Journal of Machine Learning Research 8 (2007) 1145-1164 Submitted 11/05; Revised 1/07; Published 5/07

Penalized Model-Based Clustering with Application to Variable Selection

Wei Pan WEIP@BIOSTAT.UMN.EDU
Division of Biostatistics
School of Public Health
University of Minnesota
Minneapolis, MN 55455, USA

Xiaotong Shen XSHEN@STAT.UMN.EDU
School of Statistics
University of Minnesota
Minneapolis, MN 55455, USA

Editor: Bin Yu

Abstract

Variable selection in clustering analysis is both challenging and important. In the context of model-based clustering analysis with a common diagonal covariance matrix, which is especially suitable for "high dimension, low sample size" settings, we propose a penalized likelihood approach with an L_1 penalty function, automatically realizing variable selection via thresholding and delivering a sparse solution. We derive an EM algorithm to fit our proposed model, and propose a modified BIC as a model selection criterion to choose the number of components and the penalization parameter. A simulation study and an application to gene function prediction with gene expression profiles demonstrate the utility of our method.

Keywords: BIC, EM, mixture model, penalized likelihood, soft-thresholding, shrinkage

1. Introduction

This article concerns variable selection in model-based clustering, especially for "high dimension, low sample size" data, where the data dimension greatly exceeds the number of observations. Specifically, given n P-dimensional observations $x_j = (x_{j1}, ..., x_{jP})'$ for $j = 1, ..., n$, we aim to group the data into a few, say K, clusters such that the observations in the same cluster are more similar to each other than those from different clusters. In this context, some of the attributes x_{jp}'s of x_j may not be relevant: use of such attributes only introduces noise, and may impede uncovering the clustering structure of interest. In addition, removing non-informative attributes may largely enhance interpretability. Due to lack of statistical models in many existing clustering algorithms, it is difficult to implement principled variable selection, though some promising methods have been proposed, based largely on heuristics (Friedman and Meulman, 2004; Mangasarian and Wild, 2004). In contrast, model-based clustering (McLachlan and Peel, 2002; Fraley and Raftery, 2002) assumes that data come from a finite mixture model with each component corresponding to a cluster; with such a statistical model, statistical inference, including variable selection, can be carried out. Because, to our knowledge, no formal hypothesis tests are available to assess the statistical significance of an attribute, it is unclear how to implement a sequential variable selection, such as forward additions

and/or backward eliminations of variables, in model-based clustering; an alternative is to conduct best subset selection, which however is unrealistic for high-dimensional data: for example, with $P = 1000$, there are more than 10^{300} possible models to be considered, which is prohibitive given the current standard computing power. Furthermore, even for smaller problems, as in regression, due to its discreteness, best subset selection may be unstable and may not work well in selecting relevant variables (Tibshirani, 1996); most importantly, unlike in regression or classification but unique to clustering or semi-supervised learning, best subset selection may identify a correct model which however is of no interest, as to be confirmed by our numerical example later.

With high-dimensional data, as an alternative to variable selection, one may apply dimension reduction techniques, such as principal component analysis, prior to clustering (Ghosh and Chinnaiyan, 2002; Liu et al., 2003). A possible drawback of this approach is the separation between dimension reduction and subsequent clustering; for example, as pointed out by many researchers (Chang, 1983; Yeung and Ruzzo, 2001; Raftery, 2003), using first few principal components in clustering may destroy the clustering structure of the original data.

