Model-Based Learning Using a Mixture of Mixtures of Gaussian and Uniform Distributions

Ryan P. Browne,
Paul D. McNicholas, *Member*, *IEEE*, and
Matthew D. Sparling

Abstract—We introduce a mixture model whereby each mixture component is itself a mixture of a multivariate Gaussian distribution and a multivariate uniform distribution. Although this model could be used for model-based clustering (model-based unsupervised learning) or model-based classification (model-based semi-supervised learning), we focus on the more general model-based classification framework. In this setting, we fit our mixture models to data where some of the observations have known group memberships and the goal is to predict the memberships of observations with unknown labels. We also present a density estimation example. A generalized expectation-maximization algorithm is used to estimate the parameters and thereby give classifications in this mixture of mixtures model. To simplify the model and the associated parameter estimation, we suggest holding some parameters fixed—this leads to the introduction of more parsimonious models. A simulation study is performed to illustrate how the model allows for bursts of probability and locally higher tails. Two further simulation studies illustrate how the model performs on data simulated from multivariate Gaussian distributions and on data from multivariate t-distributions. This novel approach is also applied to real data and the performance of our approach under the various restrictions is discussed.

Index Terms—Statistical computing, multivariate statistics.

1 Introduction

MODEL-BASED clustering is an unsupervised learning technique that fits mixture models to data and gives classifications based on the estimates of the parameters of these mixture models. Model-based classification is a semi-supervised version of model-based clustering. Within this latter paradigm, the observations with known group memberships are jointly modeled with the observations that need to be classified within a joint likelihood framework (see Section 2). A supervised version of this approach can be used and is sometimes called model-based discriminant analysis [11]; further discussion of this approach will be left for Section 4. When mixture models have been used for clustering or classification, the multivariate Gaussian mixture model has been predominant, e.g., [7], [10], [18]; however, other densities, such as the multivariate t-distribution, e.g., [2], [3], [15], have also been used. The density of a parametric finite mixture model can be written

$$\xi(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f(\mathbf{x} \mid \boldsymbol{\theta}_g),$$

where the $\pi_g > 0$ such that $\sum_{g=1}^{G} \pi_g = 1$ are the mixing proportions, $f(\mathbf{x} \mid \boldsymbol{\theta}_g)$ is the density of a multivariate random vector $\mathbf{X}$ with parameters $\boldsymbol{\theta}_g$, and $\boldsymbol{\vartheta} = (\pi_1, \ldots, \pi_G, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$ is the vector of parameters. Herein, we propose a mixture of a multivariate Gaussian distribution and a multivariate uniform distribution as the component density $f(\mathbf{x} \mid \boldsymbol{\theta}_g)$. This flexible model allows for bursts of probability, locally higher tails, or both (see Section 3.1). Three different model formulations are proposed herein and compared to the multivariate Gaussian approach based on simulation studies and real data analyses.

- *The authors are with the Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario N1G2W1, Canada. E-mail: {rbrowne, paul.mcnicholas, msparlin}@uoguelph.ca.*

Manuscript received 23 Dec. 2010; revised 22 June 2011; accepted 21 Aug. 2011; published online 8 Oct. 2011.
Recommended for acceptance by N. Lawrence.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference. IEEECS Log Number TPAMI-2010-12-0981.
Digital Object Identifier no. 10.1109/TPAMI.2011.199.

