10-6-2003

A Nested Unsupervised Approach to Identifying Novel Molecular Subtypes

Elizabeth Garrett
Johns Hopkins University School of Medicine, Sidney Kimmel Comprehensive Cancer Center, esg@jhu.edu

Giovanni Parmigiani
Johns Hopkins University, Departments of Oncology, Biostatistics & Pathology, gp@jhu.edu

A nested unsupervised approach to identifying novel molecular subtypes

Elizabeth S. Garrett & Giovanni Parmigiani

Sidney Kimmel Comprehensive Cancer Center
Johns Hopkins University, Baltimore, MD

Abstract

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 to find 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.

1 Introduction

The wide availability of high throughput assays in biological research is generating many high-dimensional datasets 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 limitations: 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 are 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-way latent classes, representing over- under- and typical expression (Parmigiani et al., 2002). The goals of the three-way latent class analysis are to (1) identify variables which show variation across the sample population which is not likely to be the result of measurement error, (2) choose subsets of variables which show similar patterns across observations, and (3) define population sub-classes 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 article we present a generalization of this approach. 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 sometime additional cancer types. The normal samples are used in 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-way 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 termed 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 article 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 Modeling of Latent Classes

Consider a sample of I individuals, for which we collected a vector of binary class identifiers c and a $J \times I$ matrix of predictors A with elements a_{ji}. In genomic applications, the number of predictors J is in the tens of thousands and much larger than I. 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.

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_{ji} defined by:

$$\begin{aligned} e_{ji} &= -1 \quad \text{variable } j \text{ is abnormally low in subject } i \\ e_{ji} &= 0 \quad \text{variable } j \text{ is at a typical level in subject } i \\ e_{ji} &= 1 \quad \text{variable } j \text{ is abnormally high in subject } i. \end{aligned}$$

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 an implemented clinically. Parmigiani et al (2002) and the associated discussion provide additional motivation and details.

In the unsupervised setting, all the component indicators e's are estimated using mixture modeling techniques. In the nested unsupervised setting, we propose to consider the following relationship between the e's and c:

$$\begin{array}{llll} \text{if } c_i = 0 & \text{then} & e_{ji} = 0 & \text{for } j = 1, \ldots, J \\ \text{if } c_i = 1 & \text{then} & e_{ji} \text{ is unknown} & \text{for } j = 1, \ldots, J \end{array}$$

This is motivated by the desire of ensuring 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 $e_{ji} = 0$. 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 j, the distributions of measurements in the low, normal and high class are $f_{-1,j}$, $f_{0,j}$, repectively. That is,

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

We define π_j^+ to be the population proportion of subjects who have a high value for variable j and π_j^- to be the population proportion of subjects who have a low value for variable j. The model assumes that the e_{ji}'s are independent conditional on the π's and f's.

This approach is similar to a latent class or latent profile model (Bartholomew, 1987; 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 facilitating 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 $f_{-1,j}$ and $f_{1,j}$ and a Gaussian distribution for $f_{0,j}$ (Garrett and Parmigiani, 2003). The parameterization is as follows:

$$\begin{aligned} f_{-1,j}(\cdot) &= \mathcal{U}(-\kappa_j^+ + \alpha_i + \mu_j, \alpha_i + \mu_j) \\ f_{0,j}(\cdot) &= \mathcal{N}(\alpha_i + \mu_j, \sigma_j) \\ f_{1,j}(\cdot) &= \mathcal{U}(\alpha_i + \mu_j, \alpha_i + \mu_j + \kappa_j^-). \end{aligned}$$

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

In the Gaussian distribution, $\alpha_i + \mu_j$ represents the mean of the typical expression distribution for gene j in sample i, with μ_j as the gene effect and α_i as a subject-specific effect. We include α_i to adjust for the possibility that the values in sample i might be higher or lower on average than other samples. In pre-normalized gene expression data, the main function of the α_i's is to readjust the normalization so that it only applies to the normal and not the regulated observations. σ_j is the standard deviation of the normal category in gene j. The upper and lower limits of the high and low distributions are $\alpha_i + \mu_j + \kappa_j^+$ and $\alpha_i + \mu_j - \kappa_j^-$, 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.

