Bayesian Model-Averaging in Unsupervised Learning From Microarray Data

Mario Medvedovic
University of Cincinnati Med. Ctr
Department of Environmental Health
Cincinnati, OH 45221-0056
Mario.Medvedovic@uc.edu

Junhai Guo
University of Cincinnati Med. Ctr.
Department of Environmental Health
Cincinnati, OH 45221-0056
guojs@ucmail.uc.edu

ABSTRACT

Unsupervised identification of patterns in microarray data has been a productive approach to uncovering relationships between genes and the biological process in which they are involved. Traditional model-based clustering approaches as well as some recently developed model-based mining approaches for integrating genomic and functional genomic data rely on one's ability to determine the correct number of clusters or modules in the data. In this paper we demonstrate that the performance of such methods in general can be significantly improved by accounting for uncertainties inherent to the process of identifying the optimal number of clusters in the data. We demonstrate that the Bayesian averaging approach to clustering via infinite mixture model offers a more robust performance than the traditional finite mixture model in which the optimal number of clusters is determined using the Bayesian Information Criterion. This performance improvement is demonstrated through a simulation study and by the analysis of a relatively large microarray dataset. Finally, we describe the novel heuristic modification of the Gibbs sampler used to fit the infinite mixture mode that effectively deals with issues of slow mixing.

Keywords

Guides, instructions, authors kit, conference publications.

1.INTRODUCTION

Unsupervised identification of patterns in microarray data has been a productive approach to uncovering relationships between genes and the biological process in which they are involved. Conceptually, unsupervised learning from microarray data can be done by identifying genes with similar expression patterns across different experimental conditions, identifying groups of experimental conditions or biological samples with similar expression profiles, or the two dimensional clustering that simultaneously clusters genes and biological samples. In this paper we will be talking mostly about identifying groups of genes with similar expression patterns (profiles) across different biological samples. Groups of such genes are said to be co-expressed and they define patterns of expression. The utility of identifying such groups of co-expressed genes is in the assumption that the co-expression is a reflection of a shared regulatory mechanism driving similarities of expression profiles. Consequently, such groups of genes can be used as a starting point for dissecting expression regulatory mechanisms [23], or functional annotation by assuming that functionally-related genes are most likely to be co-regulated [5].

Clustering methods used for unsupervised identification of co-expressed genes can be loosely grouped into heuristic methods based on various distance measures, and model-based methods which are based on the probabilistic model of the data generation process. Given a distance measure, various heuristic methods proceed to organize gene expression profiles in a hierarchical fashion [3] or by partitioning them into a pre-specified number of clusters of co-expressed genes (e.g. K-means algorithm and Self-organizing maps).

In a model-based approach to clustering, the probability distribution of the observed data is approximated by a probabilistic model. Parameters in such a model define clusters of similar observations and a cluster analysis is performed by estimating these parameters from the data. The Finite Mixture (FM) model is the most common model-based approach to clustering [11]. In the context of microarray data, the FM model was introduced by [24]. In this approach, similar individual profiles are assumed to have been generated by a common underlying "pattern" represented by a multivariate Gaussian random variable. Given the correct number of mixture components (clusters) one can use an EM algorithm to estimate parameters of this model and then use the parameter estimates to assign individual profiles to appropriate clusters. Recently, various generalizations of the Bayesian mixture approach in terms of sophisticated Bayesian probabilistic models have been used to integrate various pieces of additional information in the process of identifying co-expressed genes [21;22], and to identify "modules" of co-regulated genes through the integrated modeling of combinatorial regulation mechanisms and gene expressions via context-specific Bayesian networks [19].

The common denominator of the above-mentioned model-based methods is that they rely on the prior specification of the number of clusters in the data or on one's ability to determine the correct number of clusters from the data. When the correct number of clusters is determined in the data analysis (e.g. by calculating Bayesian Information Criterion – BIC for models with different number of clusters), uncertainties related to its selection are generally not taken into account in the subsequent analysis. Previously we described the Bayesian Infinite Mixture (IM) model for the clustering of gene expression profiles [12] which effectively circumvents the problem of identifying the "correct" number of clusters. In our approach, the clusters are formed based on the posterior distribution of clusterings, which is generated by a Gibbs sampler. The clusterings generated by the Gibbs sampler can vary from one cycle to the next. Consequently, posterior probabilities with various features of the posterior distribution of clusterings are obtained after averaging over models with all possible number of clusters.

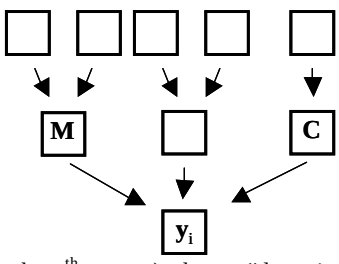
In this paper we describe a new simulated annealing-motivated algorithm for sampling from the posterior distribution of clusterings that effectively solves the severe mixing problem exhibited by Gibbs sampler in high-dimensional situations. More

importantly, we demonstrate dramatic positive effects that Bayesian averaging can have on discovering patterns in microarray data through both a simulation study and the analysis of a relevant real-world microarray dataset. These results are likely to bear on further development of model-based unsupervised learning methods that rely on either the specification of the correct number of clusters or its estimation from the data.

