A comparison of cluster analysis methods using DNA methylation data

Kimberly D. Siegmund¹,*, Peter W. Laird²,3 and Ite A. Laird-Offringa²,3

¹Department of Preventive Medicine, ²Department of Surgery and ³Department of Biochemistry and Molecular Biology, Norris Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA

Received on July 15, 2003; revised on March 1, 2004; accepted on March 2, 2004
Advance Access publication March 25, 2004

ABSTRACT
Motivation: Aberrant DNA methylation is common in cancer. DNA methylation profiles differ between tumor types and subtypes and provide a powerful diagnostic tool for identifying clusters of samples and/or genes. DNA methylation data obtained with the quantitative, highly sensitive MethyLight technology is not normally distributed; it frequently contains an excess of zeros. Established tools to analyze this type of data do not exist. Here, we evaluate a variety of methods for cluster analysis to determine which is most reliable.

Results: We introduce a Bernoulli–lognormal mixture model for clustering DNA methylation data obtained using MethyLight. We model the outcomes using a two-part distribution having discrete and continuous components. It is compared with standard cluster analysis approaches for continuous data and for discrete data. In a simulation study, we find that the two-part model has the lowest classification error rate for mixture outcome data compared with other approaches. The methods are illustrated using DNA methylation data from a study of lung cancer cell lines. Compared with competing hierarchical clustering methods, the mixture model approaches have the lowest cross-validation error for detecting lung cancer subtype (non-small versus small cell). The Bernoulli–lognormal mixture assigns observations to subgroups with the lowest uncertainty.

Availability: Software is available upon request from the authors.

Contact: kims@usc.edu
Supplementary information: http://www-rcf.usc.edu/~kims/SupplementaryInfo.html

INTRODUCTION
Cancer patients can show variable response to therapy even among subgroups with similar diagnoses. Consequently, novel methods for classifying disease are sought using molecular analysis from new technologies. A lot of attention has been given to classification based on gene expression profiles measured from mRNA microarrays (Alon et al., 1999; Eisen et al., 1998; Golub et al., 1999; Hastie et al., 2000; Khan et al., 1998; McLachlan et al., 2002; Tamayo et al., 1999; Yeung et al., 2001). An alternative molecular footprint for classification is DNA methylation profiles (Model et al., 2001; Virmani et al., 2002).

DNA methylation, the addition of a methyl group to the fifth position of cytosine in the context of a CpG dinucleotide, is a modification of DNA that is important for normal organism development. Areas of genes that are rich in CpG dinucleotides (so-called CpG islands) are usually not methylated in normal tissues but frequently become hypermethylated in cancer. This hypermethylation is associated with gene silencing (Jones and Laird, 1999) and is an important mechanism for the inactivation of tumor suppressor genes. Methylation profiles differ between cancers arising in different organs (Costello et al., 2000; Esteller et al., 2001) and even between different cancer histologies from the same organ, as exemplified by methylation analyses of different subtypes of leukemia (Model et al., 2001) and lung cancer (Virmani et al., 2002). Because methylation profiling utilizes DNA, which is much more stable than the RNA used for expression profiling, methylation signatures are detectable in samples obtained from a wide variety of conditions, including formalin-fixed material, serum and sputum (Laird, 2003; Tsou et al., 2002). DNA methylation analysis promises to become a powerful tool in cancer diagnosis, with possible applications to tailoring of treatment and prognostication. Development of mathematical models to evaluate complex methylation data is therefore of prime interest.

Cluster analysis attempts to identify novel subgroups that are distinct from one another. Such methods have been applied to gene expression data for the detection of subgroups of samples with similar expression profiles across genes or for the detection of subgroups of genes with similar expression profiles across samples. Among biologists, agglomerative hierarchical cluster analysis is a popular approach. In this approach each observation starts as its own cluster and at

*To whom correspondence should be addressed.
successive steps of the algorithm, clusters are merged until only a single cluster remains. Although the approach is appealing for visualizing the data, it can be unstable for classification because once an observation is assigned to a cluster it cannot be removed. To deal with this, methods to reallocate already assigned observations have been developed. An alternative approach is to use model-based methods that split the dataset. Such methods can be more stable when the object is to find a few important clusters (Kaufman and Rousseuw, 1990). Furthermore, they may be preferred to heuristic algorithms for the purpose of selecting the number of clusters or estimating the probability with which a given observation is assigned to a given subgroup.