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 $\mu_j + \alpha_i$, 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_j^+ and π_j^-). 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 $\kappa_j^+ > r\sigma_j$ and $\kappa_j^- > r\sigma_j$ 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 modeling applied to microarray data see (Lee et al., 2000; McLachlan et al., 2002; 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 variables by assuming that the variable-specific parameters (e.g., μ_j, π_j^+, etc.) follow additional probability distributions. This is motivated by two factors: (1) due to the high variable-to-subject ratio, there is relatively little information with which to estimate variable-specific parameters, and (2) technological aspects of the assays would affect many or all of the variables similarly.

Specifically, we use the following hierarchical distributions to describe the variation of

parameters across variables:

$$\begin{aligned}
\mu_j|\theta_\mu,\tau_\mu &\sim \mathcal{N}(\theta_\mu,\tau_\mu)\\
\sigma_j^{-2}|\gamma,\lambda &\sim \mathcal{G}(\gamma,\lambda)\\
\kappa_j^+|\theta_\kappa^+ &\sim \mathcal{E}(\theta_\kappa^+)\\
\kappa_j^-|\theta_\kappa^- &\sim \mathcal{E}(\theta_\kappa^-)\\
\text{logit}(\pi_j^+)|\theta_\pi^+ &\sim \mathcal{N}(\theta_\pi^+,\tau_\pi^+)\\
\text{logit}(\pi_j^-)|\theta_\pi^- &\sim \mathcal{N}(\theta_\pi^-,\tau_\pi^-)
\end{aligned}$$

where $\mathcal{G}$ is the gammma distribution, and $\mathcal{E}$ is the exponential distribution. We assume variable-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 variables allows for data-driven estimation. An advantage of the hierarchical model is that for variables which show little or no evidence of high or low values (i.e. $\pi_j^- \approx \pi_j^+ \approx 0$), 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_i. 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 an MCMC estimation procedure, in which the data are augmented with a trichotomous indicator, e_{ji} for each a_{ji}, with the additional constraint that $e_{ji} = 0$ if $c_i = 0$ (see also Diebolt and Robert, 1994 and 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 κ's, we used the following 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_{ji}'s), the full conditional distribution of the π_j's is a Dirichlet distribution, and the full conditional distribution of the parameters of the normal component is conjugate, with the additional constraint that $\sigma r < \min(\kappa_j^+, \kappa_j^-)$.

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

$$p_{ji}^+ = P(e_{ji} = 1|a_{ji},\omega) = \frac{\pi_j^+ f_{1,j}(a_{ji})}{\pi_j^+ f_{1,j}(a_{ji}) + \pi_j^- f_{-1,j}(a_{ji}) + (1-\pi_j^+-\pi_j^-) f_{0,j}(a_{ji})} \tag{2}$$

$$p_{ji}^- = P(e_{ji} = -1|a_{ji},\omega) = \frac{\pi_j^+ f_{-1,j}(a_{ji})}{\pi_j^+ f_{1,j}(a_{ji}) + \pi_j^- f_{-1,j}(a_{ji}) + (1-\pi_j^+-\pi_j^-) f_{0,j}(a_{ji})}, \tag{3}$$

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_{ji}^+ and p_{ji}^- 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 ji 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_{ji}^+ and p_{ji}^- by $p_{ji} = p_{ji}^+ - p_{ji}^-$. We refer to this new variable as the "poe scale", where poe is an acronym for "probability of expression." The transformation from a_{ji} to p_{ji} is useful because we have essentially made the data independent of the method with which the measurements were assayed. For example, the a_{ji} could be expression values from oligonucleotide arrays, or from cDNA arrays, or from other means for measuring genetic activity. Additionally, all variables 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 $J \times I$ matrix of p_{ji}'s can now be used in any clustering or other analytic method.

