STATISTICA, anno LXXII, n. 3, 2012

ITALIAN CONTRIBUTIONS ON SOME RECENT RESEARCH TOPICS IN CLUSTER ANALYSIS

Daniela G. Calò[1]

1. INTRODUCTION

Cluster analysis methods are among the most known and commonly applied multivariate analysis techniques. Renewed stimulus in the development of novel clustering methods has been constantly promoted by the questions arising in their numerous application domains, at the interface with many different disciplines, including pattern recognition and engineering. In the last decades, the progress in data capture technologies and data collection capabilities have lead to new research directions. They have permitted the collection of growing amounts of increasingly larger and more complex data, thus rising the need of non-traditional statistical techniques for extracting relevant information in wide range of domains. The development of adequate data analysis tools has attracted the interest of the statistical community, also in light of the parallel improvements in computational resources.

By taking the data-analytic challenges posed by modern data as a *fil-rouge*, this paper aims at providing a selective view of the main research lines currently followed by Italian statisticians in the field of quantitative data clustering. Attention is focused to a small part of the wealth of methods developed on this research field: for each one of the selected research line some representative contributions are mentioned, favouring the most recent ones in order to enable the interested reader to go back (through these references) to the previous related literature.

The presentation is organized as follows. Sections 2 and Section 3 refer to the traditional real-valued case-by-variable data matrix and focus on the issues raised by high-dimensionality and by contaminating observations, respectively. In both sections, a separate discussion of the topic is given concerning the clustering approach based on Gaussian mixture models (GMMs), due to the prominent role this approach has gained in the literature as a sound statistical framework to cluster analysis. As an example of the problem of dealing with composite informa-

[1] The paper presents the content of a presentation to the SIS 2011 Statistical Conference "Statistics in the 150 years from Italian Unification", held in Bologna, Italy, between 8 June and 10 June. Therefore, the proposed review is last-updated on mid-2011.

tion, Section 4 focuses on time series clustering, which has lately emerged as an important research trend especially in data mining applications. Section 5 is devoted to some types of data that require specifically-designed clustering methods; in particular, it focuses on functional data and interval data, as these types of data seem to have most attracted the interest of Italian researchers in the last decade, also beyond the field of cluster analysis. Some final remarks are reported in Section 6.

2. HIGH-DIMENSIONAL DATA

In recent years, growing amounts of data are automatically recorded and stored in the form of high-dimensional observations: examples range from DNA microarray data, about the expression levels of thousands of genes observed on a number of sample tissues, to supermarket scanner data, pertaining to the products purchased by each customer.

In high-dimensional settings, the clustering task is more difficult: the standard assumption that units within the same cluster are similar across all variables might be restrictive; the simultaneous use of all the observed variables may mask the effect of the variable subset that contains clustering information; finally, computational effort increases and interpretation becomes challenging with increasing dimensionality. Concerning Gaussian mixture-based clustering, the number of model parameters grows quadratically as the number of the variables, p, increases when component covariance matrices are not restricted; when p is large relative to the sample size, this may negatively affect the clustering performance of the model.

In order to cope with high-dimensional data, statistical methods that simultaneously perform clustering and dimensionality reduction have been actively investigated, according to two main approaches. On one side, the approach based on feature extraction techniques and, on the other side, the so-called "two-mode clustering" solution, which aims at reducing the set of variables by means of a suitable partition. These two approaches have been pursued by Italian researchers both in the deterministic framework of least-squares partitioning and in the probabilistic one of Gaussian mixture modelling; a focus on the latter framework is given in the next subsection.

A well known example of a feature extraction solution in the deterministic context is Vichi and Kiers' factorial *k*-means (Vichi and Kiers, 2001): it aims at simultaneously finding an optimal partitioning of the units and an optimal low-dimensional space by minimizing the within-cluster deviance of the projected data. A possible drawback of using factor/component techniques is that some observed variables may turn out to be correlated with several extracted features, with possible complications in interpretation. In this respect, the factorial *k*-means idea of clustering units on a q-dimensional space ($q<p$) has been augmented by Vichi and Saporta (2009) by adding the constraint that each one of the q spanning dimensions is a linear combination of a disjoint set of variables only

(more precisely, the linear combination with maximum variance). In this way interpretation is easier, because each observed variable can only be assigned to one of the extracted features, which additionally leads to partitioning the variables in *q* clusters; in light of this, Vichi and Saporta's contribution can also be seen as a special case in the class of two-mode clustering methods.

Two-mode clustering is mainly applied in gene expression data, with the aim of identifying subsets of genes that exhibit similar expression patterns across subsets of tissues. More generally, it is applicable whenever the local association structures between the rows and the columns of the data matrix have to be discovered (see Van Mechelen *et al.*, 2004, for an overview on two-mode clustering, including methods implying nested or overlapping row-and-column clustering). In this research area, Rocci and Vichi (2010) have proposed a generalization of the two-mode partitioning model known as double *k*-means (Vichi, 2000). While double *k*-means method specifies the same partition of the variables for each cluster of the *n* units (and vice-versa), Rocci and Vichi's method allows a different partition of the *p* variables (in, say, q_k clusters, for k=1,...,K) within each one of the *K* clusters of units (by simply transposing the data matrix, the method can also be applied for discovering a variable partition with a different unit partitioning in each cluster of variables): unit membership is specified by a binary n×K matrix, and variable partitioning is defined by *K* binary matrices, having generic dimension p×q_k.

Moreover, in text/web mining applications, where the data matrix entries denote the occurrence of a word in a document, Balbi *et al.* (2010) propose a two-mode partitioning method for discovering groups of documents that are similar across different sets of words: the clustering solution is obtained by optimizing – via a genetic algorithm (see Baragona *et al.*, 2011) – a specific index measuring predictability in contingency tables.