2.FINITE AND INFINITE MIXTURES MODEL BASED CLUSTERING FOR MICROARRAY DATA

Suppose that T gene expression profiles were observed across M experimental conditions. If y_{ij} represents the expression measurement for the ith gene under j^{th} experimental condition then $\mathbf{y}_i=(y_{i1},y_{i2},\ldots,y_{iM})$ represents the expression profile for the ith gene. In a mixture model, each gene expression profile is viewed as being generated by one out of Q different underlying expression patterns. Expression profiles generated by the same pattern form a cluster of similar expression profiles. If c_i is the classification variable indicating the pattern that generates the ith mean expression profile (c_i=q means that the ith expression profile was generated by the qth pattern), then a "dustering" is defined by a set of classification variables for all genes $\mathbf{C}=(c_1, c_2,\ldots, c_T)$. Underlying patterns generating clusters of expression profiles are represented by multivariate Gaussian random variables. Profiles clustering together are assumed to be a random sample from the same multivariate Gaussian distribution.

Figure 1: Bayesian mixture model



The hierarchical structure of the model is described in terms of a Directed Acyclic Network in **Figure 1**. Nodes (squares) in this diagram represent random variables and directed arcs (arrows) specify conditional dependences between variables in terms of the directed Markov property, which states that a variable is conditionally independent of its non-descendants given its parents in the model. $\mathbf{M}=(\boldsymbol{\mu}_1,\ldots, \boldsymbol{\mu}_Q)$ and $\boldsymbol{\Sigma}=(\sigma_1^2\mathbf{I},\ldots, \sigma_Q^2\mathbf{I})$ denote means and variance-covariance matrices of multivariate Gaussian random variables defining Q underlying patterns respectively ($\mathbf{I}$ denotes the identity matrix). Variables ($\boldsymbol{\lambda}$, τ), (β,ϕ), and α are hyper-parameters in prior distributions of model parameters $\mathbf{M}$, $\boldsymbol{\Sigma}$ and $\mathbf{C}$ respectively. In the case of an FM model, the number of mixture components (Q) is considered fixed, while the IM model represents the limiting case when $Q\rightarrow\infty$. Details of the development of IM models and their relationship to mixtures with a Dirichlet process prior [4] are described elsewhere [15;17]. We have previously described Bayesian versions of both finite and infinite mixtures and corresponding Gibbs samplers [12;14]. In this paper, finite mixtures model were treated from a frequentist perspective and estimated using the EM algorithm as implemented in the MCLUST software [6]. The Gibbs sampler for estimating the posterior distribution of clusterings in the IM model is described below. The specification of the prior distribution for classification variables ($\mathbf{C}$) determines whether the model represents finite or infinite mixtures.

2.1Gibbs Sampler

Gibbs sampler [7] is a general procedure for sampling observations from a multivariate distribution. A Gibbs sampler proceeds by iteratively drawing observations from complete posterior conditional distributions of all components. As the number of iterations approaches infinity, such a sequence describes observations from the joint multivariate distribution. In our case, we use the Gibbs sampler to estimate the joint posterior distribution of all parameters in our hierarchical model, given the data. We then use the marginal posterior distribution of clusterings to calculate posterior pairwise probabilities of coexpression (PPPC) for all pairs of expression profiles. Suppose that the sequence of clusterings ($\mathbf{C}^B$, $\mathbf{C}^{B+1}$,.., $\mathbf{C}^S$) was generated by the Gibbs sampler after B "burn-in" cycles. The pair-wise probabilities for two genes to be generated by the same pattern are estimated as:

$$P_{ij} = \frac{\#\text{ of samples after "burn-in" for which } c_i = c_j}{S-B}.$$

Using these probabilities as a similarity measure, clusters of similar expression profiles are created using a traditional agglomerative hierarchical clustering with similarities between groups of profiles being defined using the complete linkage. Complete descriptions of the posterior conditional distributions used by the Gibbs sampler can be found in [12], with the slight modification of using an independent, equal variance, covariance structure while in the original model we used the different variance elliptical model.

2.2Convergence of the Gibbs sampler

Two aspects of the Gibbs sampler convergence that generally need to be assessed are the appropriateness of the "burn-in" period, after which a Gibbs sampler has attained its stationary distribution, and the mixing of the sampler, which describes how well a finite sample obtained by Gibbs sampler approximates the target distribution. It has generally been well documented that the simple Gibbs sampler often has very poor mixing properties in both FM and IM models [2;14], probably due to the multi-modality of the posterior distribution. In such a situation, the sampler will be unable to describe the whole posterior distribution in a computationally feasible number of steps. The sampler will get trapped in a sub-optimal mode of the posterior distribution resulting in sub-optimal clustering results; or, because the sampler fails to visit all areas with significant posterior probabilities, confidence estimates in the generated clustering will be biased. Previously, we described a heuristic algorithm for "heating up" the Markov chain described by the Gibbs sampler by using "reverse annealing." The optimal annealing schedule was chosen based on running a significant number of independent chains with different maximum annealing constants. Here we describe a new heuristic algorithm that adjusts the annealing exponent dynamically. Consequently, only a single run is needed to estimate the posterior distribution.