3 Data reduction approaches in nested unsupervised analyses

3.1 Evaluating diagnostic characteristics of variables

The goals of the analyses that follow are to find a relatively small number of variables which show variation across subjects, show consistent values within normals, and possibly show evidence of subtypes within the disease class. Constistency within variables can be assessed by examining a variable's "specificity", and evidence of subtypes across subjects can be assessed by "sensitivity." We define specificity (sp_j), sensitivity (se_j), positive sensitivity (se_j^+), and negative sensitivity (se_j^-) for variable j to be

$$\begin{aligned} sp_j &= P(\text{subject } i \text{ is classified as normal} \mid \text{subject } i \text{ is normal}) \\ se_j &= P(\text{subject } i \text{ is classified as high or low} \mid \text{subject } i \text{ is diseased}) \\ se_j^+ &= P(\text{subject } i \text{ is classified as high} \mid \text{subject } i \text{ is diseased}) \\ se_j^- &= P(\text{subject } i \text{ is classified as low} \mid \text{subject } i \text{ is diseased}) \end{aligned}$$

To calculate specificity (sp_j) and sensitivities (se_j, se_j^+, se_j^-), we define a threshold on the latent class membership probabilities and assign points to the high, normal and low categories. Specifically, we estimate the true category of variable j for subject i (e_{ji}) with $\hat{e}_{ji}$, such that

$$\begin{aligned} \hat{e}_{ji} = \quad -1 \quad & \text{if } p_{ji}^- > p_0 \\ = \quad 1 \quad & \text{if } p_{ji}^+ > p_0 \\ = \quad 0 \quad & \text{else} \end{aligned}$$

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 0.5, although other cutoffs can be

chosen, ranging from 0 to 1. The effect of choosing a cutoff closer to 0 will tend to classify more values as normal, decreasing sensitivity and increasing specificity. Choosing a cutoff closer to 1 will have the opposite effect. After choosing a threshold and categorizing each variable for each subject, we can calculate sp_j, se_j, se_j^+, and se_j^- for each variable as follows:

$$\begin{aligned} sp_j &= \frac{\sum_{i:\nu_i=1}(1-|\hat{e}_{ji}|)}{\sum_{i=1}^{I}\nu_i} \\ se_j &= \frac{\sum_{i:\nu_i=0}(|\hat{e}_{ji}|)}{\sum_{i=1}^{I}(1-\nu_i)} \\ se_j^+ &= \frac{\sum_{i:\nu_i=0} I(\hat{e}_{ji}=1)}{\sum_{i=1}^{I}(1-\nu_i)} \\ se_j^- &= \frac{\sum_{i:\nu_i=0} I(\hat{e}_{ji}=-1)}{\sum_{i=1}^{I}(1-\nu_i)} \end{aligned}$$

We are interested in variables which are consistent across normal subjects. This corresponds to choosing variables (i.e., genes) that show high specificity, sp_j. If we are also interested in variables which show evidence of subtypes of disease, then we would choose variables 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 subjects tend to have low value relative to normals. If a variable has moderate levels of both se^+ and se^-, then there is a subtype of diseased subjects that show low levels and another subtype showing high levels.

We can now reduce our dataset by choosing variables which show sufficient specificity. This level of specificity will depend to some extent on the dataset under consideration, but a specificity of 0.5 will suggest that at least half of the normals are classified as normals by the variable of interest. In the case of high throughput assays, this seemingly low threshold will usually have the effect of weeding out a very significant portion of genes (i.e., greater than half). 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 althougth we are interested in variables 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 variables which show little evidence of being related to the disease process.

3.2 Creating subsets of similar variables

Estimates of class assigment probabilities can be used for mining for genes that are likely to provide interesting subgroups of the diseased category. In the application of Section 4 we use the following mining approach, 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 subjects within a variable are expected to be low and how many to be high. For

example, the pattern $\{0.05, 0.20\}$ indicates that 5% of subjects should be low, and 20% of subjects should be high for a variable. The remaining 75% would then be in the "typical" component of the mixture.

2. Sort variables according to consistency with pattern defined in step 1. For each variable, subjects are assumed to follow the pattern from step 1. Then, using the estimates of p_{ji}^+ and p_{ji}^-, we can calculate the probability that, for variable j, the subjects have the specified pattern. We sort variables by this probability.

