Cluster Analysis Using Multivariate Normal Mixture Models to Detect Differential Gene Expression With Microarray Data

Yi He¹, Wei Pan¹* and Jizhen Lin²

¹Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN
²Department of Otolaryngology, School of Medicine, University of Minnesota, Minneapolis, MN

*Corresponding author

Email addresses:
  YH: yihe@biostat.umn.edu
  WP: weip@biostat.umn.edu
  JL: linxx004@umn.edu

October 2005; Revised December 2005 and February 2006
Abstract

DNA microarrays make it possible to study simultaneously the expression of thousands of genes in a biological sample. Univariate clustering techniques have been used to discover target genes with differential expression between two experimental conditions. Because of possible loss of information due to use of univariate summary statistics, it may be more effective to use multivariate statistics. We present multivariate normal mixture model based clustering analyses to detect differential gene expression between two conditions.

Deviating from the general mixture model and model-based clustering, we propose mixture models with specific mean and covariance structures that account for special features of two-condition microarray experiments. Explicit updating formulas in the EM algorithm for three such models are derived. The methods are applied to a real dataset to compare the expression levels of 1,176 genes of rats with and without pneumococcal middle-ear infection to illustrate the performance and usefulness of this approach. About 10 genes and 20 genes are found to be differentially expressed in a 6-dimensional modeling and a bivariate modeling, respectively. Two simulation studies are conducted to compare the performance of univariate and multivariate methods. Depending on data, neither method can always dominate the other. The results suggest that multivariate normal mixture models can be useful alternatives to univariate methods to detect differential gene expression in exploratory data analysis.
1 Introduction

DNA microarrays make it possible to measure rapidly and efficiently the expression levels of all genes in a biological sample, providing a means to detect the pattern of gene expression in a cell, and thus enabling one to learn important information about the cell state (Eisen et al. 1998). However, the process of transforming microarray data into meaningful biological insights is impeded by the complexity of the data. One of the most commonly used approaches to analyzing microarray data is to apply statistical clustering methods.

In this paper we consider the use of model-based clustering in the context of detecting differentially expressed genes, which is to identify genes with altered expression levels between two experimental conditions; although the methods can be extended easily to more than two conditions (Broet et al 2004), we focus on the case with two conditions. Typically several replicated experiments are carried out under each condition. To detect differential gene expression, one derives some summary statistic for each gene, which is usually univariate for its easy applicability. Two general classes of approaches are thus taken. The first is based on formal hypothesis testing: the null distribution of the statistic is used to calculate p-values or q-values after controlling for the family-wise type I error rate or false discovery rate. However, any result critically depends on the validity of the null distribution being used, which is derived based on some approximations or modeling assumptions. As an alternative, the second is less formal and is exploratory in nature. Clustering analysis of univariate summary statistics is such an approach. Pan et al. (2002) proposed a $t$ statistic-based normal mixture model, and Broet et al. (2002) proposed a Bayesian mixture model based on the difference of average gene expression levels between the two conditions. This latter method will be called $d$ statistic-based approach in this paper; see also McLachlan et al (2005) for a use of the $d$ statistic. As an improvement to the $t$ statistic for small samples, SAM-t statistic (Tusher et al. 2001) can be used. All the three statistics are univariate, which however may lead to loss of information. Furthermore, in contrast to formal hypothesis testing, clustering multivariate statistics is feasible. Hence, as an alternative to exclusive use of univariate statistics in the current practice, we propose and evaluate the use of some multivariate statistics in model-based clustering. As to be shown, it is found that clustering multivariate statistics can be advantageous to univariate
statistics in some situations. It is emphasized that here we consider exploratory analysis using model-based clustering, in contrast to formal hypothesis testing (e.g. Allison et al 2002; Broet et al 2004).

Although we focus on model-based clustering, other clustering methods, such as hierarchical clustering (Sokal and Michener 1958) and K-means (Lloyd 1957; MacQueen 1967), can be also applied here, as is the case with general clustering analysis for expression pattern discovery. However, these methods have their own limitations, one of which is the difficulty in determining the number of clusters. The other is that there is no clear definition of what a cluster is in the first place. In contrast, model-based clustering has some advantages in these aspects (McLachlan and Basford 1988; McLachlan et al. 2002). Using model-based clustering, Yeung et al. (2001) and Ghosh et al. (2002) considered clustering genes for gene expression pattern discovery, while McLachlan et al. (2002) clustered tissue samples. Note that our purpose here is different from above general clustering analysis for pattern discovery: we restrict our attention to detecting differential gene expression between two conditions, which requires accounting for special features of two-condition microarray experiments; specifically, there are at most two unequal mean expression levels for each gene across experiments in two-condition microarray experiments, whereas the general clustering allows the mean expression level changes with each experiment.

The main purpose of this article is to study clustering multivariate statistics, as alternatives to univariate statistics, to detect differential gene expression in exploratory data analysis. As a simple example to motivate the use of multivariate statistics, Figure 1 shows a simulated dataset generated from a multivariate normal mixture distribution with three clusters; see section 4.1 for details. Most genes were in the two clusters with equal expression while the other cluster contains differentially expressed genes. The histogram (right panel) of the univariate t statistics showed only a small number of outliers, presumably differentially expressed genes; in contrast, clustering a bivariate statistic (to be given later) clearly identified a cluster of the genes, deviating from the identity line (left panel). Thus, the use of the bivariate statistic was more powerful than the univariate statistic to detect differentially expressed genes in this example.

In this article, a comparison between univariate and multivariate normal mixture model-based
clustering is illustrated with an application to a real dataset containing the expression levels of 1,176 genes of normal rats and those with pneumococcal middle-ear infection. These two types of approaches are thereafter applied to two simulated datasets for the purpose of comparison.