Other developments may be found in the paper by Augugliaro and Mineo (2011), which improves the performance of a two-mode partitioning algorithm by guiding the choice of the tuning parameters it depends on, and in the hierarchical mixture model proposed by Vicari and Alfò (2010) for partitioning customers and products on the basis of purchase data, which includes customer- or product-specific covariates to model customers' choice probabilities.

2.1. *A focus on GMM-based clustering*

Concerning the problem of over-parameterization in GMMs, one of the most popular approaches based on feature-extraction techniques is the so-called Mixtures of Factor Analyzers (MFA, McLachlan *et al.* 2007). MFA assumes that the data are randomly drawn from a population consisting of *K* subpopulations, with unknown mixing proportions, each subpopulation being described by a different (*i.e.* "local") Factor Analysis model (possibly involving a different number of factors, q_k). Alternatively, Montanari and Viroli (2010a) propose to reduce the number of parameters by using a global, rather than a local, latent variable model: by assuming a generative linear factor model, with *q* independent common factors modelled as a K-component mixture of q-variate Gaussian densities, the p observed variables turn out to be modelled as a K-component GMM as well, which is characterized by a more parsimonious parameterization, in terms of both component mean vectors and component covariance matrices.

Later, Montanari and Viroli (2010b) and Baek *et al.* (2010) have developed more flexible solutions admitting dependence among common factors. The potentialities of Montanari and Viroli's method, named Heteroscedastic Factor Mixture Analysis (HFMA), have been explored along different directions: Galimberti *et al.* (2008) have introduced a lasso penalization on factor loadings, so that variable selection is contextually performed; Calò and Viroli (2010) have proposed a finite mixture model for clustering multilevel data, in which HFMA is assumed at the lower level of the hierarchy, thus relaxing the classic "local independence assumption"; Viroli (2010) has developed an extension of MFA to non-Gaussian factor analyzers by modelling each mixture component by HFMA.

The idea of using a component/factor technique to produce a partition of the variables in addition to the partitioning of units (similarly to Vichi and Saporta, 2009) has been developed in the GMM framework by Martella *et al.* (2010). The paper is inspired by the ability of MFA to perform local dimension reduction through a different factor loading matrix in each mixture component. The Authors restrict the component-specific loading matrices to be binary and row-stochastic, which implies that the component covariance matrices are block diagonal. Thus, by introducing variable clustering in MFA, two-mode multi-partitioning purposes are addressed in the GMM framework: different variable partitions are discovered in the clusters of units identified by the mixture model.

Finally, some contributions have appeared about the idea of including variable selection in GMM clustering algorithm, along the lines of Raftery and Dean, (2006) (see also Maugis *et al.*, 2009): Raftery and Dean propose a stepwise algorithm in which the decision of including/excluding a variable is taken by comparing (in terms of BIC difference) the two models that are defined whether or not the assumption is made that the candidate variable is conditionally independent of the cluster membership given the so-far selected variables.

Moving from the well-known criticisms of stepwise selection strategies, Scrucca (2010) considers using genetic algorithms to perform "all subsets" selection, over the space of all subsets of size q ($q<p$). At this aim, the fitness value for a subset of variables is assessed by the BIC difference between two mixture models, one assuming "some" clustering structure (*i.e.* having at least two components) and the other one assuming no clustering structure (*i.e.* having a single component).

Galimberti and Soffritti (2009) have tried to extend Raftery and Dean's approach to the more complex setting in which multiple subsets of variables containing different group structures are present among observed variables, including the set of uninformative variables. In their proposal, the original assumption of a single partition of the units (with cluster membership being specified by a single latent multinomial random variable z) across all the relevant variables is replaced by assuming $q>1$ unit partitions (specified by q mutually independent latent variables, z1, ..., z_q), related to q clusters of relevant variables. Thus, a multi-partitioning procedure that intrinsically performs variable selection is obtained.

2.2. *Data visualization*

Data visualization has gained a relevant role in many applications, thanks to modern graphic capabilities, as a valuable aid to explore and interpret high-dimensional data. Visualization purposes are the main motivation of the paper by Scrucca (2010), which proposes a way to integrate dimension reduction into GMM-based clustering. Instead of imposing a latent variable model, the smallest subspace that captures most of clustering information contained in the data is searched for. The orthogonal spanning directions maximizing variation both in cluster means and cluster covariances are identified using an eigendecomposition method.

New research efforts in data visualization are being inspired by the possible combination of visual interactive tools and data analysis techniques (Palumbo *et al.* 2008). A recent contribution in this direction is given by Iodice D'Enza *et al.* (2008) in the context of association rule mining (which can be viewed as a "mode seeking" clustering problem on a very high-dimensional sparse data matrix having sales transactions in the rows and all items sold in a store in the columns). The correspondence analysis-based strategy proposed in the paper aims at detecting the most potentially interesting items; the included graphical representations of the items help the user in focusing attention towards the most relevant content in output interpretation.

3. CONTAMINATED DATA

Contaminating observations are more likely to occur in large data sets, possibly masking one another. Multivariate outliers are known to be hardly detectable as multivariate data have no "natural ordering"; in addition, it should be stressed that in a clustering perspective the term "contamination" concerns different sources of heterogeneity, that can occur simultaneously: it denotes not only observations that are distant from the bulk of the data but also unusual observations within a cluster or "bridge-points" lying between clusters.

Outliers can derail most clustering methods, including GMM-based ones, leading to poor estimates and clustering results. This has driven a special interest in multivariate outlier detection and robust clustering (see Garcìa-Escudero *et al.*, 2010) (this distinction being elusive since a relatively large group of outliers can be considered as a separate cluster, indeed). Among the recent Italian contributions on these topics, two main research lines can be distinguished: the trimming-based one (underpinning outlier identification), and the mixture-based one, in which contamination is modelled by adding components to the mixture.