There has been increasing interest in variable selection for model-based clustering, mostly within the Bayesian framework (Liu et al., 2003; Hoff, 2005, 2006; Tadesse et al., 2005; Raftery and Dean, 2006; Kim et al., 2006). An idea is to parametrize the mean of cluster k as $\mu_k = \mu + \delta_k$, where μ is the global mean. It is clear that, if some components of δ_k are 0, then the corresponding attributes are not informative to clustering, at least in terms of the means/locations. Two Bayesian approaches have been proposed based on this idea (Liu et al., 2003; Hoff, 2005, 2006). Another Bayesian approach, analogous to stepwise variable selection in regression, is to sequentially compare two nested models to determine whether an attribute should be included in or excluded from the current model based on a greedy search (Raftery and Dean, 2006), which may be computationally too time-consuming for high-dimensional data. In contrast, to our knowledge, no frequentist alternatives to subset selection are available for variable selection in model-based clustering. In light of the success of penalized regression with variable selection (Tibshirani, 1996; Fan and Li, 2001), we conjecture that penalization may be also viable to variable selection in clustering, and hence we propose an approach through penalized model-based clustering. Specifically, cluster-specific means μ_k are adaptively shrunken towards the global mean μ; with an appropriately chosen penalty function, some components of μ_k are estimated to be exactly the same as that of μ, effectively realizing variable selection. We also propose a modified BIC as a model selection criterion to adaptively determine the amount of penalization as well as the number of clusters. Note that, although there is an extensive body of literature on penalized likelihood methods, most focus on classification and regression; in particular, to our knowledge, we are not aware of any existing works on penalized likelihood particularly designed for multivariate clustering.

Recent advances in high-throughput biotechnologies, such as microarrays, have generated a large amount of high-dimensional data, and have led to routine use of clustering analyses, for example, in gene function discovery (Eisen et al., 1998) and cancer subtype discovery (Golub et al., 1999; Ghosh and Chinnaiyan, 2002; McLachlan et al., 2002; Yeung et al., 2001). In these applications, one key issue is how to select variables: although expression levels of thousands or tens of thousands of genes are measured on each microarray, corresponding to a high-dimensional observation, it is known that not all the genes are related to the phenotype of interest, for example, subtypes of a cancer; in fact, often only a small number of the genes are relevant, and identifying those genes is one of the biologically most important goals. Hence, variable selection in clustering not only improves the performance in identifying interesting clusters, as to be shown later, but also largely

facilitates interpretation of results, and even directly addresses biological questions of interest, for example, which genes are involved in the biology of a cancer or its subtypes. In this article, in addition to the promising application of cancer subtype discovery, we also apply model-based clustering to the task of gene function discovery (Li and Hong, 2001; Ghosh and Chinnaiyan, 2002). Although the human genome and many other genome sequencing projects have led to a discovery of many new genes, biological functions of many genes remain unknown; many known functions also need to be refined. It has become popular to cluster gene expression profiles to discover unknown gene functions.

In the remaining parts of this article, we first review briefly the standard model-based clustering, then we introduce a general framework for penalized model-based clustering. We propose a specific implementation with an L_1 penalty, resulting in soft-thresholding on the mean parameters, and thus realizing automatic variable selection. We derive an EM algorithm to compute the maximum penalized likelihood estimates for the model; a modified BIC is used to determine the number of components and the value of the penalization parameter in penalized model-based clustering. We compare the proposed method with the standard method using simulated data and gene expression data for tumor subtype discovery and gene function prediction; in particular, we illustrate problems associated with clustering without variable selection, and those with best subset selection, concluding that penalized clustering is an effective and simple method for variable selection. We end the article with a short discussion on some open questions.

2. Methods

We first give a brief review on model-based clustering with a finite Normal mixture model, then we introduce our penalized model-based clustering, including an EM algorithm and a modified BIC for model selection.

2.1 Model-based Clustering

In model-based clustering, it is assumed that each observation x is drawn from a finite mixture distribution $f(x;\Theta) = \sum_{k=1}^{K} \pi_k f_k(x;\theta_k)$, with the mixing proportion π_k, component-specific distribution f_k and its parameters θ_k. Denote by $\Theta = \{(\pi_k, \theta_k) : k = 1, ..., K\}$ all unknown parameters, with restriction that $0 \le \pi_k \le 1$ for any k and $\sum_{k=1}^{K} \pi_k = 1$. Each component of the mixture distribution corresponds to a cluster. The number of clusters, K, has to be determined in practice; see section 2.4.

Given data x_j, $j = 1, ..., n$, the log-likelihood is

$$\log L(\Theta) = \sum_{j=1}^{n} \log[\sum_{k=1}^{K} \pi_k f_k(x_j;\theta_k)].$$

Maximization of the above log-likelihood with respect to Θ is difficult, and it is common to use the EM algorithm (Dempster et al., 1977) by casting the problem in the framework of missing data. Define z_{kj} as the indicator of whether x_j is from component k; that is, $z_{kj} = 1$ if x_j is indeed from component k, and $z_{kj} = 0$ otherwise. If the missing data z_{kj}'s could be observed, then the log-likelihood for the complete data is:

$$\log L_c(\Theta) = \sum_k \sum_j z_{kj}[\log \pi_k + \log f_k(x_j;\theta_k)].$$

The EM algorithm can be applied to obtain the maximum likelihood estimator (MLE) of Θ; see McLachlan and Peel (2002) and Fraley and Raftery (2002) for more details.

