Clustering of Expression Data from Microarrays: a Mixture-based Approach

Francesco Bartolucci * Francesca Chiaromonte †

Abstract

A central aim of many statistical analyses of microarray data is to cluster genes according to their similarity in expression behavior. In this paper, we perform clustering based on the likelihood fit of a multivariate normal mixture. This approach has several advantages with respect to standard partitioning or hierarchical algorithms; it has an unambiguous inferential characterization, it produces soft partitions through membership probabilities, it allows one to model component mean vectors and covariance structures, and to manage anomalous and missing observations in a natural way. In particular, our mixture-based approach allows us to (i) model component mean vectors through linear reparameterizations, (ii) model component covariance structures through constraints on a special decomposition, (iii) handle outliers through the introduction of a contamination term (uniform on the hypervolume of the data), and (iv) impute missing values. The maximum likelihood estimation of parameters and membership probabilities, and the imputation of missing values, is accomplished through the EM algorithm. Concerning model selection, we employ the classical Bayesian Information Criterion, pragmatically combined with consideration of other features, such as overall membership strength, within-cluster dispersion, and weight of the contamination term. To illustrate our approach, we analyze publicly available data on the reaction of yeast cells to heat shocks. The results of our analysis suggest two alternative clustering models, which provide two different and interesting interpretations of the structure in the data.

Keywords: Clustering; EM algorithm; Maximum Likelihood Estimation; Microarray data; Multivariate Normal Distribution.

*Department of Statistics, University of Perugia, Perugia ITALY.

†Department of Statistics, Penn State University, University Park PA, 16802.

1 Introduction

The last few years have seen the rapid development of new experimental technologies to gather genomic information. Among these, *DNA Microarrays* produce so-called *global gene expression data*, i.e. the simultaneous recording of expression levels for thousands of genes, possibly the whole genome for simple organisms such as yeast. The expression of a gene corresponds to its transcription abundance, or speaking loosely, the intensity with which the gene is "copied upon" to allow, for example, the synthesis of a certain protein. Investigating expression patterns for a very large number of known or putative genes, and over a wide set of conditions, can provide invaluable insight on their roles, actions and interactions. For example, the data we analyze in Section 5 contains expression measurements for the whole yeast genome at several time points, following an abrupt change in temperature – a "heat shock".

Many analyses of microarray data have focused on clustering genes according to some measure of similarity between their expression profiles over a set of experimental conditions. The intent of gene clustering is to ascertain whether the expression behavior of a very large and seemingly complex collective of genes can be summarized through a small number of groups and characteristic expression patterns. It is believed that genes grouped together because of profile similarity are likely to be involved in similar or highly related functions, and perhaps to be co-regulated.

Standard partitioning or hierarchical clustering algorithms have been applied by several groups (see, among others, Chu *et al.*, 1998, Eisen *et al.*, 1998, Spellman *et al.*, 1998, Tamayo *et al.*, 1999, Tavazoie *et al.*, 1999, Iyer *et al.*, 1999), and have identified interesting gene groups and characteristic expression patterns in several experimental contexts. However, the heuristic basis of these algorithms is generally recognized as unsatisfactory. For example, microarray data are affected by several sources of error, and often contain missing values. The outcomes of standard clustering algorithms can be very sensitive to anomalous observations, and to the way that missing ones are imputed. Moreover, the reliability of partitions and characteristic patterns produced by these algorithms is difficult to assess, and some studies (e.g. Holter et al., 2000) suggest that expression data may not actually present well separated clusters. A second generation of studies (see, among others, Brown *et al.*, 2000, Kerr and Churchill, 2000, Hastie *et al.*, 2000, Van der Laan and Bryan, 2000, Heyer *et al.*, 1999) have brought further progress by using more sophisticated clustering strategies, employing resampling schemes, topology-constrained

and/or supervised versions of partitioning algorithms, etc.

Generally, and in particular when faced with the absence of clear-cut "natural" clusters, the analysis can benefit enormously from algorithms that are capable of generating soft partitions, and of exploiting existing information, hypotheses, and/or relevant covariates. Most recently, clustering techniques based on the likelihood fit of multivariate normal mixtures have been applied to microarray data (see Holmes and Bruno, 2000, Brasch and Friedman, 2001, Yeung et al., 2001, and for background on mixture models, McLachan and Basford, 1988, Lindsay, 1995, McLachan and Peel, 2000). Here, we explore and expand upon this approach. Its general advantages are an unambiguous inferential characterization, soft partitioning achieved through membership probabilities, control on characteristic patterns and cluster shapes implemented through the modeling of component mean vectors and covariance structures, and very importantly, the management of anomalous and missing observations.

In particular, with our approach we can (i) model component mean vectors through linear reparameterizations; (ii) model component covariance structures through constraints on a special decomposition; (iii) handle outliers through a contamination term (uniform on the hypervolume of the data; see Fraley and Raftery, 1998); and (iv) impute missing observations.

In comparison to the above referenced applications of mixture-based clustering to microarray data, we stress mean modeling and missing value imputation, and employ a different decomposition in modeling covariances. Mean modeling (see, for instance, Quandt and Ramesey, 1978, Jones and McLachlan, 1992 and Peng *et al.*, 1996) is crucial in some microarray applications. It permits one to use and/or test specific hypotheses on characteristic expression patterns, in the quest for parsimonious and effective descriptions of the data. Covariance modeling determines size, shape, and orientation of the putative clusters. The covariance decomposition we employ is of interest because it separates the correlation structure from the overall size and the coordinate-wise distribution of the variability. Finally, the imputation of missing data values is very important because microarray data often contains a large quantity of missing or unusable readings.

Another crucial advantage of mixture-based clustering is that fits involving different number of components, as well as different constraint formulations for means and covariances, can be evaluated and compared through sensible criteria such as likelihood ratios and BIC (Bayesian Information Criterion). When comparing different constraint formulations on the same number of components, we propose

a pragmatic strategy in which BIC is accompanied by consideration of overall membership strength, within-cluster dispersion, and contamination weight.

The computational challenge of fitting contaminated, high-dimensional normal mixtures under multiple constraints is considerable, but the Expectation-Maximization (EM) algorithm implementation that we have constructed manages it straightforwardly and efficiently, even when dealing with the thousands of observations in a typical microarray data set.

The remainder of the paper is organized as follows: Section 2 provides a brief description of DNA microarrays, and of how expression data are collected and preprocessed. Section 3 describes the contaminated multivariate normal mixture scheme, and gives examples of mean and covariance modeling. Section 4 reviews EM maximum likelihood, and the criteria for model evaluation. Section 5 illustrates our approach through the analysis of publicly available data on yeast reaction to heat shocks. Final comments are given in Section 6, and technical details summarized in an Appendix.

2 A brief overview of microarray data

A typical microarray is a glass slide imprinted with a regular grid of thousands of small spots or holes, each containing fragments of genetic material from lab generated DNA. Most positions in the grid are occupied by known or putative genes whose expression behavior is under study. Others contain replicates and various types of controls that may be employed in data preprocessing stages.

For "one-color" arrays, messenger RNA (mRNA), or possibly complementary DNA (cDNA), is extracted from an experimental sample, labeled with a fluorescent dye, and poured over the chip. Portions of the mRNA hybridize, or bind, preferentially to the resident DNA at different locations on the slide. In "two-color" arrays, mRNA (or cDNA) from the experimental sample is again labelled with a dye, usually red, while mRNA (cDNA) from a separate sample of "reference" cells is labeled with a different dye, usually green. Both are then allowed to hybridize to the DNA on the microarray chip.

After hybridization, excess material is removed from the chip. An image is then generated by shining light through the chip, and processed to extract a dye intensity – a separate reading is made for each dye – at each grid position. This is the bare microarray data. Each intensity reading is proportional to the amount of dye-tagged mRNA (cDNA) that has "bound" with the genetic material at a given grid

position, and thus to the expression of the corresponding gene in the original sample under consideration.

Most microarray studies involve several experimental conditions, with corresponding samples allocated to different chips. In some cases, the data captures a time course; for instance, experimental samples may be taken from the same system (an organism, or tissue culture) at discrete time points during a process such as the cell cycle, or the reaction to a shock or treatment, etc. The aim in these cases is to investigate gene expression variation along the process. In other cases, experimental conditions are of a different nature; for instance, samples may be obtained from different tissues, such as different cancers, organs, etc., with the aim of investigating differential expression of genes across them.

With two-color arrays, reference readings are obtained by using the same reference sample, dye-tagged in green, on all chips. With one-color arrays, reference readings may be obtained by allocating the reference sample to a separate chip. In other cases, reference readings are just an artificial construct to "normalize" experimental measurements, so they could be readings chosen arbitrarily from one of the experimental chips, or some average of readings from all experimental chips.

Final data from a microarray study are usually "properly normalized" log-ratios of experimental readings to reference readings, representing an expression profile over experimental conditions for each of the genes. The general aim of the various normalization procedures proposed in the literature is to make log-ratios from different chips comparable, by eliminating gross sources of non-experimental inter-chip variation in readings.