MCD estimators (Rousseeuw and Van Driessen, 1999) are popular trimming tools that require a trimming proportion to be specified; in particular, the squared

Mahalanobis distance involving such estimates is commonly used as a test-statistic for outlier testing under the normality assumption. Cerioli (2010) has found an approximation to the exact null distribution of this statistic yielding more accurate cut-off values than those based on the asymptotic χ^2 distribution.

In contrast to MCD estimators, the Forward Search methods (FS, Atkinson *et al.* 2010) provide data-dependent flexible trimming: being based on the strategy of sequentially fitting the model to data subsets, S_m, of increasing size ($m = m_0, ..., n$, starting from a set S_{m_0} of possibly uncontaminated observations), the FS lets the data decide what is best, thus preserving robustness while ensuring high efficiency. Riani *et al.* (2009) propose to use the minimum squared Mahalanobis distance (computed on S_m) among points not included in S_m in a testing procedure for the null hypothesis of "no contamination" in a normal population. This inferential tool is hopefully going to be extended to a clustering set-up. At present, a FS-based method for exploratory cluster analysis has been devised by Atkinson and Riani (2007): it provides a variety of informative plots that allow to tackle both the problem of robust clustering (with a data-driven assessment of the true number of clusters) and that of outlier identification, at the same time. Farcomeni (2009) has resorted to the FS in devising a method for coping with contamination in the class of double *k*-means methods. He presents a two-mode extension of the trimmed *k*-means procedure, involving a FS-based selection of the amount of trimming (where trimming is allowed both for the units and for the variables). The proposed method inherits from FS the benefit of robustly estimating cluster centroids, while performing outlier detection at the same time.

3.1. *A focus on GMM-based clustering*

When the number of mixture components is treated as fixed, a small proportion of outliers can dramatically affect ML estimates, as well as the corresponding clustering solutions. Two main approaches to the problem were proposed in the literature: Banfield and Raftery (1993) suggested to add a "noise component", modelled as a uniform density on the convex-hull of the data; Peel and McLachlan (2000) considered using mixtures of multivariate *t* densities. Since the appearance of this latter paper, mixtures of *t* distributions are becoming more and more popular (McLachlan *et al.*, 2007). This motivated Greselin and Ingrassia (2010) in investigating the issue of how to prevent the EM algorithm to converge to spurious solutions in fitting *t*-mixtures. For mixtures of K multivariate elliptical distributions, they prove that imposing suitable constraints on the eigenvalues of the $p \times p$ definite positive matrices Σ_k (k=1, ..., K) ensures that the likelihood function has a global minimum; then, following Ingrassia and Rocci (2007), they propose a constrained monotone EM algorithm for *t*-mixture estimation. Still in the framework of *t*-mixtures, Ingrassia *et al.* (2010) are working on using *t* mixtures in Cluster Weighted Modeling (Gershenfeld *et al.*, 1999), with the aim of studying the dependence of a response variable Y on some random vector X accounting for population heterogeneity. The idea is to use *t* densities to model, in each population group, both the conditional density of Y given X and the marginal density of

X (which are usually assumed to be normally distributed in the current literature on Cluster Weighted Modeling).

As far as Banfield and Raftery's method is concerned, Coretto and Hennig (2010) introduce two modifications (limited, at present, to the univariate setting), based on different modelling solutions for the "noise component": the former, (i), takes a uniform distribution with unknown support [*a*, *b*] (not necessarily coinciding with the range of the data); the latter, (ii), takes an improper uniform density on the whole real line (with a data-driven choice of the constant density value, *c*>0). Later, the same Authors have theoretically investigated Banfield and Raftery's original proposal (Coretto and Hennig, 2011): they show that it does not necessarily define the (global) maximum likelihood (ML) estimator for the assumed model, neither it defines a consistent estimator. On the contrary, a constrained ML estimator is shown to exist (and to be consistent) for model (i), and an algorithm for constrained ML is derived.

Lately, in the literature a shift is being observed from the idea of contamination in cluster distribution to the more general concept of "deviation from normality". Even under this wider perspective, GMM-based clustering still suffers from the problem that more components (than clusters) are needed to capture any deviation. Different solutions have been recently proposed to address this difficulty. Asymmetry (or both asymmetry and outliers) in cluster distribution can be handled by fitting mixtures of multivariate skew-normal or skew-*t* densities (see Lin, 2009 and Wang *et al*., 2010). Another way to enable the number of components to correspond to the number of clusters is to merge the Gaussian components that are not sufficiently separated to be interpreted as clusters (Baudry *et al*., 2010; Hennig, 2010; Rocci, 2010). A similar idea is to assume that each cluster is well-modelled by a Gaussian mixture, as proposed by Bartolucci (2005) in the one-dimensional setting; the contribution of Viroli (2010) can be also viewed in this latter framework.

4. TEMPORAL DATA

Temporal data arise in many application fields, ranging from time-course gene expression analysis to electricity consumption monitoring (for an example on this latter field, see Giordano *et al*., 2011). When dealing with the problem of grouping similar time series, the clustering task is made more complicated by the fact that conventional distance/dissimilarity measures ignore the dynamic structure of the series and are sensitive to possible distorsion in time axis (Corduas, 2010). Moreover, in the case of multivariate time series, data have the form of a "three-way" array and the aim is to cluster *p*-dimensional time trajectories. A review on dissimilarity indexes between multivariate time series can be found in Baragona (2010). In this Section, Italian contributions on clustering discrete-time series are distinguished according to the three main ways to establish the concept of distance/dissimilarity between time series: the model-based approach, which relies on econometric modelling; the feature-based methods, which are more akin to the data-mining framework; finally, the approach based on raw-data. Recent contributions to the topic of Hidden Markov Models for longitudinal data analysis are mentioned as well.

