$

and keep the set R^l with the $n - [n\alpha]$ observations with largest D_i's. Split R^l into $R^l = \{R_1^l, ..., R_k^l\}$ with $R_j^l = \{x_i \in R^l : p_j^l f(x_i; m_j^l, U_j^l, \sigma_j^l) = D_i\}$.

2.2. Update parameters by using:

- $p_j^{l+1} \hookleftarrow$ "$n_j^l/[n(1-\alpha)]$ with n_j^l equal to the number of data points in R_j^l".
- $m_j^{l+1} \hookleftarrow$ "The sample mean of the observations in R_j^l".
- $U_j^{l+1} \hookleftarrow$ "A matrix whose columns are equal to the d_j unitary eigenvectors associated to the largest eigenvalues of the sample covariance matrices of observations in R_j^l".

Use the sum of squared orthogonal residuals to obtain initial scatters $s_j^2 = \frac{1}{n_j^l} \sum_{x_i \in R_j^l} \|x_i - P_{H_j^l}(x_i)\|^2$ with $H_j^l \equiv \{m_j^l, U_j^l\}$. To satisfy the constrains, they must be "truncated" as:

$$[s_j^2]_t = \begin{cases} s_j^2 & \text{if } s_j^2 \in [t, ct] \\ t & \text{if } s_j^2 < t \\ ct & \text{if } s_j^2 > ct \end{cases}. \tag{8}$$

Search for $t_{\text{opt}} = \arg\max_t \sum_{j=1}^k \sum_{x_i \in R_j^l} \log f(x_i; m_j^{l+1}, U_j^{l+1}, [s_j^2]_t)$ and take

- $\sigma_j^{l+1} \hookleftarrow \sqrt{[s_j^2]_{t_{\text{opt}}}}$.

3. *Compute the evaluation function:* Perform L iterations of the process described in step 2 and compute the final associated target function (3).
4. *Repeat several times:* Draw S random starting values and keep the solution leading to the maximal value of the target function.

Determining t_{opt} implies solving a one-dimensional optimization problem that can be easily done by resorting to numerical methods. More details concerning the rationale of this algorithm can be found in [5]. We denote the previous algorithm as VE-method when the density (4) is applied and as NE-method when using (5).

4 Examples

Simulation study: Let us consider a clustering problem where observations are generated around a point, a line and a plane in $\mathbb{R}^3$. We generate uniformly-distributed points on the sets $C_1 = \{(x_1,x_2,x_3) : x_1 = x_2 = x_3 = 3\}$ (no random choice), on $C_2 = \{(x_1,x_2,x_3) : 1 \leq x_1 \leq 6, x_2 = x_3 = 3\}$, and, on $C_3 = \{(x_1,x_2,x_3) : x_1 = -2, 1 \leq x_2 \leq 6, 1 \leq x_3 \leq 6\}$. Later, we add error terms in the orthogonal of the C_j's considering the models introduced in Section 2. Finally, points are randomly drawn on the cube $[-4,6] \times [-4,6] \times [-4,6]$ as "gross errors". Fig. 2 shows the result of the proposed clustering approach for a data set drawn from that simulations scheme.

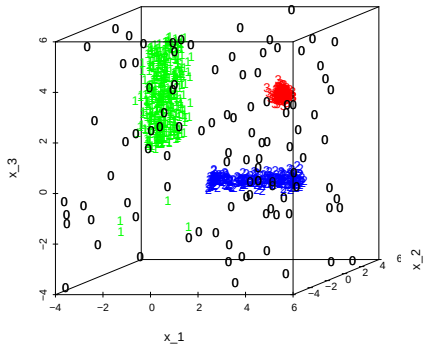

Fig. 2 Result of the NE-method with $k = 3$, d_j's $= (0,1,2)$, $c = 2$ and $\alpha = .1$.

A comparative study based on the previous simulation scheme with VE- and NE-methods has been carried out. We have also considered an alternative (Euclidean distance) ED-method where the D_i's in (7) are replaced by the more simple expressions $D_i = \inf_{j=1,\ldots,k} \|x_i - P_{H_j^l}(x_i)\|$ and no updating of the scatter parameters is done. The ED-method is a straightforward extension of the RLGA in [6].

100 random samples of size $n = 400$ from the previously described simulation schemes with VE- and NE- models for the orthogonal errors are randomly drawn and the associated results for the three clustering VE-, NE- and ED- methods are monitored. Fig. 3 shows the mean proportion of misclassified observations along these 100 random samples. The NE-model seems to have a higher complexity since higher number of random initializations is needed. Notice that the results favor the VE-method even when the true model generating the data was indeed the NE-model. We can also see that parameter S is more critical than L.

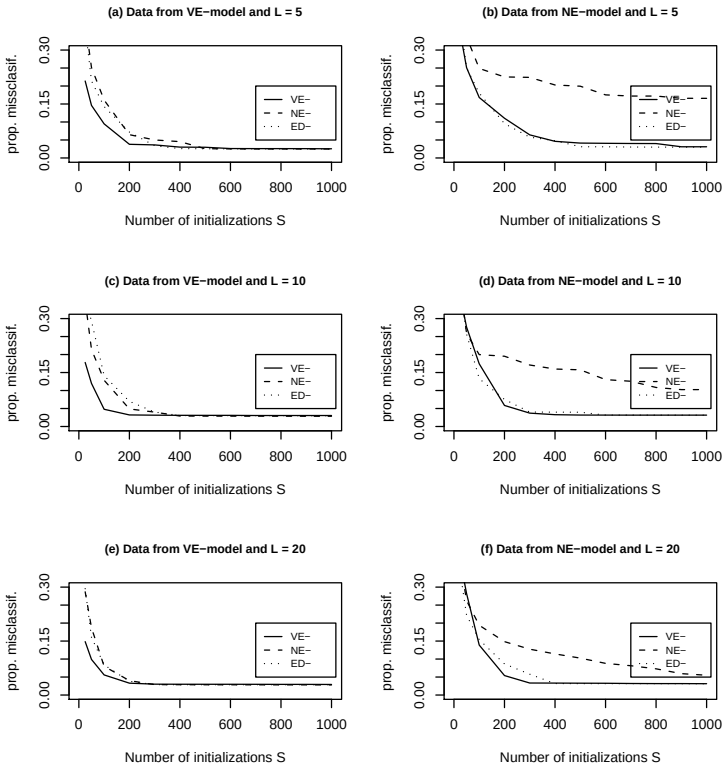

Fig. 3 Proportion of misclassified observations in the simulation study described in the text.