3. Calculate a $J \times J$ matrix of variable agreement, where r_{jk} represents agreement between variables j and k:

$$r_{jk} = \sum_{i=1}^{I} (p_{ji}^+ p_{ki}^+ + p_{ji}^- p_{ki}^- + (1 - p_{ji})(1 - p_{ki}))$$

4. Define variable "coherence" as the diagonal of the agreement matrix. Coherence for variable j is r_{jj}. Identify variables as potential "seed" variables if their coherence is above a prespecified cutoff.

5. Choose the variable with the largest score (i.e., probability) from step 2 and which is sufficiently coherent as seed variable.

6. Choose variables that show substantial agreement with the seed variable, either as a fixed agreement cutoff, or as a proportion of coherence of the seed variable. Add these variables to the "group" which is seeded by variable chosen in step 5.

7. Remove the variables in group from consideration. Repeat steps 5 and 6 to identify remaining groups.

We apply this approach to a subset of variables which have sufficient specificity and sensitivity. For each repetition of gene mining, we find homogeneous sets of variables and, for the purpose of defining molecular profiles, generally need to choose just one variable to represent the group. Some of the variables 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 measured variables are known genes but many are ESTs (expressed sequence tags) 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 variable 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 variable.

After a subset of variables has been identified using this method, we use the genes to define pattern profiles. For example, if we have chosen only two variables, then for each subject we can calculate the probability that the subject belongs to one of 8 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_{ji} and $e_{j'i}$ values for the two chosen variables, j and j'. We use the p_{ji} value for estimation of profile probabilities. For a set of k genes, there are 3^k possible patterns. Because the number of patterns get very large even for moderate k, it is generally preferred to choose relatively few predictors.

4 Identifying subclasses of lung adenocarcinoma

4.1 Data

We now illustrate the nested unsupervised methodology described so far using gene expression data that includes normal and cancerous lung samples (Bhattacharjee et al., 2001). The specimens in this dataset include 139 lung adenocarcinomas (adeno), and 17 normal lung (NL) specimens. Throughout, normal samples are indicated in figures with the symbol "*". The primary analytic goal is to indentify subgroups of adenos. 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 dataset and was chosen based on its overlap with a comparable dataset 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 artifacts, 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 4.2 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 4.2. 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 remaining, a reduction in the number of variables of about 80%.

A similarity image of samples is shown in Figure 4.2. Entries are Pearson's correlation calculated using the p_{ji} matrix of poe scores for the 1182 selected genes. We see that the subset of genes that we have chosen does a very good job of separating the normal samples from the adenocarcinomas. 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.