Reference is made to univariate series unless otherwise stated. The case of continuously varying time points is considered in Subsection 5.2.

4.1. *Model-based approach*

The observed series are assumed to be generated by some time series parametric model; thus, time trajectories are compared according to the properties of the respective underlying stochastic processes. In this framework, the idea, proposed by Piccolo (1990), of evaluating the dissimilarity between two ARIMA invertible processes by the Euclidean distance between the coefficients of their $\mathrm{AR}(\infty)$ representation has inspired numerous developments, as reviewed in Corduas and Piccolo (2008). In particular, Corduas and Piccolo (2008) obtain the asymptotic distribution of the squared Euclidean distance between the vectors of ML estimates of AR weights, and use this result to define a test to determine whether two series differ significantly or not; a partitioning method is proposed too, which considers the 0/1 distance matrix defined by the testing results and aims at ordering its rows/columns so that it best approximates a block-diagonal matrix.

Otranto (2008) adapts the same idea of comparing autoregressive approximations to the problem of identifying clusters of series with homogeneous volatility within the class of GARCH models; a further extension to a class of multivariate GARCH models is given in Otranto (2010), who presents an agglomerative algorithm for automatic detection of clusters of multivariate series having homogeneous correlation dynamics. Alternatively, De Gregorio and Iacus (2010) propose a nonparametric distance in a situation where observed data form a Markov process: by adopting an orthonormal basis estimator of the transition operator of the process, a distance between two series is established by comparing the corresponding basis coefficient estimates.

4.2. *Feature-based approach*

The so called "feature-based approach" is motivated by the fact that high dimensionality (*i.e.* the possibly large number of time points) can blur the clustering structure and slow down the clustering algorithm. It consists in extracting from each series a set of lower-dimensional features that capture the dynamic structure of the data, and in measuring the distance/dissimilarity between two series in terms of such a synthetic representation.

In this framework, D'Urso and Maharaj (2009) and Maharaj *et al.* (2010) propose to use the following features, respectively: the estimated autocorrelation coefficients for different time-lags (under the stationarity assumption), and the estimated wavelet variances associated with the different frequency bands the series is decomposed into (when one aims at distinguishing among different variability patterns). In both the papers, the Euclidean distance between two representations is then used in the fuzzy *k*-means loss function. In the same fuzzy context, Maharaj and D'Urso (2011) propose a feature-based comparison in the frequency domain, which consists in representing a stationary time series by its estimated cepstrum, *i.e.* the spectrum of the logarithm of the spectrum. The same Authors are working on extending their methods to the case of multivariate time series (D'Urso and Maharaj, *in press*).

A further contribution can be found in Giordano *et al.* (2011): after the most relevant (*i.e.* dominant) frequencies have been extracted from each series, the *C* dominant frequencies that occur most frequently are selected, and each series is represented by a *C*-dimensional binary vector (1/0 flags whether the selected frequency is/isn't one of the frequencies extracted from that series). Finally, the set of observed series is partitioned by grouping together those having identical representative vectors.

4.3. *Raw-data based approach*

In this framework, dissimilarity measures are defined directly on raw series data rather than on the corresponding model-based or feature-based representations. Among Italian contributions to this research line, we focus attention to those pertaining to multivariate time series, *i.e.* to complex data structured as 3-way arrays (units × variables × time-occasions). Two main options have been pursued in this context, depending on whether a cross-sectional or a longitudinal analysis is preferred. In the former, the emphasis is on comparing the static *p*-variate characteristics of the units. In the latter multivariate histories are compared according to the geometrical features of the trajectories, like slope or concavity/convexity (D'Urso, 2000).

Along these lines, numerous contributions and developments have appeared. In the most recent literature, an example is given by Coppi *et al.* (2010), where the problem of clustering a set of spatial units on the basis of their multivariate time trajectories is tackled. Two solutions in a fuzzy *k*-means approach are proposed (which also account for the spatial contiguities among the units), for the cross-sectional and the longitudinal analysis, respectively: in the former, dissimilarity is assessed by a weighted sum, over time, of the "instantaneous" squared Euclidean distances between units in the space of the observed variables; in the latter, dissimilarity is assessed by the sum, over time, of the squared Euclidean distances between lag 1 difference vectors.

Longitudinal and cross-sectional analysis are currently being pursued further by Vichi (2010). In the longitudinal approach, a dissimilarity measure comparing trajectories in terms of their shape is used in a T3Clus algorithm (Rocci and Vichi, 2005), so that a low-dimensional representation of the clusters of trajectories is provided too. In addition, a way to combine cross-sectional and longitudinal perspectives is presented: it consists in assuming a *k*-means clustering model for each time occasion and a Vector AutoRegression model for the dynamic evolution of each cluster centroid (in this respect, the proposal is halfway between raw-data and model-based approaches). Thus, homogeneous clusters can be identified for each time occasion and the dynamic evolution of their centroids can be studied; this methodological solution aims at discovering patterns of evolving patterns.

