Detecting Wine Adulterations Employing Robust Mixture of Factor Analyzers

Andrea Cappozzo and Francesca Greselin

Abstract An authentic food is one that is what it claims to be. Nowadays, more and more attention is devoted to the food market: stakeholders, throughout the value chain, need to receive exact information about the specific product they are commercing with. To ascertain varietal genuineness and distinguish potentially doctored food, in this paper we propose to employ a robust mixture estimation method. Particularly, in a wine authenticity framework with unobserved heterogeneity, we jointly perform genuine wine classification and contamination detection. Our methodology models the data as arising from a mixture of Gaussian factors and depicts the observations with the lowest contributions to the overall likelihood as illegal samples. The advantage of using robust estimation on a real wine dataset is shown, in comparison with many other classification approaches. Moreover, the simulation results confirm the effectiveness of our approach in dealing with an adulterated dataset.

Keywords Mixtures of factor analyzers · Food authenticity · Model-based clustering · Wine adulteration · Robust estimation · Impartial trimming

1 Introduction and Motivation

The wine segment is identified as a luxury market category, with savvy as well as non-expert customers willing to spend a premium price for a product of a specific vintage and cultivar. Therefore, in the context of global markets, analytical methods for wine identification are needed in order to protect wine quality and prevent its illegal adulteration.

A. Cappozzo (✉) · F. Greselin
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milano, Italy
e-mail: a.cappozzo@campus.unimib.it; francesca.greselin@unimib.it

F. Greselin et al. (eds.), *Statistical Learning of Complex Data*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-030-21140-0_2

In the present work we employ an approach based on robust estimation of mixtures of Gaussian Analyzers, for discriminating corrupted red wines samples from their authentic variety. In a modeling context, we assume a probability distribution function for the chemical and physical characteristics measured on the wines, considering a density in the form of a mixture, whenever the dataset presents more than a wine variety. As a consequence, the probability that a wine sample comes from one specific grape can be estimated from the model, performing classification through the Bayes rule. Robust estimation of the parameters in the model is adopted to recognize the corrupted data. Particularly, we expect that adulterated observations would be implausible under the robustly estimated model: the illegal subsample is revealed by selecting observations with the lowest contributions to the overall likelihood using impartial trimming, without imposing any assumption on their underlying density.

The rest of the paper is organized as follows: in Sect. 2 the notation is introduced and the main concepts about Gaussian Mixtures of Factor Analyzers (MFA), trimmed MFA likelihood, and the Alternating Expectation-Conditional Maximization (AECM) algorithm are summarized. Section 3 presents the *wine* dataset [7] and classification results obtained performing a robust estimation of Gaussian mixtures of factor analyzers. Section 4 reports a simulation study carried out employing parameters estimated from the model in Sect. 3, in a specific framework of contaminated dataset.

The original contribution of the present paper is given in the benchmark study on unsupervised methods, the adaptation of the robust Bayesian Information Criterion (BIC) introduced in [3] to MFA, and a first application of robust MFA in a somehow realistic adulteration scenario.

An application on real data and some simulation results confirm the effectiveness of our approach in dealing with an adulterated dataset 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 the definition and some features of the mixture of Gaussian Factor Analyzers (MFA) and its parameter estimation procedure. MFA is a powerful tool for modeling unobserved heterogeneity in a population, as it concurrently performs clustering and local dimensionality reduction, within each cluster. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample of size n on a p-dimensional random vector. An MFA assumes that each observation $\mathbf{X}_i$ is given by

$$\mathbf{X}_i = \boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g \mathbf{U}_{ig} + \mathbf{e}_{ig} \tag{1}$$

