Biostatistics (2008), **9**, 1, *pp.* 66–80
doi:10.1093/biostatistics/kxm012
Advance Access publication on April 27, 2007

Model-based clustering on the unit sphere with an illustration using gene expression profiles

JEAN-LUC DORTET-BERNADET*

Institut de Recherche Mathématique Avancée (IRMA), UMR 7501 CNRS, Université Louis Pasteur, Strasbourg, France
dortet@math.u-strasbg.fr

NICOLAS WICKER

Laboratoire de Bioinformatique et Génomique Intégratives, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), UMR 7104, Université Louis Pasteur, Strasbourg, France
wicker@igbmc.u-strasbg.fr

SUMMARY

We consider model-based clustering of data that lie on a unit sphere. Such data arise in the analysis of microarray experiments when the gene expressions are standardized so that they have mean 0 and variance 1 across the arrays. We propose to model the clusters on the sphere with inverse stereographic projections of multivariate normal distributions. The corresponding model-based clustering algorithm is described. This algorithm is applied first to simulated data sets to assess the performance of several criteria for determining the number of clusters and to compare its performance with existing methods and second to a real reference data set of standardized gene expression profiles.

Keywords: Clustering; Directional data; Microarrays; Mixture.

1. INTRODUCTION

We consider model-based clustering of data that lie on a $(p-1)$-dimensional unit sphere $\mathcal{S}_{p-1}$ of R_p. Although the statistical analysis of data on a sphere is quite common in various fields, the clustering of directional data is relatively recent; one can quote, for example, Peel *and others* (2001) for the analysis of rock fracture data in the case $p = 3$ or Banerjee *and others* (2005) for text analysis in the general case $p \geqslant 3$. Our main motivation is that, in gene expression analysis, standardized genes that have mean 0 and variance 1 across the arrays can be interpreted as directional data. This standardization is applicable when one is interested in the gene expression variation under different conditions, such as different types of gene expression profiles as given in Tavazoie *and others* (1999). The aim of the clustering of such standardized genes is to produce groups of genes that are functionally related (Eisen *and others*, 1998).

*To whom correspondence should be addressed.

There is an extensive literature on the use of different clustering techniques. Within the field of gene expression analysis, one can quote, for example, hierarchical clustering (Eisen *and others*, 1998), self-organizing maps (Tamayo *and others*, 1999), or k means (Tavazoie *and others*, 1999). Nevertheless, it is generally recognized that such methods do not offer a rigorous statistical setting and some crucial questions cannot be addressed, for example, the type or the number of clusters present in the data. The model-based clustering approach offers an alternative. It assumes that the data are generated by a mixture of underlying probability distributions such as multivariate normal distributions that allow different shapes. A general presentation of this method and some applications can be found in Banfield and Raftery (1993), Celeux and Govaert (1993), McLachlan and Basford (1988), or Titterington *and others* (1985). The estimation of the different coefficients of the mixture model is commonly done via the EM algorithm of Dempster *and others* (1977). Yeung *and others* (2001) studied the use of Gaussian mixture models for clustering gene expressions. They showed that in general the method performs well, and they explored the validity of the use of the multivariate normal distribution for different transformations of real expression data.

Within the scope of a model-based clustering of data points on a sphere, one needs to consider appropriate distributions. Although there is an extensive literature on directional statistics and distributions (e.g. Mardia and Jupp, 1999, or Fisher *and others*, 1987), the emphasis in this literature is on distributions on spheres in R_2 or R_3, and direct generalization to higher dimensions seems problematic. The well-known Fisher distribution on the sphere is defined for any dimensional space and is an extension of the von Mises distribution. Banerjee *and others* (2005) used this distribution in a model-based clustering of directional data. Nevertheless, since this Fisher distribution is only defined through a mean direction μ and a real parameter related to its spread, the isodensity lines are circular, and thus the resulting clusters tend to be spherical. Our aim is to consider a more general type of distribution that allows different shapes and orientations. The so-called Kent distributions (Kent, 1982) have this interesting feature. Nevertheless, the estimation of their parameters is problematic, and the author proposed, for the sphere in R_3, estimation by moments. These distributions have been used by Peel *and others* (2001) to form groups of fracture data via a model-based clustering. In this paper, we propose to use the inverse stereographic projections of multivariate normal distributions. We will see that, apart from some pathological cases (that in practice can be avoided), such distributions allow a clustering with various shapes and orientations.

In Section 2 of the paper, we introduce the notations and briefly describe a standardization of multivariate data points that gives directional data. In Section 3, the inverse stereographic mapping of multivariate Gaussian distributions is presented. Some features of such distributions on the sphere are given and the construction of the density function is explained. We also derive a method for estimating the parameters. In Section 4, the corresponding model-based clustering algorithm is introduced and used in Section 5 on several simulated data sets. These simulations allow us to assess the performance of common criteria for determining the number of clusters and to compare the clustering algorithm described in this paper with existing methods. Finally in Section 6, we study a real reference data set that comes from the analysis of gene expression profiles, the yeast data set studied in Cho *and others* (1998).

2. Notations—standardization of multivariate data points on a unit sphere

To introduce the statistical setting used throughout this paper, we describe a common standardization procedure of multivariate data points. This standardization is of interest for clustering when one considers as a similarity index of 2 data points the correlation of their coordinates.

Suppose that the n objects, $i = 1, \ldots, n$, to be clustered are initially represented as multivariate data points in R_{p+1},

$$x^{i*} = (x_1^{i*}, \ldots, x_j^{i*}, \ldots, x_{p+1}^{i*}),$$

where the $p+1$ coordinates correspond to the observations of $p+1$ real random variables. For each object i, $i = 1, \ldots, n$, let us denote by

$$\overline{x}^{i*} = \frac{1}{p+1} \sum_{j=1}^{p+1} x_j^{i*} \quad \text{and} \quad \widehat{\sigma}_{x^{i*}} = \left\{ \frac{1}{p+1} \sum_{j=1}^{p+1} (x_j^{i*} - \overline{x}^{i*})^2 \right\}^{1/2}$$

the empirical mean and standard deviation of its coordinates. We consider the standardization of the coordinates defined by, for $i = 1, \ldots, n$ and $j = 1, \ldots, p+1$,

$$\tilde{x}_j^i = \frac{1}{\sqrt{p+1}} \frac{x_j^{i*} - \overline{x}^{i*}}{\widehat{\sigma}_{x^{i*}}}.$$