2 Methods

2.1 Rat data

Four young and pathogen-free Sprague-Dawley rats were inoculated with pneumococcus in phosphate-buffered saline (PBS) and served as the pneumococcus group. Two other rats served as controls. The data were collected from radioactively labeled cDNA microarrays (Friemert 1998) with 1,176 genes. There were six samples with the first two corresponding to that without pneumococcal middle-ear infection while the last four corresponding to that with the disease, and the data were normalized, as in Pan et al. (2002), by median centering and inter-quartile scaling on each array; see Kafadar and Phang (2003) for a discussion on relevant issues.

2.2 Model-based clustering

The finite normal mixture model assumes that the data to be clustered are from several sub-populations or clusters with distinct normal distributions. Specifically, let \( x_1, \ldots, x_n \) denote \( n \) \( p \)-dimensional observations, possibly univariate or multivariate statistics for the \( n \) genes here; we use \( j \) to index observations while using \( i \) to index clusters. It is assumed that each observation \( x_j \) is from a mixture of a number, say \( g \), of \( p \)-variate normal densities with some unknown mixing proportions \( (\pi_1, \pi_2, \ldots, \pi_g) \):

\[
f(x_j; \Psi_g) = \sum_{i=1}^{g} \pi_i f(x_j; \mu_i, \Sigma_i) \tag{1}
\]

where \( f(x_j; \mu_i, \Sigma_i) \) denotes the \( p \)-variate normal density function with mean \( \mu_i \) and covariance matrix \( \Sigma_i \), \( i=1,\ldots,g \). Here the vector \( \Psi_g \) denotes all unknown parameters \( \{(\pi_i, \mu_i, \Sigma_i) : i = 1,\ldots,g\} \). The mixture model is typically fitted by maximum likelihood using the expectation-maximization (EM) algorithm (Dempster et al 1977), resulting in maximum likelihood estimate (MLE) \( \hat{\Psi}_g \). The
posterior probability of observation \( x_j \)'s belonging to component or cluster \( i \) is

\[
\tau_{ij} = \frac{\pi_i \phi(x_j; \hat{\mu}_i, \hat{\Sigma}_i)}{f(x_j; \hat{\Psi}_g)}. \tag{2}
\]

Finally, assign each observation to the cluster with the maximum posterior probability.

In contrast to many other approaches that fail to determine the number of components \( g \), model-based clustering provides several useful and objective selection criteria, which have been used in other model selection problems. One of the most widely used is the Bayesian Information Criterion (BIC) (Schwartz, 1978),

\[
BIC = -2 \log(\hat{\Psi}_g) + v_g \log(n) = -2 \sum_{j=1}^{n} \log f(x_j; \hat{\Psi}_g) + v_g \log(n) \tag{3}
\]

where \( v_g \) is the number of independent parameters in \( \Psi_g \).

We have assumed that observations \( x_j \)'s are independent. With microarray data, if there is any between-gene correlation (Qiu et al 2005), e.g. due to data normalization (Reilly et al 2003), under mild conditions, the MLEs of the mean parameters \( \mu_i \)'s are still consistent by the theory of estimating functions; however, the covariance parameter estimates may or may not be consistent. In addition, with a between-gene correlation, the effective sample size in BIC is no longer \( n \). Although some attempts to correct the problem are emerging (Efron 2005), as discussed by Newton et al (2004), it is in general quite difficult to model between-gene correlations. The standard treatment is to ignore such correlations in the literature; this is what we will adopt throughout this article too.

2.3 Univariate clustering

2.3.1 Univariate statistics

Suppose there are two experimental conditions with \( l \) replications in the first condition and \( m \) replications in the second condition, thus resulting in \( p = l + m \) arrays in total. For example, in the rat data \( l = 2 \) and \( m = 4 \). We wish to identify target genes which are differentially expressed between the two conditions. The most commonly used univariate summary statistics for gene expression data include the \( t \)-statistic, \( d \)-statistic (Broet et al. 2002) and SAM-\( t \) statistic (Tusher
et al. 2001). Specifically, suppose that observed expression levels of gene \( j \) are \( x_{j1}, \ldots, x_{jp} \), then

\[
\begin{align*}
t_j &= \frac{z_{j1} - z_{j0}}{\sqrt{\frac{\sum_{k=1}^{l}(x_{jk} - z_{j0})^2}{l(l-1)} + \frac{\sum_{k=l+1}^{p}(x_{jk} - z_{j1})^2}{m(m-1)}}} \quad (4) \\
d_j &= z_{j1} - z_{j0} \quad (5) \\
SAMt_j &= \frac{z_{j1} - z_{j0}}{\sqrt{\frac{\sum_{k=1}^{l}(x_{jk} - z_{j0})^2}{l(l-1)} + \frac{\sum_{k=l+1}^{p}(x_{jk} - z_{j1})^2}{m(m-1)}} + s_0} \quad (6)
\end{align*}
\]

where \( z_{j0} = \sum_{k=1}^{l} x_{jk} / l \) and \( z_{j1} = \sum_{k=l+1}^{p} x_{jk} / m \) are the sample means of gene \( j \)'s observed expression levels under the two conditions, and \( s_0 \) is the median of \( se_j^2 \)s (Efron et al. 2001) with

\[
se_j = \sqrt{\frac{\sum_{k=1}^{l}(x_{jk} - z_{j0})^2}{l(l-1)} + \frac{\sum_{k=l+1}^{p}(x_{jk} - z_{j1})^2}{m(m-1)}}.
\]

The SAM-\( t \) statistic is a modification to the \( t \) with a small constant \( s_0 \) used to stabilize the denominator of the statistic.

