Bernoulli **10**(6), 2004, 951–969

A nested unsupervised approach to identifying novel molecular subtypes

ELIZABETH S. GARRETT* and GIOVANNI PARMIGIANI**

*Division of Oncology and Biostatistics, Johns Hopkins University School of Medicine, Suite 1103, 550 North Broadway, Baltimore MD 21205, USA, and Sidney Kimmel Comprehensive Cancer Center, Suite 1100, 401 North Broadway, Baltimore MD 21231, USA. E-mail: *esg@jhu.edu; **gp@jhu.edu*

In classification problems arising in genomics research it is common to study populations for which a broad class assignment is known (say, normal versus diseased) and one seeks undiscovered subclasses within one or both of the known classes. Formally, this problem can be thought of as an unsupervised analysis nested within a supervised one. Here we take the view that the nested unsupervised analysis can successfully utilize information from the entire data set for constructing and/or selecting useful predictors. Specifically, we propose a mixture model approach to the nested unsupervised problem, where the supervised information is used to develop latent classes which are in turn used for data mining and robust unsupervised analysis. Our solution is illustrated using data on molecular classification of lung adenocarcinoma.

Keywords: Bayesian model; class discovery; gene expression; lung cancer

1. Introduction

The wide availability of high-throughput assays in biological research is generating many high-dimensional data sets that pose novel analysis questions. For example, in genomics and proteomics, a single experiment can provide information on thousands of genes or proteins from a single biological sample. One of the most challenging uses of such information is the identification of novel molecular subclasses. This task has been approached using a combination of unsupervised clustering and visualization. While these methods have led to important progress in understanding biological phenomena, especially in the area of cancer classification (Mohr *et al.* 2002), there remain at least two important limiations: first, approaches using observed RNA or protein expression levels can be overly sensitive to noise and outliers; second, approaches using constructs that depend on a large number of genetic dimensions tend to generate molecular subclasses whose interpretation is tied to a specific technological platform and is likely to be obscured from a biological standpoint.

To address these issues, we recently proposed analysis and visualization approaches for gene expression based on three-component latent classes, representing over-, under- and typical expression (Parmigiani *et al.* 2002). The goals of the three-component latent class analysis are: to identify variables which show variation across the sample population which is not likely to be the result of measurement error; to choose subsets of variables which

1350–7265

show similar patterns across observations; and to define population subclasses using a small number of non-redundant variables. Class indicators replace observed expression by a scale that is both more robust and more easily interpretable across technologies, and can facilitate expert-based dimension reduction. Classes are identified using a Bayesian hierarchical mixture model approach that searches for evidence of clustering of expression levels across biological samples.

In this paper we present a generalization of this approach where we take advantage of the knowledge that some of the biological samples are known to be normal, whereas the remaining samples are diseased. This approach differs from that previously mentioned which ignores all phenotype information for both model fit and gene selection. Here we focus on using the information from normal samples to more accurately and efficiently define what 'typical' gene expression is for each gene and on ways to assess which genes are informative by comparing their expression tendencies in normal versus cancer samples. The motivating application area is molecular classification of cancer using genomic data. Even though the focus of these analyses is the search for yet undiscovered subgroups within broad morphological classes of cancer, studies often include both cancer and normal samples (Bhattacharjee *et al.* 2001), and sometimes additional cancer types. The normal samples are used for clustering of genes, to facilitate indentification and interpretation of groups of coregulated genes. Here we pursue a more formal way of incorporating information from normal samples in the discovery of subclasses of cancers. Specifically, we use class membership on normals to improve the fit of the mixture model and the reliability of the latent class assignment. The resulting three-component scale is then used in the unsupervised analysis of the cancer, including visualization, gene mining, and profile definition. More broadly, there are many situations arising in molecular biology research where it is assumed that a population is comprised of known classes (say, normal and disease) and that within the disease class there are undiscovered disease subtypes. Formally, this problem can be thought of as an unsupervised analysis nested within a supervised one. We term this, for brevity, the 'nested unsupervised' case. The class information is useful for the nested unsupervised analysis because it allows, broadly speaking, for a better definition of predictors.

In this paper we define a latent class model for the nested unsupervised case (Section 2), discuss data reduction and data mining techniques that make use of the supervised information in the unsupervised analysis (Section 3), and demonstrate the methodology in the analysis of gene expresion data on lung adenocarcinomas (Section 4).

2. Nested unsupervised analysis via supervised latent classes

2.1. Mixture modelling of latent classes

Consider a sample of T individuals, for whom we have collected a vector of binary class identifiers c and a $G \times T$ matrix of predictors A with elements a_{gt}. In genomic applications, the number of predictors G is in the tens of thousands and much larger than T. We define the goal of a nested unsupervised analysis to be that of finding subgroups

within each of the classes $c = 1$ and $c = 0$. For concreteness, we will refer to class $c = 0$ as normal and class $c = 1$ as cancer. For simplicity of exposition we will only focus on identifying subclasses within the cancer class.

Our approach differs from the model presented by Parmigiani *et al.* (2002) in that we consider the case where a limited amount of information is known about true classes of the subjects. Because we take a nested unsupervised approach, we are able to define the normal class by the normal observations. We use a mixture model to determine the criteria for categorizing values as low, normal and high.

The basic underlying assumption from which our model arises is that the distribution of each variable (e.g., gene expression or protein expression) across individuals follows a three-component mixture model, with components indicators e_{gt} defined by:

$$e_{gt}\begin{cases} -1 & \text{variable } g \text{ is abnormally low in subject } t, \\ 0 & \text{variable } g \text{ is at a typical level in subject } t, \\ 1 & \text{variable } g \text{ is abnormally high in subject } t. \end{cases}$$

These components provide a scale that has lower resolution than the absolute measurements, but is more interpretable biologically, more likely to preserve its meaning across technologies, and more amenable to defining class memberships that can be validated and implemented clinically. Parmigiani *et al.* (2002) and the associated discussion provide additional motivation and details.

In the unsupervised setting, all the component indicators e_{gt} are estimated using mixture modelling techniques. In the nested unsupervised setting, we propose to consider the following relationship between the e_{gt} and c_t:

$$\begin{array}{ll} \text{if } c_t = 0 \text{ then } e_{gt} = 0 & \text{for } g = 1, \ldots, G; \\ \text{if } c_t = 1 \text{ then } e_{gt} \text{ is unknown} & \text{for } g = 1, \ldots, G. \end{array}$$

This parametrization assumes that all of the normal samples are from the normal component of the mixture, while the disease samples are from the overexpression uniform component ($e_{gt} = 1$), the underexpression uniform component ($e_{gt} = -1$), or the normal component ($e_{gt} = 0$), implying that for model estimation e_{gt} can take values of 1 or -1 only in the cancer samples. This is motivated by the desire to ensure that the typical level category, $e = 0$, is interpretable as the category that is expected in normal samples. In cancer samples, because of the multiplicity of mechanisms leading to cancer and the fact that many genes are not involved in carcinogenesis, we do not preclude the case where $e_{gt} = 0$ for samples where $c = 1$. This assumption generates an asymmetry in the way the unsupervised classification is nested in the supervised analysis, but also allows us to borrow strength from the class information in defining novel subtypes. The efficiency of this approach will improve with the homogeneity of a predictor within the normal samples.