with probability π_g for $g = 1, \ldots, G$. The total number of components in the mixture is denoted by G, $\boldsymbol{\mu}_g$ are $p \times 1$ mean vectors, $\boldsymbol{\Lambda}_g$ 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*, $\mathbf{e}_{ig} \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_g)$ are the *errors*, and $\boldsymbol{\Psi}_g$ are $p \times p$ diagonal matrices. Note that $d < p$, that is the p observable features are supposed to be jointly explained by a smaller number of d unobservable factors. Further, $\mathbf{U}_{ig}$ and $\mathbf{e}_{ig}$ are independent, for $i = 1, \ldots, n$ and $g = 1, \ldots G$. Unconditionally, therefore, $\mathbf{X}_i$ has a density in the form of a G-components multivariate normal mixture:

$$f_{\mathbf{X}_i}(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g \phi_p(\mathbf{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \tag{2}$$

where $\phi_p(\cdot; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ denotes the p-multivariate normal density, whose covariance matrix $\boldsymbol{\Sigma}_g$ has the following decomposition $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$.

When estimating MFA through the usual Maximum Likelihood approach, two issues arise. Firstly, departure from normality in the data may cause biased or misleading inference. Some initial attempts in the literature to overcome this issue propose to consider mixtures of t-factor analyzers [15], but the breakdown properties of the estimators are not improved [10]. The second concern is related to the unboundedness of the log-likelihood function [4], which leads to estimation issues, like the appearance of non-interesting *spurious maximizers* and degenerate solutions. To cope with this second issue, Common/Isotropic noise matrices/patterned covariances [1] and a mild constrained estimation [9] have been considered. The methodology considered here employs model estimation, complemented with *trimming* and *constrained estimation*, to provide robustness, to exclude singularities, and to reduce spurious solutions, along the lines of [8]. Therefore, with this approach, we overcome both previously mentioned issues.

A mixture of Gaussian factor components is fitted 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* [18],

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

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 $\sum_{i=1}^{n} \zeta(\mathbf{x}_i) = \lceil n(1-\alpha) \rceil$, where the less plausible observations under the currently estimated model are tentatively trimmed out at each step of the iterations that lead to the final estimate. In the specific application to wine authenticity analysis described in Sect. 3, they are supposed to be originated by wine adulteration.

Then, a constrained maximization of (3) is adopted, by imposing $\psi_{g,ll} \leq c\, \psi_{h,mm}$ 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 $\boldsymbol{\Psi}_g$, and $1 \leq c < +\infty$, to avoid the $|\Sigma_g| \to 0$ case. This constraint can be seen as an adaptation to MFA of those introduced in [11]. The Maximum Likelihood estimator of $\boldsymbol{\Psi}_g$ under the given constraints leads to a well-defined maximization problem.

The Alternating Expectation-Conditional Maximization—an extension of the Expectation-Maximization algorithm—is considered, in view of the factor structure of the model. The M-step is replaced by some computationally simpler conditional maximization (CM) steps, along with different specifications of missing data. The idea is to partition the vector of parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$, in such a way that $\mathscr{L}_{trim}$ is easy to be maximized for $\boldsymbol{\theta}_1$ given $\boldsymbol{\theta}_2$ and vice versa. Therefore, two cycles are performed at each algorithm iteration:

$1^{st} cycle$: we set $\boldsymbol{\theta}_1 = \{\pi_g, \boldsymbol{\mu}_g, g = 1, \ldots, G\}$; 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 $\boldsymbol{\theta}_1$.

$2^{nd} cycle$: we set $\boldsymbol{\theta}_2 = \{\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g, g = 1, \ldots, G\}$, here the missing data are the group labels $\mathbf{Z}$ and the unobserved latent factors $\mathbf{U}_{11}, \ldots, \mathbf{U}_{nG}$. We perform a Trimming step, then a E-step, and a constrained CM-step, i.e., a conditional exact constrained maximization of $\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g$.

A detailed description of the algorithm is given in [8].

