CLADAG2017

Book of Short Papers

Editors: Francesca Greselin,
Francesco Mola and Mariangela Zenga

via Sottoriva, 9 - 46100 Mantova (MN)
P. IVA 02346110204
E-book (PDF version) published in September 2017
ISBN 978-88-99459-71-0

This book is the collection of the Abstract / Short Papers submitted by the authors of the International Conference of The CLAssification and Data Analysis Group (CLADAG) of the Italian Statistical Society (SIS), held in Milan (Italy), University of Milano-Bicocca, September 13-15, 2017.

Euro 9,00

Contributed sessions

Classification of Multiway and Functional Data

A generalized Mahalanobis distance for the classification of functional data

Andrea Ghiglietti, Francesca Ieva, Anna Maria Paganoni

Classification methods for multivariate functional data with applications to biomedical signals

Andrea Martino, Andrea Ghiglietti, Anna M. Paganoni

A new Biclustering method for functional data: theory and applications

Jacopo Di Iorio, Simone Vantini

A leap into functional Hilbert spaces with Harold Hotelling

Alessia Pini, Aymeric Stamm, *Simone Vantini*

Sampling Designs and Stochastic models

Statistical matching under informative probability sampling

Daniela Marella, Danny Pfeffermann

Goodness-of-fit test for discrete distributions under complex sampling design

Pier Luigi Conti

Structural learning for complex survey data

Daniela Marella*, Paola Vicard*

The size distribution of Italian firms: an empirical analysis

Anna Maria Fiori, Anna Motta

Robust statistical methods

New proposal for clustering based on trimming and restrictions

Luis Angel Garcìa Escudero, Francesca Greselin, *Agustin Mayo Iscar*

Wine authenticity assessed via trimming

Andrea Cappozzo, Francesca Greselin

Robust and sparse clustering for high-dimensional data

Sarka Brodinova, Peter Filzmoser, Thomas Ortner, Maia Zaharieva, Christian Breiteneder

M-quantile regression for multivariate longitudinal data

Marco Alfo', *Maria Francesca Marino,* Maria Giovanna Ranalli, Nicola Salvati, Nikos Tzavidis

New proposals in Clustering methods

Reduced K-means Principal Component Multinomial Regression for studying the relationships between spectrometry and soil texture

Pietro Amenta, *Antonio Lucadamo*, Antonio Pasquale Leone

Comparing clusterings by copula information based distance

Marta Nai Ruscone

Fuzzy methods for the analysis of psychometric data

Isabella Morlini

Inverse clustering: the paradigm, its meaning, and illustrative examples

Jan W. Owsinski, Jaroslaw Stanczak, Karol Opara, Slawomir Zadrozny

WINE AUTHENTICITY ASSESSED VIA TRIMMING

Andrea Cappozzo [1], Francesca Greselin [1]

[1] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, (e-mail: a.cappozzo@campus.unimib.it; francesca.greselin@unimib.it)

ABSTRACT: An authentic food is one that is what it claims to be. Consumers and food processors need to be assured they receive exactly the specific product they pay for. To ascertain varietal genuinity and distinguish doctored food, in this paper we propose to employ a robust mixture estimation method. It has been shown to be a valid tool for food authenticity studies, when applied to food data with unobserved heterogeneity, to classify genuine wines and identify low proportions of observations with different origins. Our methodology models the data as arising from a mixture of Gaussian factors and employ a threshold on the multivariate density to bring apart the less plausible data under the fitted model. Simulation results assess the effectiveness of the proposed approach and yield very good misclassification rates when compared to analogous methods.

KEYWORDS: Classification; Food authenticity studies; Model-based clustering; Wine; Authenticity; Chemometrics; Robust estimation; Trimming.

1 Introduction

Wine has an economic value that is associated with a luxury product, and its consumers demand reliable information. The present work provides an application of robust estimation of mixtures of Gaussian Analyzers as a tool to employ observed chemical features to classify red wines according to their authentic variety, and to discriminate them from illegal adulteration. It is indeed generally accepted that flavours and chemical compositions of wines are not only related to genetic factors (grape variety) but also to environmental conditions in vineyards (climate, soil composition and geology, microflora) and to human practices. Throughout the growth and maturation of grape berries metabolic changes occur, in such a way that at harvest time the berries contain the major grapevine compounds conferring the wine organoleptic characteristics (da Silva *et al.* , 2005).