We compare a variety of approaches for clustering DNA methylation data obtained using the quantitative MethyLight technology. MethyLight utilizes DNA that has been treated with bisulfite, which converts unmethylated Cs to Us. This allows PCR primers to be designed so that they specifically recognize methylated versus unmethylated sequences (Eads et al., 2000). Generally, the three MethyLight primers used to analyze each locus are designed to recognize only fully methylated sequences and result in the sampling of ∼8 CpGs per locus, posing very stringent criteria for obtaining a positive methylation signal. Because of this, samples in which all DNA strands are not or only partially methylated in the tested region will be negative, while samples showing methylation will yield a number greater than zero, indicating the level of methylation compared with a fully methylated control DNA. Thus, the DNA methylation data obtained with MethyLight has the unique characteristic of being a quantitative measure with an ‘excess’ of zeros; the measurements for any single locus can follow a mixture distribution of discrete and continuous observations. Such DNA methylation profiles can be very informative for classifying samples because the proportion of methylated samples, the mean of the (positive) methylation levels or both may differ between subgroups. To capitalize on these dual characteristics, we propose to model the outcome profiles could distinguish between two subtypes of lung cancer, non-small cell and small cell. The analysis was limited to a subset of 7 out of 24 CpG regions. Each of the seven CpG regions was predictive of lung cancer subtype. Although feature selection is an important aspect of class discovery models, we focus on the comparison of methods for mixture outcome data using these same seven loci.

In a simulation study, we compare our new approach for clustering mixture outcome data when the data are not normally distributed, even after standard transformations, to standard clustering methods for continuous variables and for discrete variables. We then apply the methods to DNA methylation data from a study of lung cancer (Virmani et al., 2002).

**METHODS**

DNA methylation is measured using MethyLight, a quantitative real-time PCR technique (Eads et al., 2000). The technology measures the frequency of molecules in which a series of CpG sites (usually ∼8 sites) in a given CpG region are methylated. The individual measurements are normalized for the total amount of DNA present and the proportion of fully methylated molecules from an enzymatically methylated sample (SssI-treated sperm DNA) using two different control genes. The resulting measurement is the percentage of methylated reference (PMR). The distribution of PMR values follows a mixture of discrete and continuous observations.

A dataset consists of DNA measurements on a sample of \( N \) subjects at \( FCpG \) regions (features). The outcome is displayed in an \( N \times F \) matrix where each row denotes a subject and each column a feature.

**Lung cancer data**

Cluster analysis methods are compared using DNA methylation profiles from 87 lung cancer cell lines. The primary analysis of the data is described in a paper by Virmani et al. (2002). The goal was to demonstrate whether DNA methylation profiles could distinguish between two subtypes of lung cancer, non-small cell and small cell. The analysis was limited to a subset of 7 out of 24 CpG regions. Each of the seven CpG regions was predictive of lung cancer subtype. Although feature selection is an important aspect of class discovery models, we focus on the comparison of methods for mixture outcome data using these same seven loci.

**Agglomerative hierarchical cluster analysis**

Agglomerative hierarchical cluster analysis is a method that starts with each observation as its own cluster, and slowly merges similar clusters of observations until only one cluster remains. Heuristic and model-based approaches are available. Heuristic approaches merge clusters based on distance functions. The first merge is between the two observations with the smallest distance between them. Let \( D(y_i, y_j) \) denote the distance between two multivariate observations, \( y_i \) and \( y_j \). Several distance metrics are available. We use Euclidean distance for continuous (or mixture) outcome measures and Manhattan distance for discrete outcome. Euclidean distance measures the sum of squared deviations between two observations at each feature \( f \), \( D(y_i, y_j) = \sum_f (y_{if} - y_{jf})^2 \), and Manhattan distance sums the absolute deviations, \( D(y_i, y_j) = \sum_f |y_{if} - y_{jf}| \). For clusters with more than
one observation, distance is measured as a function of all pairs of observations from different clusters. Different options include the minimum (single linkage), maximum (complete linkage) or average distance (average linkage). In our analyses, we use average linkage. The procedure is fit using the function agnes() in S-PLUS version 6.1 (SPLUS, 2002, http://www.insightful.com/products/default.asp).

Model-based hierarchical cluster analysis merges observations to increase the classification likelihood. Let \( y_1, y_2, \ldots, y_n \) denote independent multivariate observations and \( \ell_1, \ell_2, \ldots, \ell_n \) are labels indicating the groups to which each observation belongs, e.g. \( \ell_i = k \) if the \( i \)-th observation belongs to group \( k \). Observations in group \( k \) are modeled using a multivariate normal distribution, \( f_k \sim N(\mu_k, \Sigma_k) \). Then the classification likelihood is

\[
CL(\theta_1, \ldots, \theta_K; \ell_1, \ldots, \ell_n \mid y_1, \ldots, y_n) = \prod_{i=1}^{n} f_{\ell_i}(y_i \mid \theta_{\ell_i}),
\]

where \( \theta_k \) denotes the parameters for group \( k \).