For more details on microarray experiments, image processing to obtain intensity readings, and normalization procedures, see Eisen and Brown (1999), Yang *et al.* (2000),
and *http://www.nhgri.nih.gov/DIR/VIP/Glossary/Illustration/microarray_technology.html.*

Two additional steps, viewed sometimes as data preprocessing, and sometimes as an integral part of the analysis, are filtering and missing value imputation. Filtering refers to separating out those genes whose expression profiles do (or do not) show significant variation over conditions. This often will restrict further analysis to hundreds, rather than thousands, of genes. The literature on heuristic and rigorous criteria to evaluate significant variation is rapidly growing (see for example Dudoit *et al.*, 2000, and Lee *et al.*, 2000).

Missing values occur when the experimental readings for some grid positions are lacking or unusable (inconsistent, unreliable, above or below feasible detection levels, etc.) due to problems occurring in

the hybridization or the image processing phase. Protocols for missing value imputation range from heuristic to statistically motivated, and have also been the object of growing attention (see for example Hastie *et al.*, 2001). In some studies, though, genes whose expression profiles are incomplete are simply discarded at the outset.

3 The contaminated multivariate normal mixture scheme

Let $\boldsymbol{y} \in \mathbb{R}^T$ denote the vector of expression levels (the expression profile) of a gene on T conditions of interest. We assume that $\boldsymbol{y}$ follows a mixture of C multivariate Gaussian distributions, plus a uniform contamination term; that is

$$\boldsymbol{y} \sim \pi_0 \frac{1}{V} + \sum_{c=1}^{C} \pi_c N_T(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), \tag{1}$$

where V is the hypervolume of the data region (see Fraley and Raftery, 1998, p. 583), $\pi_0, \ldots, \pi_C$ are non-negative weights adding up to 1, and $N_T(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ denotes the T-variate Gaussian distribution with mean vector $\boldsymbol{\mu}_c$ and covariance matrix $\boldsymbol{\Sigma}_c$. Equation (1) allows us to think of the expression profile of any specific gene as coming from one of $C+1$ components, or rather, as belonging to either one of C elliptically shaped clusters, or to a uniformly spread contamination class (component 0). The contamination is meant to capture anomalous observations, so as to reduce their effect on the fit of the Gaussian components. With our approach, it is possible to model both the mean vector and the covariance matrix of each Gaussian component.

Mean modeling: Each $\boldsymbol{\mu}_c$ can be rewritten as a linear function of a vector $\boldsymbol{\beta}_c$ of $p_c \leq T$ parameters; that is

$$\boldsymbol{\mu}_c = \boldsymbol{X}_c \boldsymbol{\beta}_c, \tag{2}$$

where $\boldsymbol{X}_c$ is an appropriate $T \times p_c$ design matrix of full column rank. The tth row of such a matrix contains the values of a set of auxiliary variables characterizing the tth condition. Equation (2) is different from the parameterization used by Jones and McLachlan (1992). There, following a multivariate regression approach, the mean of the cth component used to model the ith observation is given by $\boldsymbol{\mu}_{ic} = \boldsymbol{B}_c \boldsymbol{x}_i$, where $\boldsymbol{B}_c$ is a matrix of parameters and $\boldsymbol{x}_i$ a vector of covariates characterizing the observation (see also Peng *et al.*, 1996). We prefer the parameterization in Eq. (2) because we are interested in modeling mean patterns on the basis of auxiliary variables associated with the T conditions.

Moreover, in the microarray context, covariates characterizing observations (i.e. genes) are seldomly available.

A few examples will help in gauging the scope of the proposed parameterization. When $\boldsymbol{X}_c = \boldsymbol{I}_T$, $\boldsymbol{\mu}_c$ is completely unconstrained; this can be useful for exploratory purposes since the corresponding clusters will contain genes whose average expression profiles are unforeseen and hence not modeled. On the other extreme, a pattern constant over conditions is modeled through the design matrix $\boldsymbol{X}_c = \mathbf{1}_T$, where $\mathbf{1}_T$ denotes a column vector of T ones; constant patterns can be used for filtering purposes, as the corresponding clusters will contain genes whose average expression profiles do not vary over conditions. An intermediate situation is given by a mean pattern repeating itself over $s = T/u$ specified subsets of u conditions; in this situation, the design matrix is $\boldsymbol{X}_c = \mathbf{1}_s \otimes \boldsymbol{I}_u$, where $\otimes$ stands for the Kronecker product. For time course data, say given at times $\tau_1, \ldots, \tau_T$, mean patterns can represent specified time trends. A linear trend is modeled through the two-column matrix $\boldsymbol{X}_c = (\mathbf{1}_T \ \boldsymbol{\tau})$, where $\boldsymbol{\tau} = (\tau_1 \ \ldots \ \tau_T)'$, while a quadratic trend is modeled through the three-column matrix $\boldsymbol{X}_c = (\mathbf{1}_T \ \boldsymbol{\tau} \ \boldsymbol{q})$, where $\boldsymbol{q} = (\tau_1^2 \ \ldots \ \tau_T^2)'$. Also, a mean pattern can be modeled on some conditions and left unconstrained on others; for example, the block diagonal matrix $\boldsymbol{X}_c = \mathrm{diag}(\boldsymbol{I}_h, \boldsymbol{M}, \boldsymbol{I}_k)$ leaves the first h and last k entries of $\boldsymbol{\mu}_c$ unconstrained, while modeling the middle $T - (h + k)$ through $\boldsymbol{M}$.

It is worth noting that a mean pattern that is not constrained to be constant (e.g. an unconstrained pattern, or a polynomially modeled pattern) may be constant or approximately constant in its estimation, and thus can employed for filtering purposes (see Section 5 below).

Covariance modeling: Covariance matrices may also be constrained to embody knowledge relative to the within-cluster variability structure. Banfield and Raftery (1993) employ a framework for representing these matrices, based on a typical eigenvalue decomposition (see also Yeung *et al.* (2001)). There, $\boldsymbol{\Sigma}_c$ is represented as $\boldsymbol{\Sigma}_c = \lambda_c \boldsymbol{D}_c \boldsymbol{A}_c \boldsymbol{D}_c'$ where λ_c (a positive scalar) determines the volume of component c, $\boldsymbol{D}_c$ (an orthogonal matrix of eigenvectors) determines its orientation, and $\boldsymbol{A}_c$ (a diagonal matrix with entries proportional to the eigenvalues) determines its shape. In practice, these authors considered three main types of covariance matrices: (i) totally unconstrained, with no assumption on the within-cluster variability structure; (ii) diagonal with unconstrained entries, $\boldsymbol{\Sigma}_c = \mathrm{diag}(\sigma_{c1}^2, \ldots, \sigma_{cT}^2)$, representing a situation in which within-cluster variability is unrelated over conditions, but may differ in size; (iii) diagonal with equal entries, $\boldsymbol{\Sigma}_c = \sigma_c^2 \boldsymbol{I}_T$, representing a situation in which within-cluster variability

about the mean pattern is the same, and unrelated over conditions (spherical clusters).

Here, we consider an alternative decomposition to represent the covariance matrices, namely

$$\boldsymbol{\Sigma}_c = \sigma_c^2 \mathrm{diag}(\boldsymbol{\nu}_c) \boldsymbol{\Omega}_c \mathrm{diag}(\boldsymbol{\nu}_c) \tag{3}$$

where σ_c^2 s a positive scalar, $\boldsymbol{\nu}_c$ is a T-vector with positive entries normalized so that $\boldsymbol{\nu}_c' \boldsymbol{\nu}_c = T$, and $\boldsymbol{\Omega}_c$ is a correlation matrix. This decomposition has the interesting feature of separating the correlation structure ($\boldsymbol{\Omega}_c$) from the overall size (σ_c^2) and the coordinate-wise distribution ($\boldsymbol{\nu}_c$) of the variability.

Several models may be conceived on the basis of Eq. (3), depending on whether σ_c^2, $\boldsymbol{\nu}_c$ and $\boldsymbol{\Omega}_c$ are totally unconstrained or constrained is some way. More precisely, within our approach, we may require that: for each Gaussian component, the σ_c^2's are totally unconstrained and have to be estimated (U); $\sigma_c^2 = \sigma^2$, for $c = 1, \ldots, C$, with σ^2 to be estimated (E); $\sigma_c^2 = \bar{\sigma}^2$, for $c = 1, \ldots, C$, with $\bar{\sigma}^2$ fixed arbitrarily (F). Similarly, we may require that the $\boldsymbol{\nu}_c$'s and/or the $\boldsymbol{\Omega}_c$'s are U, E or F. The only rule that has to be followed in combining these constraints is that: (i) if the σ_c^2's are E, then $\boldsymbol{\nu}_c$'s must be E or F, whereas, if the σ_c^2's are F then also the $\boldsymbol{\nu}_c$'s must be F; (ii) if the $\boldsymbol{\nu}_c$'s are E, then $\boldsymbol{\Omega}_c$'s must be E or F, whereas, if the $\boldsymbol{\nu}_c$'s are F then also the $\boldsymbol{\Omega}_c$'s must be F. Among the models that may be implemented, some of the most interesting are:

UUU: σ_c^2's, $\boldsymbol{\nu}_c$'s and $\boldsymbol{\Omega}_c$'s are unconstrained; this formulation is useful in the preliminary stage of an analysis, for explanatory purposes;