2.2 Penalized Model-based Clustering

With the same motivation as in penalized regression, we propose a penalized model-based clustering approach. The general purpose of penalization is for model regularization, which in general can enhance the predictive power of a model, and may be even necessary in some situations. For example, in the univariate Normal mixture model with each $f_k(.) = \phi(.;\mu_k,\sigma_k)$, it is well-known that with $\sigma_k \to 0$, we have a degeneracy with $\log L$ being unbounded, and thus no (unrestricted) MLE exists. Ciuperca et al. (2003) proposed a penalized likelihood approach to dealing with the degeneracy: by penalizing small variance components σ_k, one circumvents the problem. In addition, as to be discussed in the next section, with an appropriate choice of penalty function, we can realize a sparse solution, resulting in automatic variable selection.

Specifically, we regularize $\log L(\Theta)$ to yield a penalized log-likelihood:

$$\log L_P(\Theta) = \sum_{j=1}^{n} \log[\sum_{k=1}^{K} \pi_k f_k(x_j;\theta_k)] - h_\lambda(\Theta),$$

where $h_\lambda()$ is a penalty function with penalization parameter λ. The choice of $h_\lambda()$ depends on the goal of the analysis; see Fan and Li (2001) for some general theory. Correspondingly, the penalized log-likelihood for the complete data is

$$\log L_{c,P}(\Theta) = \sum_{k=1}^{K}\sum_{j=1}^{n} z_{kj}[\log \pi_k + \log f_k(x_j;\theta_k)] - h_\lambda(\Theta).$$

We propose using such penalized model-based clustering as a general way to regularize parameter estimates. This can be useful for high-dimensional data, especially for situations of "large P, small n". Recently Fraley and Raftery (2005) proposed a Bayesian approach to regularizing model-based clustering; there is a large body of literature on Bayesian mixture modeling, for example, Richardson and Green (1997), Jasra et al. (2005) and references therein. There is a well known connection between penalized likelihood and Bayesian modeling (Hastie et al., 2001): it can be regarded that minus the penalty function $-h_\lambda(\Theta)$ is proportional to the log density of the prior distribution for parameters Θ, and the penalized (log) likelihood is proportional to the (log) posterior density.

Note that in contrast to Ciuperca et al. (2003), where only univariate Normal mixture models were considered, our main interest here is in multivariate clustering.

2.3 Penalizing Mean Parameters

Now we propose a specific implementation of penalized model-based clustering to realize variable selection. Consider the common case with each component f_k as Normal. We are particularly interested in "large P, small n" often encountered in genomic studies. Hence, as in naive Bayes classification, we adopt a working independence model for components of x_j. Furthermore, to facilitate variable selection for "large P, small n" settings, a common diagonal covariance matrix is used across clusters; more discussions on this choice is given in Section 4. We assume throughout this article that, prior to clustering analysis, we have standardized data so that each attribute has

sample mean 0 and sample variance 1. Specifically, we have

$$f_k(x;\theta_k) = \frac{1}{(2\pi)^{P/2}|V|}\exp\left(-\frac{1}{2}(x-\mu_k)'V^{-1}(x-\mu_k)\right),$$

where $V = diag(\sigma_1,\sigma_2,...,\sigma_P)$, and $|V| = \prod_{p=1}^{P}\sigma_p$. We propose using the L_1 penalty:

$$h_\lambda(\Theta) = \lambda\sum_k\sum_p|\mu_{kp}|,$$

though other penalty functions may be also suitable, as discussed in Fan and Li (2001) in the context of regression. The main goal is to obtain a sparse solution with many small estimates of μ_{kp}'s automatically set to 0, thus realizing variable selection.