If $\pi(.)$ is the target posterior distribution, "reverse annealing" refers to "flattening" of the posterior distribution using the transformation $\pi^{(\xi)}(x) = \frac{\pi^{\xi}(x)}{K(\xi)}$, $\xi < 1$, where $K(\xi)$ is the

normalizing constant. Based on this general idea, if $p(c_i=j|C_{-i}, \Theta)$ is the conditional posterior probability of placing the ith profile into the jth cluster then "flattened probabilities" are defined as

$$p(c_i = j \mid C_{-i}, \Theta)^{(\xi)} = \frac{p(c_i = j \mid C_{-i}, \Theta)^{\xi}}{K(\xi)}, \quad \xi < 1.$$

Since the mixing problem with the Gibbs sampler for the IM model can be particularly pronounced in its inability to generate new clusters, we keep track of the posterior probability of placing a profile in a new cluster. If this probability p_{new} is below the given threshold p_{min}, we decrease ξ by the value ξ_{step}. If p_{new} is above p_{min}, we increase ξ by ξ_{step}. Possible values of ξ are further constrained by the requirement that $0<\xi_{min}<\xi<\xi_{max}\leq 1$. Our modified Gibbs sampler now proceeds by generating n_{cold} samples from the unmodified conditional posteriors (cold cycles). It then generates a single sample using "heated" dassification probabilities (heated cycle). The p_{new} from the heated cycle is used to increase or decrease the value of ξ by ξ_{step}. However, only the sample from the last cold cycle (n_{cold} cycles after the heated cycle) is used in the estimation of the posterior distribution of clusterings. In our simulations, we used $n_{cold}=5$, $\xi_{min}=0.1$, $\xi_{max}=1$, $p_{new}=0.01$ and $p_{step}=0.1$. Due to the high computational complexity in the analysis of the cancer data, we used $n_{cold}=3$.

2.3Finite mixture model and EM algorithm

We used the MCLUST package's EMclust procedure to fit finite mixture models to our simulated and real-world data sets. The optimal number of clusters was selected by calculating the Bayesian Information Criterion (BIC) [18] for models for different number of clusters. The only model used in this study was the equal variance, independent, spherical shape (EII) covariance model [6]. The EMclust procedure fits the finite mixture model by first performing an appropriate model-based hierarchical clustering (in the case of the EII model this amounts to the Ward's clustering algorithm). Resulting parameters are used as a starting point for the EM algorithm. Given the maximum likelihood estimates of the model parameters, profiles are assigned to clusters based on the posterior probabilities being generated by different mixtures components and the Maximum A Posteriory (MAP) hypothesis. To compare the performance of the FM model to the IM model, we also calculate FM model based PPPC's as

$$P_{ij} = \sum_{k=1}^{Q} p(c_i = k) p(c_j = k),$$

where $p(c_i=k)$ is the posterior probability of the profile i being generated by component k.

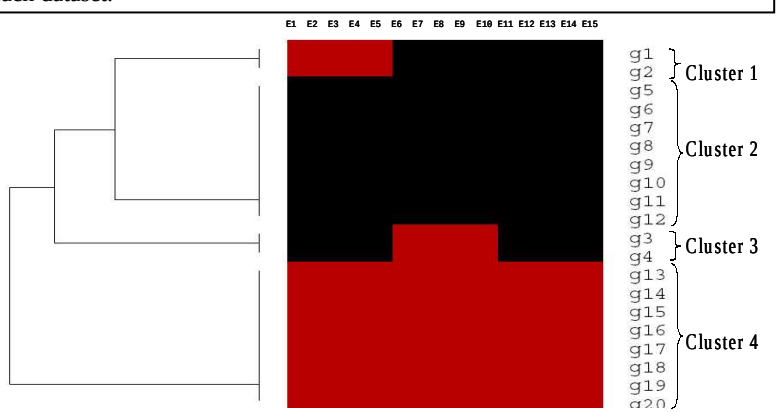
3.SIMULATION STUDY

First, we assessed the importance of Bayesian model-averaging in a simulation study. The study was designed to assess the performance of both FM and IM models in the frequentist sense. That is, we assessed the power of the two clustering methods to separate two different clusters in repeated experiments. We simulated 100 datasets each representing the clustering structure depicted in **Figure 2**. The heat map represents the values of the mean vectors for mixture components generating each profile. Red represents the value of 1 and the black represents the value of 0. For example, in each dataset, profile "g1" was randomly drawn from the 15-dimensional Gaussian random distribution whose mean vector is equal to 1 in first 5 dimensions ($e_1,\ldots,e_5$) and 0 in other 10 dimensions ($e_6,\ldots,e_{15}$). The covariance matrix $\sigma^2\mathbf{I}$ was used so that the data is compatible with our model assumptions. Data was simulated for $\sigma \in c$. This range allowed us to assess the performance of the two approaches in easy and progressively more difficult (i.e. noisier) situations.