UUE: σ_c^2's and $\boldsymbol{\nu}_c$'s are unconstrained but $\boldsymbol{\Omega}_c = \boldsymbol{\Omega}$, for $c = 1, \ldots, C$, with $\boldsymbol{\Omega}$ to be estimated; in this way we require all Gaussian components to have the same correlation structure;

UUF: σ_c^2's and $\boldsymbol{\nu}_c$'s are unconstrained but $\boldsymbol{\Omega}_c = \bar{\boldsymbol{\Omega}}$, for $c = 1, \ldots, C$, with $\bar{\boldsymbol{\Omega}}$ fixed arbitrarily; for instance we may have $\bar{\boldsymbol{\Omega}} = \boldsymbol{I}_T$ (uncorrelation), $\bar{\boldsymbol{\Omega}} = (1-\rho)\boldsymbol{I}_t + \rho \mathbf{1}_T \mathbf{1}_T'$ (equal correlation) or $\bar{\boldsymbol{\Omega}}$ with entries $\omega_{tu} = \rho^{|t-u|}$ (AR(1) structure), with $\rho \in (0, 1)$;

UEE: σ_c^2's are unconstrained but $\boldsymbol{\nu}_c = \boldsymbol{\nu}$ and $\boldsymbol{\Omega}_c = \boldsymbol{\Omega}$, for $c = 1, \ldots, C$, with $\boldsymbol{\nu}$ and $\boldsymbol{\Omega}$ to be estimated; in this way we allow Gaussian components to differ in volume, but require them to have the same shape and orientation;

UFF: σ_c^2's are unconstrained but $\boldsymbol{\nu}_c = \bar{\boldsymbol{\nu}}$ and $\boldsymbol{\Omega}_c = \bar{\boldsymbol{\Omega}}$, for $c = 1, \ldots, C$, with $\bar{\boldsymbol{\nu}}$ and $\bar{\boldsymbol{\Omega}}$ fixed arbitrarily; in particular, when $\bar{\boldsymbol{\nu}} = \mathbf{1}_T$ and $\bar{\boldsymbol{\Omega}} = \boldsymbol{I}_T$ we have the so-called *unequal volume spherical model*;

EEE: $\sigma_c^2 = \sigma^2$, $\boldsymbol{\nu}_c = \boldsymbol{\nu}$ and $\boldsymbol{\Omega}_c = \boldsymbol{\Omega}$, for $c = 1, \ldots, C$, with σ^2, $\boldsymbol{\nu}$ and $\boldsymbol{\Omega}$ to be estimated; in this case all components are bound to have the same covariance structure;

EFF: $\sigma_c^2 = \sigma^2$, $\boldsymbol{\nu}_c = \bar{\boldsymbol{\nu}}$ and $\boldsymbol{\Omega}_c = \bar{\boldsymbol{\Omega}}$, for $c = 1, \ldots, C$, with σ^2 to be estimated and $\bar{\boldsymbol{\nu}}$ and $\bar{\boldsymbol{\Omega}}$ fixed arbitrarily; in particular, when $\bar{\boldsymbol{\nu}} = \mathbf{1}_T$ and $\bar{\boldsymbol{\Omega}} = \boldsymbol{I}_T$ we have the so-called *equal volume spherical model.*

In summary, according to our approach, a model is formulated by specifying: (i) the number of components, C; (ii) the constraints on the means, through the matrices $\boldsymbol{X}_c$'s in the parameterization (2); and (iii) the constraints on the covariance structures, formulated according to the decomposition (3). The classical mixture of multivariate Gaussian distributions with unconstrained means and covariances may be obtained setting $\boldsymbol{X}_c = \boldsymbol{I}_T$, for $c = 1, \ldots, C$, and requiring covariances to be of the UUU type. We refer to this model as the *full model.* However, the hope is to obtain a more parsimonious model posing constraints on means and/or covariance matrices. In particular, since each $\boldsymbol{\Sigma}_c$ contains $T(T+1)/2$ parameters, while each mean $\boldsymbol{\mu}_c$ contains only T, constraints on covariances are likely to be more effective in reducing the number of parameters. On the other hand, constraints on means are often crucial for incorporating or testing substantive hypotheses on characteristic expression behaviors. The search for parsimonious representations, together with maximum likelihood estimation, is illustrated in the following Section.

4 Maximum likelihood estimation and model selection

We implement maximum likelihood through the EM algorithm (Dempster *et al.*, 1977), considering our data as incomplete due to both the missing expression values, and the lack of information on the component to which each observation/gene belongs.

The T-dimensional vector $\boldsymbol{y}_i$ containing expression levels for the ith gene is split into two subvectors, $\boldsymbol{y}_i^{\mathrm{o}}$ and $\boldsymbol{y}_i^{\mathrm{u}}$, which refer, respectively, to the observed and unobserved (missing) expression levels for i. Accordingly, the vector $\boldsymbol{\mu}_c$ is split into the two subvectors $\boldsymbol{\mu}_{ic}^{\mathrm{o}}$ and $\boldsymbol{\mu}_{ic}^{\mathrm{u}}$ and the matrix $\boldsymbol{\Sigma}_c$ into the four submatrices $\boldsymbol{\Sigma}_{ic}^{\mathrm{oo}}$, $\boldsymbol{\Sigma}_{ic}^{\mathrm{ou}}$, $\boldsymbol{\Sigma}_{ic}^{\mathrm{uo}}$ and $\boldsymbol{\Sigma}_{ic}^{\mathrm{uu}}$; for instance, $\boldsymbol{\Sigma}_{ic}^{\mathrm{uo}}$ has dimension $(T - o_i) \times o_i$, where o_i is the dimension of $\boldsymbol{y}_i^{\mathrm{o}}$, and is obtained by selecting the rows of $\boldsymbol{\Sigma}_c$ corresponding to $\boldsymbol{y}_i^{\mathrm{u}}$ and the columns corresponding to $\boldsymbol{y}_i^{\mathrm{o}}$. Now, let $\boldsymbol{\eta}_c$ contain the parameters in $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$, and set $\boldsymbol{\pi} = (\pi_0 \ \ldots \ \pi_C)'$. Also,

let $\boldsymbol{Y}^{\text{o}}$ be short-hand notation for all observed expression data, and $\boldsymbol{\theta}$ for all the terms in the model reparameterization ($\boldsymbol{\pi}$, plus $\boldsymbol{\beta}_c$, $c = 1, \ldots, C$ for the means, and σ_c^2, $\boldsymbol{\nu}_c$ and $\boldsymbol{\Omega}_c$, $c = 1, \ldots, C$ for the covariances). Finally, let $f_c(\boldsymbol{y}_i^{\text{o}}; \boldsymbol{\eta}_c)$ be the density of the cth component of the mixture in $\boldsymbol{y}_i^{\text{o}}$ alone, parameterized through $\boldsymbol{\eta}_c$ (in practice, this is the density of a o_i-dimensional Gaussian distribution determined by $\boldsymbol{\mu}_{ic}^{\text{o}}$ and $\boldsymbol{\Sigma}_{ic}^{\text{oo}}$), and form the vector

$$\boldsymbol{f}(\boldsymbol{y}_i^{\text{o}}; \boldsymbol{\eta}) = (1/V(\boldsymbol{y}_i^{\text{o}}) \;\; f_1(\boldsymbol{y}_i^{\text{o}}; \boldsymbol{\eta}_1) \;\; \ldots \;\; f_C(\boldsymbol{y}_i^{\text{o}}; \boldsymbol{\eta}_C))',$$

where $\boldsymbol{\eta} = (\boldsymbol{\eta}_1' \;\; \ldots \;\; \boldsymbol{\eta}_C')'$, and $V(\boldsymbol{y}_i^{\text{o}})$ is the hypervolume of the region containing the observed expression levels for gene i (see Appendix A1 for details). With this notation, the log-likelihood may be expressed as

$$l_{Y^{\text{o}}}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log[\boldsymbol{\pi}' \boldsymbol{f}(\boldsymbol{y}_i^{\text{o}}; \boldsymbol{\eta})]. \tag{4}$$

This is referred to as the *incomplete log-likelihood*, in contrast to the *complete log-likelihood*. The latter we could compute if we knew to which component each gene belonged, and all the unobserved expression levels.