3 Wine Recognition Data

The wine recognition dataset, firstly analysed in [7], reports results of a chemical and physical analysis for three different wine types, grown in the same region in Italy. Originally, 28 attributes were recorded for 178 wine samples derived from three different cultivars: Barolo, Grignolino, and Barbera. A reduced version of the original dataset with only thirteen variables is publicly available in the University of California, Irvine Machine Learning data repository, commonly used in testing the performance of newly introduced supervised and unsupervised classifiers. Particularly, in the unsupervised classification literature the wine recognition data has been considered to assess cluster analysis in information-theoretic terms via minimisation of the partition entropy [19], to prove the modelling capabilities of a generalized Dirichlet mixture [2], to evaluate the efficacy of employing distances based on non-Euclidean norms [5] and of Random Forest dissimilarity [20]. More recently, also parsimonious Gaussian mixture models have been applied to the Italian wines dataset [16].

Here our purpose is twofold: we want to explore the classification performance of a robust estimation based on mixtures of Gaussian Factors Analyzers, and we aim at obtaining realistic parameters for the subsequent simulation study. The dataset, available in the *pgmm* R package [17], contains 27 of the 28 original variables, since the sulphur measurements were not available. Initially, to perform model selection and detect the most suitable values of factors d and groups G, an adaptation to the

Table 1 *Robust BIC* [3] for different choices of the number of factors d and the number of groups G for the robust MFA model on wine data, trimming level $\alpha = 0.05$ and $c = 20$. The smallest value is obtained with $d = 4$ and $G = 2$

	G		
d	1	**2**	3
1	9082.58	8282.92	8223.46
2	8560.62	8107.62	8112.90
3	8352.26	8042.02	8199.38
4	8160.77	**7969.64**	8315.23
5	8102.77	8044.03	8456.00
6	8097.06	8165.67	8735.63

Table 2 Classification table for the robust MFA with number of factors $d = 4$, number of groups $G = 3$, trimming level $\alpha = 0.05$ and $c = 20$ on the wine data

	1	2	3
Barolo	59	0	0
Grignolino	0	71	0
Barbera	0	0	48

Trimmed observations are classified a posteriori according to the Bayes rule

MFA framework of the robust Bayesian Information Criterion, firstly introduced in [3], has been considered. That is, $BIC = -2\mathscr{L}_{trim}(x; \hat{\theta}) + v^c \log n^*$ where $v^c = (G - 1 + Gp + G(pd - d(d-1)/2) + (Gp - 1)(1 - 1/c) + 1)$ denotes the number of free parameters in the model (depending on the value of the constraint c) and $n^* = \lceil n(1-\alpha) \rceil$ the number of non-trimmed observations. Robust BIC for different choices of the number of factors d and the number of groups G are reported in Table 1, considering a trimming level $\alpha = 0.05$ and $c = 20$. The value of the robust BIC is minimized for $d = 4$ and $G = 2$, suggesting a mixture with just two components. Careful investigation on this result highlighted that robust MFA methodology tended to cluster together Barolo and Grignolino samples as arising from the same mixture component, while clearly separating Barbera observations. It is worth recalling [7] that the wines in this study were collected over the time period of 1970–1979, and the Barbera wines are predominantly from a later period than the Barolo or Grignolino wines. Therefore, considering the nature of the phenomena under study and the risks related to rigidly selecting the number of components in a mixture model only on the basis of the results provided by an information criteria, such as BIC [13], we decided to employ a robust MFA with $d = 4$, $G = 3$, and $\alpha = 0.05$, leading to the classification matrix reported in Table 2. Employing a robust MFA rather than a Gaussian mixture leads to a 60% reduction in the number of parameters to be estimated (470 against 1217). Notice, in addition, that after robust estimation, also the trimmed observations can be a posteriori classified according to the Bayes rule, i.e., assigning each of them to the component g having greater value of $D_g(\mathbf{x}, \theta) = \phi_p(\mathbf{x}; \mu_g, \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g)\pi_g$.

Results in Table 2 show that the robust MFA algorithm led to a perfect clusterization of the samples according to their true wine type.