### 2.3.2 EM algorithm

The EM is iterative: given some univariate statistics \( y_j \) for \( j = 1, \ldots, n \), and the starting values of \( \{\pi_i, \mu_i, \Sigma_i = \sigma_i^2\} \), EM algorithm iteratively updates the estimates until convergence. Suppose the estimates at \( k \)th iteration are \( \hat{\pi}_i^{(k)}, \hat{\mu}_i^{(k)}, \hat{\sigma}_i^{2(k)} \), then the \( (k+1) \)th iteration proceeds by updating (McLachlan and Peel 2000; Pan et al. 2002)

\[
\begin{align*}
\hat{\pi}_i^{(k+1)} &= \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)}}{n} \\
\hat{\mu}_i^{(k+1)} &= \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} y_j}{\sum_{j=1}^{n} \tau_{ij}^{(k)}} \\
\hat{\sigma}_i^{2(k+1)} &= \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} (y_j - \hat{\mu}_i^{(k+1)})^2}{\sum_{j=1}^{n} \tau_{ij}^{(k)}} \quad (7)
\end{align*}
\]

with

\[
\tau_{ij}^{(k)} = \frac{\hat{\pi}_i^{(k)} \phi(y_j; \hat{\mu}_i^{(k)}, \hat{\sigma}_i^{2(k)})}{f(y_j; \hat{\Psi}_g)} \quad (8)
\]

for \( i = 1, \ldots, g \).
2.4 Multivariate clustering

2.4.1 Multivariate models

As before, suppose gene expression levels are observed under two experimental conditions with \(l\) and \(m\) replications respectively. Within the same cluster, it is reasonable to assume that each gene has the same mean expression level under the same condition; that is, in cluster \(i\), the mean vector is \(\mu_i = (\eta_{i1}, \eta_{i2}, 1, m)^T\), where \(1 \_k\) is a row vector of 1's of length \(k\). Several covariance structures for \(\Sigma_i\) are going to be considered. The first is an independence model with a diagonal covariance matrix and a common within-condition variance; that is, the covariance matrix for the genes in cluster \(i\) is \(\Sigma_i = \text{diag}(\sigma_{11}^2, \ldots, \sigma_{1p}^2, \sigma_{22}^2, \ldots, \sigma_{2p}^2)\) with only two unknown parameters \(\theta_i = (\sigma_{11}^2, \sigma_{12}^2)\). This covariance model is probably simplest as well as most widely used, corresponding to the usual assumption being used in many univariate statistics, such as \(t\)- and SAM-t statistics. A slightly more general covariance matrix is an independence model allowing different variances for different arrays: \(\Sigma_i = \text{diag}(\sigma_{11}^2, \ldots, \sigma_{1p}^2)\) with \(p\) unknown parameters \(\theta_i = (\sigma_{11}^2, \ldots, \sigma_{1p}^2)\). The model works, for example, when a preceding data normalization procedure, usually focusing on the mean structure, not the dispersion, is not successful to scale each array appropriately and hence does not yield a constant within-condition variance for each gene. The third one is unstructured, which is most general by allowing both unequal within- and between-condition variances and between-array correlations. It however involves a large number of parameters, \(p(p + 1)/2\). Usually it is assumed that the array experiments are independent, as in the first two models. However, due to the way of biological samples being prepared, between-array correlations may exist (Efron 2004). Again we assume that an observed gene expression profile or its summary statistic follows a mixture of multivariate normal model (1).

2.4.2 EM algorithm

Let \(\eta_i\) be the collection of unknown mean parameters and \(\theta_i\) the collection of unknown parameters in the covariance structure for cluster \(i\). The derivations of the EM algorithm with three different covariance matrices are provided in Appendix; see Theorem 0.4 therein. In summary, the EM algorithm is
1. Initialize \( \pi_i^{(0)} \), \( \tau_i^{(0)} \) by a K-means method with a given number of clusters \( g \).

2. Update the mean parameters \( \eta_k = (\eta_{k1}, \eta_{k2})^T \):

\[
\eta_i^{(k+1)} = \frac{\left( |A_{11}| \ |A_{12}| \right)^{-1} \sum_{j=1}^{n} \tau_{ij}^{(k)} \left( |A^{(k)} y_j|_1 \right)}{\sum_{j=1}^{n} \tau_{ij}^{(k)}}
\]

(9)

where \( A^{(k)} = \sum_{i}^{-1(k)} \) and \( |.| \) is the sum of entries of (.).

3. Update the covariance parameters for the chosen covariance structure:

(a) Unrestricted covariance structure:

\[
\Sigma_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)} (y_j - \mu_i^{(k+1)}) (y_j - \mu_i^{(k+1)})^T / \sum_{j=1}^{n} \tau_{ij}^{(k)}
\]

(10)

(b) Independence covariance structure:

\[
\hat{\Sigma}_i^{(k+1)} = \text{diag}(\sigma_{i1}^{2(k+1)}, \ldots, \sigma_{ip}^{2(k+1)})
\]

\[
\sigma_{ir}^{2(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} (y_{jr} - \mu_{ir}^{(k)})^2}{\sum_{j=1}^{n} \tau_{ij}^{(k)}}
\]

(11)

(c) Independence and a common within-condition variance:

\[
\hat{\Sigma}_i^{(k+1)} = \text{diag}(\sigma_{i1}^{2(k+1)}, \ldots, \sigma_{i2}^{2(k+1)}, \sigma_{21}^{2(k+1)}, \ldots, \sigma_{22}^{2(k+1)})
\]

\[
\sigma_{i1}^{2(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} \sum_{r=1}^{l} (y_{jr} - \mu_{ir}^{(k)})^2}{l \sum_{j=1}^{n} \tau_{ij}^{(k)}}
\]

\[
\sigma_{i2}^{2(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} \sum_{r=l+1}^{p} \sum_{r'=l+1}^{p} (y_{jr} - \mu_{ir}^{(k)})^2}{m \sum_{j=1}^{n} \tau_{ij}^{(k)}}
\]

(12)

4. Update the posterior and prior probability estimates:

\[
\tau_{ij}^{(k+1)} = \frac{\tilde{\pi}_i^{(k+1)} \phi(y_j; \mu_i^{(k+1)}, \hat{\Sigma}_i^{(k+1) / f(y_j; \hat{\Psi}_g^{(k+1)})}}
\]

\[
\tilde{\pi}_i^{(k+1)} = \frac{1}{n} \sum_{j=1}^{n} \tau_{ij}^{(k+1)}
\]