Next we derive an EM algorithm for the above penalized model-based clustering; in particular, it is confirmed that the L_1-penalty yields a thresholding rule with the desired sparsity property. The derivation closely follows from that for standard model-based clustering (McLachlan and Peel, 2002) and the general methodology for penalized likelihood (Green, 1990). We use generic notation $\Theta^{(m)}$ to represent the parameter estimates at iteration m, and use $X = (x_1,...,x_n)$ to denote all the observations. It is easy to verify that the E-step yields

$$Q_P(\Theta;\Theta^{(m)}) = E_{\Theta^{(m)}}(\log L_{c,P}|X) = \sum_k\sum_j\tau_{kj}^{(m)}[\log\pi_k + \log f_k(x_j;\theta_k)] - \lambda\sum_k\sum_p|\mu_{kp}|,$$

where

$$\tau_{kj}^{(m)} = \frac{\pi_k^{(m)}f_k(x_j;\theta_k^{(m)})}{f(x_j;\Theta^{(m)})} = \frac{\pi_k^{(m)}f_k(x_j;\theta_k^{(m)})}{\sum_{k=1}^{K}\pi_k^{(m)}f_k(x_j;\theta_k^{(m)})} \tag{1}$$

is the estimated posterior probability of x_j's coming from component k.

The M-step maximizes the above Q_P to update the parameter estimates. It is easy to show that

$$\frac{\partial Q_P}{\partial\pi_k} = \sum_j(\tau_{kj}^{(m)}/\pi_k - \tau_{Kj}^{(m)}/\pi_K),$$

for any $k = 1,2,...,K-1$, and

$$\frac{\partial Q_P}{\partial\sigma_p^2} = \sum_k\sum_j\tau_{kj}^{(m)}\left[-\frac{1}{2\sigma_p^2} + \frac{(x_{jp}-\mu_{kp})^2}{2\sigma_p^4}\right],$$

for any $p = 1,...,P$. Hence

$$\hat{\pi}_k^{(m+1)} = \sum_{j=1}^{n}\tau_{kj}^{(m)}/n, \quad \text{and} \quad \hat{\sigma}_p^{2,(m+1)} = \sum_{k=1}^{K}\sum_{j=1}^{n}\tau_{kj}^{(m)}(x_{jp}-\mu_{kp}^{(m)})^2/n. \tag{2}$$

Now for the mean parameters,

$$\frac{\partial Q_P}{\partial\mu_k} = \sum_j\tau_{kj}^{(m)}V^{-1}(x_j-\mu_k) - \lambda\text{sign}(\mu_k).$$

Some algebraic manipulations yield

$$\hat{\mu}_k^{(m+1)} = \text{sign}(\tilde{\mu}_k^{(m+1)}) \left(|\tilde{\mu}_k^{(m+1)}| - \frac{\lambda}{\sum_j \tau_{kj}^{(m+1)}} V^{(m+1)} 1 \right)_+, \tag{3}$$

where $\tilde{\mu}_k^{(m+1)} = \sum_j \tau_{kj}^{(m+1)} x_j / \sum_j \tau_{kj}^{(m+1)}$ is the usual update for μ_k if no penalty is imposed; for any f, $f_+ = f$ if $f > 0$, and $f_+ = 0$ otherwise; 1 is a vector with all elements 1's. Note that all the operations in (3), including sign() and $()_+$, are component-wise. It is evident that, if $\lambda > |\sum_{j=1}^n \tau_{kj}^{(m+1)} x_{jp} / \sigma_p^{2,(m+1)}|$, then $\hat{\mu}_{kp}^{(m+1)} = 0$; otherwise, $\hat{\mu}_{kp}^{(m+1)}$ is obtained by shrinking $\tilde{\mu}_{kp}^{(m+1)}$ towards 0 by an amount $\lambda \sigma_p^{2,(m+1)} / \sum_{j=1}^n \tau_{kj}^{(m+1)}$.

The above iteration is repeated until convergence, resulting in the maximum penalized likelihood estimate (MPLE) $\hat{\Theta}$. Then we use (1) to calculate the posterior probability of any observation x's belonging to each cluster, and assign the observation to the cluster with the highest probability. Because of possible existence of multiple local maxima, we run the algorithm multiple times, and each time we use the result from a randomly started K-means algorithm as starting values for the EM. We fit a series of models with various values of K and λ, then use a model selection criterion to choose their appropriate values, as to be discussed in the next section.