3.1Results

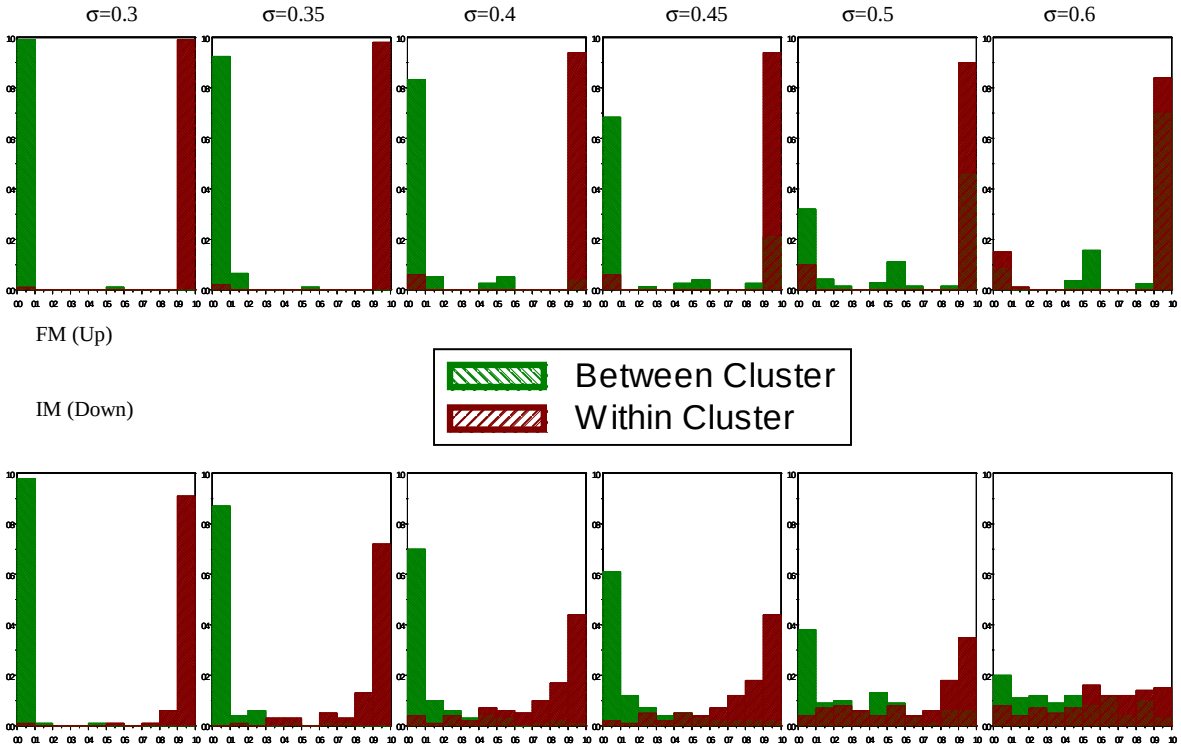
Both methods performed very well in separating two larger and most divergent clusters (Cluster2 and Cluster4) under the conditions of our simulation study. Therefore, we are focusing on the more difficult task of separating clusters 1 and 2. Profiles from these two clusters differ only within first 5 dimensions ("experiments") and Cluster1 is defi ned by only two profiles. The major question we are asking is how often can we expect the two clusters to be separated. We are assessing this question by observing the distribution of PPPC's for the two profiles in Cluster1 in relation to PPPC's between pr ofiles in Cluster1 and Cluster2. In a sense we are assessing the ability of our clustering methods to correctly conclude that profiles in Cluster1 are different from profiles with Cluster2. However, unlike traditional statistical hypothesis testing procedures, we do not supply the labels for profiles that we are comparing.

Figure 2: Heat map of the clustering structure for the simulated data. Total of 20 15-dimensional profiles belonging to 4 unbalanced clusters are generated in each dataset.



Results of this simulation study support our thesis that FM model-based clustering in which the number of clusters is chosen by the BIC criterion suffers because of its inability to incorporate in the results of the analysis the uncertainty inherent in the process of determining the number of clusters. Histograms in Figure 3 show the "over-confidence" of the FM-based PPPC's which is typical of a statistical analysis that fails to take into account all sources of uncertainty (i.e. variability). The majority of the PPPC's generated by the FM model are clustered around 0 and 1 indicating the high confidence in the separation or non-separation in all situations, even when they are wrong. For example, for the highest noise level, close to 70% of between cluster PPPC's are greater than 0.9 indicating high confidence in the false conclusion that these profiles belong to the same cluster. On the other hand, PPPC's seem to be more reflective of the level of evidence for separating the two clusters present in the data. While the level of confidence in the separation is being reduced as we move from the low-noise to high-noise data, the fraction of PPPC's offering a high confidence in the false conclusion remains low even in the noisiest situation.