(13)

5. Let \( k \leftarrow k + 1 \) and repeat the above three steps until convergence.

As before, different numbers of clusters are tried and the best model is selected by BIC.
3 Application to the rat data

3.1 Univariate normal mixture model

The corresponding mixture model clustering based on univariate statistics can be done by replacing $y_j$ by $t_j, d_j$, or SAM$f_j$. Figure 2 shows the clusters of gene expression profiles across all six experiments using the three univariate statistics. For example, the fitted mixture model for the $t$-statistics was

$$ f(\cdot; \hat{\Psi}) = 0.508 \times N(0.88, 5.58) + 0.042 \times N(6.75, 77.23) + 0.450 \times N(-0.31, 1.15). $$

Less than 5% of the genes (30 genes in total) were in the second cluster with a large variance and a mean far away from those of the other two clusters, suggesting that the second cluster be likely to contain the genes with expression changes.

[Figure 2 is about here.]

3.2 Bivariate normal mixture model

To take advantage of the special feature of two-condition microarray experiments, a bivariate summary statistic can be used. Let $z_{jo}$ and $z_{j1}$ be the sample average expression levels for gene $j$ under the two conditions respectively; that is, $z_{jo} = \sum_{k=1}^{2} x_{jk}/2$ and $z_{j1} = \sum_{k=3}^{6} x_{jk}/4$, where $x_{ij}$ is the $i$th gene expression level in $j$th sample. we fitted a series of bivariate normal mixture models with unstructured covariance matrices to $(z_{jo}, z_{j1})$, which was a special case of multivariate clustering with unstructured covariance matrices with $l = m = 1$. The resulting best model judged by BIC contained five clusters,

$$ f(\cdot; \hat{\Psi}) = 0.021 \times N(\begin{bmatrix} 0.88 \\ 1.22 \end{bmatrix}, \begin{bmatrix} 0.34 & 0.21 \\ 0.21 & 0.24 \end{bmatrix}) + 0.552 \times N(\begin{bmatrix} -0.016 \\ -0.026 \end{bmatrix}, \begin{bmatrix} 0.001 & -0.00004 \\ -0.00004 & 0.0005 \end{bmatrix}) $$

$$ + 0.268 \times N(\begin{bmatrix} 0.007 \\ 0.032 \end{bmatrix}, \begin{bmatrix} 0.001 & 0.0005 \\ 0.0005 & 0.002 \end{bmatrix}) + 0.105 \times N(\begin{bmatrix} 0.07 \\ 0.14 \end{bmatrix}, \begin{bmatrix} 0.004 & 0.0004 \\ 0.0004 & 0.007 \end{bmatrix}) $$

$$ + 0.055 \times N(\begin{bmatrix} 0.27 \\ 0.40 \end{bmatrix}, \begin{bmatrix} 0.031 & 0.012 \\ 0.012 & 0.031 \end{bmatrix}) $$

(14)

92.5% of the genes were in clusters 2, 3 and 4 showing a fairly flat expression pattern across 6 experiments with little variability (Figure 3). About 5.5% of the genes were in cluster 5 showing
some evidence for differential expression between the two experimental conditions with a mean
difference of 0.13. The other 21 genes in cluster 1 demonstrated a strong evidence for differential
expression (Figure 4). The latter two clusters displayed larger variabilities than that of the other
clusters.

[Table 1 is about here.]

3.3 Multivariate normal mixture models

[Table 2 is about here.]

Rather than using a univariate (also called 1-dimensional, i.e. 1-D) or bivariate (i.e. 2-D)
statistic, one can also use the observed expression profile across all the experiments, which is a p-D
vector; we had $p = 6$ for the rat data. A 5-component model was found to be the best based on
BIC when unstructured covariance matrices were used, wherein 10 of 1,176 genes were found to be
highly likely to have altered expression and another 23 genes were likely. The other 1,143 genes
showed no evidence of being differentially expressed.

All the three covariance structures gave similar results for this dataset. They identified the same
number of clusters with similar gene expression profiles in the clusters (Figure 4). Furthermore, the
genes identified by the methods largely overlapped with each other (Figure 4). As a comparison,
the clustering results by 1-D methods were also drawn in a similar fashion (Figure 5).

[Figures 3, 4 and 5 are about here.]

4 Simulations

To further compare the performance of various clustering methods, two synthetic data sets were
used. To mimic the rat data, we considered six experiments with two experiments for control
samples and four experiments for cases.

For any model-based clustering method, one to seven components were sequentially considered
and the best one was selected by BIC. Due to its popularity, we also considered the multivariate
K-means clustering, which is also known to be a special case of model-based clustering with a common diagonal covariance matrix $\sigma^2 I$ across all clusters; we used the same number of clusters as that of p-D model-based clustering with unstructured covariance matrices. There might not be a clear one-to-one correspondence between the true clusters and predicted clusters; for example, the number of clusters based on a clustering method might differ from the true number of clusters. Hence, to determine the differentially expressed (DE) genes based on a clustering method, we mimicked what we would do in practice. Specifically, we used z0-z1 scatter plots (as Figures 4-5): if a cluster was not too large (under the a priori assumption of only few DE genes) and it appeared to be off from the identity line, we claimed it to be a cluster containing DE genes. Admittedly, this was somewhat subjective, which however was a nature of any exploratory data analysis.