It can be seen that, if $\hat{\mu}_{kp} = 0$ for all k, then the p-th attribute does not contribute to clustering: it will be cancelled out from the numerator and the denominator of (1). In contrast, in the standard method, all attributes contribute to the posterior probability calculation.

Note that in (3), if we use $\tilde{\mu}_k^{(m+1)}$, instead of $\hat{\mu}_k^{(m+1)}$, we obtain the standard model-based clustering, which is equivalent to using $\lambda = 0$. In our numerical examples, to reduce bias, we used $\tilde{\mu}_{kp}^{(m)}$ in (2) to estimate σ_p^2, though we did not find much difference in several simulations if $\hat{\mu}_{kp}^{(m)}$ was indeed used. In addition, if we replace $\hat{\mu}_k^{(m+1)}$ by

$$\hat{\mu}_{k,H}^{(m+1)} = \tilde{\mu}_k^{(m+1)} I(\lambda > |\sum_{j=1}^n \tau_{kj}^{(m+1)} V^{-1,(m)} x_j|),$$

we obtain so-called hard-thresholding, which is in contrast to soft-thresholding in (3). In our numerical examples, we found that hard-thresholding gave results similar to those of soft-thresholding, and we will skip its further discussion.

2.4 Model Selection

In practice, we need to determine the number of components, K. This is realized by first fitting a series of models with various numbers of components, and then using a model selection criterion to choose the best one. For standard model-based clustering, it is common to use Bayesian information criterion (BIC) (Schwarz, 1978) defined as

$$BIC = -2 \log L(\tilde{\Theta}) + \log(n) d,$$

where $\tilde{\Theta}$ is the MLE, and $d = \dim(\Theta)$ is the total number of unknown parameters (Fraley and Raftery, 1998). In our proposed model, we have $d = K + P + KP - 1$, because we have three sets of parameters, π_k's, σ_p's and μ_{kp}'s, under the constraint $\sum_{k=1}^K \pi_k = 1$.

For penalized model-based clustering, in addition to K, we also have to choose an appropriate value of penalization parameter λ; a model selection criterion has to account for the adaptive choice of λ. One difficulty in using the above BIC criterion is that it is not always clear what is d in a penalized model. Although other resampling-based model selection methods, such as cross-validation or generalized degrees of freedom (Shen and Ye, 2002) can be employed, they are computationally more demanding, and even prohibitive for large and/or high-dimensional data as considered here. Following a conjecture of Efron et al. (2004) and a result of Zou et al. (2004) for L_1-penalized regression, we treat d as the number of non-zero parameter estimates, modifying BIC for penalized model-based clustering as

$$BIC = -2\log L(\hat{\Theta}) + \log(n)d_e,$$

where $\hat{\Theta}$ is the MPLE, and $d_e = K + P + KP - 1 - q$ is the effective number of parameters; we set q as the number of the MPLE mean components that equal to 0. Hence, as expected, due to thresholding, $d_e < d$ with a large penalization parameter λ.

3. Results

We first present results for simulated data, then consider clustering samples and clustering genes for two microarray data.

3.1 Simulated Data

We consider first high-dimensional data, then, to facilitate comparisons with best subset selection, we also consider low-dimensional data.

3.1.1 LARGE P

We first considered simulated data as described in Hoff (2004) and Hoff (2005). In each simulated data set, there were two clusters based on the first 150 attributes, while only one cluster based on the remaining 850 attributes; in other words, there were a total of $P = 1000$ variables with the first 150 effective while the other 850 as noise variables in forming two clusters. Specifically, there were $n = 100$ observations with 85 in one cluster and 15 in the other: the first 150 variables were iid from $N(0,1)$ for the first cluster, whereas they were iid from $N(1.5,1)$ for the second cluster; the remaining 850 variables were all iid from $N(0,1)$ for either cluster. Hence, there were 150 informative attributes and 850 noise ones.

For each of 100 simulated data sets, we fitted a series of models with the number of components $K = 1$, 2, 3 and various values of penalization parameter $\lambda = 0$, 1, 1.5, 2, 5, 7.5, 10, 12.5, 15, 17.5, 20, 25 and 30.