Banfield and Raftery (1993) develop a general framework for clustering under multivariate normal distributions. The covariance matrix is modeled using the eigenvalue decomposition, \( \Sigma_k = \lambda_k D_k A_k D_k^T \), where \( \lambda_k \) is a scalar, \( D_k \) the matrix of eigenvectors and \( A_k \) a diagonal matrix with entries proportional to the eigenvalues of \( \Sigma_k \). The components of the decomposition, \( \lambda_k, D_k \) and \( A_k \), influence the volume, orientation and shape of observations from the different groups, respectively. The most constrained model with the smallest number of parameters is given by \( \Sigma_k = \lambda I \), where \( I \) is the identity matrix. This corresponds to a spherical model with equal volume for each group (EI). The inequality volume spherical model is parameterized by \( \Sigma_k = \lambda_k I \) (VI). The most general unconstrained model is given by \( \Sigma_k = \lambda_k D_k A_k D_k^T \), where each of the \( \lambda_k, D_k \) and \( A_k \) may vary between different groups. These models can be fit using MCLUST, a function written for S-PLUS and available at http://www.stat.washington.edu/raftery/mclust (Fraley and Raftery, 1999).

Mixture models

In mixture models we assume that the data belong to one of the fixed number of clusters and the variable denoting group membership is unobserved. McLachlan and Basford (1988) and McLachlan and Peel (2000) give a general overview of these models. Within each group, the observations follow a specified probability distribution function. The likelihood is

\[
L(\theta_1, \ldots, \theta_K; \pi_1, \ldots, \pi_K \mid y_1, \ldots, y_n) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(y_i \mid \theta_k),
\]

where \( \pi_k \) is the proportion of observations belonging to group \( k (\pi_k \geq 0 \) and \( \sum_{k=1}^{K} \pi_k = 1) \). Generally, the likelihood is maximized using the EM algorithm. After convergence, clusters can be formed by classifying observations to the group for which they have the highest posterior probability of membership. For each observation, one minus the maximum of the posterior probabilities across groups gives a measure of class uncertainty.

Gaussian mixture model In the Gaussian mixture model, \( y_i \sim N(\mu_k, \Sigma_k) \). The covariance matrix \( \Sigma_k \) is modeled using the same eigenvalue decomposition of the covariance matrix described for the model-based hierarchical cluster analysis. The Gaussian mixture model is fit using the function MCLUST (2002 version). Also available for fitting mixtures of normals is the EMMIX program of McLachlan et al. (1999).

Bernoulli mixture model (latent class model) A common name for the Bernoulli mixture model is latent class model. Here, we model the dichotomous observations denoting methylation status, positive versus negative. In the latent class model \( f_k \sim \text{Bernoulli}(p_{k,f}) \), where \( p_{k,f} \) is the methylation probability at feature \( f \) in class \( k \). We denote the methylation status across the features for the \( i \)-th subject using a multivariate vector \( w_i, i = 1, \ldots, n \). For subject \( i \) at feature \( f, w_{if} \) is 1 if methylated and 0 if not methylated. We assume that conditional on class membership each feature is methylated independently of all others and write the likelihood

\[
L(p_1, \ldots, p_K; \pi_1, \ldots, \pi_K \mid w_1, \ldots, w_n) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \prod_{f=1}^{F} p_{k,f}^{w_{if}} (1 - p_{k,f})^{1-w_{if}}.
\]

Since the variance is a function of the methylation probability at each feature and these probabilities are conditionally independent within each cluster this model corresponds to a diagonal variance function with unequal variances.

We maximize the likelihood using the EM algorithm, estimating the methylation probabilities \( p_{k,f} \) on the logit scale. The EM algorithm is converged when the change in the mixture likelihood is less than \( 1 \times 10^{-6} \). Because of the problem of multiple maximum, we use 11 different starting values and report results for the solution that yields the largest log likelihood. For 10 runs, subjects are assigned initially to subgroups at random. In the eleventh run, we assign subjects to the classification obtained using agglomerative hierarchical cluster analysis.

Bernoulli–lognormal mixture model In the Bernoulli–lognormal mixture model, we model each outcome using a mixture distribution for the discrete and continuous components. Currently, we consider only the simplest variance structure where conditional on subgroup \( k \), the methylation levels are independent across loci. Letting \( z_{if} = \ln y_{if} \) if
the best model for the data (BIC). Bayesian information criterion (BIC) can be used to select such as agglomerative hierarchical cluster analysis is that the lung cancer dataset. Each dataset contains 609 observations: data for two groups using the sample sizes reflected in the feature of log-normal and discrete observations. We generate model-based clustering approaches when the data are a mix-

The Bernoulli–lognormal mixture model is fit using the EM algorithm and multiple starting values. Initial class assignments are generated using three methods: the Gaussian mixture model with $\Sigma_k = \lambda I$ and $\Sigma_k = \lambda_k I$ and the latent class model with measurements dichotomized into positive versus zero. Results are reported for the solution yielding the maximum log-likelihood of the three runs. The same convergence criterion is used as for the latent class model (change in log likelihood $< 1 \times 10^{-6}$).