By construction, it is clear that we have for each $i = 1, \ldots, n$,

$$\sum_{j=1}^{p+1} \tilde{x}_j^i = 0 \quad \text{and} \quad \sum_{j=1}^{p+1} (\tilde{x}_j^i)^2 = 1.$$

Thus, the points $\tilde{x}^i = (\tilde{x}_1^i, \ldots, \tilde{x}_{p+1}^i)$ lie at the intersection of a plane of dimension p with the unit sphere of R_{p+1}. Consequently, all these $\tilde{x}^i$'s lie on the unit sphere $\mathcal{S}_{p-1}$ of R_p defined by

$$\mathcal{S}_{p-1} = \left\{ s = (s_1, \ldots, s_p) \in R_p \colon \sum_{j=1}^{p} s_j^2 = 1 \right\}.$$

Recall that if $\langle \cdot, \cdot \rangle$ denotes the usual scalar product for R_{p+1}, then for 2 objects i and i',

$$\left\langle \tilde{x}^i, \tilde{x}^{i'} \right\rangle = \frac{1}{\widehat{\sigma}_{x^{i*}} \widehat{\sigma}_{x^{i'*}}} \sum_{j=1}^{p+1} (x_j^{i*} - \overline{x}^{i*}) \left(x_j^{i'*} - \overline{x}^{i'*} \right)$$

represents the correlation of the initial coordinates.

Subsequently, for each object $i = 1, \ldots, n$, the vector x^i that lies on the unit sphere $\mathcal{S}_{p-1}$,

$$x^i = (x_1^i, \ldots, x_p^i),$$

will correspond to the set of its standardized coordinates expressed in a given orthonormal basis of R_p. Briefly, these points of R_p can be constructed by first expressing the $\tilde{x}^i$'s in a given orthonormal basis of R_{p+1} whose first vector is proportional to $(1, \ldots, 1)'$ and then by removing the first coordinate. Note that this transformation does not affect the meaning of the scalar product, as 2 data points that are close on the sphere $\mathcal{S}_{p-1}$ have their initial coordinates correlated. Finally, we formulate the scope of this paper as the clustering of n data points x^i in R_p that lie on the unit sphere $\mathcal{S}_{p-1}$.

3. USE OF INVERSE STEREOGRAPHIC PROJECTION OF MULTIVARIATE NORMAL DISTRIBUTIONS

We consider in this section the use of inverse stereographic projection of multivariate normal distributions. Some features of such distributions on the sphere are given, after which we present the construction of the corresponding density function and derive a method for estimating its parameters.

The distributions on the sphere $\mathcal{S}_{p-1}$ that we will use in the mixture model are denoted by $\mathcal{L}_{\mu,\Sigma}$. They depend on 2 types of parameters: a direction μ on the sphere $\mathcal{S}_{p-1}$ and Σ, a $(p-1)\times(p-1)$ positive definite matrix. A distribution $\mathcal{L}_{\mu,\Sigma}$ corresponds to the image via an inverse stereographic projection of a multivariate normal distribution $\mathcal{N}_{p-1}(0,\Sigma)$ that is defined on the plane of dimension $p-1$ perpendicular to μ. Recall that, if one considers a given direction μ on the sphere $\mathcal{S}_{p-1}$, the corresponding stereographic projection of a point x that belongs to $\mathcal{S}_{p-1}$ lies at the intersection of a line joining the "antipole" $-\mu$ and x, with a given plane perpendicular to μ.

We use for convention the stereographic projection described in Figure 1 for the simple case $p=2$ and where μ is the unit vector e_1 of the horizontal axis. In this figure, the point z represents the stereographic projection on the real line of the point of the unit circle with Cartesian coordinates (x, y); it lies at the intersection of the plane generated by e_2, the unit vector of the vertical axis, with the line joining the antipole $(-1, 0)$ and (x, y). The stereographic projection of the sphere $\mathcal{S}_2$ of R_3 is described, for example, in Mardia (1972, p. 216).

By construction, the isodensity lines of a $\mathcal{L}_{\mu,\Sigma}$-distribution are inverse stereographic mappings of ellipsoids. This allows different shapes and orientations. See Figure 2 for examples of these isodensity lines for a same direction μ and for different "reasonable" values of the Σ matrix. In these cases, the $\mathcal{L}_{\mu,\Sigma}$-distributions are unimodal with modal direction μ. Note that the isodensity lines of a $\mathcal{L}_{\mu,\Sigma}$-distribution are circular for the case $\Sigma=\sigma^2 I_{p-1}$.

As the values of the Σ matrix tend to be large, a $\mathcal{L}_{\mu,\Sigma}$-distribution can become multimodal: regions of R_{p-1} far from 0 have relatively large probability mass according to a $\mathcal{N}_{p-1}(0,\Sigma)$ distribution, and they are mapped onto relatively small regions of $\mathcal{S}_{p-1}$. In this case, μ is only a local maximum of the density function and the center of symmetry of the maxima set. See Figure 3 for an example in R_3. One can verify that a necessary and sufficient condition for the density of a $\mathcal{L}_{\mu,\Sigma}$-distribution being unimodal is that the greatest eigenvalue of Σ is smaller than $1/2(p-1)$. Clearly, this feature is problematic in the setting of model-based clustering. Nevertheless, the problem can be avoided in practice: a reasonable cluster covers a reasonable portion of the sphere so that the isodensity lines of the corresponding $\mathcal{L}_{\mu,\Sigma}$-distribution are centered around a unique mean direction. We will return to this point in Section 4.