For each variable g, the distributions of measurements in the low, normal and high class are $f_{-1,g}$, $f_{0,g}$, $f_{1,g}$, respectively. That is,

$$a_{gt}|(e_{gt} = e) \sim f_{e,g}(\cdot), \qquad e \in \{-1, 0, 1\}. \tag{1}$$

We define π_g^+ to be the population proportion of subjects who have a high value for variable

g and π_g^- to be the population proportion of subjects who have a low value for variable g. The model assumes that the e_{gt} are independent conditional on the πs and fs.

This approach is similar to a latent class or latent profile model (Bartholomew and Knott 1999; McCutcheon 1987; Arminger *et al.* 1995) where the classes and subtypes are defined by patterns of the observed variables. However, in the standard latent class and latent profile models, the variables that define the latent classes are predetermined. In our case, one of the challenges is to facilitate gene mining and expert selection of a small number of relevant genes from a set of thousands.

2.2. Distributional assumptions

In our software implementation, we have used uniform ($\mathcal{U}$) distributions for $f_{-1,g}$ and $f_{1,g}$ and a Gaussian distribution for $f_{0,g}$ (Garrett and Parmigiani 2003). The parametrization is as follows:

$$f_{-1,g}(\cdot) = \mathcal{U}(-\kappa_g^- + \alpha_t + \mu_g, \alpha_t + \mu_g),$$

$$f_{0,g}(\cdot) = \mathcal{N}(\alpha_t + \mu_g, \sigma_g),$$

$$f_{1,g}(\cdot) = \mathcal{U}(\alpha_t + \mu_g, \alpha_t + \mu_g + \kappa_g^+).$$

In practice, these distributions have proven successful in capturing the categorical nature of gene expression data in both simulated and real data sets.

In the Gaussian distribution, $\alpha_t + \mu_g$ represents the mean of the typical expression distribution for gene g in sample t, with μ_g as the gene effect and α_t as a subject-specific effect. We include α_t to adjust for the possibility that the values in sample t might be higher or lower on average than other samples. In pre-normalized gene expression data, the main function of the α_t is to readjust the normalization so that it only applies to the normal and not the regulated observations. σ_g is the standard deviation of the normal category in gene g. The upper and lower limits of the high and low distributions are $\alpha_t + \mu_g + \kappa_g^+$ and $\alpha_t + \mu_g - \kappa_g^-$, respectively.

There are many choices for the distributions which would likely achieve the same goals. Our reasons for choosing the above distributions are partly mathematical convenience and partly due to to the nature of genetic and proteomic data. For example, it can be assumed in many cases that the error associated with measuring gene expression follows a Gaussian distribution, justifying our use of the the Gaussian distribution for normal expression. In our applied setting, the uniform distribution naturally lends itself to the case of differential gene expression. In cancer applications, differential expressions are thought to be caused by the failure of biological mechanisms. As a result, the observed expression levels may take a broad range of values. Although we advocate the use of the priors that we have chosen, it should be stressed that other prior distribution may be used if desired.

For estimation, choosing the uniform distributions is efficient because it requires the estimation of relatively few additional parameters. One of the limits of each of the uniform components is defined by $\mu_g + \alpha_t$, and so only one additional parameter is required.

Consider the analogous case of a mixture of three Gaussian distributions: the Gaussian mixture model would require six gene-specific parameters, whereas our model only requires four (this does not include the estimates of π_g^+ and π_g^-). This property is convenient in that stable estimates are provided even when the majority of the genes tend to fall into the normal expression case. Additionally, because of the flat shape of the uniform, no values are assigned very low densities. We have imposed an additional constraint that $\kappa_g^+ > r\sigma_g$ and $\kappa_g^- > r\sigma_g$ to ensure that the uniforms truly represent high and low values and do not have a large portion of their range overlapping with the Gaussian component. In our implementation, we generally choose a value of $r > 3$, which ensures relatively little overlap between the Gaussian and the uniform components.

Examples of normal/uniform mixtures for finding outliers and sparse clusters are discussed by Fraley and Raftery (1998). For other examples of mixture modelling applied to microarray data, see Lee *et al.* (2000), McLachlan *et al*. (2002) and Yeung *et al*. (2001).

As in Parmigiani *et al*. (2002), a Bayesian hierarchical model is used to estimate the mixture model proposed above. The estimation approach yields posterior distributions for each of the parameters of interest. We borrow strength across genes by assuming that the gene-specific parameters (e.g., μ_g, π_g^+) follow additional probability distributions. This is motivated by two factors: first, due to the high gene-to-subject ratio, there is relatively little information with which to estimate gene-specific parameters; and second, technological aspects of the assays would affect many or all of the genes similarly.

Specifically, we use the following hierarchical distributions to describe the variation of parameters across genes:

$$\mu_g|\theta_\mu, \tau_\mu \sim \mathcal{N}(\theta_\mu, \tau_\mu),$$

$$\sigma_g^{-2}|\gamma, \lambda \sim \mathcal{G}(\gamma, \lambda),$$

$$\kappa_g^+|\theta_\kappa^+ \sim \mathcal{E}(\theta_\kappa^+),$$

$$\kappa_g^-|\theta_\kappa^- \sim \mathcal{E}(\theta_\kappa^-),$$

$$\text{logit}(\pi_g^+)|\theta_\pi^+ \sim \mathcal{N}(\theta_\pi^+, \tau_\pi^+),$$

$$\text{logit}(\pi_g^-)|\theta_\pi^- \sim \mathcal{N}(\theta_\pi^-, \tau_\pi^-),$$

where $\mathcal{G}$ is the gammma distribution and $\mathcal{E}$ is the exponential distribution. We assume that gene-specific parameters are independent conditional on the hyperparameters on the right-hand side of the distributions above. Hyperparameters can be assigned dispersed, non-informative priors, as the large number of genes allows for data-driven estimation. An advantage of the hierarchical model is that for genes which show little or no evidence of high or low values (i.e., $\pi_g^- \approx \pi_g^+ \approx 0$), there is essentially no information in the data with which to estimate the parameters associated with the high and low distributions. The hierarchical model uses information from the other genes with which to estimate parameters for these variables. Notice that there is no hierarchical distribution for α_t. The model could easily be generalized to include this, but in practice it does not appear to be necessary or to affect model estimates.