Table 1 summarizes the means and standard errors of BIC for the standard clustering using all 1000 attributes, and those of BIC and penalization parameter λ of the selected models in penalized clustering. Table 2 gives the frequencies of the selected K for both the methods. Twenty out of a hundred times, standard model-based clustering incorrectly selected $K = 1$, failing to discover the existence of the two clusters of interest. In some sense, the result was reasonable and unsurprising: indeed, there were two clusters based on the first 150 attributes in the data; however, based on any of the other 850 attributes, there was only one cluster. Because standard clustering used all the attributes, noting that 850 was much larger than 150, as expected it might choose $K = 1$. In contrast, with an appropriate variable selection, penalized clustering more frequently chose the model with

	Standard	Penalized	
K	BIC	BIC	λ
1	92923 (2)	88393 (0)	1.00 (0.00)
2	92834 (12)	85738 (65)	9.60 (0.22)
3	96282 (13)	86778 (32)	9.60 (0.12)

Table 1: Mean BIC (with standard errors in parentheses) with various numbers (K) of clusters in the standard and penalized clustering for 100 simulated data sets. λ is the value of the penalization parameter minimizing BIC for the given K.

	Standard		Penalized				
K	Freq	BIC	Freq	BIC	λ	#Zero1	#Zero0
1	20	92923 (5)	0	-	-	-	-
2	80	92791 (10)	94	85679 (64)	9.57 (0.22)	1.1 (0.2)	832.5 (1.7)
3	0	-	6	86348 (43)	7.50 (0.00)	0.0 (0.0)	626.5 (5.6)

Table 2: Frequencies of the selected numbers (K) of clusters in the standard and penalized clustering from 100 simulated data sets with $P = 1000$. The corresponding means (with standard errors in parentheses) of BIC, λ, the number of the first 150 informative attributes excluded (#Zero1), and the number of the last 850 noise attributes excluded (#Zero0) are also included.

	Standard	Penalized				
K	Freq	Freq	#(1 and 2)	#(1 or 2, not both)	#Zero1	#Zero0
1	100	6	0	0	2	8
					(0)	(0)
2	0	38	30	6	0.26	5.08
					(0.09)	(0.26)
3	0	56	42	12	0.29	2.45
					(0.07)	(0.19)

Table 3: Frequencies of the selected numbers (K) of clusters in the standard and penalized clustering from 100 simulated data sets with $P = 10$. The frequencies of the corresponding models including both the two informative attributes (#(1 and 2)), or only one of the two (#(1 or 2, not both)), and the means (with standard errors in parentheses) of the number of the two informative attributes excluded (#Zero1), and the number of the other 8 noise attributes excluded (#Zero0) are also included.

$K = 2$, uncovering the interesting structure in the data. Importantly, the penalized approach can automatically select attributes: out of the total 850 noise attributes, on average, 833 attributes were correctly identified and not used in the final clustering; on the other hand, only one out of 150 informative attributes was not used.

Penalized clustering gave perfect assignments for $K = 2$: the 15 and 85 observations from two distributions/classes were correctly assigned to clusters 1 and 2 respectively. More interestingly, even when $K = 3$ was selected, the assignments were also correct; the clustering results for the 6 simulated data sets with $K = 3$ were the following: i) for two data sets, the 15 observations from class 1 were assigned to cluster 1, and 45 and 40 observations from class 2 were assigned to clusters 2 and 3 respectively, denoted as $\{(15,0,0),(0,45,40)\}$; ii) for two data sets: $\{(15,0,0),(0,49,36)\}$; iii) for the other two, $\{(15,0,0),(0,48,37)\}$ and $\{(15,0,0),(0,46,39)\}$ respectively.

It would be interesting to compare our method with best subset selection, which however was computationally prohibitive: with 1000 attributes, there were $2^{1000} \approx 10^{300}$ possible subsets/models! Note that, because there is no formal significance test for each individual attribute in model-based clustering, the commonly used sequential variable selection in regression is not applicable here. Below, we considered a problem with a much smaller P so that a comparison with best subset selection was possible.