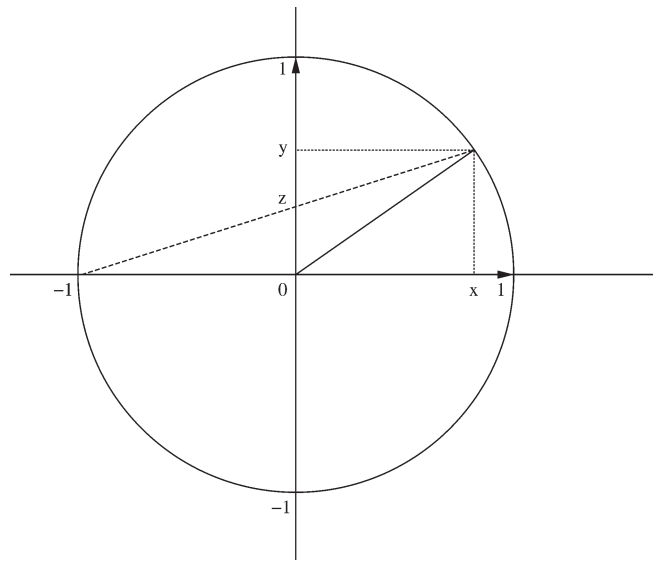

Fig. 1. Case $\mathcal{S}_1$ of R_2, $\mu=e_1$; z stereographic mapping of (x, y).

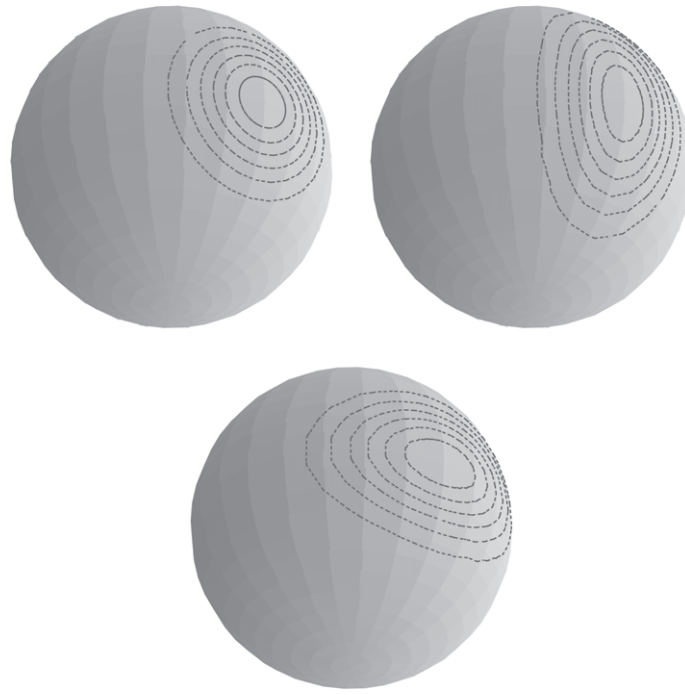

Fig. 2. Case $\mathcal{S}_2$ of R_3; examples of isodensity lines for different Σ's.

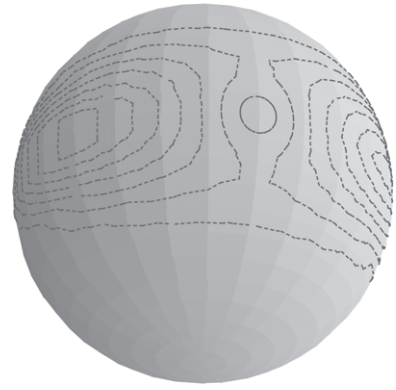

Fig. 3. Case $\mathcal{S}_2$ of R_3; example of isodensity lines for a multimodal distribution (large Σ).

We denote by $\{e_1, \ldots, e_p\}$ the canonical basis of R_p that corresponds to the construction of the coordinates of x^i. In order to define a density function associated with $\mathcal{L}_{\mu,\Sigma}$, we first consider the case where μ coincides with the direction e_1. Thus, the stereographic projection maps $\mathcal{S}_{p-1}$ on the space generated by $\{e_2, \ldots, e_p\}$. We call this projection as $P(\cdot)$. If $y = (y_2, \ldots, y_p)$ represents the stereographic projection of x according to P expressed in the basis $\{e_2, \ldots, e_p\}$, we have, for $i = 2, \ldots, p$,

$$y_i = \frac{x_i}{1 + x_1}.$$

If $d\mathcal{S}$ denotes the surface element on the sphere $\mathcal{S}_{p-1}$, we find that the expression of the density of $\mathcal{L}_{e_1,\Sigma}$ with respect to $d\mathcal{S}$ is given by

$$f_{e_1,\Sigma}(x_1,\ldots,x_p) = \frac{1}{(2\pi)^{(p-1)/2}}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}P(x)'\Sigma^{-1}P(x)\right\}\frac{1}{(1+x_1)^{p-1}},$$

where $x = (x_1,\ldots,x_p)$ is a point on $\mathcal{S}_{p-1}$ and A' denotes the transpose of A. The calculation of this density function, using the polar coordinates of x, is described in Section A of the supplementary material available at *Biostatistics* online (http://www.biostatistics.oxfordjournals.org).

In order to define the density of $\mathcal{L}_{\mu,\Sigma}$ in the general case where μ is a given direction on $\mathcal{S}_{p-1}$, we consider a given rotation $R_\mu(\cdot)$ in R_p such that $R_\mu(e_1) = \mu$. The density function of $\mathcal{L}_{\mu,\Sigma}$ with respect to $d\mathcal{S}$ is

$$f_{\mu,\Sigma}(x) = f_{e_1,\Sigma}(R_\mu^{-1}(x))$$

or, more explicitly,

$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{(p-1)/2}}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}P(R_\mu^{-1}(x))'\Sigma^{-1}P(R_\mu^{-1}(x))\right\}\frac{1}{(1+\langle x,\mu\rangle)^{p-1}}$$

since the Jacobian of the transformation $R_\mu(\cdot)$ is equal to 1.