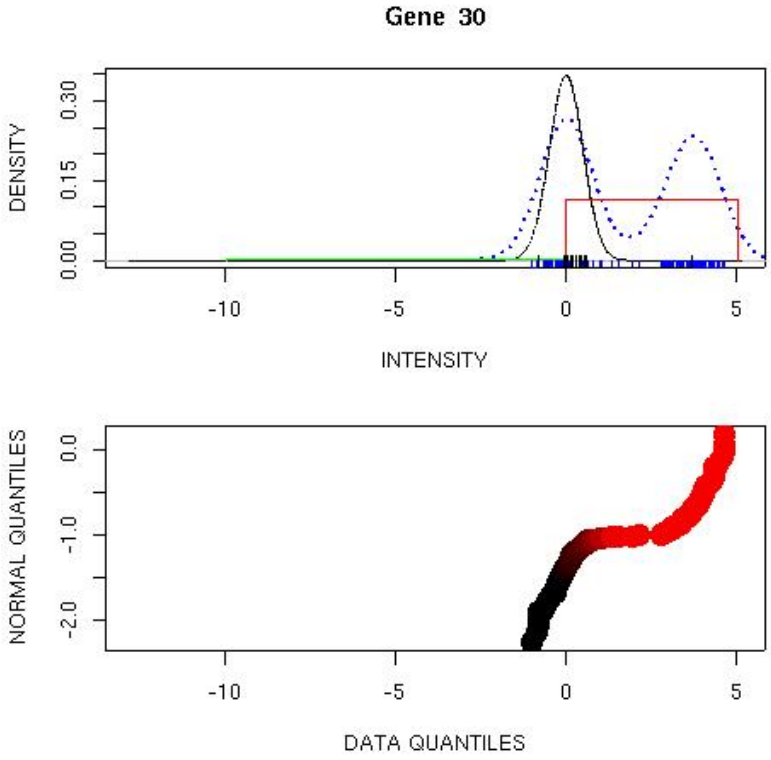

Figure 1: Estimated mixture components for gene 30. Blue vertical marks are the estimated residuals $a_{ij} - \mu_j - \alpha_i$, for the cancer samples. Black vertical marks are the corresponding residuals for normal samples. The dotted line is a kernel density estimate of the distribution of the residuals. The solid lines correspond to the best fitting uniform and normal components of the mixture, multiplied by the corresponding mixture weights (green = low, black = typical, red = high). The bottom row displays the normal quantile plot, with gray shades proportional to the probability $1 - \hat{p}_{ji}$ of being from the normal.

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 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), respectively. The genes are (1) BRCA1 (breast cancer 1), a tumor suppressor gene that is related to the famililal breast/ovarian cancer syndrome (Szabo and King, 1997) as well as other cancers; (2) Meis1 (myeloid ecotropic viral integration), which is a transcription factor known to be related to oncogenesis (Moskow et al., 1995); and (3) FGF7 (fibroblast growth factor 7), which is related to lung development (Ware and Matthay, 2002).

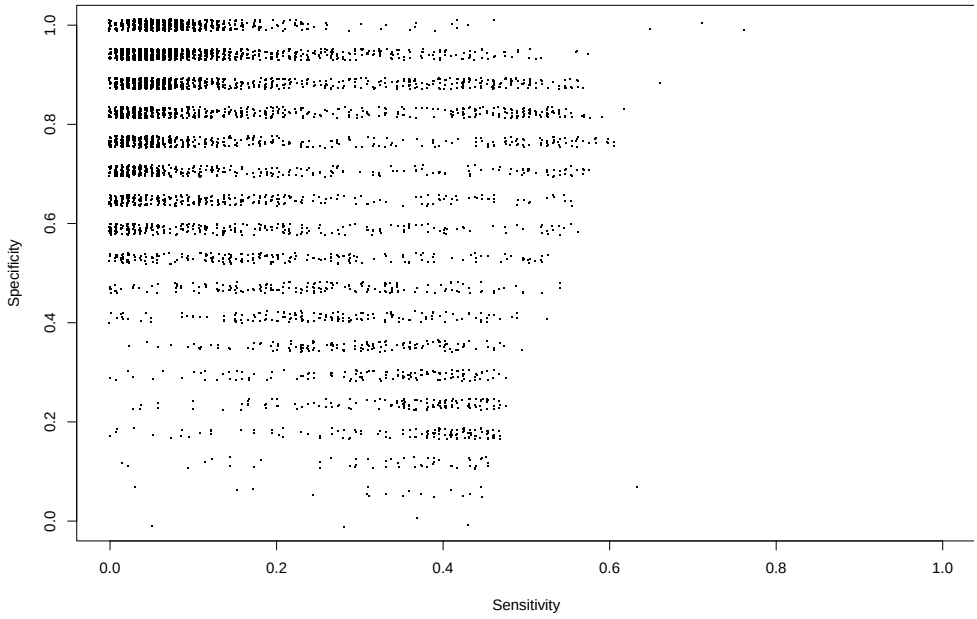

Figure 2: Scatterplot of sensitivity and overall specificity of the genes analyzed. Specificity can take on only 18 values, as there are 17 normal samples. For the purpose of this scatterplot, vertical cooridinates have been sligltly perturbed.