We fit this model using a Markov chain Monte Carlo estimation procedure, in which the data are augmented with a trichotomous indicator, e_{gt} for each a_{gt}, with the additional constraint that $e_{gt} = 0$ if $c_t = 0$ (see also Diebolt and Robert 1994; West and Turner 1994). The constraint has important implications in the interpretation of results. In the gene expression data that we will examine in the next section, there are 139 cancer samples and only 17 normal samples. If there are genes which clearly delineate the cancers from the normals, we would expect that only the normal samples would have expression values consistent with $e = 0$, and the cancer samples would appear to have $e = -1$ or $e = 1$.

As in the unsupervised version, to facilitate sampling of the κs, we used the sampling sequence $[\kappa|\omega^*]$, $[e|\kappa, \omega^*]$, $[\omega^*|\kappa, e]$. Symbols refer to parameter vectors or matrices, brackets refer to posterior distributions. We use ω as shorthand for the full set of parameters, and ω^* for ω with κ removed. Given the class indicators (e_{gt}), the full conditional distribution of the π_g is a Dirichlet distribution, and the full conditional distribution of the parameters of the normal component is conjugate, with the additional constraint that $\sigma r < \min(\kappa_g^+, \kappa_g^-)$.

For each point in the predictor matrix, the probability of latent class membership is

$$p_{gt}^+ = P(e_{gt} = 1|a_{gt}, \omega) = \frac{\pi_g^+ f_{1,g}(a_{gt})}{\pi_g^+ f_{1,g}(a_{gt}) + \pi_g^- f_{-1,g}(a_{gt}) + (1 - \pi_g^+ - \pi_g^-) f_{0,g}(a_{gt})}, \quad (2)$$

$$p_{gt}^- = P(e_{gt} = -1|a_{gt}, \omega) = \frac{\pi_g^+ f_{-1,g}(a_{gt})}{\pi_g^+ f_{1,g}(a_{gt}) + \pi_g^- f_{-1,g}(a_{gt}) + (1 - \pi_g^+ - \pi_g^-) f_{0,g}(a_{gt})}. \quad (3)$$

The quantities in equations (2) and (3) can be interpreted as measures of the distance between observed measurements and measurements that would be expected in normal subjects. Values of p_{gt}^+ and p_{gt}^- that are close to 0 indicate similarity to normal subjects, while values close to 1 indicate levels that are either high or low as compared to what is seen in normal subjects.

A point gt can only have high positive probability of belonging to the high or to the low category, but not both, as the two categories are not overlapping. Exploiting this fact, we can combine p_{gt}^+ and p_{gt}^- by $p_{gt} = p_{gt}^+ - p_{gt}^-$. We refer to this new variable as the 'poe scale', where poe is an acronym for 'probability of expression'. The transformation from a_{gt} to p_{gt} is useful because we have essentially made the data independent of the method with which the measurements were assayed. For example, the a_{gt} could be expression values from oligonucleotide arrays, or from cDNA arrays, or from other means for measuring genetic activity. Additionally, all genes are now measured on the same scale so we can directly compare variables across subjects. We present some specific tools for data reduction in the next section. However, the $G \times T$ matrix of p_{gt} values can now be used in any clustering or other analytic method.

3. Data reduction approaches in nested unsupervised analyses

We now shift our focus to the application of the above model where the variables of interest are genes, and the subjects are referred to as biologic samples.

3.1. Evaluating diagnostic characteristics of variables

The goals of the analyses that follow are to find a relatively small number of genes which show variation across samples, show consistent values within normal samples, and possibly show evidence of subtypes within the disease class. To do this, we assign each expression value to one of the components of the mixture model, based on the estimated p_{gt} values. Specifically, we estimate the true category of gene g for subject t (e_{gt}) with $\hat{e}_{gt}$, such that

$$\hat{e}_{gt} = \begin{cases} -1 & \text{if } p_{gt} < -p_0, \\ 1 & \text{if } p_{gt} > p_0, \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where p_0 is a fixed threshold. Because the high and low class probability are strongly negatively correlated, a natural choice is a threshold of $p_0 = 0.5$, although other cut-offs can be chosen, ranging from 0 to 1. Note that this is also done for the normal samples: in the model estimation, we fixed the $e_{gt} = 0$ for normal samples. However, to determine whether or not the normal samples do tend to fall within the normal component of the mixture, we assign their expression values to one of the three categories based on their fitted p^+_{gt} and p^-_{gt} values. We then use the matrix of $\hat{e}_{gt}$ values to determine which genes accurately allocate normal samples to the normal component and cancer samples to the uniform components.

The allocation of normal samples to the normal component can be assessed by examining a variable's 'specificity', and evidence of allocation of cancerous samples to the uniform components can be assessed by 'sensitivity'. We define specificity (sp_g), sensitivity (se_g), positive sensitivity (se^+_g), and negative sensitivity (se^-_g) for gene g as follows:

$$sp_g = P \text{ (sample } t \text{ is classified as normal by gene } g | \text{sample } t \text{ is normal)},$$

$$se_g = P \text{ (sample } t \text{ is classified as high or low by gene } g | \text{sample } t \text{ is diseased)},$$

$$se^+_g = P \text{ (sample } t \text{ is classified as high by gene } g | \text{sample } t \text{ is diseased)},$$

$$se^-_g = P \text{ (sample } t \text{ is classified as low by gene } g | \text{sample } t \text{ is diseased)}.$$

To calculate specificity (sp_g) and sensitivities (se_g, se^+_g, se^-_g), we use the estimates of $\hat{e}_{gt}$ found above where samples have been assigned to the high, normal and low categories. Note that the specificity and sensitivities are calculable in this nested unsupervised approach due to the use of the normal phenotype information. In the previous implementations of the poe model where sample phenotype was not included in the model, sensitivity and specificity could not be estimated.

The effect of choosing a cut-offs (i.e., p_0) closer to 0 will tend to classify more samples as normal, decreasing sensitivity and increasing specificity. Choosing a cut-off closer to 1 will have the opposite effect. After choosing a threshold and categorizing each expression value for each sample, we can calculate sp_g, se_g, se^+_g, and se^-_g for each gene as follows:

$$sp_g = \frac{\sum_{t:v_t=1} (1 - |\hat{e}_{gt}|)}{\sum_{t=1}^{T} (1 - c_t)},$$

$$se_g = \frac{\sum_{t:v_t=0} (|\hat{e}_{gt}|)}{\sum_{t=1}^{T} c_t},$$

$$se_g^+ = \frac{\sum_{t:v_t=0} I(\hat{e}_{gt} = 1)}{\sum_{t=1}^{T} c_t},$$

$$se_j^- = \frac{\sum_{t:v_i=0} I(\hat{e}_{gt} = -1)}{\sum_{t=1}^{T} c_t},$$

where $c_t = 0$ if the sample is normal and 1 if the sample is cancerous. Based on (4), we can see that the p_{gt}^+ and p_{gt}^- are reflected strongly in the sensitivities and specificities. But, what the sensitivities and specificities provide is a summary of the p_{gt} information in each of the two groups (i.e., the normals and cancers) for each gene. Hence, they are statistics that summarize how well the genes separate the cancers into the uniform components of the mixture and the normal samples into the normal component of the mixture.