4.4. *Hidden Markov Models*

New methods for classifying individuals according to the evolution of a latent individual characteristic of interest have been developed in the framework of Hidden Markov Models (HMMs) for longitudinal data (Vermunt *et al.*, 2008). Maruotti and Ryden (2009) have considered HMMs for longitudinal count data, with Poisson distributions in the conditional part of the hidden Markov model: besides including covariates in the generalized linear predictor modelling the Poisson parameter, they add individual-specific random effects in order to account for the unobserved individual heterogeneity not captured by the available covariates. Since the maximum likelihood estimate of the random term distribution, which is left unspecified, is given by a discrete distribution, their approach yields a finite mixture of homogeneous HMMs. Concerning non-homogeneous HMMs, Maruotti and Rocci (2010) adopt an analogous parameterization for the hidden part of the model (*i.e.* in the transition probabilities among the Markov model latent states); following the same nonparametric maximum likelihood approach described above, they show that a finite mixture of non-homogeneous HMMs is obtained. Mixtures of HMMs have been applied in different research fields: an interesting example is given in De Angelis (2011), where the model introduced by Vermunt *et al.* (2008) is applied to the study of the poverty phenomenon in Italy, providing insights both on its dynamic behaviour through time and on its heterogeneity among Italian households.

5. NON-STANDARD DATA

Research efforts are being attracted also by the analysis of specific types of data, whose nature requires that specifically-designed methods are defined. Particularly active research lines in Italy are those devoted to uncertainty-affected data and to functional data, as it will be illustrated in the following subsections.

5.1. *Uncertainty-affected data*

The classical representation of a statistical unit by means of a single (crisp) value for each one of the p considered variables may be indeed reductive or inconsistent in case of imprecision (due to the difficulty of accurate measurement) or to vagueness in the definition of what is being observed. Common ways to describe the uncertainty affecting observed values is to represent the data by means of fuzzy numbers or intervals (of the real line). Along these two approaches, two examples of recent contributions in cluster analysis include Coppi *et al.* (2011) and Irpino and Verde (2008), respectively.

In the general class of LR_2 fuzzy numbers (each number being described by 4 quantities: the pair of Left and Right centres and the pair of Left and Right spreads), Coppi *et al.* propose to assess the dissimilarity between two fuzzy *p*-dimensional objects by taking a weighted sum of the squared Euclidean distances between the centres and between the spreads of the objects; by adopting this dissimilarity measure into the fuzzy *k*-means loss function, they provide a method that is able to deal also with the additional uncertainty pertaining to cluster assignment. The paper contains also a first attempt of robust clustering of multivariate fuzzy data, which is based on the "possibilistic *k*-means" approach (Yang and Wu, 2006).

In the last decades, the analysis of interval-valued data has lately attracted a great deal of interest in Italy, within the methodological setting of Symbolic Data Analysis (Diday and Noirhomme, 2008). In particular, as far as cluster analysis is concerned, Irpino and Verde (2008) have introduced a distance measure for interval-valued data or set-valued data (in the multivariate setting): by interpreting a generic interval as the support of a uniform density, the distance between the respective quantile functions is considered, and then employed as an inertia criterion in a classical iterative partitioning algorithm.

The same Authors have extended the above-mentioned proposal to histogram-valued data, which represent data with further complexity: in database aggregation and synthesis, after the values of a variable have been aggregated over a set of lower-level individual observations, histograms have the attractive property of preserving distributive information. In Verde and Irpino (2008), a distance measure between histograms is proposed, which is shown to satisfy the decomposition property in "between-clusters" and "within-clusters" components. It is employed in association with Dynamic Clustering methods.

5.2. *Functional data*

In functional data analysis, the generic observation is given by the values of a smooth random function, measured (with error) on a fine discrete grid: examples are earthquake waveforms (Adelfio *et al.*, 2010) and the surfaces obtained by modern image analysis tools. On these data, smooth function estimation is usually performed by means of Fourier or B-spline basis functions; examples of alternative estimation approaches can be found in Pigoli and Sangalli (2010) and (Di Battista *et al.*, 2011). A problem peculiar to functional data is curve misalignment, which can act as a confounding factor when trying to cluster the curves (Morlini, 2007). To avoid this risk, Sangalli *et al.* (2010) propose a procedure that simultaneously performs clustering and alignment on a set of *n* functional observations. The aim is to find *k* template curves, one for each cluster, and *n* aligning functions such that the overall similarity between each aligned curve and the most similar template curve is maximized: this optimization problem is tackled through a *k*-means-like algorithm, alternating (at each iteration) a template estimation step and an alignment-assignment step.

Moreover, clustering applications on geographically referenced functional data (like meteorological data recorded over a period by sensors located in different sites) create the need for methods that take spatial dependence among the curves into account. In this context, Romano *et al.* (2010) are exploring how the spatially constrained clustering methods proposed in the literature can be integrated in the functional framework.

6. FINAL REMARKS

The leading thread followed in the paper forced us to leave aside contributions concerning specific problems in the selected topics.

It is the case of GMM likelihood unboundedness, which has been deeply studied by Ingrassia and Rocci (2011). In light of the results obtained on the convergence behaviour of the EM algorithm towards degeneracy, they have observed that the risk of unboundedness can be prevented by putting a numerical constraint in EM iterations: the specification of this constraint does not require any a priori information about mixture components, unlike what happens in other methods already presented in the literature (Ingrassia and Rocci, 2007).

Other types of non-standard and complex data with relevant clustering applications could have been mentioned as well. It is the case of dissimilarity data matrices, which represent complex objects describing different classification structures of a set of units. The issue of partitioning a set of dissimilarity matrices (concerning the same set of units) into homogeneous clusters has been addressed by Vicari and Vichi (2009); the same idea of classification comparison has motivated Morlini and Zani (2010) in studying an index for comparing two hierarchical clusterings. Another example is image segmentation, which has inspired the contribution by Alfò *et al.* (2009) on the use of a spatial model for the cluster membership process in a finite mixture on geographical units. Furthermore, it should be mentioned the case of the highly evolving multiple streams of data, emerging continuously over time, on the web or in financial applications: they are the object of ongoing research by Balzanella *et al.* (2011) on incremental clustering methods, in order to cope with the need of "on the fly" methods of data analysis.