Figure 3: Histograms of PPPC's for pairs of profiles belonging to the same cluster (Within Cluster) and pairs of profiles belonging to different clusters (Between Cluster) in 100 simulated datasets for 6 different noise levels.



We can further drive the analogy with traditional statistical hypothesis testing procedures by constructing Receiver Operating Characteristic (ROC) curves that assess the ability of a clustering method to correctly separate profiles from different clusters. We are again focusing of ability to separate profiles in Cluster1 from profiles in Cluster2. For a fixed cut-off point X, we consider that the clustering procedure is correctly concluding that a profile i from Cluster1 does not belong to Cluster2 if $\max\{p(c_i=c_j$ for all profiles j from Cluster2$\}<X$. We consider that the clustering procedure is incorrectly concluding that profile 1 and profile 2 from Cluster1 do not belong in the same cluster if $p(c_i=c_j)<X$. The true positive rate (TPR) is the proportion of times that a correct decision is made and the false positive rate (FPR) is the proportion of times that an incorrect decision is made. As the cut-off X is increased from 0 to 1, both TPR and the FPR will increase. The area under the curve relating the TPR and FPR as X is increased from 0 to 1 describes the efficiency of a statistical procedure with the random decision-making having an area of 0.5 while the ideal statistical procedure would have an area equal to 1. ROC's for the FM and IM models for different noise levels are given in Figure 4. It seems that for each, except the lowest noise level, the IM model significantly outperforms the FM procedure.

4.CANCER DATA ANALYSIS

While the simulation study seems to indicate the importance of model averaging in the model-based cluster analysis, the question remains whether these advantages make any difference in the context of analyzing real-world data. Demonstrating advantages of one clustering method over another in the context of real-world data is complicated by the uncertainties related to the "correct" clustering which is generally not known. To address this question we reproduced the analysis described by [10]. They demonstrated how human cancer databases of microarray data could be used to study a molecular mechanism of cancer induction. In their study, they first identified 21 cyclin D1 target genes in *in-vitro* laboratory experiments. They followed up with an investigation of the relationship between CD1 and these 21 genes in a cancer gene expression database [16]. The statistical significance of that association in the cancer data was established by showing that the distribution of Euclidian distances between

expression profiles of these gene and CD1 were higher then expected by chance (p-value=0.048 using a resampling version of the Kolomogorov-Smirnov test). The conclusion was that the in-vitro signature of the CD1 overexpression is preserved in primary human tumors. We clustered the cancer expression data using the Euclidian distance, IM and FM models with the optimal number of clusters (56), elected by the BIC as described before (Figure 5). Based on results of these cluster analyses, two important points can be made: (1) just by a visual inspection of heat maps, it is apparent that model-based clustering approaches (both FM and IM) created "deaner" groupings of genes with similar expression patterns than the Euclidian distance-based hierarchical clustering procedure. (2) the over-confidence of the FM model, noted in the analysis of the simulated data, is evident in this analysis as well. The consequence of such an over-confidence is that the FM model identifies only 2 of the 21 genes of interest to be significantly associated with CD1. On the other hand, the IM model identifies 7 of them.

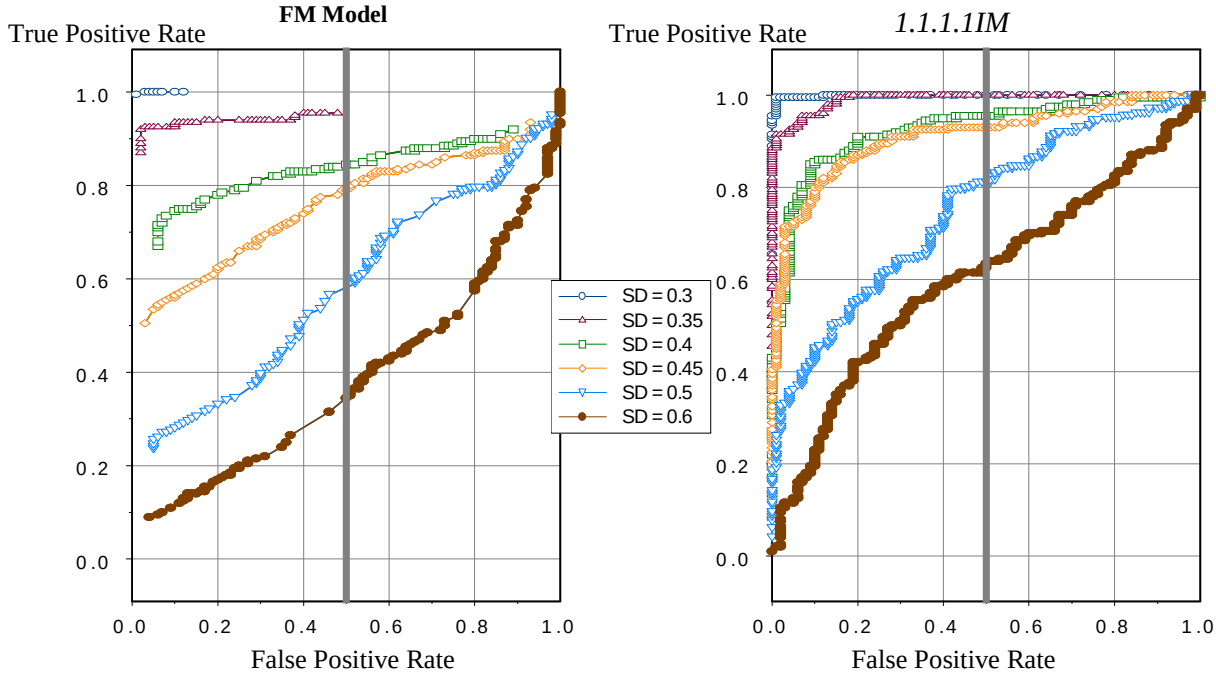
The distribution of the Euclidian distances also seems to suggest that only the associations between CD1 and three of the 21 genes of interest are above the experimental noise. In the context of the Kolmogorov-Smirnov (KS) analysis of Euclidian distances (Figure 3A in Lamb et al. 2003), there are indications that actually 7 to 10 of the 21 genes are contributing to the significance of the association. However, the statistical significance of this observation is impossible to assess within their framework. These results suggest that the IM model is capable of identifying most biologically meaningful relationships in the data by integrating the power of the model-based approach to pull information from the whole dataset while accounting for the uncertainty introduced by not knowing the number of clusters in the data.