For each of the samples, we estimate the 3^3 profile probabilities and show this graphically in Figure 4.3 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 4.3. 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 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

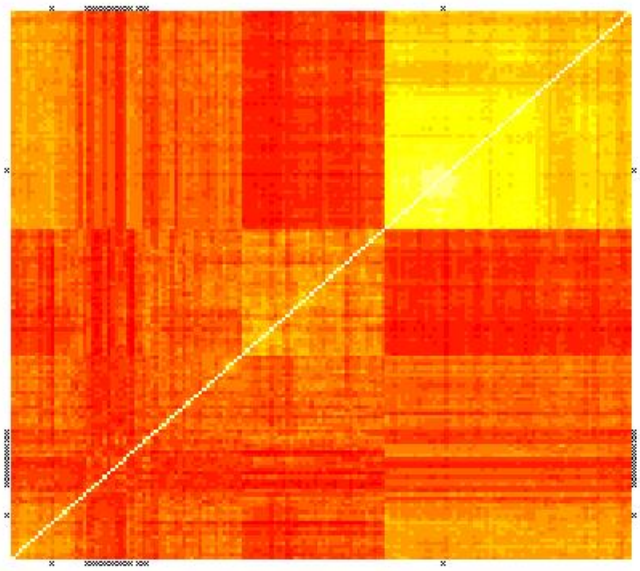

Figure 3: Pairwise Pearson's correlations of samples. Normal samples are indicated by the symbol "*" on the axes. Rows and columns have been sorted using a divisive clustering algorithm.

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 molecular profiles. This requires integration of complex modeling, 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 modeling 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

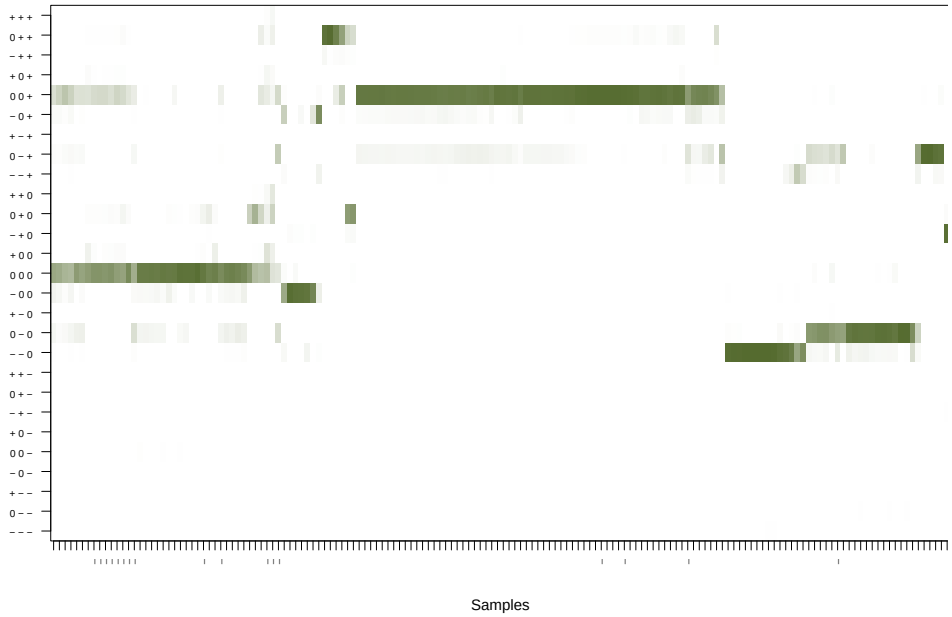

Figure 4: Molecular profiles probabilities. Each row corresponds to one of the 27 molecular profiles defined by the expression status of genes BRCA1, MEIS1, and FGF7. Each column corresponds to a sample. For example, the point for row (1,-1,0) for tumor 79 is the probability that the true expression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

screening of genes as predictors, using simple and interpretable measures such as sensitivity ans 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 is in Parmigiani et al., 2002.

Acknowledgment

Work partly supported by NCI grants P50CA88843, P50CA62924-05, DK-58757, 5P30 CA06973-39 and NIH grant HL 99-024.