In a probabilistic modeling approach for wine authentication, we assume a probability distribution function for the measurements in wine samples, e.g.

log-proportions of anthocyanin or concentrations of any compound. Whenever more than a variety could appear in the sample, we may adopt a density in the form of a mixture. Afterwards, the probability that a wine sample comes from a specific grape variety can be estimated from the model, and each wine sample is assigned to the grape variety with higher probability, using the Bayes rule. As the model arises from measurements belonging to authentic wines, we expect that observations coming from different sources would be unplausible under the estimated model. By selecting observations with the lowest contributions to the overall likelihood, we are confident to discriminate the illegal subsample. Instead of hypothesizing a component of the mixture to model a few observations as forged data, using impartial trimming we are able to detect such data without any strong assumption on their density. Our simulation results confirm the effectiveness of this approach, when compared to analogous methods, such as partition around medoids and non robust mixtures of Gaussian and mixtures of patterned Gaussian factors.

2 Mixtures of Gaussian Factors Analyzers

In this section we briefly recall definition and features of the mixtures of Gaussian Factor Analyzers (MFA). MFA are powerful tools for modeling unobserved heterogeneity in a population, as they concurrently performs clustering and local dimensionality reduction, within each cluster.

An MFA assumes that the observation $\mathbf{X}_i$, for $i = 1,\dots,n$, is given by

$$\mathbf{X}_i = \mu_g + \Lambda_g \mathbf{U}_{ig} + \mathbf{e}_{ig} \tag{1}$$

with probability π_g for $g = 1,\dots,G$, where Λ_g are the $p \times d$ matrices of *factor loadings*, $\mathbf{U}_{ig} \overset{iid}{\sim} \mathcal{N}(\mathbf{0},\mathbf{I}_d)$ are the *factors*, $\mathbf{e}_{ig} \overset{iid}{\sim} \mathcal{N}(\mathbf{0},\Psi_g)$ are the *errors*, and Ψ_g are $p \times p$ diagonal matrices. Further, $\mathbf{U}_{ig}$ and $\mathbf{e}_{ig}$ are independent, for $i = 1,\dots,n$ and $g = 1,\dots G$. Unconditionally, therefore, $\mathbf{X}_i$ is a mixture of G normal densities

$$f(\mathbf{X}_i;\theta) = \sum_{g=1}^{G} \pi_g \phi_p(\mathbf{X}_i;\mu_g,\Sigma_g) \tag{2}$$

where the g-th component-covariance matrix Σ_g has the form $\Sigma_g = \Lambda_g \Lambda_g' + \Psi_g$.

When estimating MFA through the usual Maximum Likelihood approach, two issues arise.

Firstly, departure from normality in the data (noise, contamination, outliers,...) may cause biased or misleading inference. Some initial attempts in the literature to overcome this issue, propose to consider mixtures of t-factor analyzers (McLachlan *et al.* , 2003). The heavier - than normal - tails of the t-distribution allow to incorporate mild outliers, but the breakdown properties of the estimators are not improved (Hennig, 2004).

The second issue is related to the unboundedness of the log-likelihood function (Day, 1969), that causes estimation issues, even when ML estimation is applied to artificial data drawn from a given finite mixture model, i.e. without adding any kind of contamination. It favors the appearance of non-interesting local maximizers (called *spurious maximizers*) and degenerate solutions. To overcome this second issue, Common/Isotropic noise matrices/patterned covariances have been considered (Baek *et al.* , 2010), or a mild constrained estimation (Greselin & Ingrassia, 2015). Here, we employ model estimation, complemented with *trimming* and *constrained estimation*, to provide robustness, to exclude singularities and to reduce spurious solutions, along the lines of García-Escudero *et al.* (2016).

2.1 Trimmed mixture log-likelihood

We fit a mixture of Gaussian factor components to a given dataset $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ in $\mathbb{R}^p$ by maximizing a *trimmed mixture log-likelihood* (Neykov *et al.* , 2007),

$$\mathcal{L}_{trim} = \sum_{i=1}^{n} \zeta(\mathbf{x}_i) \log \left[\sum_{g=1}^{G} \phi_p(\mathbf{x}_i; \mu_g, \Lambda_g, \Psi_g) \pi_g \right] \tag{3}$$

where $\zeta(\cdot)$ is a 0-1 trimming indicator function, that tells us whether observation $\mathbf{x}_i$ is trimmed off or not. If $\zeta(\mathbf{x}_i)$=0 $\mathbf{x}_i$ is trimmed off, otherwise $\zeta(\mathbf{x}_i)$=1.