An advantage of a model- over rule-based approaches such as agglomerative hierarchical cluster analysis is that the Bayesian information criterion (BIC) can be used to select the best model for the data (BIC $= -2 \times$ log-likelihood + number of parameters $\times$ ln(number of observations)). The BIC permits the comparison of models that have different numbers of clusters and are not nested. The best model gives the lowest BIC value. Convention dictates that if the difference in BIC values for two models is greater than 2 there is positive evidence for model differences. The evidence is considered strong for differences between 6 and 10 and very strong for differences greater than 10 (Fraley and Raftery, 1998; Kass and Raftery, 1995).

**Simulation study**

We designed a simulation study to evaluate the different model-based clustering approaches when the data are a mixture of log-normal and discrete observations. We generate data for two groups using the sample sizes reflected in the lung cancer dataset. Each dataset contains 609 observations: 87 subjects ($n_1 = 46$, $n_2 = 41$) measured at seven CpG regions. We assume that within each group, the measurements at each CpG region are independent of those at other regions. This is consistent with what is observed for the CpG regions studied in the lung cancer dataset.

We generate data under two scenarios. In the first, the within-group methylation values follow the same distribution at each CpG region and in the second, the within-group distributions across loci are patterned after those observed in the lung cancer dataset. For each scenario, we start by creating a dataset with log-normal distributions at each CpG region and then create mixture distributions by reassigning to zero with a given probability the measurements from the log-normal distribution to zero. In the first scenario, data are generated from a log-normal distribution with SD of 1 and difference in mean value of 1.5 between the two groups for data on the natural-log scale. Four different mixture outcome models are created using this data. Let $p_1$ denote the probability a measurement in group 1 is assigned a value zero and $p_2$ the corresponding probability in group 2. We assign zeros with the following probabilities for the four mixture distributions: (i) $p_1 = p_2 = 0.2$, (ii) $p_1 = p_2 = 0.4$, (iii) $p_1 = 0.2$, $p_2 = 0.6$ and (iv) $p_1 = 0.6$, $p_2 = 0.2$. Distributions (i) and (ii) have equal frequency of zeros in the two unobserved subgroups. Distributions (iii) and (iv) have the same mean number of zeros as situation (ii), but their frequency differs in the two groups. The distribution of zeros across the two subgroups either enhances [situation (iii) or diminishes situation (iv)] the difference in the group mean at a locus compared to the difference computed using the positive values only.

In the simulations patterned after the lung cancer dataset, we create only two models. The first uses only the log-normal distributions with means and SDs presented in Table 1. These distributions correspond to the seven features measured for two classes of lung cancer, small-cell and non-small cell. In the second we create mixture outcome using discrete and log-normal distributions, assigning observations the value of zero using the probabilities of non-negative PMR values. For each model, we generate 100 replicate datasets.

**Data transformations**

We compare results from three different approaches for dichotomizing the data for the discrete variable methods: (1) zero versus positive values, (2) above and below the median value and (3) above and below the median of the positive values. For the continuous data approaches, the measurements are analyzed on the log-scale. The data are transformed by adding one to the measurement and taking the natural log. For the outcomes following a mixture distribution, transforming

---

### Table 1. Distribution of log-transformed PMR values for seven CpG regions in lung cancer cell lines

<table>
<thead>
<tr>
<th>CpG region</th>
<th>Non-small cell lung cancer</th>
<th>Small cell lung cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% with Positive PMR</td>
<td>Mean</td>
</tr>
<tr>
<td>ESRI</td>
<td>74%</td>
<td>2.590</td>
</tr>
<tr>
<td>APC</td>
<td>74%</td>
<td>3.556</td>
</tr>
<tr>
<td>MGMT-M1</td>
<td>83%</td>
<td>0.786</td>
</tr>
<tr>
<td>PTGS2</td>
<td>78%</td>
<td>-0.084</td>
</tr>
<tr>
<td>MTHFR</td>
<td>98%</td>
<td>3.803</td>
</tr>
<tr>
<td>MYOD1</td>
<td>65%</td>
<td>1.998</td>
</tr>
<tr>
<td>CALCA</td>
<td>76%</td>
<td>2.929</td>
</tr>
</tbody>
</table>

Pooled estimate of SD = 2.479.

*a Mean and SD of positive PMR values on natural-log scale [ln(PMR)].
the data helps brings in the right tail, but still suffers from having an excess of zeros. After transformation each feature is standardized giving them equal importance in the analysis. In the analysis using the Bernoulli–lognormal mixture model, the positive measures are log-transformed and standardized ignoring the zeros.