With regard to the maximum likelihood estimation of μ and Σ, suppose now that we observe a number $n > p$ of independent realizations of the distribution $\mathcal{L}_{\mu,\Sigma}$. The values of μ and Σ are unknown and are to be estimated. We denote by $l(x^1,\ldots,x^n;\mu,\Sigma)$ the corresponding log-likelihood function:

$$l(x^1,\ldots,x^n;\mu,\Sigma) = \sum_{i=1}^{n}\log f_{\mu,\Sigma}(x^i).$$

It is clear that, for a given μ on $\mathcal{S}_{p-1}$, the matrix Σ that maximizes this function is the one that maximizes the expression

$$\sum_{i=1}^{n}\log\left[\frac{1}{(2\pi)^{(p-1)/2}}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}P(R_\mu^{-1}(x^i))'\Sigma^{-1}P(R_\mu^{-1}(x^i))\right\}\right].$$

We denote by $\widehat{\Sigma}(\mu)$ the positive definite matrix solution of this maximization problem. Standard results on multivariate normal distributions (Anderson, 1958) give

$$\widehat{\Sigma}(\mu) = \frac{1}{n}\sum_{i=1}^{n}P(R_\mu^{-1}(x^i))P(R_\mu^{-1}(x^i))',$$

that is the empirical covariance matrix of the stereographic projections of the x^i's on the plane perpendicular to μ. By replacing Σ by $\widehat{\Sigma}(\mu)$ in the log-likelihood function, it appears that the maximum likelihood estimate $\widehat{\mu}_{\mathrm{MLE}}$ of μ maximizes the expression, defined on $\mathcal{S}_{p-1}$,

$$\mathrm{Expr}(\mu) = -\frac{1}{2}n\log|\widehat{\Sigma}(\mu)| - (p-1)\sum_{i=1}^{n}\log\{1+\langle x^i,\mu\rangle\}.$$

When this expression has been maximized, the maximum likelihood estimate $\widehat{\Sigma}_{\mathrm{MLE}}$ of Σ is then given by

$$\widehat{\Sigma}_{\mathrm{MLE}} = \widehat{\Sigma}(\widehat{\mu}_{\mathrm{MLE}}).$$

To our knowledge, there is no closed expression for $\widehat{\mu}_{\mathrm{MLE}}$. Nevertheless, in practice, it can be easily found via a heuristic search algorithm.

4. MODEL-BASED CLUSTERING WITH $\mathcal{L}_{\mu,\Sigma}$-DISTRIBUTIONS

We detail in this section some features of the model-based clustering algorithm. We describe an EM algorithm for the maximum likelihood estimation of the parameters of a mixture of $\mathcal{L}_{\mu,\Sigma}$-distributions. We consider also the use of an extra component to model the noise. Finally, the problem of dimension reduction is discussed.

4.1 *The EM algorithm*

We briefly describe here the algorithm needed to obtain the maximum likelihood estimates of the unknown parameters of a mixture of G distributions $\mathcal{L}_{\mu_1,\Sigma_1},\ldots,\mathcal{L}_{\mu_G,\Sigma_G}$ from the observation of n data points $x^1,\ldots,x^n$ on $\mathcal{S}_{n-1}$. This estimation is classically done by an EM-type algorithm. For more details on this technique, see McLachlan and Basford (1988) or Banfield and Raftery (1993).

If $\tau_1,\ldots,\tau_G$ represent the weights of the G different distributions in the mixture and the x^i's are considered as independent realizations, the observed likelihood is expressed as

$$\mathcal{L}_M(\mu_1,\ldots,\mu_G,\Sigma_1,\ldots,\Sigma_G,\tau_1,\ldots,\tau_G|x^1,\ldots,x^n)=\prod_{i=1}^{n}\sum_{g=1}^{G}\tau_g f_{\mu_g,\Sigma_g}(x^i).$$

One considers the observations x^i as being incomplete. The component labels z_g^i are then introduced with the convention

$$z_g^i=\begin{cases}1, & \text{if } x^i \text{ belongs to cluster } g,\\ 0, & \text{if not.}\end{cases}$$

The vectors z^i are assumed to be independent realizations from a multinomial distribution consisting of one draw on G categories with respective probabilities $\tau_1,\ldots,\tau_G$. The "complete" log-likelihood is then given by

$$l((\mu_g,\Sigma_g,\tau_g),g=1,\ldots,G|(x^i,z_g^i),i=1,\ldots,n,g=1,\ldots,G)=\sum_{i=1}^{n}\sum_{g=1}^{G}z_g^i\log\left\{\tau_g f_{\mu_g,\Sigma_g}(x^i)\right\}.$$

The EM algorithm requires an iterative use of 2 steps, the E-step and the M-step. Briefly, in the E-step, one calculates the conditional expectations of the variables z_g^i given the observed data and the current estimate of the unknown parameters. This step comes down to calculating the values

$$\widehat{z}_g^i=\frac{\tau_g f_{\mu_g,\Sigma_g}(x^i)}{\sum_{g'=1}^{G}\tau_{g'}f_{\mu_{g'},\Sigma_{g'}}(x^i)}.$$

The M-step corresponds to the determination of the maximum likelihood estimates of all the unknown parameters, estimation using all the observations weighted by the z_g^i's. More precisely, we calculate for each $g=1,\ldots,G$,

$$\widehat{\tau}_g=\frac{1}{n}\sum_{i=1}^{n}\widehat{z}_g^i,$$

then we obtain the updated values of $\widehat{\mu}_{g\mathrm{MLE}}$ and $\widehat{\Sigma}_{g\mathrm{MLE}}$ as maximum likelihood estimates from all the x^i's weighted by $\widehat{z}_g^i/\left(\sum_{i'=1}^{n}\widehat{z}_g^{i'}\right)$. The procedure is reiterated until the convergence to a (local) maximum of the log-likelihood function.

One can also consider the Classification EM algorithm (Celeux and Govaert, 1992) where the E-step is replaced by a discrete classification of the z_g^i's.

4.2 *Addition of a noise component*

In practice, the data set can contain a significant amount of noise. As in Banfield and Raftery (1993), this noise can be modeled with an extra component in the mixture model, which then allows the data to determine the amount of noise.

In our setting, it seems natural to model the noise with a uniform distribution on the sphere. Such a procedure does not need to define a hypervolume of the data as in a classical mixture of multivariate normal distributions. The density with respect to the surface element $\mathrm{d}\mathcal{S}$ of this uniform distribution on $\mathcal{S}_{p-1}$ is defined by

$$f(x) = \frac{\Gamma(p/2)}{2\pi^{p/2}},$$

where $\Gamma(\cdot)$ denotes the gamma function. Peel *and others* (2001) used this uniform component in their mixture of Kent distributions for the case $p = 3$.

Note that, with regard to the potential problem of multimodality mentioned in Section 2, the possibility of avoiding the allocation of noisy observations to "good" clusters appears to be essential in our setting: a cluster corrupted by noisy data can have an artificially large Σ_g so that the corresponding $\mathcal{L}_{\mu,\Sigma}$-distribution is multimodal.