Let $z_{ic} = 1$ if the ith gene belongs to the cth component, and 0 otherwise, and set $\boldsymbol{z}_i = (z_{i0} \;\; \ldots \;\; z_{iC})'$. Also, let $\boldsymbol{Y}$ be short-hand notation for all expression data (including unobserved values), and $\boldsymbol{Z}$ for all the z_{ic}'s. Finally, let $f_c(\boldsymbol{y}_i; \boldsymbol{\eta}_c)$ be the density of the cth component in the whole $\boldsymbol{y}_i$ (this is a T-dimensional Gaussian), and form the vector

$$\boldsymbol{f}(\boldsymbol{y}_i; \boldsymbol{\eta}) = (1/V(\boldsymbol{y}_i) \;\; f_1(\boldsymbol{y}_i; \boldsymbol{\eta}_1) \;\; \ldots \;\; f_C(\boldsymbol{y}_i; \boldsymbol{\eta}_C))',$$

where the hypervolume $V(\boldsymbol{y}_i)$ envelops the whole $\boldsymbol{y}_i$. Then, the complete log-likelihood may be expressed as

$$l_{Y,Z}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \boldsymbol{z}_i' \log(\boldsymbol{\pi}) + \sum_{i=1}^{N} \boldsymbol{z}_i' \log[\boldsymbol{f}(\boldsymbol{y}_i; \boldsymbol{\eta})]. \tag{5}$$

In order to find the vector value $\widehat{\boldsymbol{\theta}}$ which maximizes $l_{Y^{\text{o}}}(\boldsymbol{\theta})$, the EM algorithm iterates the following two steps:

E-step: Using the current $\widehat{\boldsymbol{\theta}}$, compute the expected value of $l_{Y,Z}(\boldsymbol{\theta})$ given the incomplete data $\boldsymbol{Y}^{\text{o}}$; that is $\text{E}[l_{Y,Z}(\boldsymbol{\theta}) | \boldsymbol{Y}^{\text{o}}, \widehat{\boldsymbol{\theta}}]$. In practice, this reduces to computing $\widehat{\boldsymbol{z}}_i = \text{E}(\boldsymbol{z}_i | \boldsymbol{y}_i^{\text{o}}, \widehat{\boldsymbol{\theta}})$, for $i = 1, \ldots, n$, and $\widehat{\boldsymbol{y}}_{ic} = \text{E}(\boldsymbol{y}_i | z_{ic} = 1, \boldsymbol{y}_i^{\text{o}}, \widehat{\boldsymbol{\theta}})$ and $\widehat{\boldsymbol{S}}_{ic} = \text{E}(\boldsymbol{y}_i \boldsymbol{y}_i' | z_{ic} = 1, \boldsymbol{y}_i^{\text{o}}, \widehat{\boldsymbol{\theta}})$, for $i = 1, \ldots, n$ and $c = 1, \ldots, C$.

M-step: Update $\widehat{\boldsymbol{\theta}}$ with the vector maximizing $\mathrm{E}[l_{Y,Z}(\boldsymbol{\theta})|\boldsymbol{Y}^{\mathrm{o}}, \widehat{\boldsymbol{\theta}}]$.

Appendix A2 and A3 contain details on how the *E-* and *M-steps* are implemented under the constraints defining the class of models in Section 3.

Membership probabilities and partitions: In addition to the parameter estimates, we obtain, for all genes, estimated conditional probabilities of belonging to each of the mixture components. The probability that the ith gene belongs to the cth component is estimated by the entry $\widehat{z}_{ic}$ of $\widehat{\boldsymbol{z}}_i$, and provides a measure of the confidence with which the gene may be assigned to a cluster, or to the contamination class. Thought of as membership probabilities, the $\widehat{z}_{ic}$'s express a *soft* or *fuzzy* partition (see Gordon, 1999, sec. 5.1), in which genes may not be unequivocally assigned to one component. However, it is possible to derive a standard (hard) partition on the basis of these probabilities by assigning each gene i to the component $c_i = c(\widehat{\boldsymbol{z}}_i)$, where $c(\widehat{\boldsymbol{z}}_i)$ is the index of the largest entry of the vector $\widehat{\boldsymbol{z}}_i$. Alternatively, we could construct a hard partition with an additional class (the $(C+2)$nd, after the C regular clusters and the contamination class). This class would contain all genes whose attribution is not straightforward; more precisely, we would assign gene i to component $c_i = c(\widehat{\boldsymbol{z}}_i)$ when $m(\widehat{\boldsymbol{z}}_i) \geq \gamma$, and to the additional class otherwise, where $m(\widehat{\boldsymbol{z}}_i)$ is the value of the largest entry of $\widehat{\boldsymbol{z}}_i$, and γ is a threshold level between 0 and 1. Whether or not the additional class is created, we can calculate the percentage of genes confidently attributed to the mixture components; that is the percentage of the n quantities $m(\widehat{\boldsymbol{z}}_i)$ greater or equal to γ. This provides a measure of the overall membership strength characterizing the fit; that is, of the natural "lumpiness" or parceability of the data points in the context of the model under consideration (i.e. relative to the choice of C, $\boldsymbol{X}_c$'s, and $\boldsymbol{\Sigma}_c$'s type).

Missing values: Finally, for each gene, we obtain an estimate of the unobserved expression levels. When there is strong evidence that gene i belongs to component c, the unobserved levels $\boldsymbol{y}_i^{\mathrm{u}}$ may be estimated through $\widehat{\boldsymbol{y}}_{ic}^{\mathrm{u}}$, if $c > 1$, and through the center of the region of the unobserved data, $\widehat{\boldsymbol{y}}_{i0}^{\mathrm{u}}$, otherwise (see Appendix A1 and A2 for details). When the attribution of gene i is not straightforward, $\boldsymbol{y}_i^{\mathrm{u}}$ may be estimated as a linear combination of the vectors $\widehat{\boldsymbol{y}}_{ic}^{\mathrm{u}}$, with coefficients $\widehat{z}_{i0}, \ldots, \widehat{z}_{iC}$.

Initialization and model selection: We conclude this section outlying a sequential strategy that may be used to initialize the EM algorithm in a sensible way, and perform model selection. EM initialization is crucial because the type of log-likelihood we are considering encompasses several local maxima. Let

$\widehat{\boldsymbol{\mu}}_0$ and $\widehat{\boldsymbol{\Sigma}}_0$ be the mean and covariance matrix for the data completed assuming that all genes belong to the contamination, and, on the basis of (3), decompose $\widehat{\boldsymbol{\Sigma}}_0$ in terms of $\widehat{\sigma}_0^2$, $\widehat{\boldsymbol{\nu}}_0^2$ and $\widehat{\boldsymbol{\Omega}}_0$. The strategy we propose consists in fitting the full model for an increasing number of components, starting from the contamination alone ($C = 0$, as above), and then passing to $C = 1, 2$, etc. (it must be noted that the log-likelihood of each full model is unbounded, so that the fit identifies a local maximum – see for example McLachlan and Basford, 1988). When fitting, say, the contamination plus $\tilde{C}$ Gaussian components, the starting value for $\boldsymbol{\pi}$ is set to

$$\frac{1}{\tilde{C}+1}(\ \tilde{C}\widehat{\boldsymbol{\pi}}'\ ,\ 1\ )',$$

where $\widehat{\boldsymbol{\pi}}$ is the estimate obtained at the previous stage (i.e. fitting the contamination plus $\tilde{C}-1$ Gaussian components). Moreover, the starting values for $\boldsymbol{\mu}_{\tilde{C}}$, $\sigma^2_{\tilde{C}}$, $\boldsymbol{\nu}^2_{\tilde{C}}$ and $\boldsymbol{\Omega}_{\tilde{C}}$ are set to, respectively, $\widehat{\boldsymbol{\mu}}_0$, $\widehat{\sigma}_0^2$, $\widehat{\boldsymbol{\nu}}_0^2$ and $\widehat{\boldsymbol{\Omega}}_0$, whereas the starting values for the corresponding parameters of the other components are simply set to the estimates obtained at the previous stage.

As one proceeds increasing the number of components, the *Bayesian Information Criterion* (BIC) can be calculated for the corresponding full models. The BIC of a model is defined as

$$\text{BIC} = 2l_{Y^\circ}(\widehat{\boldsymbol{\theta}}) - r\log(n)$$

where r is the overall number of parameters (Schwarz, 1978). Additional components increase r, and usually increase (or leave unchanged) the achieved maximum log-likelihood value. In most cases, this produces a non-monotone behavior of the BIC (first increasing and then decreasing as components are added) which permits a meaningful selection of the component number C.

In general, BIC allows for direct comparison between any two models and is advantageous with respect to the likelihood ratio criterion which, in the mixture context, requires numerical estimation of the null distribution of the deviance via bootstrap methods (see for instance McLachlan, 1987).

After identifying the C whose full model provides maximal BIC, one can use mean and covariance estimates obtained from the full model and/or substantive hypotheses to formulate restricted models as described in Section 3. When fitting these models, parameter estimates from the full model are used for initialization.

To compare restricted models to one another, and to the full model, it may be reasonable to take into account other features of their fit, besides BIC. For example, one could consider the overall membership

strength, within-cluster dispersion, and the estimated weight of the contamination term. These are, in a sense, geometric features (how crowded are the spaces "between" clusters, how "tight" are the clusters, how crowded is the "background" to the cluster structure) whose relative importance with respect to BIC in selecting a final model will to a large extent depend on biological knowledge and the specific aims of the clustering exercise.

5 An application: the response of yeast to heat shocks

Gasch *et al.* (2001) performed a large microarray study to investigate the time-course of expression in yeast cells, as they respond to various types of environment changes. The original data is comprised of expression levels for the entire identified yeast genome ($n = 6152$ known and putative genes) on a total of over 140 conditions. Most of these conditions are a collection of time-courses, each spanning the process of adaptation of yeast to a different environmental stress. The study aimed at identifying characteristic expression patterns during adaptation, and their corresponding gene clusters. Using hierarchical clustering techniques, the authors discuss universal responses to stress, as well as responses that are specific to different stressors.