We compare the classification error rate of the different clustering methods for the different data distributions in a simulation study. In addition, we present parameter estimates for the Bernoulli–lognormal mixture model. For the analysis of the lung cancer dataset, we compute the leave-one-out cross-validation error. This is obtained by running the cluster analysis $n$ times in each instance omitting a different observation. For the mixture models the omitted observation is assigned to a cluster based on the highest posterior probability. For the hierarchical models, it is assigned based on the smallest distance from the cluster mean. Distance is computed by the sum across features of the absolute deviation of the observation from the cluster mean. The proportion of total observations misclassified gives the cross-validation error.

### RESULTS

We applied the different clustering routines to the simulated data, assigning samples into two subgroups. Table 2 shows the misclassification rates. The Bernoulli–lognormal model showed the lowest classification error rate across all designs. When the data were log-normally distributed, it performed as well as the Gaussian mixture model (on the log-transformed data). When zeros were added to the data, it showed a lower error rate than the other methods. The classification error rate increased for all methods including the Bernoulli–lognormal model when the data included zeros that were non-informative for class membership reflecting a decrease in the number of informative observations. The increase in error rate was greater for the continuous data approaches, presumably due to model misspecification.

When the distribution of zeros was informative for the subgroup, further differences among the methods were observed. Although the Bernoulli–lognormal (two-part) model was not influenced by the direction of the differences in the proportion of zeros between the two subgroups and the difference in mean levels among the positive values, the continuous data approaches were. The greatest difference in misclassification rates between the Bernoulli–lognormal model and the competing continuous data methods was observed when the proportion of zeros was higher in the group with the higher average methylation value among positive values (Design 1 mix 4). For this case, the difference in overall means from the two groups is smaller than the difference in the means from just the positive values. In this event, the two-part model that models the difference in proportions and difference in means separately was superior in assigning observations to classes. When the proportion of zeros was lower in the group with the higher methylation values among positive values (Design 1 mix 3), the Bernoulli–lognormal approach was only slightly more accurate than the competing mixture–model approaches. This time the difference in overall means from the two groups is larger than the difference in means of the positive values for the continuous data approaches and this increase in signal seemed to overcome the model misspecification.

<table>
<thead>
<tr>
<th>Simulated distribution</th>
<th>Agglomerative HCA average linkage</th>
<th>Gaussian HCA-EI</th>
<th>Gaussian HCA-VI</th>
<th>Bernoulli MM-EI</th>
<th>Bernoulli MM-VI</th>
<th>Bernoulli log-normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design 1$^b$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log-normal ($p_1 = p_2 = 0$)</td>
<td>13.3</td>
<td>4.7</td>
<td>4.8</td>
<td>2.6</td>
<td>2.5</td>
<td>6.9</td>
</tr>
<tr>
<td>Mix 1 ($p_1 = p_2 = 0.2$)</td>
<td>45.9</td>
<td>19.6</td>
<td>16.1</td>
<td>20.3</td>
<td>11.8</td>
<td>15.6</td>
</tr>
<tr>
<td>Mix 2 ($p_1 = p_2 = 0.4$)</td>
<td>47.2</td>
<td>32.1</td>
<td>23.9</td>
<td>35.5</td>
<td>20.2</td>
<td>39.7</td>
</tr>
<tr>
<td>Mix 3 ($p_1 = 0.2; p_2 = 0.6$)</td>
<td>20.4</td>
<td>7.2</td>
<td>7.1</td>
<td>5.7</td>
<td>5.1</td>
<td>15.6</td>
</tr>
<tr>
<td>Mix 4 ($p_1 = 0.6; p_2 = 0.2$)</td>
<td>48.5</td>
<td>43.3</td>
<td>40.1</td>
<td>45.2</td>
<td>42.3</td>
<td>16.2</td>
</tr>
<tr>
<td>Design 2$^c$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log-normal ($p_1 = p_2 = 0$)</td>
<td>47.6</td>
<td>27.5</td>
<td>27.4</td>
<td>27.5</td>
<td>30.8</td>
<td>31.6</td>
</tr>
<tr>
<td>Mix ($p_1 = p_1; p_2 = p_2$)</td>
<td>48.1</td>
<td>29.1</td>
<td>30.8</td>
<td>32.9</td>
<td>36.7</td>
<td>43.6</td>
</tr>
</tbody>
</table>

*Seven CpG regions measured on each observation. Sample sizes are 46 observations in class 1 and 41 observations in class 2.*

*Data are simulated from a log-normal distribution with mean $\mu_1 = 3.0$ (class 1), $\mu_2 = 1.5$ (class 2), SD = 1.0 on the log-scale. Proportions of zeros are listed for each distribution.*

*Methods: Agglomerative hierarchical cluster analysis (HCA) (Euclidean distance, average linkage); HCA-EI model-based hierarchical cluster analysis, equal spherical variance; HCA-VI, model-based hierarchical cluster analysis, unequal spherical variance; MM-EI, Gaussian mixture model, equal spherical variance; MM-VI, Gaussian mixture model, unequal spherical variance; positive versus 0, data dichotomized into positive versus negative; >Med. versus ≤Med., data dichotomized into above and below median; >Med.(pos) versus ≤Med.(pos), data dichotomized into above and below the median of the positive values.*
in the two cell types (78% versus 75%), the pattern of DNA
measurements averaged over the seven CpG regions is similar
cancer cell lines. Although, the proportion of positive PMR
at seven CpG regions in non-small cell and small cell lung
model misspecification.