References

Arminger G, Clogg CC, Sobel ME (eds.) (1995). *Latent Class Models*, chapter 6. New York: Plenum Press

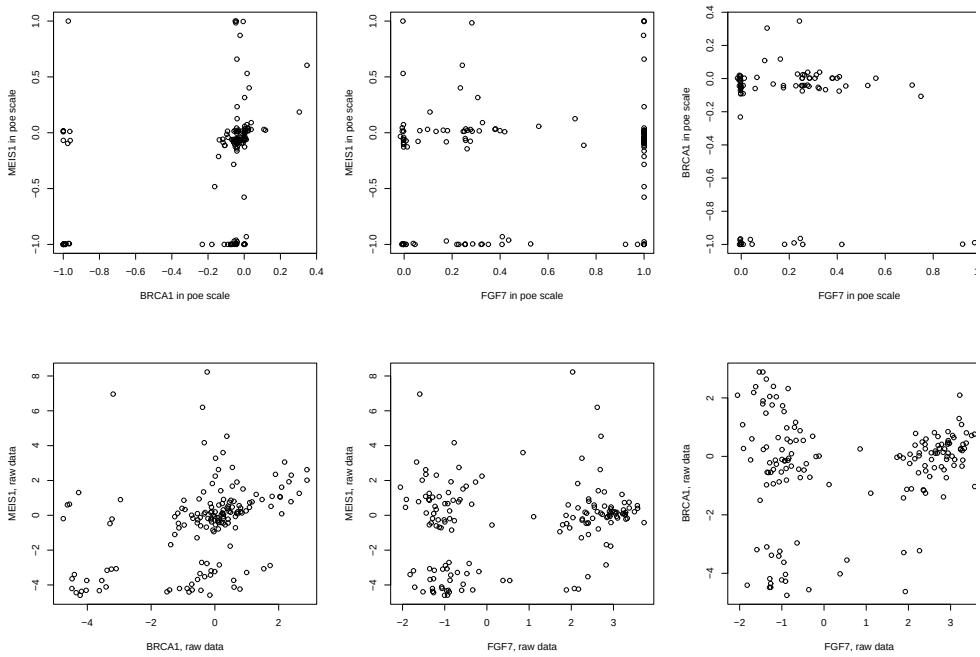

Figure 5: Scatterplots of poe scale (top row) continuous untransformed scale (bottom row) for the three genes selected for profiling (BRCA1, FGF7, and MEIS1).

Bartholomew D (1987). *Latent Variable Models and Factor Analysis*. London: Charles Griffin

Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences USA* 98:13790–13795

Diebolt J, Robert C (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Ser. B* 56:163–175

Fraley C, Raftery AE (1998). How many clusters? Which clustering method? – Answers via model-based cluster analysis. *Computer Journal* 41:578–588

Garrett ES, Parmigiani G (2003). POE: Statistical tools for molecular profiling. In: *The analysis of gene expression data: methods and software*. New York: Springer

Lee ML, Kuo FC, Whitmore GA, Sklar J (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences USA* 97(18):9834–9839

McCutcheon AL (1987). *Latent Class Analysis*. Quantitative Applications in the Social Sciences. London: Sage Publications

McLachlan GJ, Bean RW, D P (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* 18:413–422

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

Table 1: Profile assignments for 156 lung tissue samples. Profiles represent the under expression (1), normal expression (0) and under expression (-1) of genes BRCA1, MEIS1, and FGF7, respectively.

Mohr S, Leikauf GD, Keith G, Rihn BH (2002). Microarrays as Cancer Keys: An Array of Possibilities. *J Clin Oncol* 20(14):3165–3175

Moskow JJ, Bullrich F, Huebner K, Daar IO, M BA (1995). Meis1, a PBX1-related homeobox gene involved in myeloid leukemia in BXH-2 mice. *Molecular and Cellular Biology* 15:5434–43

Parmigiani G, Garrett ES, Anbazhagan R, Gabrielson E (2002). A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society, Series B* 64:717–736

Szabo CI, King MC (1997). Population genetics of BRCA1 and BRCA2. *Am J Hum Genet* 60:1013–1020

Ware LB, Matthay MA (2002). Keratinocyte and hepatocyte growth factors in the lung: roles in lung development, inflammation, and repair. *American Journal of Physiology - Lung Cellular and Molecular Physiology* 28:L924–40

West M, 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, Ruzzo W (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* 17:977–987