A fixed fraction α of observations, the *trimming level*, is unassigned by setting $\sum_{i=1}^{n} \zeta(\mathbf{x}_i) = [n(1-\alpha)]$, by selecting the less plausible observations under the currently estimated model, at each step of the iterations that lead to the final estimate. In the specific application on wine authenticity, they are supposed to be originated by wine adulterations.

2.2 Constrained maximization

We adopt a constrained maximization of $\mathcal{L}_{trim}$, to avoid its unboundedness, by imposing $\psi_{g,ll} \leq c_{noise} \psi_{h,mm}$ for $1 \leq l \neq m \leq p$ and $1 \leq g \neq h \leq G$ where $\{\psi_{g,ll}\}_{l=1,\ldots,p}$ are the diagonal element of the noise matrices Ψ_g, and $1 \leq c_{noise} < +\infty$, to avoid the $|\Sigma_g| \to 0$ case. This constraint can be seen as an adaptation to MFA of those introduced in Ingrassia & Rocci (2007). We will look for the ML estimators of Ψ_g under the given constraints, yielding a well-defined maximization problem.

2.3 Specific implementation of the EM algorithm

The Alternating Expectation - Conditional Maximization (AECM) is an extension of the EM algorithm, needed by the factor structure of the model, which employs different specifications of missing data at each stage. The idea is to partition the vector of parameters $\theta = (\theta_1', \theta_2')'$, in such a way that $\mathcal{L}_{trim}$ is easy to be maximized for θ_1 given θ_2 and viceversa. The M-step is also replaced by some computationally simpler conditional maximization (CM) steps.

1^{st}*cycle* : we set $\theta_1 = \{\pi_g, \mu_g, g = 1, \ldots, G\}$, here the missing data are the unobserved group labels $\mathbf{Z} = (\mathbf{z}_1', \ldots, \mathbf{z}_n')$. After applying a step of Trimming, by assigning to the observations with lowest likelihood a null value of the "posterior probabilities", we get one E-step, and one CM-step for obtaining parameters in θ_1.

2^{nd}*cycle* : we set $\theta_2 = \{\Lambda_g, \Psi_g, g = 1, \ldots, G\}$, here the missing data are the group labels $\mathbf{Z}$ and the unobserved latent factors $\mathbf{U} = (\mathbf{U}_{11}, \ldots, \mathbf{U}_{nG})$. We perform a Trimming step, then a E-step, and a constrained CM-step, i.e. a conditional exact constrained maximization of Λ_g, Ψ_g.

3 Simulation Study

The purpose of this simulation study is to show the effectiveness of estimating a robust MFA on a set of observations drawn from two luxury wines, Barolo and Grignolino, and to distinguish observations not belonging to such grapes. We generate 200 observations from a MFA as in (2) where $G = 3$, $p = 27$ and $d = 4$. A first subset of 95 data are drawn with parameters μ_1, Λ_1, Ψ_1 corresponding to Barolo data, and the second subset of 95 observations are drawn with μ_2, Λ_2, Ψ_2 corresponding to Grignolino data, estimated on the *wine* dataset (available within the *R* package *pgmm*, McNicholas *et al.* 2015). The "contamination" is created by a further subset of 10 observations drawn with μ_3, Λ_3, Ψ_3 from Barbera data. The problem of distinguishing adulterated observations from the real mixture components is addressed, together with the algorithm performance in correctly classifying the authentic units. Hence we will estimate a MFA with $G = 2$, $p = 27$, $d = 4$, and trimming level $\alpha = 0.05$, and we compare results with other popular methods: Partiton around medoids, Gaussian mixtures estimated via Mclust, and Mixtures of patterned Gaussian factors estimated by *pgmm*. To perform each of the $B = 1000$ estimations, algorithms have been initialized following the indications of their respective authors, say 10 random starts at each run of *AECM*; default setting for the "build phase" of *pam*, as in Maechler *et al.* (2017); applying model-based hierarchical clustering as per default setting in Fraley *et al.* (2012) for *Mclust*, and 10 random starts at each run, as suggested in McNicholas & Murphy (2008) for *pgmm*.