For the discrete data methods, dichotomizing at the median
of the positive values resulted in a lower misclassification rate
than dichotomizing at the median of all values when the distri-
bution of zeros was not informative of class. When the zeros
were informative and the proportion of zeros was higher in
the group with the higher average methylation value among
positive measures (Design 1 mix 4), it was best to categorize
the data into no methylation versus positive methylation. This
suggests that the signal comparing the proportion of zeros was
greater than the signal coming from the difference in mean
values.

In general, the model-based approaches had lower misclas-
sification rates than the heuristic hierarchical cluster analysis
approach using average linkage; however, this difference was
less striking for the discrete data compared with the continu-
ous data (data not shown). Among the model-based Gaussian
approaches, when zeros were added to the dataset the error
rate was lower under the unequal spherical variance structure
\( \Sigma_k = \lambda_k I \) compared to the equal spherical variance structure (\( \Sigma_k = \lambda I \)). We interpret this as a method for absorbing
model misspecification.

Table 1 shows the distribution of methylation measurements
at seven CpG regions in non-small cell and small cell lung
cancer cell lines. Although, the proportion of positive PMR
measurements averaged over the seven CpG regions is similar
in the two cell types (78% versus 75%), the pattern of DNA
methylation across loci varies by cell type. Some CpG regions
are more frequently methylated in small cell lung cancer than
in non-small cell (PTGS2) while for other CpG regions the
opposite is true (MYOD1). The absolute difference in the
proportion of positive PMR varies from 2% to 26% with an
average of 13%. Differences in levels of methylation are also
observed between cell types. Among the methylated samples,
PTGS2, CALCA, MTHFR and ESR1, all show higher methyl-
ation levels in small cell lung cancer (Wilcoxon \( p < 0.05 \)).
MGMT-M1 on the other hand shows higher average methyla-
tion levels in non-small cell lung cancer (Wilcoxon \( p = 0.04 \)).
For these seven loci, the cell type having the greater propor-
tion of positive PMR values at a given locus also has the higher
mean of the positive measures. Exceptions have been observed
for other CpG regions (Virmani et al., 2002).

In the simulations using the distributions from the lung can-
cer data, all the mixture-model approaches roughly showed
similar classification error rates (Table 2, Design 2). This is
consistent with the previous results where the direction of
effects is the same for the discrete and continuous data com-
ponents and may reflect the distribution of the features selected
for modeling.

Table 3 shows the parameter estimates from the Bernoulli–
lognormal model. In Table 3, the distributions reflect the
estimates from all features combined for the models in
Design 1 and feature-specific distributions of the models in

### Table 3. Parameter estimates for Designs 1 and 2, using the Bernoulli–lognormal mixture model on 100 datasets (mean ± SD). For Design 1, estimates are averaged across all seven features