2 Methodology

2.1 Model-Based Classification

Model-based classification is receiving renewed attention, eg., [3], [17], but despite being the more general case, remains the "poor cousin" to model-based clustering when it comes to coverage within the literature. Consider n p-dimensional observations of which k are known to belong to one of G groups. Without loss of generality, we order the observations so that the first k have known group memberships: $\mathbf{x}_1, \ldots, \mathbf{x}_k, \mathbf{x}_{k+1}, \ldots, \mathbf{x}_n$. Let $\mathbf{z}_i$ denote the group membership of observation i, where $z_{ig} = 1$ if observation i belongs to group g and $z_{ig} = 0$ otherwise. The likelihood for these data can then be written

$$\begin{aligned}\mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x}, \mathbf{z}) = &\prod_{i=1}^{k} \prod_{g=1}^{G} \left[\pi_g f(\mathbf{x}_i \mid \boldsymbol{\theta}_g)\right]^{z_{ig}} \\ &\times \prod_{j=k+1}^{n} \sum_{h=1}^{H} \pi_h f(\mathbf{x}_j \mid \boldsymbol{\theta}_h),\end{aligned} \tag{1}$$

for $H \geq G$; we consider the common $H = G$ case herein, but the approach is analogous for any $H \geq G$. Note that if we set $k = 0$ in (1), we have the model-based clustering likelihood.

2.2 A Mixture of Mixtures

In this paper, we consider a mixture of mixtures of Gaussian and uniform distributions so that our component density is of the form

$$f(\mathbf{x} \mid \boldsymbol{\theta}_g) = w_g \phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) + (1 - w_g)\, u(\mathbf{x} \mid \boldsymbol{\zeta}_g), \tag{2}$$

where $\phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is the density of a multivariate Gaussian random vector with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$, $w_g \in [0, 1]$ is the inner mixing proportion, and $u_g(\mathbf{x} \mid \boldsymbol{\zeta}_g)$ is a multivariate uniform distribution with the range for each dimension denoted by elements of the $2p$-vector $\boldsymbol{\zeta}_g$. That is, the range for the jth element of observation i in component g is given by $(\zeta_{gj}, \zeta_{g2j})$ for $j = 1, \ldots, p$. Our modeling framework arises from using this component density (2) in the likelihood in (1); for $k > 0$, this is a model-based classification approach, but for $k = 0$ this reduces to model-based clustering. Herein, we focus on the former case ($k > 0$) and we propose three variations by holding some parameters fixed during estimation (see Section 2.3).

Note that work has previously been carried out on mixtures of mixtures, e.g., [22], [27]. Such work, however, has been largely confined to the case where all densities are of the same type; most often a mixture of mixtures of Gaussian distributions.

2.3 Model Fitting

We describe an algorithm for fitting the full model (i.e., without constraints) because the others, which will be described later, are special cases. For model fitting, we use a variant of the expectation-maximization (EM) algorithm [8], which is a natural approach for maximum likelihood estimation when data are missing. In our case, there are two sources of missing data; one arises from the fact that we do not know some of the group memberships, i.e., we do not know the z_{ig} for $i = k+1, \ldots, n$; the other arises from the fact that we do not know whether an observation in group g is part of the Gaussian or the uniform component. To denote this second source of missing data, we introduce v_{ig} so that $v_{ig} = 1$ if observation i belongs to the multivariate Gaussian component in group g and $v_{ig} = 0$ if observation i belongs to the multivariate uniform component in group g. The likelihood of these missing data together with the observed data, collectively called the

TABLE 1
The Models under Consideration Herein

Model	Description
I	Set $w_g = 1$; this gives a mixture of multivariate Gaussian distributions.
II	Set the ranges of each multivariate uniform distribution $\boldsymbol{\zeta}_g$ to the minimum and maximum of each variable and set $w_g = 0.95$.
III	Set the ranges of each multivariate uniform distribution $\boldsymbol{\zeta}_g$ to the minimum and maximum of each variable.
IV	No constraints; the full model.

complete-data, is known as the complete-data likelihood. The EM algorithm iterates between two steps—an expectation (E-) step and a maximization (M-) step—until convergence.