We are interested in genes which are consistent across normal samples. This corresponds to choosing genes that show high specificity, sp_g. If we are also interested in genes which show evidence of subtypes of disease, then we would choose genes that also had one of the levels of sensitivity away from the extremes, which suggests that for individuals who have disease only a fraction of them show high or low expression. If a variable has very low se^+ and relatively high se^-, the diseased samples tend to have low values relative to normals. If a variable has moderate levels of both se^+ and se^-, then there is a subtype of diseased samples that show low levels and another subtype showing high levels.

We can now reduce our data set by choosing genes which show sufficient specificity. Recall that the specificity of gene g refers to the probability that expression values gene for g in a normal sample come from the normal component of the mixture distribution described in Section 2. In general, we would expect the specificities to be high because the model is estimated assuming $e_{gt} = 0$ for all normal samples. However, for some genes, some of the expression values in the normal samples may be more consistent with expression values from the cancer samples. It is these genes that we are not interested in

exploring further. Hence, setting a threshold for specificity will allow us to weed out genes which will not be useful for distinguishing normals from cancers. This level of specificity will depend to some extent on the data set under consideration, but as an example, setting a specificity of 0.6 ensures that at least 60% of the normals are classified as normals by the gene of interest. For sensitivity, we use the overall sensitivity value (*se*), where we choose a much lower threshold due to the hypothesis that there are subtypes within the disease categories. For example, a threshold of 0.10 for *se* will sufficiently eliminate variables that show almost no evidence of association with disease versus normal status. Note that although we are interested in genes with high specificity and low sensitivity, we tend to not be interested in variables with low specificity and high sensitivity. These variables would tend to categorize normal subjects into the disease class.

By setting thresholds for specificity and sensitivity, we can effectively eliminate genes which show little evidence of being related to the disease process. A cautionary note is that the precision of the sensitivities and specificities will depend on the number of samples. In the example that is presented in the next section, the number of normal samples is relatively small and so the threshold should be chosen conservatively.

3.2. Creating subsets of similar variables

Estimates of class assigment probabilities can be used to mine for genes that are likely to provide interesting subgroups of the diseased category. The 'mining' method for creating subsets is primarily exploratory in that figures are provided to give the user a sense of how well combinations of genes are able to distinguish subclasses within the diseased subjects.

Before describing the approach for finding subsets, we define two statistics which are critical for selecting genes: gene coherence and gene agreement. We calculate the $G \times G$ matrix

$$r_{gk} = \sum_{t=1}^{T} (p_{gt}^{+} p_{kt}^{+} + p_{gt}^{-} p_{kt}^{-} + (1 - p_{gt})(1 - p_{kt})),$$

where coherence is measured by the diagonal of the agreement matrix, and agreement by the off-diagonal elements. Specifically, the coherence of gene g is represented by r_{gg}, which measures how 'cleanly' gene g is able to discriminate between over-, typically and underexpressed genes. Genes with values of p_{gt}^{+} and p_{gt}^{-} close to 0 and 1 will have high coherence, while those with values closer to 0.50 will have low coherence. Gene agreement between two genes g and k is defined by r_{gk}, which measures the expected proportion of agreements in defining samples as over-, typically and underexpressed. After calculating the gene coherence and gene agreement, some exploratory analysis of the values should be performed to determine what are sufficient agreement and coherence values. For example, the 75th or 90th percentile of these distributions might be chosen. However, the choice should depend on the values within the data set of interest: if overall coherence appears to be high, then a threshold close to the 50th percentile might be warranted, whereas if coherence is generally low, then setting a threshold closer to the 90th percentile might be more appropriate. For defining sufficient gene agreement, we have used two types of measures: a

fixed level of gene agreement, as just described; or agreement as a proportion of coherence of the 'seed' variable, which will be discussed in more detail below.

The mining method is algorithmic, and user inputs guide the resulting profiles. In step 1, the user defines a level of differential expression (both under and over, e.g. 15% underexpressed and 0% overexpressed) which determines what types of genes will be chosen. The differential expression pattern will have a strong influence on which genes are selected, and, as such, it is expected that users will try a variety of patterns when searching for subsets using our approach. We do not see the dependence of profiles on chosen pattern of expression as a weakness: instead we see this approach as providing a flexible way of looking at a variety of gene expression patterns. This approach was designed with the expectation that users would indeed repeat the process multiple times with differing input patterns in step 1. And, as the algorithm is computationally simple and fast, the repetition of this exploratory approach is not time-consuming. While we note that the resulting genes chosen will depend strongly on the 'low–high' pattern chosen, the algorithm is not overly sensitive to the pattern. For example, the patterns $\{0.05, 0.20\}$ and $\{0.10, 0.25\}$ will generally yield very similar results.

The mining algorithm for finding homogeneous subsets in the application of Section 4 is shown below, and is described in more detail in Parmigiani *et al*. (2002) and Garrett and Parmigiani (2003):

1. Choose an expression pattern of interest. The idea is to state a target for how many samples are expected to show low expression and how many to show high expression for a gene. For example, the pattern $\{0.05, 0.20\}$ indicates that 5% of samples should be low and 20% should be high for a gene. The remaining 75% would then be in the 'typical' component of the mixture.
2. Sort genes according to consistency with 'low–high' distribution defined in step 1. Using the estimates of p^{+}_{gt} and p^{-}_{gt}, we can calculate, for each gene g, the probability that the distribution of over- and underexpression among the samples is the same as in the specified low–high distribution. We sort genes by this probability.
3. Choose as the 'seed' gene the one with the largest probability from step 2 which is sufficiently coherent (i.e., $r_{gg} > r^{c}$, where r^{c} is the cut-off for gene coherence).
4. Choose genes that show substantial agreement with the seed gene, either as a fixed agreement cut-off, or as a proportion of coherence of the seed variable. Add these genes to the 'group' which is seeded by gene chosen in step 3.
5. Remove the genes in the group defined in step 4 from further consideration. Repeat steps 3 and 4 to identify remaining groups.

We apply this approach to a subset of genes which have sufficient specificity and sensitivity. For each repetition of gene mining, we find homogeneous sets of genes and, for the purpose of defining molecular profiles, generally need to choose just one gene to represent the group. Some of the genes within a set may be more appealing to scientists or clinicians in terms of describing classes among subjects. In the gene expression setting, many of the rows of the gene expression matrix are truly known genes but many are expressed sequence tags (ESTs), which are small portions of the active parts of genes and usually of unknown biological function. If given a choice as to whether to define disease

subtypes using known genes or ESTs, the known genes are generally preferable. As a result, we can scan each gene group for the one that makes the most sense clinically or biologically. In the settings where the idea of 'preferred' variables does not apply, it is most logical to choose the seed variable from a group as the group's representative.