Here, we concentrate on data related to a heat shock experiment, in which the temperature is raised suddenly from 25 to 37 degrees Celsius. This data is comprised of expression levels for the 6152 genes on a $T = 8$ point time-course corresponding to 5, 10, 15, 20, 30, 40, 60 and 80 minutes after the heat shock. The data are log-ratios obtained from two-color arrays. The reference sample was constructed by pooling equal amounts of all experimental samples. The profiles of 2509 genes (40.78% of the total) have missing values.

We begin our analysis by fitting full models with increasing number of components, in order to select a reasonable value for C. Here, maximal BIC is used as a selection criteria. Figure 1 plots the BIC against component number C for the full models, as well as for some restricted models. From the figure, we see that beyond the sharp jump from $C = 1$ to 2, the BIC has a broad peak around the maximum at $C = 7$. For the full model with $C = 7$, the first row of Table 1 shows its maximum likelihood $l(\widehat{\boldsymbol{\theta}})$, the BIC, and the estimated weight of the contamination term, $\widehat{\pi}_0$. Henceforth, we will focus on contaminated mixtures with seven Gaussian components, and by posing constraints try to identify more

parsimonious models.

Table 1: model evaluation ($C = 7$)

	type	$l(\widehat{\boldsymbol{\theta}})$	BIC	$\widehat{\pi}_0$
	full model	-21787	-46322	0.0010
I	free, EEE	-24404	-49201	0.1445
II	free, EFF(sph)	-34643	-69599	0.0166
III	modeled, EEE	-25561	-51681	0.0166
IV	free, UUE	-22356	-45994	0.0003

We first consider models in which the covariance structure of each Gaussian component is the same. This assumes that all clusters have a common size, shape and orientation, and differ only by their location in the $T = 8$ dimensional space of the data. The most general such model allows the seven cluster locations and the common cluster size, shape, and orientation to be completely unconstrained; this is the *free means, EEE covariances* model (Model I). Figure 1 and Table 1 show that the BIC of Model I is quite low with respect to the full model with seven components. Nonetheless, this model has several attractive aspects. It presents good overall membership strength, "tight" clusters, and estimated mean patterns that have straightforward interpretations.

The left panel of Fig. 2 shows the histogram of maximal membership probabilities (the $m(\hat{\boldsymbol{z}}_i)$'s) for Model I. This histogram displays a remarkable concentration at high values, indicating good membership strength, and suggests that the soft partition in output from the fit is amenable to being turned into a hard partition. The left panel of Fig. 3 shows the estimated mean patterns for Model I. Before turning to their description, consider the legend numbers at the right of the panel. They tabulate, for each cluster, the percentage of data points attributed to it that have a maximal membership probability equal to or larger than $\gamma = 0.8$ (the first number), and those with a maximal membership probability less than 0.8 (the second number). For most clusters, the strongly attributed data points constitute the vast majority. The left panel of Fig. 4 shows the two up-regulated clusters obtained from the fit (corresponding to the red patterns in the left panel of Fig. 3). These are representative of the "tight" clusters (relatively small within-cluster dispersion) produced by this fit.

Now consider the estimated mean patterns of Model I (the left panel of Fig. 3). As observed by

Gasch *et al.*, the transcription response to heat is transient; all patterns, except possibly the solid blue one, converge to a stable expression level by about 60 minutes after the shock, regardless of whether they presented an up- (e.g. red patterns) or down-regulated (e.g. green patterns) response. Evidence is also seen in favor of their hypothesis that very similar but "reciprocal" patterns manifest themselves in up and down regulation (solid-red and solid-green patterns, respectively). There are also other, different patterns of up and down-regulation. The pair of red patterns (both up-regulated) and the pair of green patterns (both down-regulated) show relative temporal shifts in their transcription responses. For example, while both red patterns ascend simultaneously, the dash-red pattern shows a more sustained elevation, and a delayed decrease to its stable expression level. For the green patterns, the dash-green pattern shows a rather oscillatory route to a downward peak that is delayed, and considerably smaller in size, than that of the solid-green pattern. A delayed down-regulated pattern was also identified by Gasch *et al.* Finally, we find non-concave behaviors in the down-regulated patterns; the small upward bump in the dash-green pattern may or may not be relevant to its interpretation, but both the solid and the dash-blue patterns show distinct "oscillating" behaviors below the zero-line.

In this fit, the red, green and blue patterns all correspond to relatively small clusters (see percentages in the figure legend), and confirm the intuition that only a small fraction of the genome (here about 10.26%) is involved in the response to the heat shock. The solid black pattern, which shows practically no variation along the time course, corresponds to a cluster containing about 88.36% of the genes. For all practical purposes, this cluster is performing a filtering, accumulating all those profiles that do not vary over time except with the type of fluctuations allowed by the estimated within-cluster covariance matrix. A modest remainder (less than 2%) of the genes is attributed to the contamination class – the estimated weight of the contamination term is 0.1445.

Remaining within the context of a common component covariance structure, we attempted further reductions to see if a properly devised contraction in parameters might lead to an increase in the BIC. We fitted a model with free means and common spherical covariance for the seven components, which further assumes that the within-cluster variability is isotropic and uncorrelated along the time course; this is the *free means, EFF covariances (spherical)* model, i.e., the *equal volume spherical model* discussed in Section 3. Its BIC is so low (see Table 1 and Fig. 1), indeed much lower that of the full model with one component, that the hypothesis of equal size spherical clusters is readily rejected. The estimated

mean patterns produced by this fit are very similar to the centroids obtained by applying a standard k-means algorithm to the data (not shown). They show parabolic-like behaviors before the stabilization at $t = 60$, consistent with a simplistic hypothesis in which, during the transient response to heat, expression patterns present different but constant (over time) degrees of convexity (if up-regulated), or concavity (if down-regulated). Thus, we fitted a model with unconstrained common covariance, but modeled means following parabolae up to $t = 60$, and with the last value ($t = 80$) equal to the one in $t = 60$; this is the *modeled means, EEE covariances* model. Its BIC is better than that of the free means, EFF covariances model, but still much lower than that of the free means, EEE model and that of the full model in seven components. Also the hypothesis of "constant convexity (concavity)" during the response to heat can be readily rejected.

Although the free means, EEE covariances model is appealing in many respects, its BIC might be considered unsatisfactory. We attempt to improve the BIC by considering models that allow some differentiation among the component covariances (i.e. in their size, shape and orientation). The increased number of parameters might be overcome by gains in maximum likelihood. Free means and UEE covariances, as well as the further restricted free means and UEF covariances with $\bar{\mathbf{\Omega}}$ following an AR(1) structure (see Section 3) both provided unsatisfactory values of BIC.

Finally, we find that the model with *free means and UUE covariances* (Model IV) produces a larger BIC than that of the full model. This model postulates that each component has a common (but not fixed) correlation structure, while allowing for differences in overall variability volume (the σ_c^2's) and distribution over the time course (the $\boldsymbol{\nu}_c$'s).

If compared in terms of their BIC values, Model IV is clearly to be preferred to Model I (the free means, EEE covariances one). Model IV also presents a much different picture of the data in terms of its overall membership strength, within-cluster dispersion, and estimated mean patterns. As seen in the right panel of Fig. 2, the histogram of the maximal membership probabilities is *not* concentrated on high values; the soft partition in output from this fit is not easily converted into a hard one, as a very large number of gene expression profiles actually fall "between" clusters. This impression is confirmed by the percentages in the legend of the right panel of Fig. 3. For most clusters, weakly attributed data points are in numbers comparable to, or exceeding, those of strongly attributed data points. The right panel of Fig. 4 shows the two up-regulated clusters obtained from the fit (corresponding to the red patterns

in the right panel of Fig. 3). Allowing for different component covariances results in different amounts of within-cluster dispersion: some of the clusters are not "tight" in the traditional sense, and thus can contain very different profiles, seemingly inconsistent with one another.

The estimated mean patterns of Model IV (right panel of Fig. 3) have some similarities and some noticeable differences with those of Model I (the left panel). The transcription response to heat for Model IV still appears to be transient, and we still find an up-regulated and a down-regulated pattern that "mirror" one another (solid red and solid green). However, the strongly up-regulated, red-dash pattern of Model I seems to be replaced by a weakly up-regulated one in Model IV (dash-red in right panel). Also, the down-regulated patterns (solid and dash-green) of Model IV are generally weaker, nor do they show the strong relative temporal shift in time to peaking (compare the dash-green patterns in the left and right panels of Fig. 3). Finally, most of the evidence for non-concave down-regulated patterns in Model I has disappeared for Model IV: the oscillating blue patterns in the left panel are replaced by the gray ones in the right panel, which fluctuate about the zero line, and less markedly. In fact, the solid and dash gray patterns may be performing a filtering here, together with the almost-constant solid black pattern. The black pattern now accumulates only about 7.22% of the genes, while the two gray patterns combine to an additional 65.58%. The latter may represent small-scale expression variations, possibly not related to the response to heat. Indeed, such a scenario is not unlikely in microarray data analysis, as it is extremely difficult to collect experimental samples in such a way that, at a whole-genome scale, no expression variation occurs except for that related to the phenomenon under study; there is always the possibility that systematic but unrelated patterns will manifest themselves in the data, though hopefully on a scale smaller than that of the patterns of interest. Note, too, that the contamination class in Model IV absorbs a very small percentage of genes, less than 0.02% (the estimated weight of the contamination term is a mere 0.0003).