Sensitivity, specificity and accuracy were used to compare various methods. For completeness, we briefly review their definitions below. Based on the truth and clustering results, a classification table can be constructed as

\[
\begin{array}{c|cc}
   & + & - \\
\hline
\text{Prediction} & a & b \\
\hline
\text{Truth} & c & d \\
\end{array}
\]

where, for example, $a$ is the number of the genes that are both true and predicted DE genes. Then

\[
\text{Sensitivity} = P(\text{Prediction}=+|\text{Truth}=+) \approx \frac{a}{a+c} \quad (15)
\]

\[
\text{Specificity} = P(\text{Prediction}=-|\text{Truth}=-) \approx \frac{d}{b+d} \quad (16)
\]

\[
\text{Accuracy} = P(\text{Prediction}=\text{Truth}) \approx \frac{a+d}{a+b+c+d} \quad (17)
\]

### 4.1 Set-up 1

The first synthetic dataset was generated from a mixture model containing three normal components, one of which consisted of DE genes while the other two were for non-DE genes. The mean expression levels of a non-DE gene were the same across six experiments, while a DE gene had a lower expression level under condition one and a higher expression level under the other condition.
In general, there may exist a within-gene or between-experiment correlation between any two observed expression levels of the same gene. Specifically, there are three different correlations corresponding to within condition 1 (\(\rho_1\)), within condition 2 (\(\rho_2\)) and between condition 1 and condition 2 (\(\rho_3\)). Therefore the general covariance structure for each cluster is given by

\[
\Sigma = \begin{bmatrix}
\tau_1^2 & \rho_1 \tau_1 \tau_2 & \rho_3 \tau_1 \tau_2 & \rho_3 \tau_1 \tau_2 & \rho_3 \tau_1 \tau_2 \\
\rho_1 \tau_1^2 & \tau_1^2 & \rho_3 \tau_1 \tau_2 & \rho_3 \tau_1 \tau_2 & \rho_3 \tau_1 \tau_2 \\
\rho_3 \tau_1 \tau_2 & \rho_3 \tau_1 \tau_2 & \tau_2^2 & \rho_2 \tau_2^2 & \rho_2 \tau_2^2 \\
\rho_3 \tau_1 \tau_2 & \rho_3 \tau_1 \tau_2 & \rho_2 \tau_2^2 & \tau_2^2 & \rho_2 \tau_2^2 \\
\rho_3 \tau_1 \tau_2 & \rho_3 \tau_1 \tau_2 & \rho_2 \tau_2^2 & \rho_2 \tau_2^2 & \tau_2^2 \\
\rho_3 \tau_1 \tau_2 & \rho_3 \tau_1 \tau_2 & \rho_2 \tau_2^2 & \rho_2 \tau_2^2 & \rho_2 \tau_2^2 & \tau_2^2
\end{bmatrix}
\]  (18)

We specified the values of \(\rho_1, \rho_2, \rho_3\) to obtain specific covariance structures. For example, \(\rho_1 = \rho_2 = \rho_3\) implied a compound symmetry correlation structure. Note that an advantage of multivariate clustering is its ability to account for a possible within-gene correlation whereas using a univariate statistic typically ignores such a correlation. The specific parameters used are listed in Table 3.

[Table 3 is about here.]

[Table 4 is about here.]

The results in Table 4 show that the bivariate and multivariate clustering methods, with higher sensitivities, specificities and accuracy rates, were systematically better than the three univariate clustering methods. The performances among all multivariate methods were close; in particular, the bivariate method was one of the winners.

4.2 Set-up 2

Because the data generation mechanism of set-up 1 favored multivariate mixture models, to be complete, we also simulated a second dataset such that the use of a univariate statistic was optimal. Specifically, we generated simulated data based on the \(t\)-statistics of the genes in the rat data:

1. Calculate the \(t\)-statistic for each of 1176 genes in the rat data.
2. Assume that the top 5% genes with the largest absolute values of t-statistics are target genes.

3. For any DE gene, generate its first two expression levels from a normal distribution \(N(z_{j0}, \text{var}_{j0})\), and the last four expression levels from \(N(z_{j1}, \text{var}_{j1})\), where the normal mean and variance parameters equal to the sample mean and sample variance under the corresponding condition of the rat data.

4. For any non-DE gene, generate its all six expression levels from a normal distribution with mean and variance equal to the sample mean and sample variance across the six experiments in the rat data.

[Table 5 is about here.]

The numerical results in Table 5 show that the best multivariate mixture model was the one with an unrestricted covariance matrix. Because the data were generated based on the t-statistics, this simulation set-up favored univariate methods, and as expected the SAM-t worked best. However, some of the multivariate clustering methods also worked reasonably well. In particular, among all the multivariate methods, in terms of the accuracy, the bivariate method was the second best with the performance close to the best; more impressively, the bivariate clustering even had a higher accuracy rate than the univariate \(d\)- and t-statistics.

5 Discussion

As an alternative to univariate clustering, multivariate normal mixture models are a useful exploratory data analysis tool to detect differential gene expression. Compared with univariate mixture modeling, multivariate clustering provides practitioners more flexibility in modeling microarray data: both the mean vector and the covariance structure can be appropriately chosen under given circumstances. For example, a multivariate statistic can explicitly model between-array (or within-gene) correlations; in contrast, the three univariate statistics all ignore such correlations. As discussed by Efron (2004), modeling of such correlations may be necessary for microarray data. As
a consequence of more flexible modeling, it may gain with increased power to detect differentially expressed genes, as shown in our simulation study.

Among the multivariate methods studied for two-condition microarray experiments, based on the overall performance in the simulation study and some theoretical considerations, we recommend the use of bivariate clustering, as a useful tool complementary to univariate clustering. Although other multivariate clustering can be also applied with comparable performance, bivariate clustering seems to have a reasonable modeling assumption with fewer parameters while its results are not only competitive, but also easy to be visualized, such as in a scatter-plot.

Because in multivariate clustering, there is a homogeneity assumption that all the genes in a cluster share the same covariance matrix, application of a variance stabilization transformation (Geller et al 2003; Durbin and Rocke 2004) prior to clustering is presumably to improve the performance of the model-based clustering. Finally, we comment on that an issue with t-type univariate statistics is probably due to their use of sample variances, which are unstable when the number of replicated microarrays is small, as demonstrated in Table 4, and by some recent work (e.g. Dean and Raftery 2004). For example, even though the second simulated dataset was ideal for the t-statistic, the t-statistic did not work as well as the SAM-t; the difference between the SAM-t and the t-statistic is that the former uses a small constant in its denominator to stabilize its variance estimate. We conjecture that, with some variance stabilization transformation for the second simulated dataset, multivariate clustering and univariate clustering using the d statistic will have improved performance.