4.3 *Starting values for the algorithm*

For running the EM algorithm, one needs to provide some starting values of the parameters. As the algorithm does not converge systematically to a global maximum, this initialization is performed several times.

In a classical mixture of multivariate normal distributions, such starting values are usually given by the use of a k-means algorithm or a hierarchical clustering. Different versions of such algorithms, that depend on the similarity measure or the distance measure that is used, can be considered in our particular context. An example of a k-means algorithm on a sphere that uses the angle between vectors as a distance measure is briefly described in Peel *and others* (2001). We have used this version for the algorithm presented in this paper because it appears to work well in practice. Another example is the version "spkmeans" of Dhillon and Modha (2001) where the scalar product of the vectors is used as a similarity measure.

4.4 *Dimension reduction*

It is clear that the sizes of the clusters have to be sufficiently large relatively to $p - 1$ to avoid any singular estimates of the Σ matrices. In the setting of this work, with $n \gg p$, this condition is generally satisfied (it is here the case for the study of a real data set of gene expression profiles in Section 6).

One way to tackle high-dimensional data is to consider a more parsimonious model. In the case of a Gaussian mixture model, a common recommendation for this problem is to use spherical or diagonal models (Fraley and Raftery, 1999). In our context, the use of a spherical model is similar to the use of model-based clustering via a mixture of Fisher distributions. The restriction to diagonal Σ matrices for data on $\mathcal{S}_{p-1}$ is hard to advocate since it is related to the particular choices of the basis on the planes perpendicular to the mean directions μ of the clusters. We encounter the same problem if we want to use in our context the idea of mixtures of factor models, as described, for example, in McLachlan *and others* (2002).

Another common way to handle high-dimensional data is to use principal components analysis (Anderson, 1958) for dimension reduction. Before the standardization, one looks for the linear combinations of variables with maximal variance that are uncorrelated; then the standardization is conducted on the objects that are expressed in the space of the first principal components.

5. COMPARISONS ON SIMULATED DATA SETS

The algorithm described in Section 4.1 has been tested on several simulated data sets. The goal of these simulations was to assess the performance of several criteria for determining the number of clusters and to compare the algorithm described in this paper with existing clustering methods. We give below some details on the criteria and the clustering methods that have been considered, before giving a brief description of the results of the simulations.

5.1 *Criteria for the number of clusters*

For determining the number of clusters in a data set, we first consider 2 common criteria in a mixture model setting: Akaike information criterion (AIC) (Akaike, 1973) and Bayesian information criterion (BIC) (Schwarz, 1978). Both belong to the class of penalized likelihood criteria. The use of the AIC criterion in our context consists of choosing the number of clusters as the G that maximizes the quantity

$$\mathrm{AIC}_G = \log L_G - k_G,$$

where $\log L_G$ is the estimated maximum of the log-likelihood and k_G is the number of free parameters for the mixture model with G different $\mathcal{L}_{\mu,\Sigma}$-distributions (and the noise component). In general, such a procedure is known to overestimate the true number of groups. Note that this criterion is used in Peel *and others* (2001) for their mixture of Kent distributions on the sphere $\mathcal{S}_2$ of R_3. The use of the BIC criterion consists of choosing the number of groups as the G that maximizes the quantity

$$\mathrm{BIC}_G = \log L_G - \frac{1}{2} k_G (\log n).$$

In general, this BIC criterion is known to underestimate the true number of groups. It is commonly used in the setting of Gaussian mixtures, for example, in the Mclust algorithm of Fraley and Raftery (1999).

The third criterion that we consider here is the gap statistic of Tibshirani *and others* (2001). This technique is applicable to any clustering algorithm. It compares an observed internal index, a within-cluster dispersion, to its expectation under an appropriate reference null distribution. Let W_G denote the within-cluster dispersion associated with the output of a clustering of the data into G groups. If $W_G^{*1}, \ldots, W_G^{*B}$ denote the values of the same index on B data sets generated under the reference null distribution, the gap statistic is given by

$$\mathrm{Gap}_G = \frac{1}{B} \sum_{b=1}^{B} \log(W_G^{*b}) - \log(W_G).$$

If sd_G is the standard deviation of $\log(W_G^{*b})$, $b = 1, \ldots, B$, and if $s_G = \mathrm{sd}_G \sqrt{1 + 1/B}$, then the gap criterion chooses the number of clusters as the smallest G such that

$$\mathrm{Gap}_G \geqslant \mathrm{Gap}_{G+1} - s_{G+1}.$$

In our context, we have considered as the index of within-cluster dispersion the sum of the angles to the cluster means. The uniform distribution on the sphere appears to be a natural reference null distribution.

The fourth criterion considered in this paper is a prediction-based resampling method called Clest that has been introduced in Dudoit and Fridlyand (2002). As for the gap criterion, this method is applicable to any clustering algorithm. The authors have compared its performance with 6 existing criteria using data from 4 published data sets from cancer microarray studies, in the case of the clustering of arrays. The clustering method that was considered is the partitioning around medoids method of Kaufman and Rousseeuw (1990), closely related to the k means. Briefly, the Clest procedure first consists, for each

possible number G, in repeated random splits of all the observations in a learning set and a test set. The classifications obtained on each of the test sets from the classifier constructed on the learning sets are compared to the classifications constructed on the test sets alone by using an external index. An observed similarity statistic is then computed and compared to its expected value under a reference null distribution. The number of clusters G that is chosen corresponds to the largest significant evidence against the null hypothesis. See Dudoit and Fridlyand (2002) for a more detailed description of the method.

5.2 *The clustering methods*

We compare the clustering algorithm presented in this paper with the k-means method adapted to the sphere, using the angle between vectors as a distance measure. This method is here associated with the use of the gap and the Clest criteria. We also consider a comparison with a classical mixture of multivariate normal distributions. For this purpose, we use the Mclust algorithm of Fraley and Raftery (1999), more precisely the EMclustN function that considers noisy data sets. This algorithm uses the BIC criterion to estimate the number of clusters.

Concerning the k-means method, recall that this algorithm is similar in our context to the use of a mixture of Fisher distribution and tends to produce spherical clusters. Thus, it is expected to perform poorly when there are elongated clusters in the data. Note moreover that this algorithm cannot provide a determination of the "noisy" genes.