After a subset of genes has been identified using this method (i.e., a set of 'representative' genes have been defined from several gene groups), we use the genes to define pattern profiles. For example, if we have chosen only two variables, for each subject we can calculate the probability that the subject belongs to one of nine possible profiles ((−1, −1), (0, −1), (1, −1), (−1, 0), (0, 0), (1, 0), (−1, 1), (0, 1), (1, 1)). Here the −1, 0 and 1 are the true e_{gt} and e_{kt} values for the two chosen variables, g and k. We use the p_{gt} value for estimation of profile probabilities. For a set of m genes, there are 3^m possible patterns. Because the number of patterns grows very large even for moderate m, it is generally preferred to choose relatively few genes.

Using the sample-specific profile probabilities, we can then create a 'heatmap' showing which samples tend to cluster together and what their gene expression profiles look like. Specifically, with profile on the y-axis and sample on the x-axis, we plot the profile probabilities using a colour or grey scale. This provides a graphical tool to assess how many subclassess appear and how well the disease subgroups are differentiated from the normal samples. This will be clearer when an example is seen in the next section.

We use the above plot as an exploratory tool showing subtypes of samples based on gene expression profiles of two or more genes. The question is whether what is seen in the plot can be determined to be 'real' subtypes or not. The way to determine if the subtypes are real subtypes, or perhaps have arisen due to chance, can be approached in two different ways. The first is to use statistical validation, which could be done using either another gene expression data set, or using some standard validation approaches within the data set (for a discussion of cross-validation in gene expression analyses, see Simon *et al*. 2003). The second approach is to use biological validation, where the identified profiles are interpreted and make biological sense, and, further, using more sensitive assays, such as reverse transcription polymerase-chain reaction (RT-PCR) which is known to be a very accurate way of assessing gene expression, we can find out if these profiles truly exist in the samples. However, even by using RT-PCR, we may find that the gene expressions are validated, but it is still for further investigation to discern whether the identified subtype behaves in a distinct and predictable way from the other cancer subtypes.

4. Identifying subclasses of lung adenocarcinoma

4.1. Data

We now illustrate the nested unsupervised methodology described so far using a gene expression data set that includes normal and cancerous lung samples (Bhattacharjee *et al*. 2001). The specimens in this data set include 139 lung adenocarcinomas (adeno), and 17 normal lung (NL) specimens. Throughout, normal samples are indicated in figures with varying symbols. The primary analytic goals are to indentify subgroups of adenos and compare the cancer samples to

the normal samples. Affymetrix arrays were used to obtain gene expression data on the 156 samples for 5665 genes. This set of 5665 genes is a subset of the original data set and was chosen based on its overlap with a comparable data set from another institution. We used all 5665 genes instead of choosing a smaller set through filtering to show the useful properties of data reduction based on our methods and software. More detailed information about the experimental processes can be found in Bhattacharjee *et al.* (2001). Data were preprocessed to remove experimental artefacts, and a cube root transformation was performed.

4.2. Sensitivity and specificity of genes

We used the R library POE (Garrett and Parmigiani 2003) to fit the mixture model described in Section 2. POE can be obtained at http://astor.som.jhmi.edu/poe. Figure 1 illustrates the fit of the mixture model for gene 30. There is evidence of two subgroups in the data. Most of the normal samples cluster in correspondence with the subgroup with lower expression, although one belongs to the high-expression component. Because the subgroups are of similar size, a completely unsupervised analysis may have identified either class as the 'typical' class. The additional information from normal samples permits us to attribute a more reliable interpretation to the classes.

Sensitivity and specificity of all genes can also be computed using tools in the POE library. Results are shown in Figure 2. We can see that there are many genes with high specificity, indicating that the normal samples do in fact tend to show similar expression patterns in many of the genes. Sensitivity ranges from 0 to approximately 0.8, with 75% of the genes having sensitivities less than 0.25. We filtered our 5665 genes by taking only genes with specificities above 0.8 and sensitivities above 0.10. This left us with 1182 genes, a reduction in the number of genes of about 80%.

A similarity image (i.e., 'heatmap') of samples is shown in Figure 3. Similarity entries are Pearson's correlation coefficient calculated using the p_{gt} matrix of poe scores for the 1182 selected genes. The rows and columns of the matrix have been sorted using a divisive hierarchical clustering algorithm to find groups of genes. The grey-scale intensity represents the correlation: white is perfect positive correlation and black is perfect negative correlation. We see that the subset of genes that we have chosen does a very good job of separating the normal samples from the adenocarcinomas (note that the adenocarcinomas are all very close to one another). While we could have estimated the correlation matrix and performed the divisive clustering using the entire set of 5665 genes, including genes that are not related to the phenotype of interest would have been likely to add more noise than signal to our clustering. Generally, it can be more efficient to only include meaningful variables in a cluster analysis, so that spurious clusters are not formed due to chance associations in the data.

4.3. Gene mining

We then used the procedure described in Section 3.2 to find a small number of genes which will provide molecular profiling information. The target pattern sizes used for mining genes