In the E-step, we compute the expected value of the complete-data log-likelihood, which is given by

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\vartheta} \mid \mathbf{x}, \mathbf{z}, \mathbf{v}) &= \log\Bigg\{\prod_{i=1}^{k}\prod_{g=1}^{G}\left[\pi_g\phi(\mathbf{x}_i|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)^{\hat{v}_{ig}}u(\mathbf{x}_i|\boldsymbol{\zeta}_g)^{1-\hat{v}_{ig}}\right]^{z_{ig}} \\ &\times \prod_{j=k+1}^{n}\prod_{h=1}^{G}\left[\pi_h\phi(\mathbf{x}_j|\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)^{\hat{v}_{jh}}u(\mathbf{x}_j|\boldsymbol{\zeta}_h)^{1-\hat{v}_{jh}}\right]^{\hat{z}_{jh}}\Bigg\},\end{aligned} \tag{3}$$

where $\hat{z}_{jh}$ and $\hat{v}_{ig}$ denote the conditional expected values given by

$$\hat{z}_{jh} = \frac{\hat{\pi}_h f(\mathbf{x}_j \mid \hat{\boldsymbol{\theta}}_h)}{\sum_{g=1}^{G}\hat{\pi}_g f(\mathbf{x}_j \mid \hat{\boldsymbol{\theta}}_g)},$$

for $j = k+1, \ldots, n$ and

$$\hat{v}_{ig} = \frac{\hat{w}_g\phi(\mathbf{x}_i|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)}{\hat{w}_g\phi(\mathbf{x}_i|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) + (1-\hat{w}_g)\phi(\mathbf{x}_i|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)},$$

for $i = 1, \ldots, n$, respectively.

In the M-step, we maximize $\mathcal{Q}(\boldsymbol{\vartheta} \mid \mathbf{x}, \mathbf{z}, \mathbf{v})$ with respect to the model parameters to get

$$\begin{aligned}\hat{\pi}_g &= n_g/n, \\ \hat{w}_g &= \frac{\sum_{i=1}^{k}\hat{v}_{ig}z_{ig} + \sum_{j=k+1}^{n}\hat{v}_{jg}\hat{z}_{jg}}{n_g}, \\ \hat{\boldsymbol{\mu}}_g &= \frac{1}{n_g}\left(\sum_{i=1}^{k}\hat{v}_{ig}z_{ig}\mathbf{x}_i + \sum_{j=k+1}^{n}\hat{v}_{jg}\hat{z}_{jg}\mathbf{x}_j\right), \\ \hat{\boldsymbol{\Sigma}}_g &= \frac{1}{n_g}\left(\sum_{i=1}^{k}\hat{v}_{ig}z_{ig}\mathbf{x}_i\mathbf{x}_i' + \sum_{j=k+1}^{n}\hat{v}_{jg}\hat{z}_{jg}\mathbf{x}_j\mathbf{x}_j'\right),\end{aligned}$$

where $n_g = \sum_{i=1}^{k} z_{ig} + \sum_{i=k+1}^{n}\hat{z}_{ig}$. The ranges $\boldsymbol{\zeta}_g$ cannot be estimated in closed form and so we use m iterations of simulated annealing using the "SANN" method given within the `optim` function in the R software [23]. Note that because we estimate the ranges in this fashion, the algorithm is formally a generalized-EM (GEM) algorithm (see [16, Section 1.5.5]). However, we also use constrained models where the ranges do not need to be estimated; the algorithm is an EM algorithm in these cases. Our three proposed models and the reduced model (a mixture of multivariate Gaussian distributions) are described in Table 1.

2.4 Convergence of EM Algorithms

The Aitken acceleration [1] is used to estimate the asymptotic maximum of the log-likelihood at each iteration. Based on this estimate, we can make a decision about whether or not the algorithm has converged. The Aitken acceleration at iteration t is given by

$$a^{(t)} = \left[l^{(t+1)} - l^{(t)}\right] / \left[l^{(t)} - l^{(t-1)}\right],$$

where $l^{(t+1)}$, $l^{(t)}$, and $l^{(t-1)}$ are log-likelihood values from iterations $t+1$, t, and $t-1$, respectively. The asymptotic estimate of the log-likelihood at iteration $t+1$ is given by

$$l_\infty^{(t+1)} = l^{(t)} + \frac{1}{1-a^{(t)}}\left[l^{(t+1)} - l^{(t)}\right],$$

as discussed by [5]. An EM algorithm can be considered to have converged when $l_\infty^{(t+1)} - l^{(t)} < \epsilon$ (see [14], [19]).