For completeness, the robust MFA algorithm was also applied to the more common thirteen variable subset of the wine data and comparison with the existing literature is reported in Table 3. The clustering performance with respect to the true

Table 3 Comparison of performance metrics for different methodologies on the thirteen variable subset of the wine data

	Performance metric	
Methodology	Class recovery accuracy	Adjusted Rand index
Partition entropy [19]	0.977	–
Mixture of generalized Dirichlet [2]	0.978	–
Neural gas [5]	0.954	–
Random Forest predictors [20]	–	0.93
Parsimonious Gaussian mixture [16]	0.927	0.79
Robust MFA [8]	0.994	0.98

Reported metrics come from the original articles

wine labels reports an *Adjusted Rand Index* equal to 0.98 with just one Grignolino sample wrongly assigned to the cluster identifying Barolo wines. Again then, the robust MFA methodology outperforms the results currently present in the literature for unsupervised learning on this specific dataset.

4 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 identifying units presenting an adulteration. Considering the parameters estimated obtained in Sect. 3, the artificial dataset is generated simulating 100 observations each, from Barolo and Grignolino components. Afterwards, the "contamination" is created decreasing by 15% the values of Fixed Acidity, Tartaric Acid, Malic Acid, Uronic Acids, Potassium, and Magnesium for 5 Barolo and for 5 Grignolino observations. This procedure resembles the illegal practice of adding water to wine [12]. The problem of distinguishing adulterated observations from the real mixture components is addressed, together with the algorithm performance in correctly classifying the authentic units.

We estimate a robust MFA with $G = 2$, $p = 27$, $d = 4$ and trimming level $\alpha = 0.05$. We compare our results with other popular methods: Partition around medoids, Gaussian mixtures estimated via Mclust, and Mixtures of patterned Gaussian factors estimated by $pgmm$. To perform each of the $B = 1000$ simulations, 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 [14], applying model-based hierarchical clustering as per default setting in [6] for $Mclust$ and 10 random starts at each run as suggested in [16] for $pgmm$.

Table 4 Average misclassification errors and ARI (percent average values on 1000 runs)

	AECM	*pam*	*Mclust*	*pgmm*
Misclassification error	0.0309	0.2935	0.2073	0.2314
Adjusted Rand Index	0.9362	0.5466	0.7184	0.6959

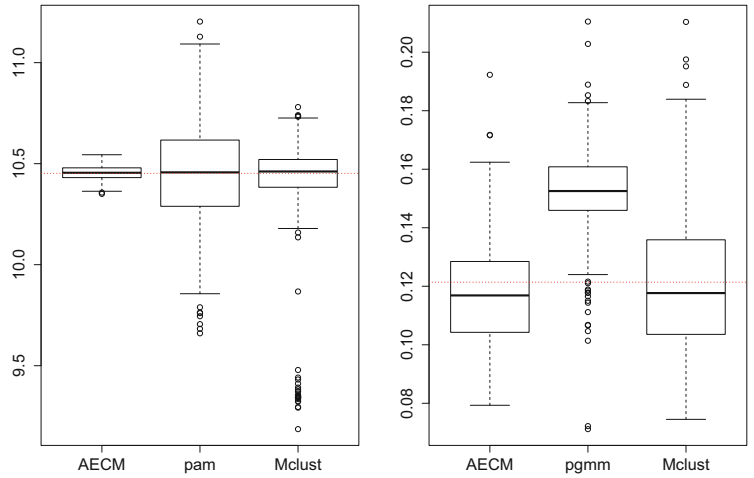

Fig. 1 Boxplots of the simulated distributions of $\hat{\boldsymbol{\mu}}_1[1]$, estimator for $\boldsymbol{\mu}_1[1] = 10.45$ (left panel); $\hat{\Sigma}_1[1, 1]$, estimator for $\Sigma_1[1, 1] = 0.1214$ (right panel)

Table 4 reports the average misclassification error and Adjusted Rand Index: the AECM algorithm reports a superb classification rate, with smaller variability of the simulated distributions for the estimated quantities, as shown in Fig. 1.