<table>
<thead>
<tr>
<th>Design 1</th>
<th>Proportion positive in Group 1</th>
<th>Proportion positive in Group 2</th>
<th>Group 1 mean</th>
<th>Group 2 mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lognormal</td>
<td>1.0</td>
<td>1.0</td>
<td>3.00 ± 0.15</td>
<td>1.50 ± 0.16</td>
<td>0.98 ± 0.06</td>
</tr>
<tr>
<td>Mix 1</td>
<td>0.82 ± 0.06</td>
<td>0.82 ± 0.07</td>
<td>3.00 ± 0.18</td>
<td>1.51 ± 0.19</td>
<td>0.98 ± 0.07</td>
</tr>
<tr>
<td>Mix 2</td>
<td>0.6 ± 0.08</td>
<td>0.6 ± 0.08</td>
<td>3.00 ± 0.21</td>
<td>1.49 ± 0.24</td>
<td>0.97 ± 0.06</td>
</tr>
<tr>
<td>Mix 3</td>
<td>0.81 ± 0.07</td>
<td>0.40 ± 0.08</td>
<td>3.00 ± 0.17</td>
<td>1.50 ± 0.28</td>
<td>0.97 ± 0.07</td>
</tr>
<tr>
<td>Mix 4</td>
<td>0.40 ± 0.07</td>
<td>0.81 ± 0.07</td>
<td>3.00 ± 0.26</td>
<td>1.50 ± 0.18</td>
<td>0.97 ± 0.09</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Design 2</th>
<th>Proportion positive in Group 1</th>
<th>Proportion positive in Group 2</th>
<th>Group 1 mean</th>
<th>Group 2 mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lognormal</td>
<td>Feature 1</td>
<td>1.0</td>
<td>1.0</td>
<td>0.29 ± 0.21</td>
<td>−0.48 ± 0.30</td>
</tr>
<tr>
<td>Feature 2</td>
<td>1.0</td>
<td>1.0</td>
<td>0.29 ± 0.23</td>
<td>−0.51 ± 0.28</td>
<td></td>
</tr>
<tr>
<td>Feature 3</td>
<td>1.0</td>
<td>1.0</td>
<td>0.33 ± 0.14</td>
<td>−0.69 ± 0.44</td>
<td></td>
</tr>
<tr>
<td>Feature 4</td>
<td>1.0</td>
<td>1.0</td>
<td>0.15 ± 0.14</td>
<td>−0.33 ± 0.37</td>
<td></td>
</tr>
<tr>
<td>Feature 5</td>
<td>1.0</td>
<td>1.0</td>
<td>0.13 ± 0.19</td>
<td>0.20 ± 0.34</td>
<td></td>
</tr>
<tr>
<td>Feature 6</td>
<td>1.0</td>
<td>1.0</td>
<td>0.18 ± 0.18</td>
<td>0.33 ± 0.34</td>
<td></td>
</tr>
<tr>
<td>Feature 7</td>
<td>1.0</td>
<td>1.0</td>
<td>0.29 ± 0.24</td>
<td>0.42 ± 0.23</td>
<td></td>
</tr>
</tbody>
</table>

For other features, see Table 3.
Table 4. Error rates for different clustering algorithms using data on lung cancer cell lines

<table>
<thead>
<tr>
<th>Y coding</th>
<th>Methoda</th>
<th>Classification error</th>
<th>Cross-validation error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardized ln(PMR + 1)</td>
<td>Average linkage</td>
<td>42 (48.3%)</td>
<td>41 (47.1%)</td>
</tr>
<tr>
<td></td>
<td>HCA-EI</td>
<td>15 (17.2%)</td>
<td>22 (25.3%)</td>
</tr>
<tr>
<td></td>
<td>HCA-VI</td>
<td>22 (25.3%)</td>
<td>24 (27.6%)</td>
</tr>
<tr>
<td></td>
<td>Mixture model-EI</td>
<td>17 (19.5%)</td>
<td>17 (19.5%)</td>
</tr>
<tr>
<td></td>
<td>Mixture model-VI</td>
<td>18 (20.7%)</td>
<td>17 (19.5%)</td>
</tr>
<tr>
<td>0, standardized ln(PMR) &gt; 0</td>
<td>Bernoulli–lognormal mixture</td>
<td>20 (23.0%)</td>
<td>21 (24.1%)</td>
</tr>
<tr>
<td>0,1b</td>
<td>Average linkage</td>
<td>41 (47.1%)</td>
<td>41 (47.1%)</td>
</tr>
<tr>
<td></td>
<td>Mixture model (LCA)</td>
<td>19 (21.8%)</td>
<td>21 (24.1%)</td>
</tr>
</tbody>
</table>

aHCA, hierarchical cluster analysis; EI, equal spherical variance; VI, unequal spherical variance; and LCA, latent class analysis.
b0, less than or equal to median of positive values; and 1, greater than median of positive values.

Design 2. The estimates appear unbiased for both the proportion of positive values and their means. As the proportion of zeros at a locus is increased, the variability of the estimated mean of the positive values increases. Due to standardization of the data prior to analysis, the mean estimates in Table 3 are different from the true simulation values. Nonetheless, the pattern of differences between the two groups is consistent with the simulation values. Large variability in the parameter estimates is consistent with the high classification error observed for Design 2.

Table 4 presents the classification and cross-validation error rates for the different cluster analysis methods using the lung cancer dataset. Using the BIC criterion, both the Gaussian mixture model and latent class model give very strong evidence in favor of the two-cluster model. In general, the model-based approaches perform similarly having classification error rates around 20% and cross-validation error rates around 25%. The Gaussian mixture models have the lowest cross-validation error (17/87 observations, 19.5%). The Gaussian equal variance model estimated 80% sensitivity (33/41) and 80% specificity (37/46) for identifying small cell lung cancer. Using the Bernoulli–lognormal model, the mean of the positive values was larger for the subgroup having more zeros in two of the seven CpG regions. This is consistent with the Bernoulli–lognormal model detecting different clusters than the Gaussian models. The heuristic approaches perform extremely poorly on these data misclassifying nearly half of the observations.

Figure 1 shows a heatmap of the data. The CpG regions are ordered by the listing in Table 1. Subjects are ordered by the posterior probability from the Bernoulli–lognormal mixture-model; subjects 1–40 are assigned to cluster one and subjects 41–87 to cluster two. The raw data are provided in Supplemental Table 1.