2.5 Classification and Performance Assessment

The posterior expected values $\hat{z}_{jg}$, for $j = k+1, \ldots, n$, may not lead to hard classifications, i.e., they will often take values other than 0 or 1. In our analyses (Sections 3.2 and 3.3), we use maximum a posteriori (MAP) classifications for the observations with unknown group memberships, where

$$\text{MAP}\{\hat{z}_{jg}\} = \begin{cases} 1, & \text{if } \max_g\{\hat{z}_{jg}\} \text{ occurs at component } g, \\ 0, & \text{otherwise,}\end{cases} \tag{4}$$

for $j = k+1, \ldots, n$.

The adjusted Rand index [12] has become the performance assessment technique of choice within the model-based learning literature; it has several desirable properties, as discussed by [26]. In the analyses in Sections 3.2 and 3.3, we use the adjusted Rand index to assess classification performance for the observations with unknown group memberships. The Rand index [24] is given by

$$(\text{number of pairwise agreements})/(\text{total number of pairs}).$$

A well-known shortcoming of the Rand index is that its expected value is greater than 0 under random classification; this makes values of the Rand index, and small values in particular, difficult to interpret. The adjusted Rand index compensates for chance by allowing for the fact that classification performed randomly will probably correctly classify some cases; it has an expected value of 0 under random classification and gives a value of 1 for perfect classification.

3 Applications

3.1 1D Simulation

In Section 1, we mentioned that a mixture of uniform and Gaussian densities allows for bursts of probability and locally higher tails. To illustrate this feature, we generated data from a univariate, two-component Model IV with component densities

$$\begin{aligned}f(\mathbf{x} \mid \boldsymbol{\theta}_1) &= 0.8\phi(\mathbf{x} \mid -2, 1^2) + 0.2u(\mathbf{x} \mid (-0.6, -0.5)), \text{and} \\ f(\mathbf{x} \mid \boldsymbol{\theta}_2) &= 0.8\phi(\mathbf{x} \mid 2, 1^2) + 0.2u(\mathbf{x} \mid (3, 8)),\end{aligned}$$

respectively, so that the first component exhibits a probability burst over $(-0.5, -0.6)$ and the second component has a locally higher tail over $(3, 8)$. We then fitted two different models, Model I (a mixture of Gaussian distributions) and Model IV, treating all of the observations as known; this is an example of the application of our model for density estimation.

From Fig. 1, we can see that our fitted model (Model IV) accounts for the burst of probability in the first component and the locally higher tail in the second component. The fitted mixture multivariate of Gaussian distributions (Model I), however, cannot capture these features. The difference in the fit of the fitted Gaussian components in each case is very clear; the fact that Model IV accounts for the burst of probability in the first component and the locally higher tail in the second component leads to better fitting Gaussian components than in Model I.

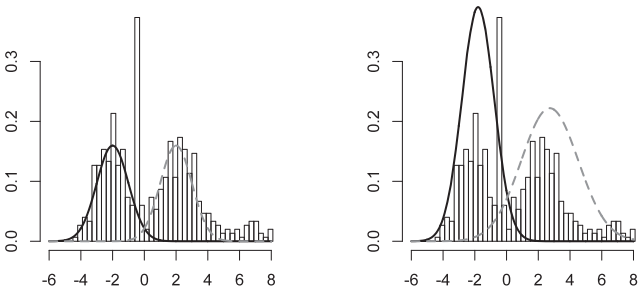

Fig. 1. Histograms of the generated data from our one-dimensional simulation, overlaid with the fitted Gaussian component densities from Models IV (left) and I (right).