For a fair comparison of the performance of the algorithms, we consider 3 clusters for *pam*, *Mclust*, and *pgmm*; whereas we consider only 2 clusters for AECM, because in this approach the adulterated group should ideally be captured by the trimmed units. A value of $c = 20$ allows to discard singularities and to reduce spurious solutions [8]. The effects of the trimming procedure are shown in Fig. 2, where the different colours and shapes represent the obtained classification. Table 5 reports the average bias and MSE for the mixture parameters (computed element-wise for every component). While an R package is under construction, R scripts containing the employed routines are available from the authors upon request.

The present simulations show initial promising results in adopting robust MFA as a tool for identifying wine adulteration. Future research regards a novel approach for semi-supervised robust clustering, allowing for impartial trimming on both labelled and unlabelled data partitions. The aim is to jointly address methodological issues in robust statistics and clustering, as well as providing consistent statistical tools required in the increasingly demanding food authenticity domain.

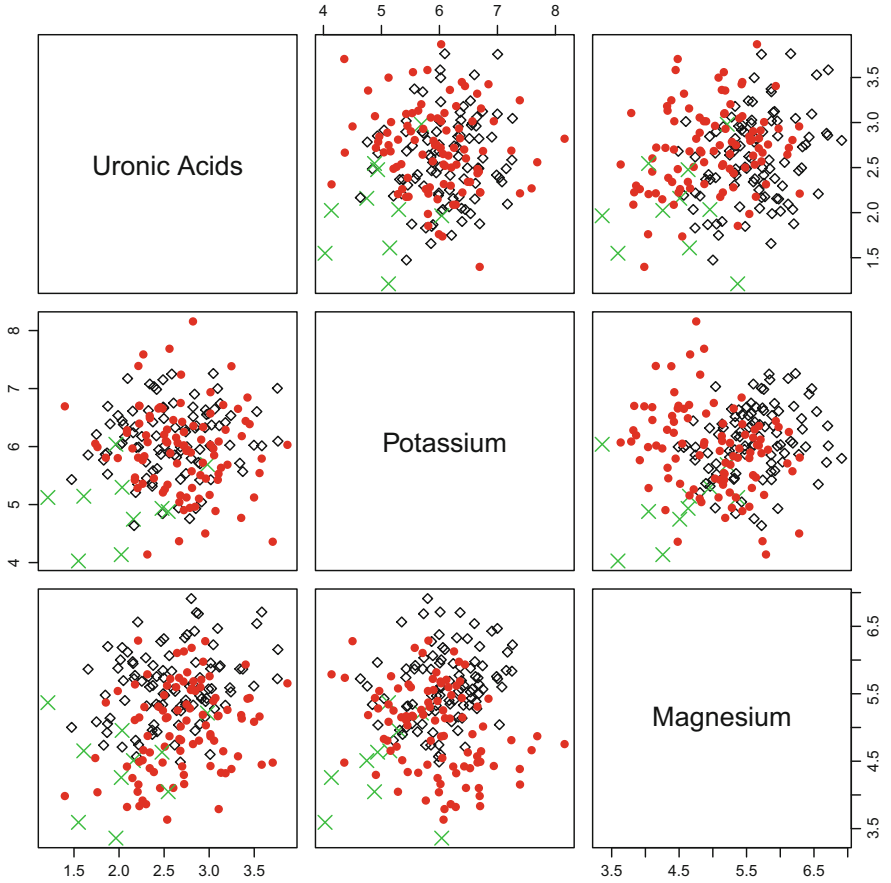

Fig. 2 Clustering of the simulated data with fitted trimmed and constrained MFA. Trimmed observations are denoted by "×"

Table 5 Bias and MSE (in parentheses) of the parameter estimators $\hat{\boldsymbol{\mu}}_g$ and $\hat{\Sigma}_g$

	AECM	*Mclust*	*pam*		*AECM*	*Mclust*	*pgmm*
$\boldsymbol{\mu}_1$	−0.0019	−0.0194	0.0069	Σ_1	0.0001	−0.001	0.0257
	(0.0029)	(0.0421)	(0.1022)		(0.0004)	(0.0022)	(0.0079)
$\boldsymbol{\mu}_2$	−0.0011	0.1522	−0.0025	Σ_2	−0.0156	−0.0164	0.0113
	(0.0042)	(0.2376)	(0.1380)		(0.0043)	(0.0043)	(0.0077)