Concerning the Mclust algorithm, one would expect that it performs poorly in our setting since it does not take into account the particular geometry of the space where the data lie. This feature can be easily verified on simple data sets. When one simulates some data from one single $\mathcal{L}_{\mu,\Sigma}$-distribution with relatively large values for Σ, so that the resulting cluster covers a relatively large portion of the sphere, the Mclust algorithm systematically overestimates the number of groups, whereas the algorithm presented in this paper gives the correct answer with any of the criteria that are considered here. See Figure 4 for such an example, for the case $p = 3$, where 60 data points have been simulated from a single $\mathcal{L}_{\mu,\Sigma}$-distribution with Σ diagonal, $\sigma_1^2 = 0.025$, and $\sigma_2^2 = 0.012$, and where a noise cluster has been simulated with 15 realizations from the uniform distribution on $\mathcal{S}_2$; for this example, the Mclust algorithm gives one additional group.

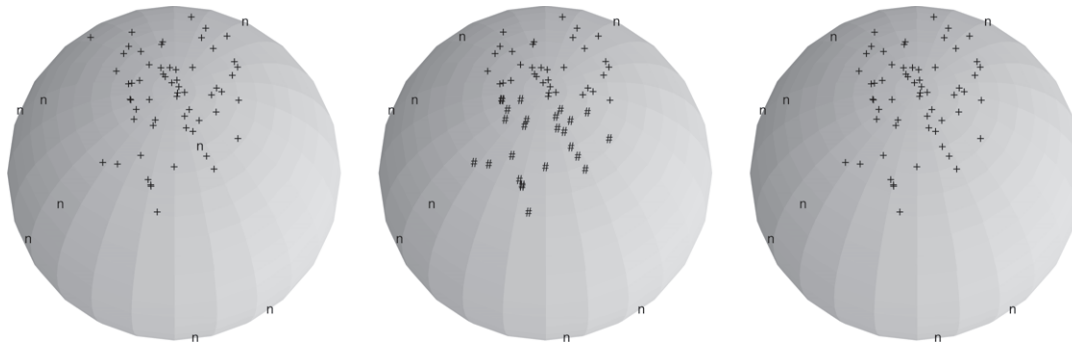

Fig. 4. One cluster from a single $\mathcal{L}_{\mu,\Sigma}$-distribution: simulated data set (left), results with Mclust (center), and results with the algorithm presented in this paper (right); the symbol "+" represents the data points from the $\mathcal{L}_{\mu,\Sigma}$-distribution, "n" represents the noisy observations, and the symbol "#" represents the data points from the additional cluster proposed by Mclust.

5.3 *Results of the simulations*

The different methods and criteria have been compared using 10 simulations of 5 types of data set. These different types of data set correspond to different sizes n, different values for the dimension p, different proportions of noisy data, and different distributions used to simulate the components of each cluster. Three types of data sets use simulations from $\mathcal{L}_{\mu,\Sigma}$-distributions, another type of data set uses Gaussian simulations mapped onto the sphere via a simple scaling, and a last type of data set is the so-called "cyclic data" described in Yeung *and others* (2001).

A detailed description of the data sets and a discussion of the results of the simulations can be found in Section B of the supplementary material available at *Biostatistics* online (http://www.biostatistics.oxfordjournals.org). Briefly, it appears that the mixture of $\mathcal{L}_{\mu,\Sigma}$-distributions associated with the AIC or the gap criterion generally performs well, compared to the other associations of methods and criteria, with a small advantage for the AIC criterion. We consider this criterion for the study of a real example in Section 6.2.

6. APPLICATION TO THE CLUSTERING OF GENE EXPRESSION PROFILES

6.1 *Clustering of gene expression profiles*

The standardization considered in Section 2 is commonly used for clustering n genes described by their expressions observed on $p + 1$ microarray experiments; see, for example, Tavazoie *and others* (1999) or Yeung *and others* (2001). As noted by these last authors, standardizing gene expression data so that the genes have mean 0 and variance 1 across the arrays is helpful when the goal of the clusters is to capture the general patterns across experiments, without considering the absolute expression levels. For example, this standardization is appropriate when one considers the fluctuation of gene expressions over cell cycles. However, each microarray experiment should correspond to a different condition; in the opposite extreme case where all the microarray experiments are replicates of the same condition, clearly the standardization is not appropriate as it comes down to a random projection of the gene expressions onto the sphere. In intermediate cases, it is possible to replace, for each gene, the expressions corresponding to the replicates of a same condition by a single mean or median value.

The scaling across the arrays maps all the gene profiles onto the surface of a sphere, in particular the profiles of some "uninteresting" genes that present low expressions and/or low variability across the experiments. Thus, a preprocessing of the raw data removing from the analysis all uninteresting genes has to be conducted before the standardization of the expressions. Many such procedures have been proposed in the microarray literature, and it is not the focus of this paper to give such recommendations.

Finally, a related but different processing of the data consists in standardizing the gene expression data so that the arrays have mean 0 and variance 1 across the genes. This standardization procedure is particularly helpful when the goal is to cluster arrays (Dudoit and Fridlyand, 2002) so as to prevent the expression levels in one particular array from dominating the average expression levels. As noted by a referee of a previous version of this paper and an associate editor, the fact that the expression levels in a few arrays dominate the average expression levels can introduce spurious correlations between genes; in this case, when the goal is to cluster genes with similar expression profiles, one can first use a standardization of the arrays across the genes, followed by a standardization of the genes across the arrays. McLachlan *and others* (2002) also used this 2 types of standardizations before clustering arrays.

6.2 *Real example: the yeast data set*

The real example that we consider here is the yeast microarray data set of gene expressions described in Cho *and others* (1998). This has been the subject of several studies recently and is generally accepted as

a reference. Briefly, for these data, thermosensible cdc28 mutant cells have been synchronized in early G1 phase by heat treatment. Then, temperature G1 block was further released and cells were harvested every 10 min covering almost 2 entire cell cycles. The complete data set contains the expression profiles of approximately 6220 genes observed over 17 time points.

Tavazoie *and others* (1999) carried out the reference clustering analysis on the most variable 2945 genes of the 6220 genes of the data set. They discovered distinct expression patterns using a k-means cluster analysis. They fixed by trial and error the number of clusters to 30 and suspected that this number may have overestimated the underlying diversity of biological expression classes in the data. The k-means method was applied to the selected 2945 genes which were previously standardized as in Section 2 and with time points at 90 and 100 min removed, so that the resulting gene expression profiles are over 15 time points. The density of points clustering criterion proposed by Wicker *and others* (2002) found 35 clusters for this data set and Mclust estimates 28 clusters.