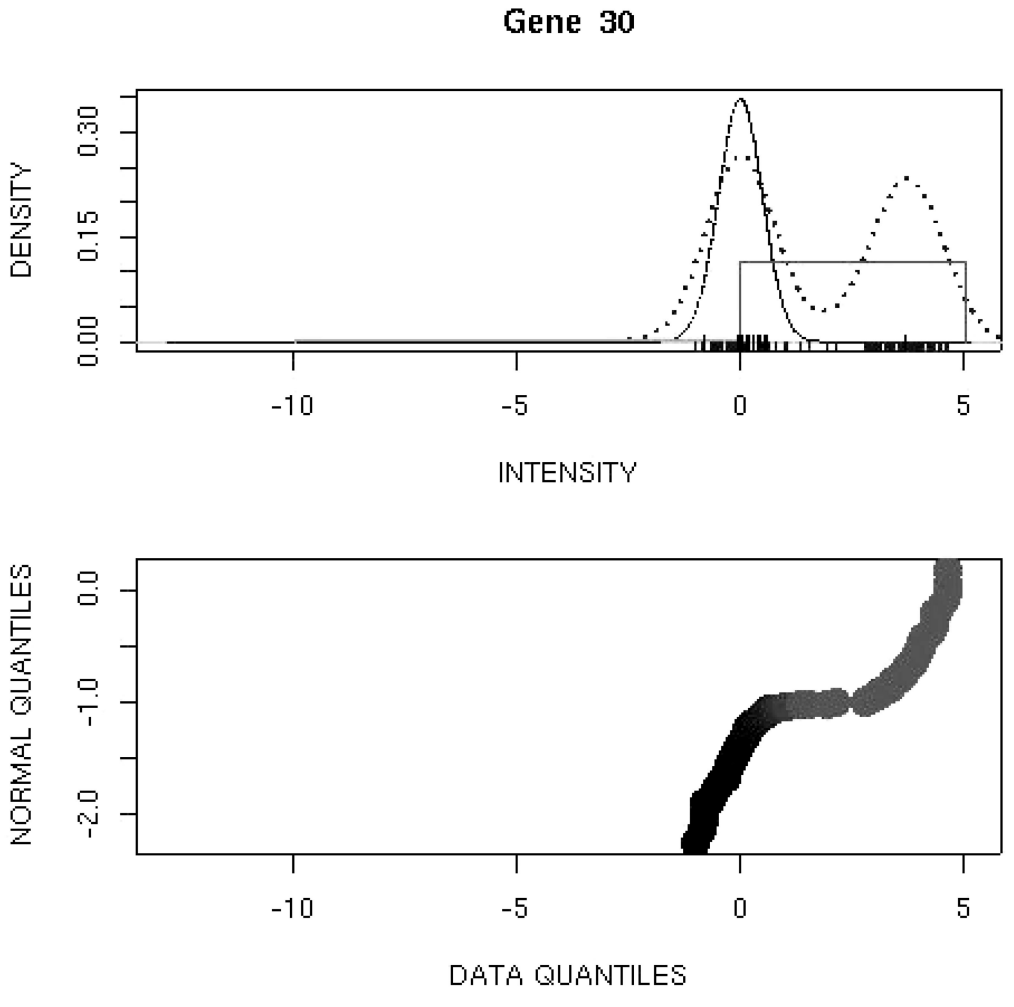

Figure 1. Estimated mixture components for gene 30. Short vertical marks along the x-axis are the estimated residuals $a_{gt} - \mu_g - \alpha_t$, for the cancer samples. Tall vertical marks are the corresponding residuals for normal samples. The dotted line is a kernel density estimate of the distribution of the residuals. The solid lines correspond to the best-fitting uniform and normal components of the mixture, multiplied by the corresponding mixture weights. The underexpression uniform ranges from -10 to 0 and the overexpression uniform ranges from 0 to 5. The bottom panel displays the normal quantile plot, with dark to lighter grey shades proportional to the probability $1 - \hat{p}_{gt}$ of being from the normal component.

were (0.1, 0.5), (0.5, 0.1), (0.1, 0.25), (0.25, 0.1), (0.2, 0.05), (0.05, 0.20), and (0.3, 0.3). After successively grouping genes for these patterns of expression in the data, we selected three genes that represented different partitions of the sample space and also had high specificities (1.00, 0.88 and 0.94) and moderate sensitivities (0.15, 0.33 and 0.58),

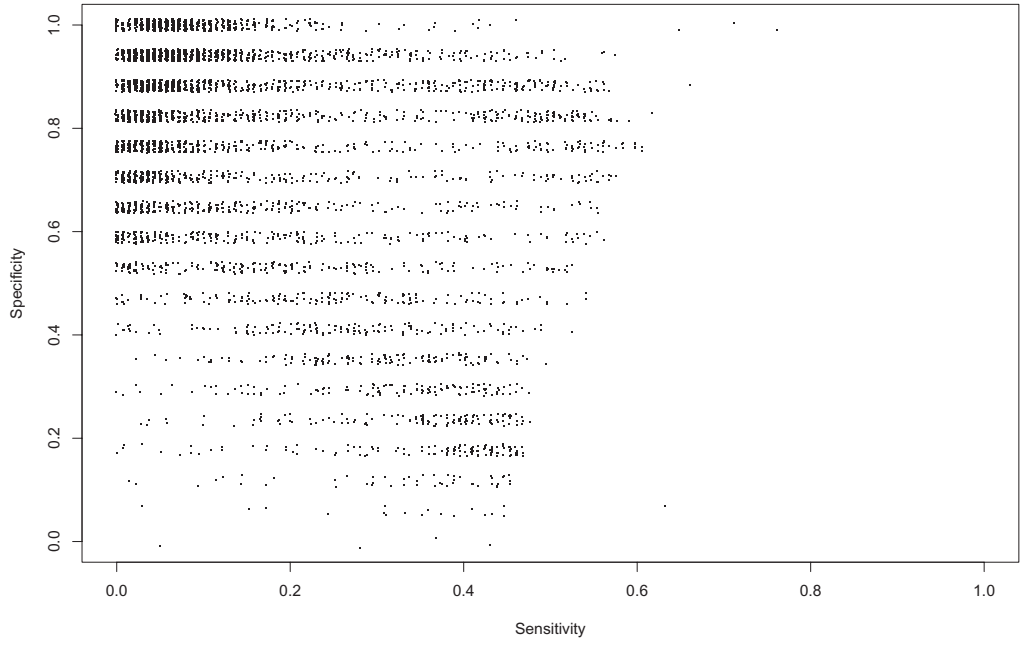

Figure 2. Scatterplot of sensitivity and overall specificity of the genes analysed. Specificity can take on only 18 values, as there are 17 normal samples. For the purpose of this scatterplot, vertical coordinates have been slightly perturbed.

respectively. The genes are: *BRCA1* (breast cancer 1), a tumour suppressor gene that is related to the famililal breast/ovarian cancer syndrome (Szabo and King 1997) as well as other cancers; *MEIS1* (myeloid ecotropic viral integration), which is a transcription factor known to be related to oncogenesis (Moskow *et al*. 1995); and *FGF7* (fibroblast growth factor 7), which is related to lung development (Ware and Matthay 2002).

For each of the samples, we estimate the 3^3 profile probabilities and show this graphically in Figure 4, with darker values representing higher probabilities. The four profiles (0,0,1), (0,0,0), (0,−1,0), and (−1,−1,0) receive relatively high probablity in a large number of samples. Profiles (0,1,1), (−1,0,0) and (0,−1,1) also receive high probability in some of the samples. As expected, many normal samples belong to the normal profile (0,0,0) with high probability, although some give high probability to other classes, as the sensitivity and specificity of the classifier genes are not 100%.

The nonlinear transformation from the expression scale to the poe scale can be thought of as a denoising transformation. The effects of denoising are illustrated in Figure 5. There tend to be tighter clusters of points in the poe-scaled data and more scatter in the raw data. The poe scale also carries information about the uncertainty with which the trichotomization can be applied.