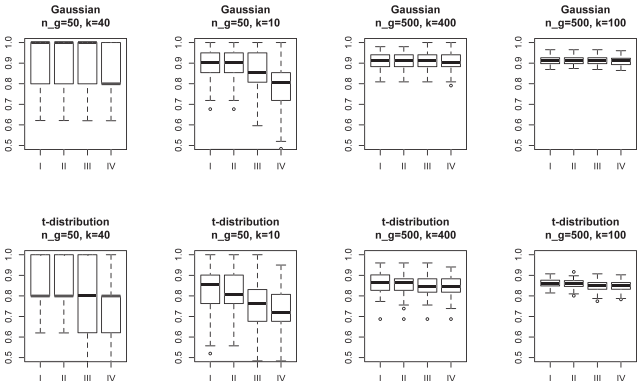

Fig. 2. Boxplots of 100 adjusted Rand indices for Models I-IV for data from each of the multivariate Gaussian and multivariate t-distributions used in our 2D simulations.

3.2 2D Simulations

To explore the classification efficacy of the proposed models, we generated data from both multivariate Gaussian and multivariate t-distributions under a variety of settings. We consider two-component mixture models ($G = 2$), set the number of observations allocated to each group n_g ($g = 1, 2$) to be either 50 or 500, and the number of observations with known group memberships in each group to be either $0.2n_g$ or $0.8n_g$.

For both the multivariate Gaussian and t-distributions, the parameters used to generate the data were

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2 & 0.112 \\ 0.112 & 1.7 \end{bmatrix}, \ \boldsymbol{\mu}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \ \boldsymbol{\mu}_2 = \begin{bmatrix} 5 \\ 6 \end{bmatrix}.$$

For the multivariate t-distribution, we used 10 degrees of freedom. The results from these simulations (Fig. 2) suggest that all four models give excellent classification performance for $n_g = 500$ and that Models I and II are more effective for $n_g = 50$. Our models are applied to real data in Section 3.3.

3.3 Real Data

We consider two real data sets, one of which is relatively easy to classify (wine) and another of which is much more difficult (yeast).

Forina et al. [9] reported 28 chemical and physical properties of three varieties of wine (Barolo, Grignolino, Barbera) from the Piedmont region of Italy. Thirteen of these variables (Table 2) are available, for 178 samples, within the `gclus` package [13] in R. We classify these wines based on variety.

The yeast data [21] were collected to help develop a system to predict protein localization sites in Gram-negative bacteria based on amino acid sequence information. We consider seven variables (Table 3) on 836 proteins and we try to classify them by type: nuclear, mitochondrial, or membrane with no N-terminal signal.

Models I, II, III, and IV were applied to these two real data sets with $k = 0.2n$ and $k = 0.8n$, respectively. We generated 100 random subsets of size k for each value of k and for each data set. The results from both data sets (Fig. 3) show Models II, III, and IV consistently give strong classification performance. Model I, the mixture of multivariate Gaussian distributions, gives strong performance in three of the four scenarios, but gives poor classification performance on the yeast data with $k = 0.2n$. These analyses demonstrate that the novel mixture of mixtures models introduced herein (Models II, III, and IV) can outperform the popular mixture of multivariate Gaussian distributions in classification problems concerning real data.

TABLE 2
The 13 Physical and Chemical Properties of Wine That Are Available in `gclus`

Alcohol	Malic acid
Alcalinity of ash	Ash
Magnesium	Total phenols
Flavonoids	Nonflavonoid phenols
Color Intensity	Proanthocyanins
Hue	Proline
OD_{280}/OD_{315} of diluted wines	

TABLE 3
The Seven Attributes of the Yeast Data That Were Used in Our Analysis

Variable	Description
mcg	McGeoch's method for signal sequence recognition.
alm	Score of the ALOM membrane spanning region prediction program.
mit	Score of discriminant analysis of the amino acid content of the N-terminal region (20 residues long) of mitochondrial and non-mitochondrial proteins.
erl	Presence of "HDEL" substring (thought to act as a signal for retention in the endoplasmic reticulum lumen).
pox	Peroxisomal targeting signal in the C-terminus.
vac	Score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins.
nuc	Score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins.