5.DISCUSSION

In this paper we demonstrated the utility of Bayesian model averaging in model-based clustering of microarray data in a simulation study and as it is applied to answer a relevant biological question using a relatively large microarray dataset. We demonstrated that the performance of the traditional finite mixture clustering approach in which the optimal number of clusters is chosen using the BIC suffers from over-confidence in false conclusions probably due its inability to account for uncertainties related to the choice of the right number of clusters. The significantly better performance of the equivalent IM model in both the simulation study and the analysis of the real-world data is most likely due to its ability to estimated the posterior distribution of clusterings by effectively averaging over models with all possible number of clusters. The consistency and the precision of the results obtained by the IM approach also suggest that our heuristic modification of the Gibbs sampler effectively alleviates the problem of slow mixing.

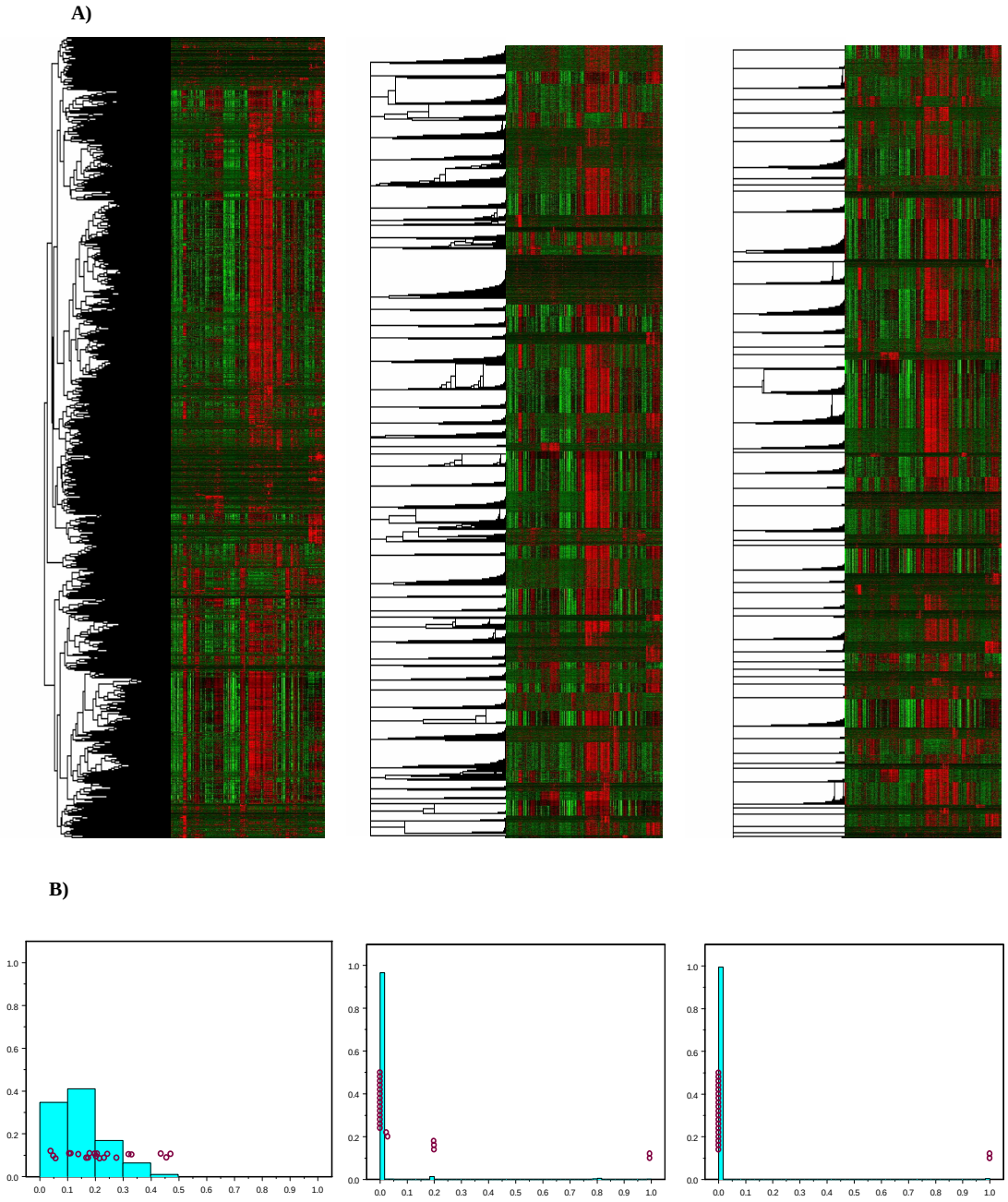
We have previously demonstrated the advantages of using model-based clustering approaches over the traditional distance-based heuristic algorithms [14;25]. Model-based methods allow for the precise treatment of the statistical characteristics of the data under investigation, such as replicated observations. Furthermore, when compared to traditional distance measure-based hierarchical clustering algorithms, they are more efficient in using the information from the whole datasets instead of using two vectors of observations at a time. This advantage has been nicely demonstrated in our analysis of cancer data in Figure 5 as well. Additionally, when compared to partitioning heuristic algorithms, such as the K-mean algorithm and the SOM's, they allow for estimation of the number of clusters by assessing the relative fit of models with different numbers of clusters.