We have run the algorithm presented in this paper on this data set of the 2945 genes considered by Tavazoie *and others* (1999). The AIC criterion gave 26 clusters of different sizes and a "noise" cluster with 114 genes. So, our estimated number of clusters is of the same order as the one estimated by Tavazoie *and others* (1999), but slightly smaller. This result is not surprising as the model has a noise component and the $\mathcal{L}_{\mu,\Sigma}$-distributions allow different shapes for the clusters. Note that the gap criterion associated with the algorithm presented in this paper gave 43 clusters and that the Clest criterion gave here only 4 clusters. The results of our clustering procedure are summarized in Figure 5, where each cluster is represented by its mean expression profile over the 15 time points.

On the whole, one can distinguish 3 major types of clusters. The first type is "cyclic gene clusters" like clusters numbered 1–4 in which one can recognize the 2 yeast cycles in the mean expression profiles. The second type includes the "stress gene clusters." In fact, due to the synchronization of the cell cycles, an important stress has been put on the cells from which they must recover when cell cycles are restarted. This can be observed in clusters 5–11. The last type includes clusters that are in between: roughly, they concern genes that are neither expressed cyclically nor sensitive to the induced stress (clusters in Figure 5 numbered 12–26).

More specifically, cluster 2 contains many known genes involved in the cell cycle, in particular in phase G1 (CLN1, CLN2, PCL1, and PCL2). Similarly, cluster 1 contains other known cell cycle genes involved in phase G2 (CLB1 and CLB2). Clusters 5 and 11, which have the same type of declining expression profile, are characteristics of genes which somatize the yeast cell stress, namely heat shock proteins (HSP12, HSP26, etc.) and oxidative heat-induced proteins (DHA2, GPD1, etc.). Genes in cluster 10 are known to respond to the cell stress by reactivating the lipid metabolism and by DNA repair.

The noise cluster (cluster numbered 27 in Figure 5) appears to gather various expression profiles for the cycles that cannot be included in other clusters. Its role is to prevent the corruption of clusters by outliers and to highlight stand-alone genes. It is noticeable that the mean profile of this noise cluster over the time points is flat. This feature is not surprising as the genes from this noise cluster are expected to be distributed uniformly on the sphere.

Note that the $\mathcal{L}_{\mu,\Sigma}$ mixture model associated with the AIC criterion has recovered nearly all the clusters that are highlighted in Tavazoie *and others* (1999). In this way, the 3 cyclic clusters numbered 2, 7, and 14 in Tavazoie *and others* (1999) match with the clusters numbered, respectively, 2, 1, and 4 in this paper. As detailed by the authors, these clusters mostly concern some genes implied in the replication and DNA synthesis (cluster here numbered 2), the budding and cell polarity (cluster here numbered 1), and the organization of the centrosome (cluster here numbered 4). Concerning the noncyclic clusters highlighted in this reference study, clusters numbered 1, 3, 4, and 8 in Tavazoie *and others* (1999) correspond to clusters here numbered, respectively, 12, 17, 7, and 8. These clusters are known to contain genes involved with ribosomal proteins (cluster here numbered 12), with RNA metabolism and translation (cluster here