The free means, EEE covariances model, and the free means, UUE covariances model provide very different descriptions of the data. Which may be more appropriate rests essentially on one's interpretation of overall membership strength and within-cluster dispersion. Through the first model – free means, EEE covariances – the data can be well parsed into "tight" clusters with a common size, shape and orientation, though one obtains a poor value of the BIC. In seeking to do better on this criterion, one is led to the second model – free means, UUE covariances – with a very fuzzy partition, and clusters

whose variability is characterized by different volumes and distributions over the time course. Either might serve as a reasonable interpretation of the biological data. In terms of characterizing the mean patterns, these two fits also lead to some common, and also some different behaviors, as discussed above. Finally, while the first fit attributed a very modest percentage of the data to the contamination term, the second is remarkable in that it makes essentially no use of it: almost all genes can be attributed, albeit weakly, to one of the multivariate normal components.

6 Discussion

The clustering approach described in this paper provides advantages that are crucially important for the analysis of microarray data. While making a clear case for this claim, the application we presented in Section 5 barely scratches the surface in illustrating the versatility and power of this approach. For a related study, we refer the reader to Yeung *et al.* (2001), in which several actual microarray data sets and some artificial data are analyzed with a similar methodology.

Using a contaminated multivariate normal mixture, gene clustering can be performed in a compact, highly informative and flexible fashion, within the unambiguous inferential setting provided by likelihood methods. Since it replaces traditional cluster memberships with membership probabilities, the approach is suitable for the analysis of data that do not present clear-cut "natural" clusters. Moreover, the approach allows us to model cluster means and covariance structures to various degrees of specificity, handle anomalous observations and impute missing values. Last, but certainly not least, it allows us to employ statistically motivated criteria, such as BIC, to aid in the selection of number of clusters and constraint formulations for means and covariances.

The scheme we presented in Section 3 is amenable to several extensions, which can further expand the scope of this approach. One very relevant and straightforward extension would formalize relationships among mean patterns through the sharing of some elements in their linear parameterizations – for instance, if two components, say c_1 and c_2, employ the same design matrix expressing a linear trend ($\boldsymbol{X}_{c_1} = \boldsymbol{X}_{c_2} = (\mathbf{1}_T \ \boldsymbol{\tau})$) it may be interesting to impose a common slope $\beta_{c_1 2} = \beta_{c_2 2} = \tilde{\beta}$. Also, a Bayesian extension of the scheme, introducing prior distributions on the parameters, would provide an additional way to exploit prior knowledge and hypotheses.

As with many sophisticated partitioning techniques, computational and initialization issues are crucial here. With our EM implementation, the computational burden is relatively mild, even in high dimension and with thousands of observations. The computation time depends linearly on number of components and number of observations, and is also affected by the patterns of missing values in the data (loosely speaking, expression profiles missing the same entries decrease the computation time, while expression profiles missing different entries increase it). Each of the fits presented in Section 5 required a time between 10 and 20 minutes using Matlab functions (available from the authors). Concerning initialization, we describe an iterative strategy in Section 4.

Finally, mixtures of multivariate Gaussian distributions may not be appropriate for some data. Yeung *et al.* (2001), describe diagnostic procedures to assess whether data are compatible with the structure postulated in Eq. (1), and discuss how to use data transformations to approximate such a structure. It should be noted, too, that taking logarithms of the ratios between experimental and reference readings, which is standard practice in many microarray studies, often suffices to mitigate strong departures from Eq. (1). This was the case, for example, in the application we presented here.

Appendix: technical details

A1: Center and volume of data range

For $i = 1, \ldots, n$ and $t = 1, \ldots, T$, let $\mathcal{L}_t$ and $\mathcal{O}_i$ be two sets of indices such that, if the entry y_{it} of $\boldsymbol{y}_i$ has been observed, then $i \in \mathcal{L}_t$ and $t \in \mathcal{O}_i$. So, letting $\bar{y}_t = \sum_{i \in \mathcal{L}_t} y_{it}/l_t$, where l_t is the cardinality of $\mathcal{L}_t$, $y_t^- = \min_{i \in \mathcal{L}_t} y_{it}$ and $y_t^+ = \max_{i \in \mathcal{L}_t} y_{it}$, the entries of the vector $\widehat{\boldsymbol{y}}_{i0}^{\mathrm{u}}$ are $\bar{y}_t$, for $t \in \mathcal{O}_i$, whereas $V(\boldsymbol{y}_i^{\mathrm{o}})$ is given by $\prod_{t \in \mathcal{O}_i}(y_t^+ - y_t^-)$.

A2: E-step, estimation of membership probabilities and of missing data

$\widehat{\boldsymbol{z}}_i$ is given by

$$\widehat{\boldsymbol{z}}_i = [\widehat{\boldsymbol{\pi}}' \boldsymbol{f}(\boldsymbol{y}_i^{\mathrm{o}}; \widehat{\boldsymbol{\eta}})]^{-1} \mathrm{diag}(\widehat{\boldsymbol{\pi}}) \boldsymbol{f}(\boldsymbol{y}_i^{\mathrm{o}}; \widehat{\boldsymbol{\eta}}); \tag{6}$$

moreover, using an obvious notation, $\widehat{\boldsymbol{y}}_{ic}$ is given by the two subvectors

$$\widehat{\boldsymbol{y}}_{ic}^{\mathrm{o}} = \boldsymbol{y}_i^{\mathrm{o}} \quad \text{and} \quad \widehat{\boldsymbol{y}}_{ic}^{\mathrm{u}} = \mathrm{E}(\boldsymbol{y}_i^{\mathrm{u}} | z_{ic} = 1, \boldsymbol{y}_i^{\mathrm{o}}, \widehat{\boldsymbol{\theta}}) = \widehat{\boldsymbol{\mu}}_c^{\mathrm{u}} + \widehat{\boldsymbol{\Sigma}}_{ic}^{\mathrm{uo}} (\widehat{\boldsymbol{\Sigma}}_{ic}^{\mathrm{oo}})^{-1} (\boldsymbol{y}_i^{\mathrm{o}} - \widehat{\boldsymbol{\mu}}_{ic}^{\mathrm{o}}) \tag{7}$$

and $\widehat{\boldsymbol{S}}_{ic}$ by the following submatrices

$$\widehat{\boldsymbol{S}}_{ic}^{\mathrm{oo}} = \boldsymbol{y}_i^{\mathrm{o}}(\boldsymbol{y}_i^{\mathrm{o}})', \quad \widehat{\boldsymbol{S}}_{ic}^{\mathrm{ou}} = \boldsymbol{y}_i^{\mathrm{o}}(\widehat{\boldsymbol{y}}_{ic}^{\mathrm{u}})', \quad \widehat{\boldsymbol{S}}_{ic}^{\mathrm{uo}} = \widehat{\boldsymbol{y}}_{ic}^{\mathrm{u}}(\boldsymbol{y}_i^{\mathrm{o}})', \quad \widehat{\boldsymbol{S}}_{ic}^{\mathrm{uu}} = \widehat{\boldsymbol{\Sigma}}_{ic}^{\mathrm{uu}} - \widehat{\boldsymbol{\Sigma}}_{ic}^{\mathrm{uo}}(\widehat{\boldsymbol{\Sigma}}_{ic}^{\mathrm{oo}})^{-1}\widehat{\boldsymbol{\Sigma}}_{ic}^{\mathrm{ou}} + \widehat{\boldsymbol{y}}_{ic}^{\mathrm{u}}(\widehat{\boldsymbol{y}}_{ic}^{\mathrm{u}})'.$$

A3: M-step

Let $\widehat{z}_{\cdot c} = \sum_{i=1}^n \widehat{z}_{ic}$, $\boldsymbol{m}_c = \sum_{i=1}^n (\widehat{z}_{ic}/\widehat{z}_{\cdot c})\widehat{\boldsymbol{y}}_{ic}$ and $\boldsymbol{T}_c = \sum_{i=1}^n (\widehat{z}_{ic}/\widehat{z}_{\cdot c})\widehat{\boldsymbol{S}}_{ic}$. Then, set $\widehat{\boldsymbol{\pi}} = \sum_{i=1}^n \widehat{\boldsymbol{z}}_{ic}/n$ and, for the full model (unconstrained modes and variances), perform the following steps, for $c = 1, \ldots, C$:

(i) set $\widehat{\boldsymbol{\mu}}_c = \boldsymbol{m}_c$ and $\boldsymbol{V}_c = \boldsymbol{T}_c - \widehat{\boldsymbol{\mu}}_c\widehat{\boldsymbol{\mu}}_c'$;

(ii) set $\widehat{\boldsymbol{\Omega}}_c = \mathrm{diag}(\boldsymbol{v}_c)^{-1}\boldsymbol{V}_c\mathrm{diag}(\boldsymbol{v}_c)^{-1}$, where $\boldsymbol{v}_c = \mathrm{diag}(\boldsymbol{V}_c)^{1/2}$;

(iii) set $\widehat{\boldsymbol{\nu}}_c = \boldsymbol{v}_c\sqrt{T/\boldsymbol{v}_c'\boldsymbol{v}_c}$;

(iv) set $\widehat{\sigma}_c^2 = \mathrm{tr}[\mathrm{diag}(\widehat{\boldsymbol{\nu}}_c)^{-1}\widehat{\boldsymbol{\Omega}}_c^{-1}\mathrm{diag}(\widehat{\boldsymbol{\nu}}_c)^{-1}\boldsymbol{V}_c]/T$.