In Table 1 we have assigned each sample to the most likely of the 27 possible profiles. We find that there is good specificity of this classification, with 14 of the 17 normal samples belonging to the normal profile (0,0,0). There is also strong evidence that other subclasses of adenocarcinoma

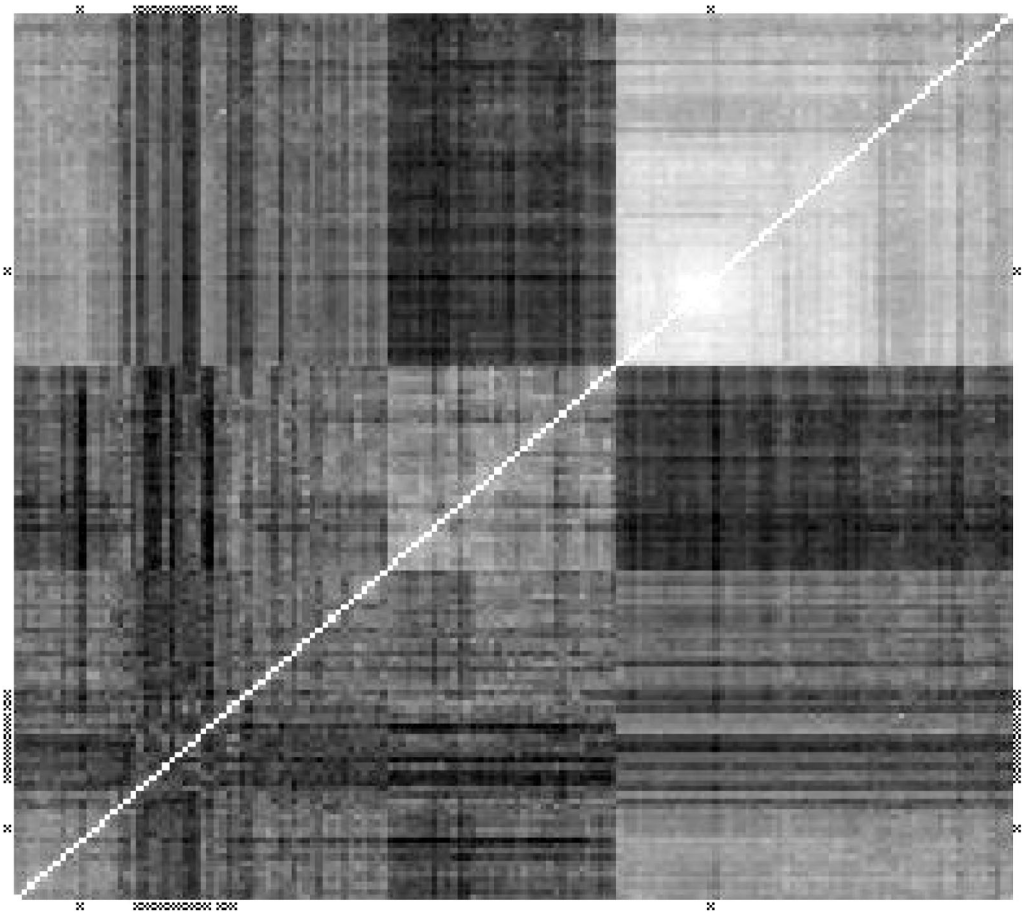

Figure 3. Pairwise Pearson's correlation matrix of $p_{gt} = p^{+}_{gt} - p^{-}_{gt}$. White indicates perfect positive correlation and black indicates perfect negative correlation. Normal samples are indicated by the symbol '×' on the axes. Rows and columns have been sorted based a divisive hierarchical clustering algorithm.

exist: 63 samples very strongly show the pattern (0,0,1), and 17 adenocarcinoma samples are classified into (0,−1,0) and another 12 into (−1,−1,0). There is some evidence that the profiles (0,1,1), (0,−1,1), and (−1,0,0) might be meaningful, due to the high probability that several adenocarcinomas (5, 7 and 4, respectively) exhibit these patterns.

5. Discussion

Genomic data analysis is posing novel challenges to high-dimensional classification. Among the most critical is to develop methods for discovery of novel biological subtypes using

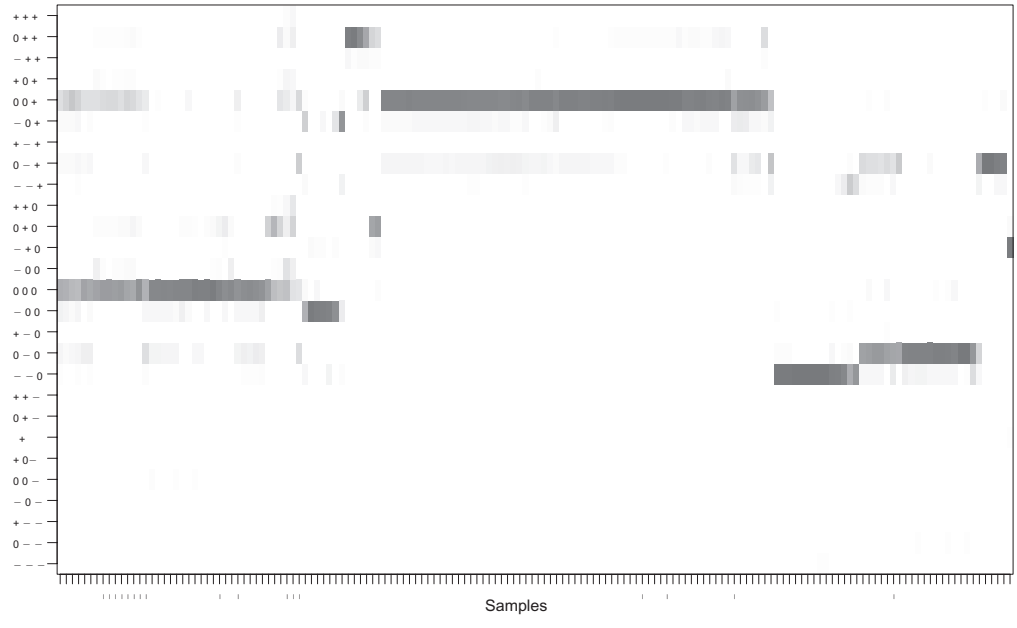

Figure 4. Molecular profiles probabilities. Probabilities are plotted in grey scale, where 0 corresponds to white, 1 corresponds to black. Each row corresponds to one of the 27 molecular profiles defined by the expression status of genes *BRCA1*, *MEIS1*, and *FGF7*. Each column corresponds to a sample. For example, the point for row $(1, -1, 0)$ for tumour 79 is the probability that the true expression indicators for tumour 79 are $(1, -1, 0)$ with regard to the genes in question. Marks on the horizontal scale identify normal samples.

molecular profiles. This requires integration of complex modelling, to properly capture sources of variation, with intuitive and interpretable visualization, to support dimension reduction with reliably elicited biological knowledge.

One of the most promising directions for dimension reduction in unsupervised analysis in genomics is to use known class assignment information involving the same predictors in a similar context. In this paper we formally explore statistical modelling of this principle. We define a nested unsupervised analysis to be the discovery of subclasses within a known class, and we discuss a mixture-based approach that builds on earlier work on unsupervised molecular profiling. We have extended the R library POE to handle this case and illustrated its use.