Real data example: As in [9], we consider position data on some earthquakes in the New Madrid seismic region from the CERI. We include all earthquakes in that catalog from 1974 to 1992 with magnitude 2.25 and above. Fig. 4 shows a scatter plot of the earthquakes positions and a nonparametric kernel based density estimation suggesting the existence of a linear tectonic fault and three main point focuses.

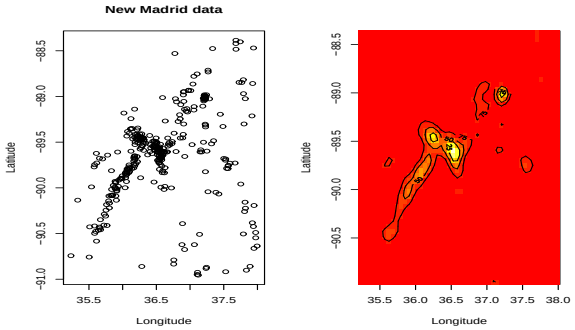

Fig. 4 Earthquake positions in the New Madrid seismic region.

Fig. 5 shows the clustering results when $k = 4$ and dimensions $(1,0,0,0)$. We have considered a high trimming level $\alpha = .4$ which allows discarding earthquakes taking place in regions where the earthquakes are not spatially concentrated.

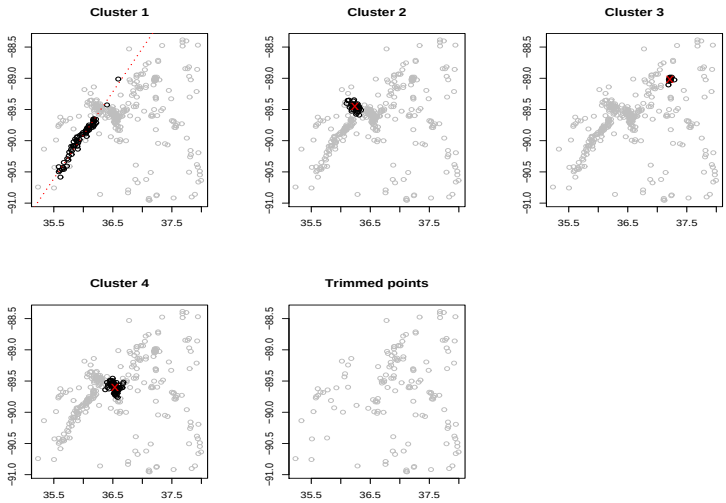

Fig. 5 Clustering results of the VE-method for $k = 4$, d_j's $= (1,0,0,0)$, c =2 and α =.4.

5 Future research directions

The proposed methodology needs to fix parameters k, d_j's, α and c. Sometimes they are known in advance but other times they are completely unknown. "Split and merge", BIC and geometrical-AIC concepts could be then applied. Another important problem is how to deal with remote observations wrongly assigned to higher dimensional linear subspaces due to their "not-bounded" spatial extension. A further second trimming or nearest neighborhood cleaning could be tried.

References

1. Banfield, J.D. and Raftery, A.E. (1993) "Model-based Gaussian and non-Gaussian clustering". *Biometrics*, **49**, 803–821.
2. Celeux, G. and Govaert, A. (1992). "Classification EM algorithm for clustering and two stochastic versions". *Comput. Statit. Data Anal.,* **13**, 315-332.
3. Dasgupta, A. and Raftery, A.E. (1998) "Detecting features in spatial point processes with clutter via model-based clustering." *J. Amer. Statist. Assoc.*, **93**, 294-302.
4. Gallegos, M.T. and Ritter, G. (2005), "A robust method for cluster analysis," *Ann. Statist.*, **33**, 347-380.
5. García-Escudero, L.A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. (2008a). "A general trimming approach to robust clustering", *Ann. Statist.*, **36**, 1324-1345.
6. García-Escudero, L.A., Gordaliza, A., San Martín, R., Van Aelst, S. and Zamar, R. (2008b). "Robust Linear Grouping". *J. Roy. Statist. Soc. B*, **71**, 301-319.
7. Hathaway, R.J. (1985), "A constrained formulation of maximum likelihood estimation for normal mixture distributions," *Ann. Statist*, **13**, 795-800.
8. Kotz, S., 1975. "Multivariate distributions at a cross-road". In *Statistical Distributions in Scientific Work*, G.P. Patil, S. Kotz, J.K. Ord, eds., **1**, 247270.
9. Standford, D.C. and Raftery, A.E. (2000). "Finding curvilinear features in Spatial point patterns: Principal Curve Clustering with Noise". *IEEE Trans. Pattern Recognition*, **22**, 601-609.
10. Van Aelst, S., Wang, X., Zamar, R. H. and Zhu, R. (2006) "Linear grouping using orthogonal regression". *Comput. Statit. Data Anal.*, **50**, 1287-1312.