We compare the posterior probabilities of group membership for three methods: the Gaussian mixture model with equal spherical variances (EI), the latent class model dichotomizing the median of the positive values and the Bernoulli–lognormal mixture model. The three approaches give the same predictions for all but 10 observations (six non-small cell,
four small cell lung cancer). For 5 of the 10 observations, the uncertainties are >20% for all three clustering approaches (uncertainty is defined as one minus the maximum posterior probability of class membership). This shows a general consistency between the results from the three methods. Summing the uncertainty values over all observations we find the lowest total uncertainty using the Bernoulli–lognormal model. All three approaches assign 62–63% of the observations with uncertainty below 5%. The Bernoulli–lognormal model assigns the fewest observations with uncertainties >20% (9 out of 87 observations compared to 14 or 15 out of 87 observations for the competing methods).

**DISCUSSION**

We introduce a new model for clustering data that resembles a mixture distribution of a continuous variable with an excess of zeros and compare it with the several standard clustering methods. In a simulation study, we found the Bernoulli–lognormal mixture model consistently yielded a lower classification error than competing approaches for data with an excess of zeros. When both the distribution of the zeros and the positive values were informative of group membership, the error rate of the dichotomous and continuous data methods was strongly influenced by the direction of the differences. If including the zeros gave a larger difference in group means than a comparison of the positive values only, the continuous data methods performed well despite model misspecification. On the other hand, if the difference in group means was attenuated by including the zeros, the methods performed extremely poorly. Only the two-part model that modeled the means and proportions separately was not affected by the direction of the differences in means and frequency of zeros between the two groups. This finding is consistent with previous work by Lachenbruch comparing two-sample test statistics (Lachenbruch, 2001). When the continuous measures are themselves highly variable, such as for the data simulated to resemble the lung cancer data, the addition of zeros to the distribution had less impact on the classification error rate for any of the methods studied.

For the analysis of the lung cancer data, we found the error rates from the model-based approaches were similar to the error rate reported by Virmani et al. (2002). In that paper, the authors used agglomerative hierarchical cluster analysis with average linkage, but categorized the measurements at each CpG region into three groups instead of two. The similarity in these results may be due to our analysis of features having differences in the means and the proportions of positive values in the same direction. Superiority of the Bernoulli–lognormal model for classification would be expected with the inclusion of features where the differences in means and proportions occur in opposite directions. Data on more CpG regions and more types of samples will tell us how common such distributions might be.

A possible improvement to the two-part model might be to reduce the number of parameters. One approach would be to use a truncated log-normal distribution where the data are considered to be truncated below some lower threshold of detection. A second method would be to model the proportion of zeros as a function of the mean of the positive values. Mixing of the component submodels has been proposed in the regression model context by researchers fitting models that mix zeros with a gamma distribution (Feuerverger, 1979) or mix zeros with a Poisson (Lambert, 1992). Currently, we are pursuing both these alternate approaches.

It is well known that feature selection will affect the clusters identified. When the goal is to identify novel subgroups of disease using large numbers of features, methods of data reduction will be needed. Our approach is to select loci that are differentially methylated in tumor and normal tissue taken from the margin. These may be most informative for the disease process. If we do not have normal tissue available, as is the case in the lung cancer study, we may select variables that show the highest coefficient of variation. As an alternate approach to variable selection, we are developing a two-dimensional cluster analysis of samples and features. Clustering features will allow data reduction without losing the mixture distributions that characterize the measurements.

Heterogeneity from the unobserved clusters induces correlation among measures within an individual. In our model-based approaches we assume conditional independence of the multivariate measurements within an individual given cluster membership. Although this appears reasonable for the lung cancer data, it may not always be appropriate. For the continuous data models, different covariance structures permit the modeling of within group correlation. For the discrete data models, the assumption can be relaxed with the addition of random effects or variable cross-products to the model (Hadgu and Qu, 1998; Yang and Becker, 1997).

For any clustering result, we will always be faced with the challenge of model validation. When studying patient tissue samples, validation can be confirmed by studying individual exposures and outcomes. Exposures such as smoking history might be predictors of DNA methylation profiles. Further DNA methylation profiles may predict outcomes such as survival or response to treatment. An advantage of a model-based clustering is the possibility for adding exposures and outcomes for the identified subgroups.

We propose a two-part model for clustering data with an excess of zeros. In an analysis of DNA methylation data from lung cancer cell lines, the method shows the lowest overall uncertainty in class assignment compared to standard approaches that either misspecify the distribution of the data, or reduce the information content. The approach is model-based, allowing us to estimate parameters and probabilities of group assignment, and to apply the BIC to select a number of clusters in the data.
ACKNOWLEDGMENTS
This work was supported by NIH CA097346.

REFERENCES