In gene expression data analysis, a practical advantage of our approach is to help in the screening of genes as predictors, using simple and interpretable measures such as sensitivity and specificity. Presecreening of predictors is normally done based on overall expression variability, which is prone to outliers and not sufficiently sensitive to clustering of samples. A second advantage of incorporating the information from the normal is a more reliable interpretation of the latent classes used in classification. Additional discussion of three-way mixture models in molecular profiling can be found in Parmigiani *et al.* (2002).

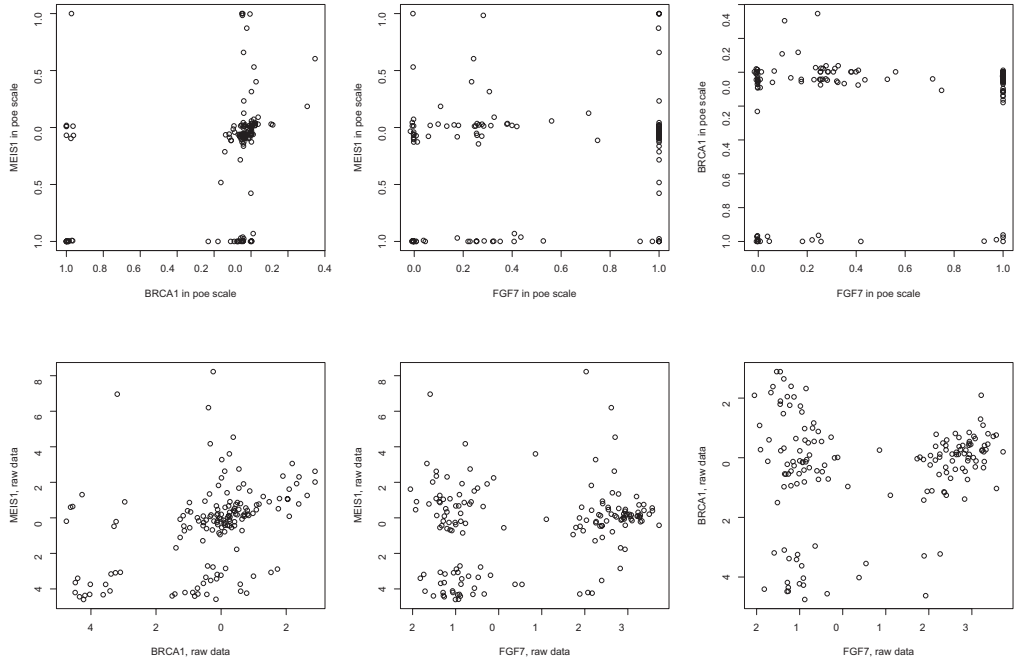

Figure 5. Scatterplots of poe scale (top row) and continuous untransformed scale (bottom row) for the three genes selected for profiling (*BRCA1*, *FGF7* and *MEIS1*).

Table 1. Profile assignments for 156 lung tissue samples. Profiles represent the overexpression (1), normal expression (0) and underexpression (−1) of genes *BRCA1*, *MEIS1* and *FGF7*, respectively

Profile	Adeno	Normal
(−1,−1,0)	12	0
(0,−1,0)	17	1
(−1,0,0)	4	0
(0,0,0)	23	14
(−1,1,0)	1	0
(0,1,0)	2	1
(−1,−1,1)	2	0
(0,−1,1)	7	0
(−1,0,1)	3	0
(0,0,1)	63	1
(0,1,1)	5	0
Total	139	17

Acknowledgement

Our work was partly supported by National Cancer Institute grants P50CA88843, P50CA62924-05, DK-58757, 5P30 CA06973-39 and National Institutes of Health grant HL 99-024.

References

Arminger, G., Clogg, C.C. and Sobel, M.E. (eds) (1995) *Latent Class Models*, Chapter 6. New York: Plenum Press.

Bartholomew, D.J. and Knott, M. (1999) *Latent Variable Models and Factor Analysis*, 2nd edn. London: Arnold.

Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E.J., Lander, E.S., Wong, W., Johnson, B.E., Golub, T.R., Sugarbaker, D.J. and Meyerson, M. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. USA*, **98**, 13 790–13 795.

Diebolt, J. and Robert, C.P. (1994) Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. Ser. B*, **56**, 363–375.

Fraley, C. and Raftery, A.E. (1998) How many clusters? Which clustering method? – Answers via model-based cluster analysis. *Comput. J.*, **41**, 578–588.

Garrett, E.S. and Parmigiani, G. (2003) POE: Statistical tools for molecular profiling. In G. Parmigiani, E.S. Garrett, R.A. Irizarry and S.L. Zeger (eds), *The Analysis of Gene Expression Data: Methods and Software*. New York: Springer-Verlag.

Lee, M.L. Kuo, F.C., Whitmore, G.A. and Sklar, J. (2000) Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. USA*, **97**, 9834–9839.

McCutcheon, A.L. (1987) *Latent Class Analysis.* London: Sage.

McLachlan, G.J., Bean R.W. and Peel, D. (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413–422.

Mohr, S., Leikauf, G.D., Keith, G. and Rihn, B.H. (2002) Microarrays as cancer keys: an array of possibilities. *J. Clinical Oncology*, **20**, 3165–3175.

Moskow, J.J., Bullrich, F., Huebner, K., Daar, I.O. and Buchberg, A.M. (1995) Meis1, a PBX1-related homeobox gene involved in myeloid leukemia in BXH-2 mice. *Mol. Cellular Biol.*, **15**, 5434–5443.

Parmigiani, G., Garrett, E.S., Anbazhagan, R. and Gabrielson, E. (2002) A statistical framework for expression-based molecular classification in cancer. *J. Roy. Statist. Soc. Ser. B*, **64**, 717–736.

Simon, R.M., Korn, E., McShane, L., Radmacher, M., Wright, G.W. and Zhao, Y. (2003) *Design and Analysis of DNA Microarray Investigations*. New York: Springer-Verlag.

Szabo, C.I. and King, M.C. (1997) Population genetics of BRCA1 and BRCA2. *Amer. J. Hum. Genetics*, **60**, 1013–1020.

Ware, L.B. and Matthay, M.A. (2002) Keratinocyte and hepatocyte growth factors in the lung: roles in lung development, inflammation, and repair. *Amer. J. Physiol. – Lung Cellular Mol. Physiol*, **282**, L924–940.

West, M. and Turner, D. (1994) Deconvolution of mixtures in analysis of neural synaptic transmission. *The Statistician*, **43**, 31–43.

Yeung, K., Fraley, C., Murua, A., Raftery, A. and Ruzzo, W. (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.

Received December 2002 and revised March 2004