Although we have restricted our attention to two-condition microarray experiments, the methods can be extended to more general cases with more than two experimental conditions (Broet et al. 2004), to time-course data (Luan and Li, 2004), or to incorporating biological knowledge (Pan 2006; Huang et al 2006). In addition, rather than using model-based clustering, it would be interesting to see whether some new clustering ideas, such as tight clustering (Tseng and Wong, 2005) and robust clustering (Liu et al, 2003), can be applied to the current context, and how they perform. These topics are worth future investigation.
Appendix: Derivation of (9)-(12)

Theorem 0.1 (Graybill 1983) Denote \( \frac{dl}{dx} = (\frac{dl_1}{dx_1}, \ldots, \frac{dl_n}{dx_n})^T \). Suppose \( l_1(x) = a'x = x'a \), where \( a = (a_1, \ldots, a_k)^T \) and \( a_i \) are any constants. \( l_2(x) = x'Ax \), where \( A = [a_{ij}] \) is a \( k \times k \) square matrix of constants, and \( l_3(x) = (bx + z)^T A(bx + z) \). Then

\[
\frac{dl_1}{dx} = a \tag{19}
\]
\[
\frac{dl_2}{dx} = (A + A^T)x \tag{20}
\]
\[
\frac{dl_3}{dx} = b^2(A + A^T)x + b(A + A^T)z \tag{21}
\]

Proof: The \( t \)-th element of \( \frac{dl_1}{dx} \) is, by definition, equal to \( \frac{dl_1}{dx_t} \) and it is clearly \( a_t \). Write \( l_2(x) = \sum_{i=1}^k \sum_{j=1}^k a_{ij}x_i x_j \). Note \( \frac{dl_2}{dx_t} = \sum_{j=1}^k x_j a_{jt} + \sum_{i=1}^k x_i a_{it} = A_t x + (A^T)_t x \), where \( A_t \) and \((A^T)_t\) are the \( t \)-th row vector of matrix \( A \) and \( A' \), respectively. If we express \( l_3(x) = b^2 x^T A x + bx^T Az + bz^T Ax + z^T Az \) and use (19) and (20), (21) can be proved. □

Theorem 0.2 If \( x = x^T 1_k \), where \( 1_k \) is a vector of 1’s of length \( k \) and \( x \) is a scalar. Let \( l_1(x) = x^T z = z^T x, l_2(x) = x^T Ax, \) and \( l_3(x) = (bx + z)^T A(bx + z) \). Then

\[
\frac{dl_1}{dx} = |z| \tag{22}
\]
\[
\frac{dl_2}{dx} = 2x|A| \tag{23}
\]
\[
\frac{dl_3}{dx} = 2b^2 x|A| + b(|Az| + |A^Tz|) \tag{24}
\]

where \(|.|\) is the sum of all entries in (.)

Proof: Equation (22)-(23) can be easily proved by noting that \( l_1(x) = x|z| \) and \( l_2(x) = x^2 1_k^T A 1_k = x^2 |A| \). (24) can be obtained by the results of (22) and (23). □

Theorem 0.3 If \( x = (x^T 1_m^T, y^T 1_m^T)^T \), where \( x \) and \( y \) are scalars. Let \( \theta = (x, y)^T \). Let \( l_1(\theta) = x^T z = z^T x, l_2(x) = x^T Ax, \) and \( l_3(x) = (bx + z)^T A(bx + z) \). Then

\[
\frac{dl_1}{d\theta} = \begin{bmatrix} |z|_1 \\ |z|_m \end{bmatrix} \tag{25}
\]
\[
\frac{dl_2}{d\theta} = (M + M^T)\theta \tag{26}
\]
\[
\frac{d\ell_3}{d\theta} = \nu^2(M + M^T)\theta + b \begin{pmatrix}
|A\mathbf{z}_i|_1 \\
|A\mathbf{z}_{i\mathbf{m}}|
\end{pmatrix} + \begin{pmatrix}
|A^T\mathbf{z}_i|_1 \\
|A^T\mathbf{z}_{i\mathbf{m}}|
\end{pmatrix}
\]

where \( \mathbf{u} = \sum_{i=1}^1 \mathbf{u}_i \), \( \mathbf{u}_{i\mathbf{m}} = \sum_{i=1}^{i+m} \mathbf{u}_i \) for a vector \( \mathbf{u} \), \( M = \begin{bmatrix} |A_{11}| & |A_{12}| \\
|A_{21}| & |A_{22}|
\end{bmatrix} \) and express \( A = \begin{bmatrix} (A_{11})_{l\times l} & (A_{12})_{l\times m} \\
(A_{21})_{m\times l} & (A_{22})_{m\times m}
\end{bmatrix} \).

Proof: Rewrite \( \mathbf{z} \) as \((\mathbf{z}_i^T, \mathbf{z}_m^T)^T\) and apply Theorem 0.2. \(\Box\)

**Theorem 0.4** For model (1) with the corresponding complete log likelihood function (28) and the expectation (29), if the mean structure is \( \eta = (\eta_1, \eta_2)^T \), then the update of the mean structure is (9). The covariance structure is updated by (10) - (12) for the corresponding situation.