When there is a linear constraint on the modes of some components, perform, for any of these components, the following step instead of step (i):

(ia) set $\widehat{\boldsymbol{\beta}}_c = (\boldsymbol{X}_c'\widehat{\boldsymbol{\Sigma}}_c^{-1}\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c'\widehat{\boldsymbol{\Sigma}}_c^{-1}\boldsymbol{m}_c$, $\widehat{\boldsymbol{\mu}}_c = \boldsymbol{X}_c\widehat{\boldsymbol{\beta}}_c$ and $\boldsymbol{V}_c = \boldsymbol{T}_c - \widehat{\boldsymbol{\mu}}_c\boldsymbol{m}_c' - \boldsymbol{m}_c\widehat{\boldsymbol{\mu}}_c' + \widehat{\boldsymbol{\mu}}_c\widehat{\boldsymbol{\mu}}_c'$.

Moreover, if the model for the $\boldsymbol{\Omega}_c$'s is E, perform the following step, instead of step (ii):

(iia) set $\widehat{\boldsymbol{\Omega}} = \mathrm{diag}(\boldsymbol{w})^{-1}\boldsymbol{W}\mathrm{diag}(\boldsymbol{w})^{-1}$, with $\boldsymbol{W} = \sum_{c=1}^C (\widehat{z}_{\cdot c}/\widehat{z}_{\cdot\cdot})\mathrm{diag}(\widehat{\boldsymbol{\nu}}_c)^{-1}\boldsymbol{V}_c\mathrm{diag}(\widehat{\boldsymbol{\nu}}_c)^{-1}$, $\widehat{z}_{\cdot\cdot} = \sum_{c=1}^C \widehat{z}_{\cdot c}$ and $\boldsymbol{w} = \mathrm{diag}(\boldsymbol{W})^{1/2}$, and then set $\widehat{\boldsymbol{\Omega}}_c = \widehat{\boldsymbol{\Omega}}$, $c = 1, \ldots, C$;

if the model for the $\boldsymbol{\Omega}_c$'s is F, jump step (ii). Similarly, when the model the $\boldsymbol{\nu}_c$'s is either U or E and the model for the $\boldsymbol{\Omega}_c$'s is either E or F, perform the following step, instead of step (iii):

(iiia) if the model for the $\boldsymbol{\nu}_c$'s is U, then find, for $c = 1, \ldots, C$, the vector $\boldsymbol{r}_c$ with positive entries such that $\boldsymbol{r}_c = \boldsymbol{A}_c\mathrm{diag}(\boldsymbol{r}_c)^{-1}/\widehat{\sigma}_c^2$, where $\boldsymbol{A}_c$ is obtained by multiplying element-by-element $\boldsymbol{V}_c$ and $\widehat{\boldsymbol{\Omega}}_c^{-1}$, and set $\widehat{\boldsymbol{\nu}}_c = \boldsymbol{r}_c\sqrt{T/\boldsymbol{r}_c'\boldsymbol{r}_c}$; if the model for the $\boldsymbol{\nu}_c$'s is E, then find the vector $\boldsymbol{r}$ with positive entries such that $\boldsymbol{r} = [\sum_{c=1}^C (\widehat{z}_{\cdot c}/\widehat{z}_{\cdot\cdot})\boldsymbol{A}_c]\boldsymbol{r}$ and set $\widehat{\boldsymbol{\nu}}_c = \widehat{\boldsymbol{\nu}}$, where $\widehat{\boldsymbol{\nu}} = \boldsymbol{r}\sqrt{T/\boldsymbol{r}'\boldsymbol{r}}$;

regardless of the model for the $\boldsymbol{\Omega}_c$'s, skip step (iii) if the model the $\boldsymbol{\nu}_c$'s is F. Finally, if the model for the $\widehat{\sigma}_c^2$'s is E, use the matrix $\boldsymbol{V} = \sum_{c=1}^C (\widehat{z}_{\cdot c}/\widehat{z}_{\cdot\cdot})\boldsymbol{V}_c$ instead of $\boldsymbol{V}_c$ at step (iv); if the model for the $\widehat{\sigma}_c^2$'s is F, skip step (iv).

References

Banfield J.D., Raftery A.E. (1993), Model-Based Gaussian and Non-Gaussian Clustering, *Biometrics*, **49**, pp. 803–821.

Brown M.P.S., Grundy W.N., Lin D., Cristianini N., Sugnet C.W., Furey T.S., Ares M., Haussler D. (2000), Knowledge-Based Analysis of Microarray Gene Expression Data by Using Support Vector Machines, *Proceedings of the National Academy of Sciences*, **97**, pp. 262–267.

Chu S., DeRisi J., Eisen M.B., Mulholland J., Botstein D., Brown P.O., Herskowitz I. (1998), The Transcriptional Program of Sporulation in Budding Yeast, *Science*, **282**, pp. 699–705.

Dasgupta A., Raftery A.E. (1998), Detecting Features in Spatial Point Processes with Clutter via Model-Based Clustering, *Journal of the American Statistical Association*, **93**, pp. 294–302.

Dempster A.P., Laird N. M., Rubin, D. B. (1977), Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society, Series B*, **39**, pp. 1-22.

Dudoit S., Yang Y.H., Callow M.J., Speed T.P. (2000), Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments. Technical Report, Department of Statistics, UC Berkeley.

Eisen M.B., Spellman P.T., Brown P.O., Botstein D. (1998), Cluster Analysis and Display of Genome-Wide Expression Patterns, *Proceedings of the National Academy of Sciences, USA*, **95**, pp. 14863–14868.

Eisen M.B., Brown P.O. (1999), DNA Arrays for Analysis of Gene Expression, *Methods Enzymology*, **303**, pp. 179–205.

Fraley C., Raftery A.E. (1998), How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis, *The Computer Journal*, **41**, pp. 570–588.

Gasch A.P., Spellman P.T., Kao C.M., Carmel-Harel O., Eisen M.B., Storz G., Botstein D., Brown P.O. (2001), Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes, *Molecular Biology of the Cell*, **11**, pp. 4241-4257.

Gordon A.D. (1999), *Classification*, Chapman & Hall, London.

Heyer L.J., Kruglyak S., Yooseph S. (1999), Exploring Expression Data: Identification and Analysis of Co-expressed Genes, *Genome Research*, **9**, pp. 1106–1115.

Hastie T., Tibshirani R., Eisen M.B., Brown P.O. Ross D., Scherf U., Weinstein J., Alizadeh A., Staudt L., Botstein D. (2000), Gene Shaving: a New Class of Clustering Methods for Expression Arrays, Technical Report, Department of Statistics, Stanford.

Hastie T., Alter O., Sherlock G., Eisen M.B., Tibshirani R., Botstein D., Brown P.O. (2000), Imputation of Missing Values in DNA Microarrays, Technical Report, Department of Statistics, Stanford.

Holter N.S., Mitra M., Maritan A., Cieplak M., Banavar J.R., Fedoroff N.V. (2000), Fundamental Patterns Underlying Gene Expression Profiles: Simplicity from Complexity, *Proceedings of the National Academy of Sciences*, **97**, pp. 8409–8414.

Iyer V.R., Eisen M.B., Ross D.T., Schuler G., Moore T., Lee J.F.C., Trent J.M., Staudt L.M., Hudson J., Boguski M.S., Lashkari D., Shalon D., Botstein D., Brown P.O. (1999), The Transcriptional Program in the Response of Human Fibroblast to Serum, *Science*, **283**, pp. 83–87.

Kerr M.K., Churchill G.A. (2000), Bootstrapping Cluster Analysis: Assessing the Reliability of Conclusions from Microarray Experiments – submitted.

Lee M.L.T., Kuo F.C., Whitmore G.A., Sklar J. (2000), Importance of Replication in Microarray Gene Expression Studies: Statistical Methods and Evidence from Repetitive cDNA Hybridizations, *Proceedings of the National Academy of Sciences, USA*, **97**, pp. 9834–9839.

Lindsay B.G. (1995), *Mixture models: theory, geometry and applications*, NFS-CBMS Regional Conference Series in Probability and Statistics.

Mardia K.V., Kent J.T., Bibby J. M. (1979), *Multivariate Analysis*, Academic Press, London.

McLachlan G.J., Basford K.E. (1988), *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York.

McLachlan G.J., Peel D. (2000), *Finite Mixture Models*, Wiley, New York.