These brief final notes only serve to remark that the challenges posed by the data nowadays arising in an increasingly wider range of domains are one of the main drivers of new developments in cluster analysis and, more generally, in data analysis. It is reasonable to expect this trend will continue in the future since, as John Tukey is reported as having said, "*the best thing about being a statistician is that you get to play in everyone's back yard*" (Hand, 2009).

Department of Statistical Sciences DANIELA G. CALÒ
University of Bologna

REFERENCES

G. ADELFIO, M. CHIODI, A. D'ALESSANDRO, D. LUZIO, (2010), *Clustering of waveforms-data based on FPCA direction*, in "Proceedings of COMPSTAT 2010", Physica-Verlag.

M. ALFÒ, L. NIEDDU, D. VICARI, (2009), *Finite mixture models for mapping spatially dependent disease counts*, "Biometrical Journal", 51, pp. 84-97.

A.C. ATKINSON, M. RIANI, (2007), *Exploratory tools for clustering multivariate data*, "Computational Statistics and Data Analysis", 52, pp. 272-285.

A.C. ATKINSON, M. RIANI, A. CERIOLI (2010), *The Forward Search: theory and data analysis*, "Journal of the Korean Statistical Society", 39, pp. 117-134.

L. AUGUGLIARO, A. MINEO, (2011), *Plaid model for microarray data: an enhancement of the pruning step*, in B. Fichet *et al.* (eds.) "Classification and multivariate analysis for complex data structures", pp. 447-456. Springer, Heidelberg.

J. BAEK, G.J. MCLACHLAN, L. FLACK, (2010), *Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualisation of high-dimensional data*, "IEEE Transactions on Pattern Analysis and Machine intelligence", 32, pp. 1298-1309.

S. BALBI, R. MIELE, G. SCEPI, (2010), *Clustering of documents from a two-way viewpoint*, in "JADT 2010: 10 th international Conference on Statistical Analysis of Textual Data".

A. BALZANELLA, Y. LECHEVALLIER, R. VERDE, (2011), *Clustering multiple data streams*, in S. Ingrassia, *et al.* (eds.) "New Perspectives in Statistical Modeling and Data Analysis", Springer.

J.D. BANFIELD, A.E. RAFTERY, (1993), *Model-based Gaussian and non-Gaussian clustering*, "Biometrics", 49, pp. 803-821.

R. BARAGONA, (2010), *Dissimilarity indexes for clustering multivariate time series*, available at http://w3.uniroma1.it/statstsmeh/download/dissimilarity_index.pdf.

R. BARAGONA, F. BATTAGLIA, I. POLI, (2011), *Evolutionary Statistical Procedures*, Springer-Verlag, Heidelberg.

F. BARTOLUCCI, (2005), *Clustering univariate observations via mixtures of unimodal normal mixtures*, "Journal of Classification", 22, pp. 203-219.

J.P. BAUDRY, A.E. RAFTERY, G. CELEUX, K. LO, R. GOTTARDO, (2010), *Combining mixture components for clustering*, "Journal of Computational and Graphical Statistics", 19, pp. 332-353.

D.G. CALÒ, C. VIROLI, (2010), *A dimensionally reduced finite mixture model for multilevel data*, "Journal of Multivariate Analysis", 101, pp. 2543-2553.

A. CERIOLI, (2010), *Multivariate outlier detection with high-breakdown estimators*, "Journal of the American Statistical Association", 105, pp. 147-156.

R. COPPI, P. D'URSO, P. GIORDANI, (2010), *A fuzzy clustering model for multivariate spatial time series*, "Journal of Classification", 27, pp. 54-88.

R. COPPI, P. D'URSO, P. GIORDANI, (2011), *Fuzzy and possibilistic clustering for fuzzy data*, "Computational Statistics & Data Analysis", doi: 10.1016/j.csda.2010.09.013.

M. CORDUAS, (2010), *Mining time series data: a selective survey*, in F. Palumbo *et al.* (eds.) "Data Analysis and Classification", pp. 355-362. Springer, Heidelberg.

M. CORDUAS, D. PICCOLO, (2008), *Time series clustering and classification by the autoregressive metric*, "Computational Statistics & Data Analysis", 52, pp. 4685-4698.

P. CORETTO, C. HENNIG, (2010), *A simulation study to compare robust clustering methods based on mixtures*, "Advances in Data Analysis and Classification", 4, pp. 111-135.

P. CORETTO, C. HENNIG, (2011), *Maximum likelihood estimation of heterogeneous mixtures of Gaussian and uniform distributions*, "Journal of Statistical Planning and inference",141, pp. 462-473.

L. DE ANGELIS, (2011), *The multidimensional measurement of poverty: a longitudinal analysis*, in "JOCLAD2011 - Book of Abstract", pp. 49-52.

A. DE GREGORIO, S.M. IACUS, (2010), *Clustering of discretely observed diffusion processes*, "Computational Statistics & Data Analysis", 54, pp. 598-606.

T. DI BATTISTA, S.A. GATTONE, A. DE SANCTIS, (2011), *Dealing with FDA estimation methods*, in S. Ingrassia, *et al.* (eds.) "New Perspectives in Statistical Modeling and Data Analysis", Springer.

E. DIDAY, M. NOIRHOMME, (2008), *Symbolic Data Analysis*, Wiley, New York.

P. D'URSO, (2000), *Dissimilarity measures for time trajectories*, "Statistical Methods & Applications", pp. 53-83.

P. D'URSO, E.A. MAHARAJ, (2009), *Autocorrelation-based fuzzy clustering of time series*, "Fuzzy Sets and Systems", 160, pp. 3565-3589.