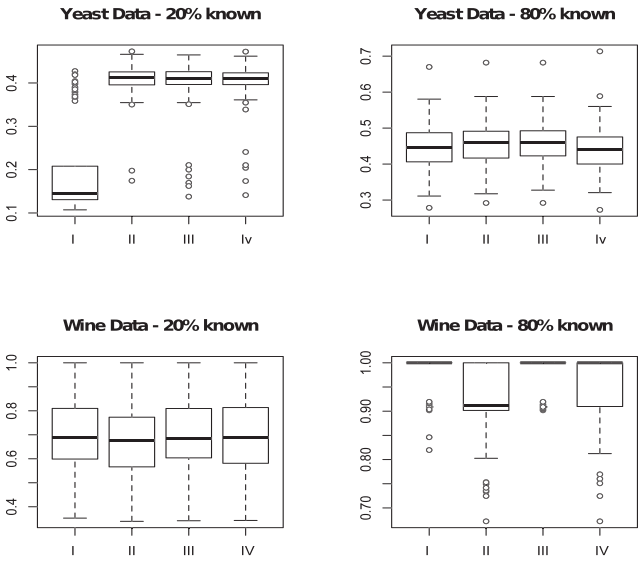

Fig. 3. Boxplots of adjusted Rand indices for wine and yeast data with 20 and 80 percent known observations.

4 Discussion and Future Work

A flexible mixture of mixtures modeling paradigm has been introduced whereby a mixture of a multivariate Gaussian distribution and a multivariate uniform distribution is used for each component density. Two different types of constraints were

considered for these models—fixing the uniform parameter and fixing the inner mixing proportion—and a total of three different models ensued. These models were applied to simulated and real data where they gave good classification performance and, in some cases, superior performance to the well-established multivariate Gaussian mixture model.

In this paper, we have demonstrated that the mixture of mixtures approach can be effective. Future work will investigate different mixture models for the component densities and will consider the use of the mixture of mixtures approach for clustering. In the clustering paradigm, model selection will become an issue; several approaches to overcoming this problem will be studied, e.g., [4], [6], [20], [25]. Work on model-based discriminant analysis [11] and further work on density estimation using these models will also be conducted.

ACKNOWLEDGMENTS

This work was supported by a Discovery Grant and a Collaborative Research and Development Grant from the Natural Sciences and Engineering Research Council of Canada, by a grant-in-aid from Compusense Inc., by a Collaborative Research Grant from the Ontario Centres of Excellence, and by the University Research Chair in Computational Statistics at the University of Guelph. Equipment used in this work was purchased with support from the Canada Foundation for Innovation's Leaders Opportunity Fund and from the Ontario Ministry for Research and Innovation's Small Infrastructure Fund.