Figure 4: ROC curves describing the ability of the two models to separate the Cluster1 from Cluster2.



Recently introduced generalizations of the traditional mixture models, based on the context-specific Bayesian networks [20], allow for identifying more complex relationships between different genes as well as incorporating other types of the data in the analysis [21;22]. In this respect, our analysis strongly suggests that the general practice of fixing the number of clusters, components, or modules in terms of [19], before fitting appropriate models might need some modifications. One possible solution for this problem could be the adaptation of the IM paradigm for such complex models. Another possible solution could go along the line of averaging results obtained by fitting models with different number of components in a post-hoc analysis.

Figure 5: A) Cluster analysis based on the Euclidian distances (left), IM model PPPC's (middle), and FM model PPPC's (right). B) Histograms of corresponding similarity measures for all genes with CD1. Circles represent the similarity measures for the 21 genes identified in the laboratory experiments.



It is important to notice that the uncertainties in the process of identifying the correct number of clusters are not necessarily the only source of uncertainties that are not taken into account by

the traditional FM approach. In high-dimensional situations, such as the cancer data analyzed in this paper, the log-likelihood maximized by the EM algorithm is almost certainly multi-modal and using any kind of strategy for choosing the "optimal" starting position will not guarantee that the solution will be globally optimal. Since the BIC calculation is based on results of the EM algorithm, these types of computational inadequacies will contribute to the overall uncertainty in the selection of the "optimal" number of clusters. Furthermore, such computa tional problems can result in sub-optimal clustering given the "optimal" number of cluster. Using different variants of the EM algorithm designed to alleviate this problem [13] can sometimes help, but the convergence to the globally optimal solution is still never guaranteed. In this respect, a properly mixing Gibbs sampler can offer another advantage due to its ability to describe the whole posterior distribution instead of searching for the highest mode of the likelihood function. We performed a limited evaluation of the convergence properties of the overall estimation approach (data not shown) and determined that EM convergence issues were probably not a factor in our simulation study due to the relatively simple clustering structure, but they were likely an additional source of uncertainty in the analysis of the cancer data.

The purpose of our analysis was not to disparage the BIC as the criterion for choosing the right number of clusters, but rather to demonstrate the problem of the whole approach in which the right model is chosen based on a preliminary analysis of the data, and where the uncertainties inherent in this process are not propagated into the final estimates of uncertainties about conclusions made based on the whole analysis process. Empirical studies have shown that the criterion works quite well in identifying the correct number of mixture components [1]. On the other hand some recent evaluations showed that an alternative approach of statistical hypothesis testing-based determination of the number of clusters [8] is more robust with respect to the deviation from the assumption of the models for individual mixture components. Unfortunately, these evaluations were made assuming only the simplest possible model for the calculation of the BIC, as implied by the K-means algorithm. It remains unclear if these advantages persist after using the complete FM approach for choosing the right covariance structure as well as the right number of clusters as proposed by the authors of MCLUST [6], or in the situation when the basic covariance structure implied by the K-means algorithm is correct, as was the case in our simulation study. Altogether, the BIC approach remains one of the dominant criteria for choosing models in statistical practice, and it is not clear that any alternative method for choosing the right number of clusters will significantly improve the overall FM performance. On the other hand, we showed that the IM model offers an elegant way around the issue of selecting the right number of clusters in the context of model-based clustering.