Table 1. *Average misclassification errors (percent average values on 1000 runs)*

Algorithm	*AECM*	*pam*	*Mclust*	*pgmm*
Misclassification error	0.0044	0.2885	0.0968	0.0054

Table 1 reports the average misclassification error: the AECM algorithm reports a superb classification rate, with smaller variability of the simulated distributions for the estimated quantities, as shown in Figure 1 for the first component of the Barolo mean $\mu_1[1] = 10.45$ and variance $\Sigma_1[1,1] = 0.12$. To evaluate the algorithms performance we consider 3 clusters for pam, Mclust and pgmm; whereas we consider 2 clusters for AECM, because in this approach the adulterated group should ideally be captured by the trimmed units. Only in 275 and 34 simulations respectively a Barolo and a Grignolino observation were wrongly trimmed out: the adulterated group was greatly identified through the trimming process.

Table 2. *Bias and MSE (in parentheses) of the parameter estimators $\hat{\mu}_g$ and $\hat{\Sigma}_g$*

	AECM	*Mclust*	*pam*		*AECM*	*Mclust*	*pgmm*
μ_1	-0.0001	-0.0038	0.0010	Σ_1	-0.0002	-0.0004	0.0150
	(0.0030)	(0.0181)	(0.1053)		(0.0004)	(0.0018)	(0.0040)
μ_2	0.0001	0.0709	-0.0115	Σ_2	-0.0161	-0.0128	0.0007
	(0.0039)	(0.1077)	(0.1396)		(0.0044)	(0.0043)	(0.0039)

Table 2 reports the average bias and MSE (in parenthesis) for the mixture parameters (computed element-wise for every component).

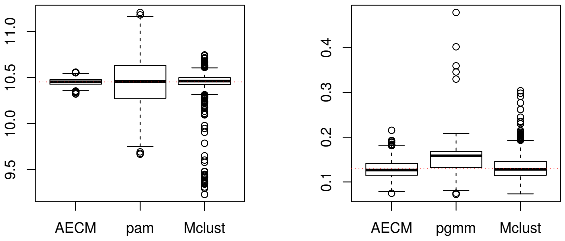

Figure 1. *Boxplots of the simulated distributions of* $\hat{\mu}_1[1]$, *estimator for* $\mu_1[1] = 10.45$ *(left panel);* $\hat{\Sigma}_1[1,1]$, *estimator for* $\Sigma_1[1,1] = 0.12$ *(right panel).*

The present first simulation results assess the effectiveness of adopting robust MFA as a tool for discriminating specific luxury wines from less expensive ones. Further investigation and applications to real datasets are currently ongoing.

References

BAEK, J., MCLACHLAN, G.J., & FLACK, L.K. 2010. Mixtures of Factor Analyzers with Common Factor Loadings: Applications to the Clustering and Visualization of High-Dimensional Data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **32**(7), 1298 –1309.

DA SILVA, F.G., IANDOLINO, A., *et al.* . 2005. Characterizing the grape transcriptome. Analysis of expressed sequence tags from multiple Vitis species and development of a compendium of gene expression during berry development. *Plant physiology*, **139**(2), 574–597.

DAY, N.E. 1969. Estimating the components of a mixture of normal distributions. *Biometrika*, **56**(3), 463–474.

FRALEY, C., RAFTERY, A.E., MURPHY, T.B., & SCRUCCA, L. 2012. *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation.*

GARCÍA-ESCUDERO, L.A., *et al.* . 2016. The joint role of trimming and constraints in robust estimation for mixtures of Gaussian factor analyzers. *Computational Statistics & Data Analysis*, **99**, 131–147.

GRESELIN, F., & INGRASSIA, S. 2015. Maximum likelihood estimation in constrained parameter spaces for mixtures of factor analyzers. *Statistics and Computing*, **25**(2), 215–226.

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

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

MAECHLER, M., ROUSSEEUW, P., STRUYF, A., HUBERT, M., & HORNIK, K. 2017. *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.6.

MCLACHLAN, G.J., PEEL, D., & BEAN, R.W. 2003. Modelling high dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis*, **41**, 379–388.

MCNICHOLAS, P.D., & MURPHY, T.B. 2008. Parsimonious Gaussian mixture models. *Statistics and Computing*, **18**(3), 285–296.

MCNICHOLAS, P.D., ELSHERBINY, A., MCDAID, A.F., & MURPHY, T.B. 2015. *pgmm: Parsimonious Gaussian Mixture Models.*

NEYKOV, N., FILZMOSER, P., DIMOVA, R., & NEYTCHEV, P. 2007. Robust Fitting of Mixtures Using the Trimmed Likelihood Estimator. *Computational Statistics & Data Analysis*, **52**(1), 299–308.