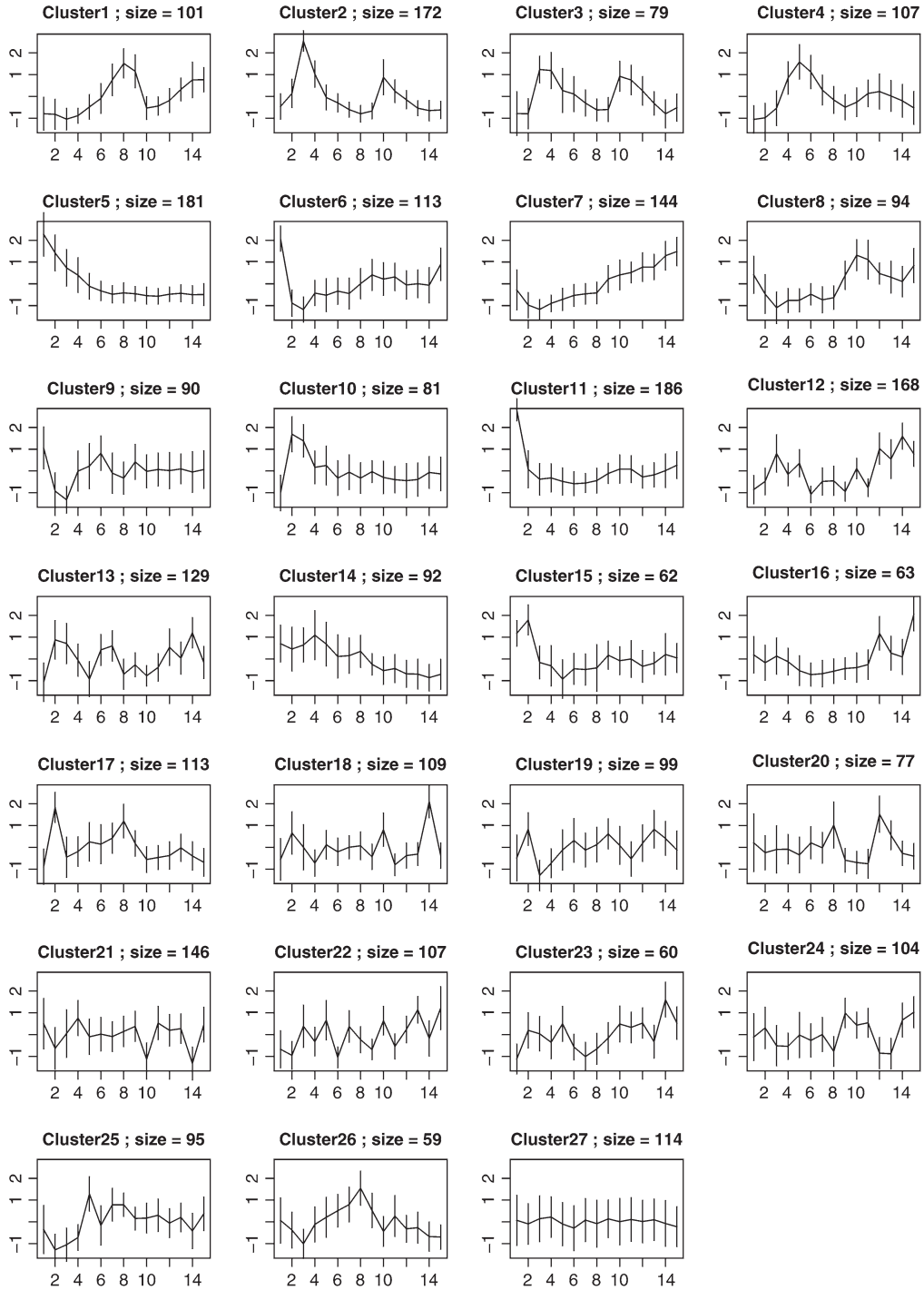

Fig. 5. Yeast data set—mean expression profiles of the 27 clusters found by the algorithm: "cyclic genes" (clusters 1–4), "stress genes" (clusters 5–11), "miscellaneous genes" (clusters 12–26), and noise cluster (cluster 27).

numbered 17), with mitochondrial organization (cluster here numbered 7), and with the carbohydrate metabolism (cluster here numbered 8).

In conclusion, our reanalysis of this data set has given interpretable results that are broadly consistent with the previous analysis by Tavazoie *and others* (1999).

7. Conclusion

This paper has considered model-based clustering of data points that lie on a unit sphere. We have described and discussed a common standardization of multivariate data points that gives directional data. The inverse stereographic projection of a multivariate normal distribution has been considered as a directional distribution. Apart from some pathological cases that can be avoided in practice, such distributions allow a clustering with various shapes and orientations. We have described a procedure to obtain the maximum likelihood estimates of the related parameters with the incorporation of a noise component. This algorithm, by comparison with other clustering methods such as the k-means algorithm on the sphere or the Mclust algorithm, appeared to perform well on simulated data sets, while the AIC and the gap criteria seem to provide good estimates of the number of clusters. Finally, we studied a real reference data set of gene expression profiles for which clustering method provided reasonable clusters with apparent biological meaning.

Acknowledgments

The authors thank an associate editor and the 2 referees of a previous version of the paper for their helpful comments. They also thank Frédéric Bertrand and Vincent Dortet-Bernadet for helpful discussions at the start of the project, Olivier Poch and Agatha Schlüter for interpreting the yeast data set results, and Adeline Legrand for providing the cluster profiles with the help of the Fasabi program. *Conflict of Interest:* None declared.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Petrov, B. N. and Csaki, F. (eds), *Second International Symposium on Information Theory (Tsahkadsor, 1971)*. Budapest, Hungary: Akademiai Kiado, pp. 267–281.

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley and Sons, Inc.

Banerjee, A., Dhillon, I., Ghosh, J. and Sra, S. (2005). Clustering on the unit hypersphere using Von Mises-Fisher distributions. *Journal of Machine Learning Research* **6**, 1345–1382.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821.

Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Journal of Computational Statistics and Data Analysis* **14**, 315–332.

Celeux, G. and Govaert, G. (1993). Comparison of the mixture and the classification maximum likelihood in cluster analysis. *Journal of Statistical Computation and Simulation* **47**, 127–146.

Cho, R., Campbell, M., Winzeler, E., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T., Gabrielian, A., Landsman, D., Lockhart D. *and others* (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell* **2**, 65–73.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 249–282.

DHILLON, I. S. AND MODHA, D. S. (2001). Concept decompositions for large sparse text using clustering. *Machine Learning* **42**, 143–175.

DUDOIT, S. AND FRIDLYAND, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology* **3**, research0036.1–21.

EISEN, M. B., SPELLMAN, P. T., BROWN, P. O. AND BOTSTEIN, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14863–14868.

FISHER, N. I., LEWIS, T. AND EMBLETON, B. J. J. (1987). *Statistical Analysis of Spherical Data*. Cambridge: Cambridge University Press.

FRALEY, C. AND RAFTERY, A. E. (1999). Mclust: software for model-based clustering. *Journal of Classification* **16**, 297–306.

KAUFMAN, L. AND ROUSSEEUW, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

KENT, J. T. (1982). The Fisher-Bingham distribution on the sphere. *Journal of the Royal Statistical Society, Series B* **44**, 71–80.

MARDIA, K. V. (1972). *Statistics of Directional Data*. New York: Academic Press, Inc.

MARDIA, K. V. AND JUPP, P. E. (1999). *Directional Statistics*, Wiley Series in Probability and Statistics. Chichester, UK: Wiley.

MCLACHLAN, G. J. AND BASFORD, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*, Statistics: Textbooks and Monographs, Volume 84. New York: Marcel Dekker, Inc.

MCLACHLAN, G. J., BEAN, R. W. AND PEEL, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* **18**, 413–422.

PEEL, D., WHITEN, W. J. AND MCLACHLAN, G. J. (2001). Fitting mixtures of Kent distributions to aid in joint set identification. *Journal of the American Statistical Association* **96**, 56–63.

SCHWARZ, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **78**, 461–464.

TAMAYO, P., SLONIM, D., MESIROV, J., ZHU, Q., KITAREEWAN, S., DMITROVSKY, E., LANDER, E. S. AND GOLUB, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 2907–2912.

TAVAZOIE, S., HUGHES, J. D., CAMPBELL, M. J., CHO, R. J. AND CHURCH, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genetics* **22**, 281–285.

TIBSHIRANI, R., WALTHER, G. AND HASTIE, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society, Series B* **63**, 411–423.

TITTERINGTON, D. M., SMITH, A. F. M. AND MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distribution*. John Wiley and Sons.

WICKER, N., DEMBELE, D., RAFFELSBERGER, W. AND POCH, O. (2002). Density of points clustering, application to transcriptomic data analysis. *Nucleic Acids Research* **30**, 3992–4000.

YEUNG, K. Y., FRALEY, C., MURUA, A., RAFTERY, A. E. AND RUZZO, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17**, 977–987.

[*Received November 22, 2005; first revision September 14, 2006; second revision November 24, 2006; third revision January 19, 2007; accepted for publication March 9, 2007*]