P. D'URSO, E.A. MAHARAJ, *Wavelet-based clustering of multivariate time series*, "Fuzzy Sets and Systems", in press.

A. FARCOMENI, (2009), *Robust double clustering*, "Journal of Classification", 26, pp. 77-101.

G. GALIMBERTI, A. MONTANARI, C. VIROLI, (2008), *Penalized factor mixture analysis for variable selection in clustered data*, "Computational Statistics & Data Analysis", 53, pp. 4301-4310.

G. GALIMBERTI, G. SOFFRITTI, (2009), *Discovering multidimensional unobserved heterogeneity through model-based cluster analysis*, available at http://www.statssa.gov.za/isi2009/Scientific Programme/IPMS/0120.pdf.

L.A. GARCÌA-ESCUDERO, A. GORDALIZA, C. MATRÁN, A. MAYO-ISCAR, (2010), *A review of robust clustering methods*, "Advances in Data Analysis and Classification", 4, pp. 89-109.

N. GERSHENFELD, B. SCHONER, F. METOIS, (1999), *Cluster-weighted modelling for time-series analysis*, "Advances in Data Analysis and Classification", 397, pp. 329-332.

F. GIORDANO, M. LA ROCCA, M.L. PARRELLA, (2011), *Clustering complex time series databases*, in B. Fichet *et al.* (eds.) "Classification and multivariate analysis for complex data structures", pp. 417-426. Springer, Heidelberg.

F. GRESELIN, S. INGRASSIA, (2010), *Constrained monotone EM algorithms for mixtures of multivariate t distributions*, "Statistics and Computing", 20, pp. 9-22.

D.J. HAND, (2009), *Modern statistics: the myth and the magic*, "Journal of the Royal Statistical Society", A, 172, pp. 287-306.

C. HENNIG, (2004), *Breakdown points for maximum likelihood-estimators of location-scale mixtures*, "Annals of Statistics", 32, pp. 1313-1340.

C. HENNIG, (2010), *Methods for merging Gaussian mixture components*, "Advances in Data Analysis and Classification", 4, pp. 3-34.

S. INGRASSIA, R. ROCCI, (2007), *Constrained monotone EM algorithms for finite mixture of multivariate Gaussians*, "Computational Statistics & Data Analysis", 51, pp. 5339-5351.

S. INGRASSIA, R. ROCCI, (2011), *Degeneracy of the EM algorithm for the MLE of multivariate Gaussian mixtures and dynamic constraints*, "Computational Statistics & Data Analysis", 55, pp. 1715-1725.

S. INGRASSIA, C. MINOTTI, G. VITTADINI, (2010), *Cluster Weighted Modelling wit Student-t components*, available at http://homes.stat.unipd.it/mgri/SIS2010/Program/8-SSVIII_Cladag/881-1507-1-RV.pdf.

A. IODICE D'ENZA, F. PALUMBO, M. GREENACRE, (2008), *Exploratory data analysis leading towards the most interesting simple association rules*, "Computational Statistics & Data Analysis", 52, pp. 3269-3281.

A. IRPINO, R. VERDE, (2008), *Dynamic clustering of interval data using a Wasserstein-based distance*, "Pattern Recognition Letters", 29, pp. 1648-1658.

T.I. LIN, (2009), *Maximum likelihood estimation for multivariate skew normal mixture models*, "Journal of Multivariate Analysis", 100, pp. 257-265.

E.A. MAHARAJ, P. D'URSO, (2011), *Fuzzy clustering of time series in the frequency domain*, "Information Sciences", 181, pp. 1187-1211.

E.A. MAHARAJ, P. D'URSO, D.U.A. GALAGEDERA, (2010), *Wavelet-based fuzzy clustering of time series*, "Journal of Classification", 27, pp. 231-275.

F. MARTELLA, M. ALFÒ, M. VICHI, (2010), *Biclustering of gene expression data by an extension of mixtures of factor analyzers*, "The international Journal of Biostatistics", 4, doi: 10.2202/1557-4679.1078.

A. MARUOTTI, R. ROCCI, (2010), *A semiparametric approach to mixed non-homogeneous hidden Markov models*, avalilable at http://homes.stat.unipd.it/mgri/SIS2010/Program/6- SVI_Vicari/851-1532-1-DR.pdf.

A. MARUOTTI, T. RYDEN, (2009), *A semiparametric approach to hidden Markov models under longitudinal observations*, "Statistics and Computing", 19, pp. 381-393.

C. MAUGIS, G. CELEUX, M.L. MARTIN-MAGNIETTE, (2009), *Variable selection for clustering with Gaussian mixture models*, "Biometrics", 65, pp. 701-709.

G.J. MCLACHLAN, R.W. BEAN, L. BEN-TOVIM JONES, (2007), *Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution*, "Computational Statistics & Data Analysis", 51, pp. 5327-5338.

A. MONTANARI, C. VIROLI, (2010), *Heteroscedastic factor mixture analysis*, "Statistical Modelling", 10, pp. 441-460.

A. MONTANARI, C. VIROLI, (2010), *The independent factor analysis approach to latent variable modeling*, "Statistics", 44, pp. 397-416.

I. MORLINI, (2007), *Searching for structure in measurements of air pollutant concentration*, "Environmetrics", 18, pp. 823-840.

I. MORLINI, S. ZANI, (2010), *A dissimilarity measure between two hierarchical clusterings*, in "CLADAG 2010- Book of Abstract", pp. 219-210.

E. OTRANTO, (2008), *Clustering heteroscedastic time series by model-based procedures*, "Computational Statistics & Data Analysis", 52, pp. 4685-4698.