Finally, although our heuristic Gibbs sampler modification has been performing very well in all situations we encountered so far, it is not clear how closely does the modified sampler approximate the posterior distribution defined by the IM model. This is problematic since some of the nice conceptual features of the Bayesian IM framework depend on being able to sample from the true posterior distribution defined by the model. For example, the meaning of the posterior pairwise probabilities is not clear unless we can claim that they are derived from the hierarchical statistical model in Figure 1. We can still use them as a high-quality distance measure, but their direct probabilistic interpretation is lost. Some work has been done on developing alternative MCMC methods for fitting conjugate infinite mixture models [9]. However, to the best of our knowledge, alternative MCMC samplers for non-conjugate models, such as the model described here, have not yet been developed.

6.ACKNOWLEDGMENTS

This work has been supported by the NHGRI research grant 1R21HG002849-01.

7.REFERENCES

[1] C. Biernacki and G. Govaert. Choosing models in model-based clustering and discriminant analysis. J. Statis. Comput. Simul., 64 (1999), 49-71.

[2] G. Celeux, M. Hurn, and C. P. Robert. Computational and Inferential Difficulties With Mixture Posterior Distributions. JASA, 95 (2000), 957-970.

[3] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. U. S. A, 95 (Dec.1998), 14863-14868.

[4] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. The Annals of Statistics, 1 (1973), 209-230.

[5] S. Fessele, H. Maier, C. Zischek, P. J. Nelson, and T. Werner. Regulatory context is a crucial part of gene function. Trends Genet., 18 (Feb.2002), 60-63.

[6] C. Fraley and A. E. Raftery. Model-Based Clustering, Discriminant Analysis, and Density Estimation. JASA, 97 (2002), 611-631.

[7] E. A. Gelfand and F. M. A. Smith. Sampling-based approaches to calculating marginal densities. Journal of The American Statistical Association, 85 (1990), 398-409.

[8] G. Hamerly and C. Elkan. Learning the k in k-means. Proceedings of the Seventeenth Annual Conference on Neural Information Processing Systems (NIPS' 03) (2003),-

[9] S. Jain and R. Neal. A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model. Technical Report No. 2003, Department of Statistics, University of Toronto (2000),-

[10] J. Lamb, S. Ramaswamy, H. L. Ford, B. Contreras, R. V. Martinez, F. S. Kittrell, C. A. Zahnow, N. Patterson, T. R. Golub, and M. E. Ewen. A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. Cell, 114 (Aug.2003), 323-334.

[11] J. G. McLachlan and E. K. Basford, *Finite Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker, 1987.

[12] M. Medvedovic and S. Sivaganesan. Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics, 18 (Sept.2002), 1194-1206.

[13] M. Medvedovic, P. Succop, R. Shukla, and K. Dixon. Clustering mutational spectra via classification likelihood and Markov Chain Monte Carlo Algorithm. Journal of Agricultural, Biological and Environmental Statistics, 6 (2001), 19-37.

[14] M. Medvedovic, Yeung K.Y., and R. E. Bumgarner. Bayesian mixture model based clustering of replicated microarray data. Bioinformatics (In Press) (2004),-

[15] R. M. Neal. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational and Graphical Statistics, 9 (2000), 249-265.

[16] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander, and T. R. Golub. Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. U. S. A, 98 (Dec.2001), 15149-15154.

[17] C. A. Rasmussen. The Infinite Gaussian Mixture Model. Advances in Neural Information Processing Systems, 12 (2000), 554-560.

[18] G. Schwarz. Estimating the dimension of a model. Ann. Stat., 6 (1978), 461-464.

[19] E. Segal, M. Shpira, A. Regev, D. Pe'er, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet., 34 (2003), 166-176.

[20] E. Segal, B. Taskar, A. Gasch, N. Friedman, and D. Koller. Rich probabilistic models for gene expression. Bioinformatics, 17 (2001), S243-S252.

[21] E. Segal, H. Wang, and D. Koller. Discovering molecular pathways from protein interaction and gene expression data. Bioinformatics, 19 (2003), I264-I272.

[22] E. Segal, R. Yelensky, and D. Koller. Genome-wide discovery of trancriptional modules from DNA sequence and gene expression. Bioinformatics, 19 (2003), I273-I282.

[23] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. Nat. Genet., 22 (July1999), 281-285.

[24] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo. Model-based clustering and data transformations for gene expression data. Bioinformatics., 17 (Oct.2001), 977-987.

[25] K. Y. Yeung, M. Medvedovic, and R. E. Bumgarner. Clustering Gene Expression Data with Repeated Measurements. Genome Biology, 4 (2003), R34-