References

1. Baek, J., McLachlan, G.J., Flack, L.K.: Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualization of high-dimensional data. IEEE Trans. Pattern Anal. Mach. Intell. **32**(7), 1298–1309 (2010)
2. Bouguila, N., Ziou, D.: A powerful finite mixture model based on the generalized Dirichlet distribution: unsupervised learning and applications. In: Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, vol. 1, pp. 280–283. IEEE, Piscataway (2004)
3. Cerioli, A., García-Escudero, L.A., Mayo-Iscar, A., Riani, M.: Finding the number of normal groups in model-based clustering via constrained likelihoods. J. Comput. Graph. Stat. **27**(2), 404–416 (2018)
4. Day, N.E.: Estimating the components of a mixture of normal distributions. Biometrika **56**(3), 463–474 (1969)
5. Doherty, K.A.J., Adams, R.G., Davey, N.: Unsupervised learning with normalised data and non-Euclidean norms. Appl. Soft Comput. **7**(1), 203–210 (2007)
6. Fop, M., Murphy, T.B., Raftery, A.E.: mclust 5: clustering, classification and density estimation using gaussian finite mixture models. R J. **XX**(August), 1–29 (2016)
7. Forina, M., Armanino, C., Castino, M., Ubigli, M.: Multivariate data analysis as a discriminating method of the origin of wines. Vitis **25**(3), 189–201 (1986)
8. García-Escudero, L.A., Gordaliza, A., Greselin, F., Ingrassia, S., Mayo-Iscar, A.: The joint role of trimming and constraints in robust estimation for mixtures of Gaussian factor analyzers. Comput. Stat. Data Anal. **99**, 131–147 (2016)
9. Greselin, F., Ingrassia, S.: Maximum likelihood estimation in constrained parameter spaces for mixtures of factor analyzers. Stat. Comput. **25**(2), 215–226 (2015)
10. Hennig, C.: Breakdown points for maximum likelihood estimators of location-scale mixtures. Ann. Stat. **32**(4), 1313–1340 (2004)
11. Ingrassia, S.: A likelihood-based constrained algorithm for multivariate normal mixture models. Stat. Methods Appl. **13**(2), 151–166 (2004)
12. Jackson, R.S.: Wine Science: Principles and Application. Academic press, Elsevier (2008)
13. Lee, S.X., McLachlan, G.J.: Finite mixtures of canonical fundamental skew t-distributions: the unification of the restricted and unrestricted skew t-mixture models. Stat. Comput. **26**(3), 573–589 (2016)
14. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.: cluster: Cluster analysis basics and extensions, R package version 2.1.0 – For new features, see the 'Changelog' file (in the package source) (2019)
15. McLachlan, G.J., Bean, R.W., Ben-Tovim Jones, L.: Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution. Comput. Stat. Data Anal. **51**(11), 5327–5338 (2007)
16. McNicholas, P.D., Murphy, T.B.: Parsimonious Gaussian mixture models. Stat. Comput. **18**(3), 285–296 (2008)
17. McNicholas, P.D., ElSherbiny, A., McDaid, A.F., Murphy, T.B.: pgmm: Parsimonious Gaussian mixture models, R package version 1.2.3. https://CRAN.R-project.org/package=pgmm (2018)
18. Neykov, N., Filzmoser, P., Dimova, R., Neytchev, P.: Robust fitting of mixtures using the trimmed likelihood estimator. Comput. Stat. Data Anal. **52**(1), 299–308 (2007)
19. Roberts, S.J., Everson, R., Rezek, I.: Maximum certainty data partitioning. Pattern Recognit. **33**(5), 833–839 (2000)
20. Shi, T., Horvath, S.: Unsupervised learning with random forest predictors. J. Comput. Graph. Stat. **15**(1), 118–138 (2006)