E. OTRANTO, (2010), *Identifying financial time series with similar dynamical conditional correlation*, "Computational Statistics & Data Analysis", 54, pp. 1-15.

F. PALUMBO, D. VISTOCCO, A. MORINEAU, (2008), *Huge multidimensional data visualization: back to the virtue of principal coordinates and dendrograms in the new computer age*, in C. Chun-Houh *et al.* (eds.) "Handbook of Data Visualization", pp. 349-387. Springer, Heidelberg.

D. PEEL, G. MCLACHLAN, (2000), *Robust mixture modeling using the t-distribution*, "Statistics and Computing", 10, pp. 339-348.

D. PICCOLO, (1990), *A distance measure for classifying ARIMA models*, "Journal of Time Series Analysis", 11, pp. 153-164.

D. PIGOLI, L.M. SANGALLI, (2010), *Wavelet smoothing for curves in more than one dimension*, available at http://homes.stat.unipd.it/mgri/SIS2010/Program/contributedpaper/646-1440-1-DR.pdf.

A.E. RAFTERY, N. DEAN, (2006), *Variable selection for model-based cluster analysis*, "Journal of the American Statistical Association", 101, pp. 168-178.

M. RIANI, A.C. ATKINSON, A. CEROLI, (2009), *Finding an unknown number of multivariate outliers*, "Journal of the Royal Statistical Society B", B, 71, pp. 447-466.

R. ROCCI, (2010), *Mixing mixtures of Gaussians*, GfKl-CLADAG 2010 Book of Abstracts, pp. 27-28.

R. ROCCI, M. VICHI, (2005), *Three-mode component analysis with crisp or fuzzy partition of units*, "Psychometrika", 70, pp. 715-736.

R. ROCCI, M. VICHI, (2010), *Two-mode multi-partitioning*, "Computational Statistics & Data Analysis", 52, pp. 1984-2003.

E. ROMANO, A. BALZANELLA, R. VERDE, (2010), *A new regionalization method for spatially dependent functional data based on local variogram models: an application on environmental data*, available at http://homes.stat.unipd.it/mgri/SIS2010/Program/16-SSXVI_Dibattista/906-1575-1-RV.pdf.

P.J. ROUSSEEUW, K. VAN DRIESSEN, (1999), *A fast algorithm for the minimum covariance determinant estimator*, "Technometrics", 41, pp. 212-223.

L.M. SANGALLI, P. SECCHI, S. VATINI, V. VITELLI, (2010), *k-mean alignment for curve clustering*, "Computational Statistics & Data Analysis", 54, pp. 1219-1233.

L. SCRUCCA, (2010), *Genetic algorithms for subset selection in model-based clustering*, available at http://homes.stat.unipd.it/mgri/SIS2010/Program/contributedpaper/590-1296-1-DR.pdf.

L. SCRUCCA, (2010), *Dimension reduction for model-based clustering*, "Statistics and Computing", 20, pp. 471-484.

I. VAN MECHELEN, H.-H. BOCK, P. DE BOECK, (2004), *Two-mode clustering methods: a structured overview*, "Statistical Methods in Medical Research", 13, pp. 363-394.

R. VERDE, A. IRPINO, (2008), *Comparing histogram data using a Mahalanobis Wasserstein distance*, in P. Brito (ed.), "COMPSTAT 2008", pp. 77-89. PhysicaVerlag, Berlin.

J.K. VERMUNT, B. TRAN, J. MAGIDSON, (2008), *Latent class models in longitudinal research*, in S. Menard (ed.), "Handbook of Longitudinal Research: Design, Mesurement, and Analysis", pp. 373-385. Burlington, MA.

D. VICARI, M. ALFÒ, (2010), *Clustering discrete choice data*, in Y. LECHEVALLIER, G. SAPORTA (eds.) *Proceedings of COMPSTAT2010*, pp. 369-378. Physica-Verlag, Heidelberg.

M. VICHI, (2000), *Double k-means clustering for simultaneous classification of objects and variables*, in S. Borra *et al.* (eds.), "Advances in Classification and Data Analysis", pp. 43-52. Springer, Berlin.

M. VICHI, (2010), *Clustering longitudinal multivariate observations*, Personal communication, http://sfc2010.univ-reunion.fr/sfc2010/images/stories/pdf/sfc2010_vichi.pdf

M. VICHI, H.A.L. KIERS, (2001), *Factorial k-means analysis for two-way data*, "Computational Statistics & Data Analysis", 37, pp. 49-64.

D. VICARI, M. VICHI, (2009), *Structural classification analysis of three-way dissimilarity data*, "Journal of Classification", 26, pp. 121-154.

M. VICHI, G. SAPORTA, (2009), *Clustering and disjoint principal component analysis*, "Computational Statistics & Data Analysis", 53, pp. 3194-3208.

C. VIROLI, (2010), *Dimensionally reduced model-based clustering through mixtures of factor mixture analyzers*, "Journal of Classification", 27, pp. 363-388.

K. WANG, S.-K.NG, G.J. MCLACHLAN, (2010), *Multivariate Skew-t Mixture Models*, in "DICTA '09", doi:10.1109/DICTA.2009.88.

M.S. YANG, K.L. WU, (2006), *Unsupervised possibilistic clustering*, "Pattern Recognition", 39, pp. 5-21.

SUMMARY

Italian contributions on some recent research topics in cluster analysis

The paper presents a selective view of the issues that are attracting the interest of Italian statisticians working on clustering methods and applications. It does not aim at providing a comprehensive overview of the wealth of methods developed in Italy on the selected topics: indeed, it focuses on methods dealing with quantitative data and, in this context, only on the most recent literature. The *fil rouge* is given by the developments which have been inspired in quantitative data clustering by the complex nature of the data nowadays arising in a broad range of applications.