Peng F., Jacobs R.A., Tanner M.A. (1996), Bayesian Inferenc in Mixture-of-Experts and Hierarchical Mixture-of-Experts Models with an Application to Speech Recognition, *Journal of the American Statistical Association*, **91**, pp. 953–960.

Tamayo P., Slonim D., Mesirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E.S., Golub T.R. (1999), Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and an Application to Hematopoietic Differentiation, *Proceedings of the National Academy of Sciences, USA*, **96**, pp. 2907–2912.

Tavazoie S., Hughes J.D., Campbell M.J., Cho R.J., Church G.M. (1999), Systematic Determination of Genetic Network Architecture, *Nature Genetics*, **22**, pp. 281–285

Spellman P.T., Sherlock G., Zhang M.Q., Vishwanath R.I., Anders K., Eisen M.B., Brown P.O., Botstein D. (1998), Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces Cerevisiae* by Microarray Hybridization, *Molecular Biology of the Cell*, **9**, pp. 3273–3297.

Yang Y.H., Dudoit S., Luu P., Speed T.P. (2000), Normalization for cDNA Microarray Data, Technical Report, Department of Statistics, UC Berkeley.

Yeung K.Y., Fraley C., Murua A., Raftery A.E. and Ruzzo W. L. (2001), Model-Based Clustering and Data Transformation for Gene Expression Data, Technical Report 396, Department of Statistics, University of Washington, Seattle, *Bioinformatics* – to appear.

Van der Laan M.J., Bryan J.F. (2000), Gene Expression Analysis with the Parametric Bootstrap, Technical Report 86, Biostatistics Group, University of California.

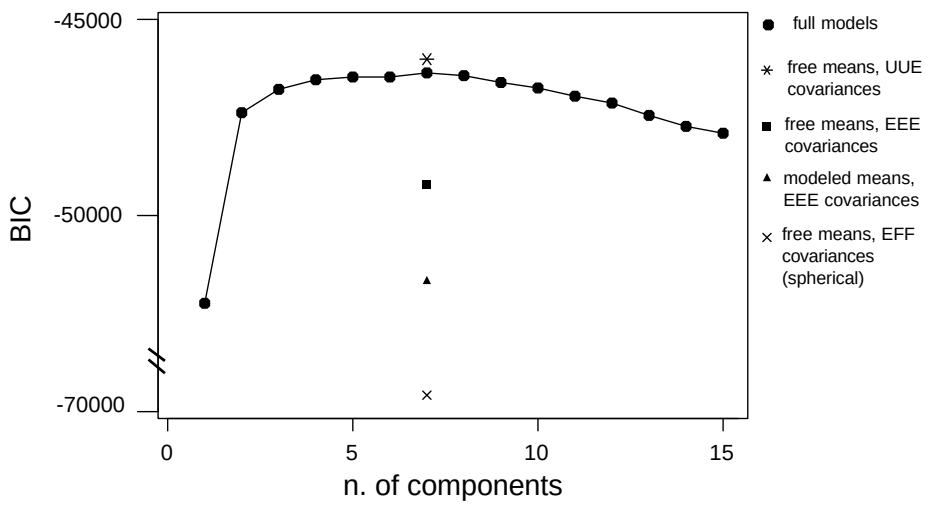

Figure 1: The BIC for model selection in the yeast heat shock data.

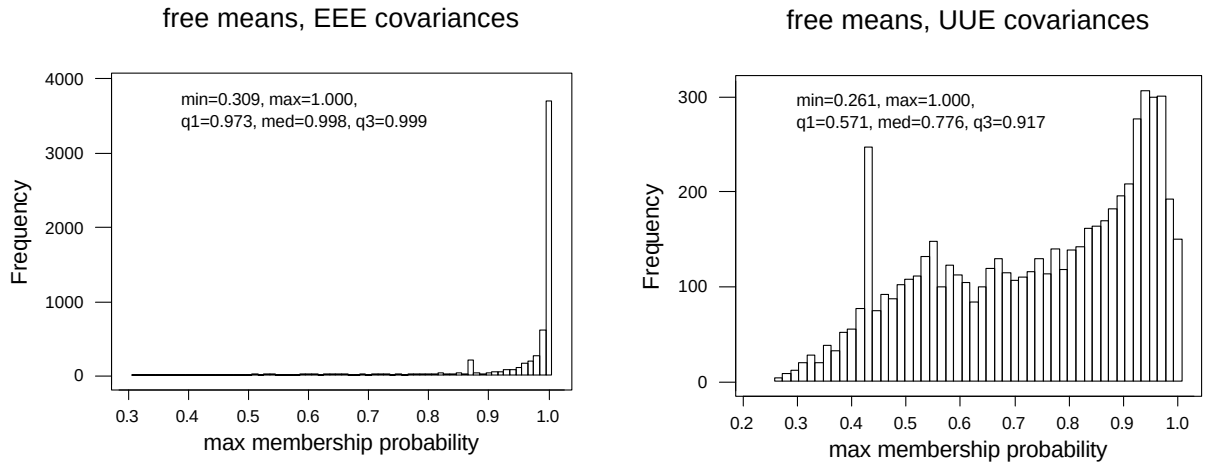

Figure 2: Histograms of maximal membership probabilities from the free means, EEE covariances fit (left) and the free means UUE fit (right), with some summary statistics.

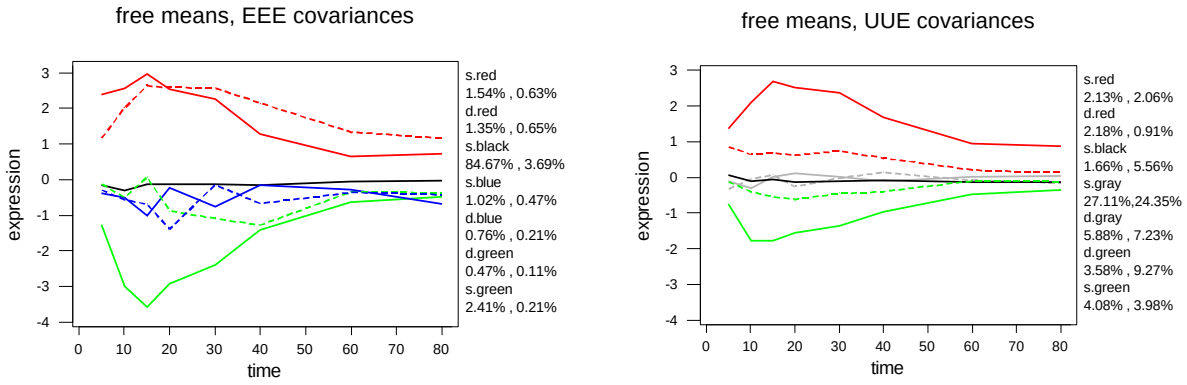

Figure 3: estimated mean patterns for the free means, EEE covariances model (left) and the free means, UUE covariances model (right). The legends contain the percentages of genes attributed to the clusters corresponding to each pattern. The first number represents the percentage of strongly attributed genes (maximal membership probability ≥ 0.8), while the second number represents the percentage of weakly attributed genes (< 0.8).

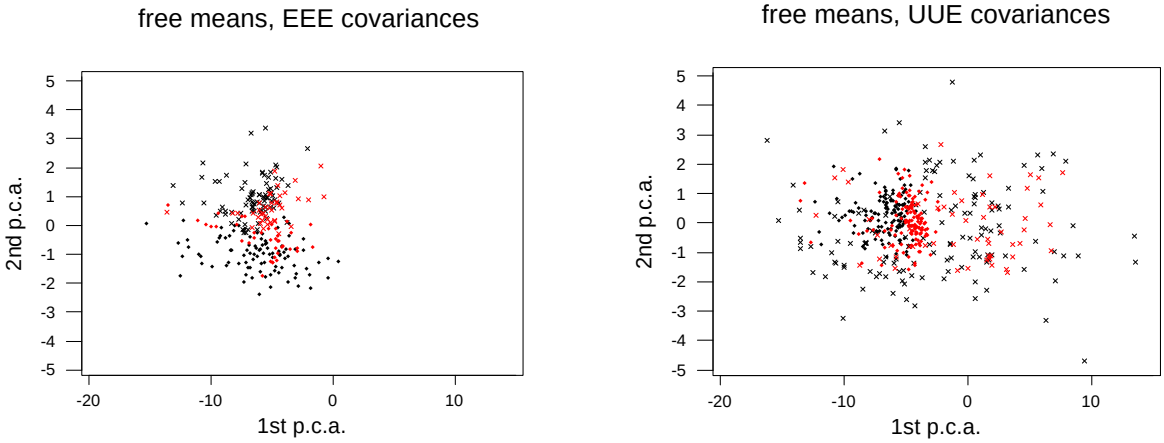

Figure 4: Two up-regulated clusters from the free means, EEE covariances fit (left) and the free means UUE fit (right), shown as data points projected on the first principal component plane. Dots are members of the cluster corresponding to the solid red pattern, and crosses members of the cluster corresponding to the dash red pattern (see Fig. 3). Here, black symbols represent strongly attributed genes (maximal membership probability ≥ 0.8), and red symbols weakly attributed ones (< 0.8).