Proof: In cluster analysis, \( (\mathbf{y}_j, \mathbf{z}_j) \) are treated as complete data, where \( \mathbf{y}_j \) is gene \( j \)'s expression levels or, more generally, a univariate or multivariate statistic for gene \( j \), and \( \mathbf{z}_j = (\mathbf{z}_{1j}, ..., \mathbf{z}_{d_j})' \) is an unknown vector of indicators of gene \( j \)'s belonging to each of the clusters. The log-likelihood function for complete data is

\[
\log L_c(\Psi_g) = \sum_{i=1}^g \sum_{j=1}^n \mathbf{z}_{ij} \{\log \pi_i + \log \phi(\mathbf{y}_j; \mu_i, \Sigma_i)\}.
\]

The E-step is simply to calculate the conditional expectation

\[
Q(\Psi_g; \hat{\Psi}_g^{(k)}) = E_{\hat{\Psi}_g^{(k)}}\{\log L_c(\Psi_g) \mid \mathbf{y}\}
\]

\[
= \sum_{i=1}^g \sum_{j=1}^n \tau_i(\mathbf{y}_j; \hat{\Psi}_g^{(k)}) \{\log \pi_i + \log \phi(\mathbf{y}_j; \mu_i, \Sigma_i)\}
\]

\[
= \sum_{i=1}^g \sum_{j=1}^n \tau_{ij}^{(k)} \{\log \pi_i - \frac{p}{2} \log 2\pi - \frac{1}{2} \log \det(\Sigma_i) - \frac{1}{2}(\mathbf{y}_j - \mu_i)^T \Sigma_i^{-1}(\mathbf{y}_j - \mu_i)\}
\]

where \( \tau_{ij}^{(k)} = \tau_i(\mathbf{y}_j; \hat{\Psi}_g^{(k)}) \) is the posterior probability of observation \( \mathbf{y}_j \) in cluster \( i \) at the \( k \)th step.

If the \( \mathbf{z}_{ij} \) were observable, then the complete-data MLE of \( \pi_i \) would be the average of \( \mathbf{z}_{ij} \)’s across \( j \), i.e. \( \hat{\pi}_i = \frac{1}{n} \sum_{j=1}^n \mathbf{z}_{ij} \) for \( i = 1, \cdots, g \). Therefore the estimate of \( \pi_i \) in the \( (k+1) \)th step is the
average of the posterior probabilities:

\[ \bar{\pi}_i^{(k+1)} = \frac{1}{n} \sum_{j=1}^{n} \pi_{ij}^{(k)} \]

Apply (27), then

\[ \frac{\partial Q}{\partial \eta} = -\sum_{j=1}^{n} \tau_{ij} \left[ -2 \left( \frac{|A_{ij}|}{|A_{ij}|} \right) + 2 \left( \frac{|A_{11}|}{|A_{21}|} \right) \eta \right] = 0 \]

Solve the above equation, (9) can be obtained. If the covariance is unrestricted, then (10) follows by MacLachlan et al. (1997). If the covariance is diagonal and each measurement has different variability, i.e. \( \Sigma_i = \text{diag}(\sigma_{i1}^2, \ldots, \sigma_{ip}^2) \), then

\[ Q_i = C + \sum_{j=1}^{n} \tau_{ij} \left\{ -\frac{1}{2} \sum_r \log \sigma_{ir}^2 - \frac{1}{2} \sum_r \frac{1}{\sigma_{ir}^2} (y_{jr} - \mu_{ir})^2 \right\} \]

\[ \frac{\partial Q_i}{\partial \sigma_{ir}^2} = \sum_{j=1}^{n} \tau_{ij} \left\{ -\frac{1}{2} \cdot \frac{1}{\sigma_{ir}^2} - \frac{1}{2} \cdot (-1) \frac{1}{\sigma_{ir}^2} (y_{jr} - \mu_{ir})^2 \right\} \]

\[ = \frac{1}{2} \sum_{j=1}^{n} \tau_{ij} \left\{ \frac{1}{\sigma_{ir}^4} (y_{jr} - \mu_{ir})^2 - \frac{1}{\sigma_{ir}^2} \right\} \]

Thus the MLE (11) is obtained.

If the covariance has only two parameters, i.e. \( \Sigma_i = \text{diag}(\sigma_{i1}^2 1_i^T, \sigma_{i2}^2 1_m^T) \), then \( \text{det}(\Sigma_i) = \sigma_{i1}^4 \sigma_{i2}^m \)

and

\[ Q_i = C + \sum_{j=1}^{n} \tau_{ij} \left\{ -\frac{1}{2} (l \log \sigma_{i1}^2 + m \log \sigma_{i2}^2) - \frac{1}{2} \frac{1}{\sigma_{i1}^2} \sum_r (y_{jr} - \mu_{ir})^2 - \frac{1}{2} \frac{1}{\sigma_{i2}^2} \sum_{r-l+1}^i (y_{jr} - \mu_{ir})^2 \right\} \]

\[ \frac{\partial Q_i}{\partial \sigma_{i1}^2} = \sum_{j=1}^{n} \tau_{ij} \left\{ -\frac{1}{2} \cdot \frac{l}{\sigma_{i1}^2} + \frac{1}{2} \cdot \frac{1}{\sigma_{i1}^4} \sum_r (y_{jr} - \mu_{ir})^2 \right\} \]

\[ \frac{\partial Q_i}{\partial \sigma_{i2}^2} = \sum_{j=1}^{n} \tau_{ij} \left\{ -\frac{1}{2} \cdot \frac{m}{\sigma_{i2}^2} + \frac{1}{2} \cdot \frac{1}{\sigma_{i2}^4} \sum_{r-l+1}^i (y_{jr} - \mu_{ir})^2 \right\} \]

Then the MLE (12) is obtained. □

**Acknowledgements**

WP was supported by NIH grant HL65462. We thank the two reviewers for many helpful and constructive comments.
References


Figure 1: Bivariate clustering and univariate clustering with the t statistic for the first synthetic data with 3 clusters.
Figure 2: Gene expression profiles in the clusters identified by univariate clustering with the rat data: the first two experiments without subacute pneumococcal middle-ear infection and the other four with infection. From top to bottom: \( t \), SAM-\( t \) and \( d \) statistics are used respectively.
Table 1. Selection of the number of components $g$. $logL$ is the maximized log-likelihood, and $\Delta(-2logL)$ is the change of $logL$ from that of the model with one fewer component.