REFERENCES

[1] A.C. Aitken, "On Bernoulli's Numerical Solution of Algebraic Equations," *Proc. Royal Soc. Edinburgh,* vol. 46, pp. 289-305, 1926.
[2] J.L. Andrews and P.D. McNicholas, "Extending Mixtures of Multivariate t-Factor Analyzers," *Statistics and Computing,* vol. 21, no. 3, pp. 361-373, 2011.
[3] J.L. Andrews, P.D. McNicholas, and S. Subedi, "Model-Based Classification via Mixtures of Multivariate t-Distributions," *J. Computational Statistics and Data Analysis,* vol. 55, no. 1, pp. 520-529, 2011.
[4] C. Biernacki, G. Celeux, and G. Govaert, "Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 7, pp. 719-725, July 2000.
[5] D. Böhning, E. Dietz, R. Schaub, P. Schlattmann, and B. Lindsay, "The Distribution of the Likelihood Ratio for Mixtures of Densities from the One-Parameter Exponential Family," *Annals Inst. of Statistical Math.,* vol. 46, pp. 373-388, 1994.
[6] G. Bouchard and G. Celeux, "Selection of Generative Models in Classification," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 4, pp. 544-554, Apr. 2006.
[7] C. Bouveyron, S. Girard, and C. Schmid, "High-Dimensional Data Clustering," *Computational Statistics and Data Analysis,* vol. 52, no. 1, pp. 502-519, 2007.
[8] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Soc. B,* vol. 39, pp. 1-38, 1977.
[9] M. Forina, C. Armanino, M. Castino, and M. Ubigli, "Multivariate Data Analysis as a Discriminating Method of the Origin of Wines," *Vitis,* vol. 25, pp. 189-201, 1986.
[10] C. Fraley and A.E. Raftery, "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *J. Am. Statistical Assoc.,* vol. 97, pp. 611-631, 2002.
[11] T. Hastie and R. Tibshirani, "Discriminant Analysis by Gaussian Mixtures," *J. Royal Statistical Soc. B,* vol. 58, pp. 155-176, 1996.
[12] L. Hubert and P. Arabie, "Comparing Partitions," *J. Classification,* vol. 2, pp. 193-218, 1985.
[13] C. Hurley, "Clustering Visualizations of Multivariate Data," *J. Computational and Graphical Statistics,* vol. 13, no. 4, pp. 788-806, 2004.
[14] B.G. Lindsay, "Mixture Models: Theory, Geometry and Applications," *Proc. NSF-CBMS Regional Conf. Series in Probability and Statistics,* vol. 5, 1995.
[15] G.J. McLachlan, R.W. Bean, and L. Ben-Tovim Jones, "Extension of the Mixture of Factor Analyzers Model to Incorporate the Multivariate t-Distribution," *Computational Statistics and Data Analysis,* vol. 51, no. 11, pp. 5327-5338, 2007.
[16] G.J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions,* second ed. John Wiley and Sons, 2008.
[17] P.D. McNicholas, "Model-Based Classification Using Latent Gaussian Mixture Models," *J. Statistical Planning and Inference,* vol. 140, no. 5, pp. 1175-1181, 2010.
[18] P.D. McNicholas and T.B. Murphy, "Model-Based Clustering of Microarray Expression Data via Latent Gaussian Mixture Models," *Bioinformatics,* vol. 26, no. 21, pp. 2705-2712, 2010.
[19] P.D. McNicholas, T.B. Murphy, A.F. McDaid, and D. Frost, "Serial and Parallel Implementations of Model-Based Clustering via Parsimonious Gaussian Mixture Models," *Computational Statistics and Data Analysis,* vol. 54, no. 3, pp. 711-723, 2010.
[20] M. Meila, "Comparing Clusterings—An Information Based Distance," *J. Multivariate Analysis,* vol. 98, no. 5, pp. 873-895, 2007.
[21] K. Nakai and M. Kanehisa, "Expert System for Predicting Protein Localization Sites in Gram-Negative Bacteria," *Proteins,* vol. 11, no. 2, pp. 95-110, 1991.
[22] P. Orbanz and J.M. Buhmann, "SAR Images as Mixtures of Gaussian Mixtures," *IEEE Int'l Conf. Image Processing,* vol. 2, pp. 209-212, 2005.
[23] R Development Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, 2010.
[24] W.M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," *J. Am. Statistical Assoc.,* vol. 66, pp. 846-850, 1971.
[25] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics,* vol. 6, pp. 461-464, 1978.
[26] D. Steinley, "Properties of the Hubert-Arabie Adjusted Rand Index," *Psychological Methods,* vol. 9, no. 3, pp. 386-396, 2004.
[27] M. Di Zio, U. Guarnera, and R. Rocci, "A Mixture of Mixture Models for a Classification Problem: The Unity Measure Error," *Computational Statistics and Data Analysis,* vol. 51, no. 5, pp. 2573-2585, 2007.