<table>
<thead>
<tr>
<th># clusters</th>
<th>logL</th>
<th>$\Delta(-2logL)$</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1392.27</td>
<td>-</td>
<td>-2749.19</td>
</tr>
<tr>
<td>2</td>
<td>3217.97</td>
<td>3651.41</td>
<td>-6358.17</td>
</tr>
<tr>
<td>3</td>
<td>3423.18</td>
<td>410.42</td>
<td>-6726.18</td>
</tr>
<tr>
<td>4</td>
<td>3489.50</td>
<td>132.62</td>
<td>-6816.38</td>
</tr>
<tr>
<td>5</td>
<td>3510.94</td>
<td>42.89</td>
<td>-6816.86</td>
</tr>
<tr>
<td>6</td>
<td>3524.59</td>
<td>27.29</td>
<td>-6801.73</td>
</tr>
<tr>
<td>7</td>
<td>3534.11</td>
<td>19.05</td>
<td>-6778.36</td>
</tr>
</tbody>
</table>

Table 2. Clustering results using various methods.

<table>
<thead>
<tr>
<th>Covariance Matrix</th>
<th># clusters</th>
<th>#target genes</th>
<th>#possible target genes</th>
<th>#genes w/o diff. exp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bivariate</td>
<td>5</td>
<td>21</td>
<td>63</td>
<td>1,092</td>
</tr>
<tr>
<td>6-D w/ unres. cov</td>
<td>5</td>
<td>10</td>
<td>23</td>
<td>1,143</td>
</tr>
<tr>
<td>6-D w/ diag. cov</td>
<td>5</td>
<td>11</td>
<td>37</td>
<td>1,128</td>
</tr>
<tr>
<td>6-D w/ 2-param. cov</td>
<td>5</td>
<td>10</td>
<td>26</td>
<td>1,140</td>
</tr>
</tbody>
</table>

Table 3. Parameters used in the first simulated data.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>mean $\mu$</th>
<th>correlation</th>
<th>Variance</th>
<th>$\pi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>[ 1 1 1 1 1 1 ]'</td>
<td>$\rho_1 = \rho_2 = \rho_3 = .3$</td>
<td>$\tau_1 = \sqrt{2}, \tau_2 = \sqrt{3}$</td>
<td>53%</td>
</tr>
<tr>
<td>B</td>
<td>[ 5 5 5 5 5 5 ]'</td>
<td>$\rho_1 = .3, \rho_2 = .5, \rho_3 = .2$</td>
<td>$\tau_1 = \sqrt{10}, \tau_2 = \sqrt{20}$</td>
<td>42%</td>
</tr>
<tr>
<td>C</td>
<td>[ 5 5 15 15 15 15 ]'</td>
<td>$\rho_1 = .3, \rho_2 = .5, \rho_3 = .2$</td>
<td>$\tau_1 = \sqrt{10}, \tau_2 = \sqrt{20}$</td>
<td>5%</td>
</tr>
</tbody>
</table>
Table 4. Numerical comparison of clustering methods with first synthetic data.

<table>
<thead>
<tr>
<th>Method</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>d</td>
<td>0.860</td>
<td>0.997</td>
<td>0.990</td>
</tr>
<tr>
<td>SAM-t</td>
<td>0.840</td>
<td>0.988</td>
<td>0.981</td>
</tr>
<tr>
<td>t</td>
<td>0.880</td>
<td>0.853</td>
<td>0.854</td>
</tr>
<tr>
<td>6-D K-means</td>
<td>.900</td>
<td>1.000</td>
<td>0.995</td>
</tr>
<tr>
<td>Bivariate</td>
<td>0.920</td>
<td>1.000</td>
<td>0.996</td>
</tr>
<tr>
<td>6-D w/ unres. cov</td>
<td>0.920</td>
<td>1.000</td>
<td>0.996</td>
</tr>
<tr>
<td>6-D w/ diag. cov</td>
<td>0.940</td>
<td>0.997</td>
<td>0.994</td>
</tr>
<tr>
<td>6-D w/ 2-param. cov</td>
<td>0.920</td>
<td>1.000</td>
<td>0.996</td>
</tr>
</tbody>
</table>

Table 5. Numerical comparison of clustering methods with second synthetic data.

<table>
<thead>
<tr>
<th>Method</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>d</td>
<td>0.762</td>
<td>0.954</td>
<td>0.945</td>
</tr>
<tr>
<td>SAM-t</td>
<td>0.847</td>
<td>0.992</td>
<td>0.986</td>
</tr>
<tr>
<td>t</td>
<td>1.000</td>
<td>0.953</td>
<td>0.955</td>
</tr>
<tr>
<td>6-D K-means</td>
<td>0.288</td>
<td>0.973</td>
<td>0.939</td>
</tr>
<tr>
<td>Bivariate</td>
<td>0.288</td>
<td>0.991</td>
<td>0.956</td>
</tr>
<tr>
<td>6-D w/ unres. cov</td>
<td>0.390</td>
<td>0.993</td>
<td>0.963</td>
</tr>
<tr>
<td>6-D w/ diag. cov</td>
<td>0.271</td>
<td>0.978</td>
<td>0.942</td>
</tr>
<tr>
<td>6-D w/ 2-param. cov</td>
<td>0.288</td>
<td>0.974</td>
<td>0.940</td>
</tr>
</tbody>
</table>
Figure 3: From top to bottom: quadrant $G$ with an unrestricted covariance, $G$ with a diagonal covariance, and $G$ with a diagonal covariance.
Comparisons of various clustering methods: Rat Data

Figure 4: Multivariate clustering results. $z_0$ and $z_1$ are the average expression levels in the two conditions with the rat data.
Figure 5: Univariate and $K$-means clustering results: $z0$ and $z1$ are the average expression levels in the two conditions with the rat data.
Figure 6: Gene expression profiles in the true and estimated clusters with simulated data. From up to bottom: truth, t, 6-D model with a diagonal covariance, 6-D model with a diagonal 2-parameter covariance, bivariate model, and 6-D with an unrestricted covariance structure.