3.1.2 SMALL P

We considered simulated data similar to those in the previous section, but with much fewer attributes. There were only $P = 10$ attributes, among which the first two were informative while the other eight were noise attributes; all other aspects remained the same. In best subset selection, first, each of the 1023 (non-null) candidate models containing all possible combinations of the 10 attributes was fitted using `Mclust()` in R with K ranging from 1 to 3 (and a common diagonal covariance matrix); for each model, the value of K was selected to give the minimum BIC; finally, we chose the final model from the 1023 fitted models as the one with the smallest BIC.

	Any attributes			≥ 2 attributes		
K	Freq	#(1 and 2)	#(1 or 2, not both)	Freq	#(1 and 2)	#(1 or 2, not both)
1	79	0	0	56	11	6
2	19	0	7	41	21	11
3	2	0	1	3	2	0

Table 4: Frequencies of the selected numbers (K) of clusters by best subset selection from 100 simulated data sets with $P = 10$. The frequencies of the corresponding models including both the two informative attributes (#(1 and 2)), or only one of the two (#(1 or 2, not both)). Two searches were conducted: all possible models including any combinations of the attributes, and only models including at least two attributes.

Table 3 gives the results for standard and penalized clustering. Again, in presence of the eight noise attributes, standard clustering always chose $K = 1$; in contrast, penalized clustering with automatic variable selection tended to chose $K > 1$, and the two informative attributes were most often retained. However, penalized clustering did not work as well as in the previous set-up with $P = 1000$: it chose $K = 3$ most often while keeping most of the noise attributes; the reason could be that with fewer informative attributes, this was a more difficult problem than that of the previous section. Nevertheless, compared with best subset selection (Table 4), it still worked much better. In best subset selection, if we considered all possible non-null models (i.e., all possible non-null combinations of attributes), it most often selected $K = 1$; in all cases, only one noise attribute was included in the selected model. Because there was indeed only one cluster according to any noise attribute, the choice of $K = 1$ based on any noise attribute was correct; in other words, the selected model was correct, though of no interest. This highlights a unique point that in variable selection for clustering, unlike in regression, a correct model for the data based on a subset of attributes (e.g., noise attributes here) may be of no interest!

In addition, we considered only the models containing at least two attributes in best subset selection (Table 4). It turned out that all the selected models included only two attributes; it was still more likely to choose $K = 1$, most often with two noise attributes, which was again caused by the the complication of having so many correct models of no interest: there was indeed only one cluster based on any two of the noise attributes. In summary, we concluded that subset selection was not suitable at all for variable selection in clustering, whereas penalized clustering was much more effective.

3.2 Tumor Subtype Discovery Using Gene Expression Profiles

Golub et al. (1999) studied discovering two subtypes of human acute leukemia, acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL), using microarray gene expression data. Distinguishing the two subtypes is clinically important because, for example, the same chemotherapy applied to ALL patients may not be suitable for AML patients. They used Affymetrix microarrays, each containing 7129 genes. We applied model-based clustering to their data with 38 patients, among which 21 were ALL while the other 11 were AML patients; each patient was treated as an observation while the genes were treated as attributes. Because most of the 7129 genes were not believed to be informative to discriminating between ALL and AML, and in fact, many of the genes

	Standard	Penalized	
K	BIC	BIC	λ
1	76966	69691	> 0
2	73802	68504	5
3	*71104*	66630	3
4	72232	65378	3
5	-	64034	2
6	-	62912	2
7	-	*61950*	2
8	-	63626	3

Table 5: BIC values for various numbers (K) of clusters in standard and penalized clustering for Golub' gene expression data.

	Standard			Penalized						
Samples/clusters	1	2	3	1	2	3	4	5	6	7
ALL	4	0	23	0	1	1	1	7	8	9
AML	0	4	7	6	0	0	0	4	1	0

Table 6: Clustering results for Golub's data.

were not even expressed in any sample, we filtered out most genes: we ranked the genes based on their sample variances across all 38 samples, and used only the top 2000 ones. For each method, we started from $g = 1$ and increased g until a minimum BIC was reached. The standard clustering chose $K = 3$ while the penalized one selected $K = 7$ (Table 5).

The clustering results are detailed in Table 6. The penalized method performed better than the standard clustering: the former incorrectly assigned five while the latter misclassified seven AML samples into the clusters with the ALL samples as the majority. In penalized clustering, although 35% of the mean parameter estimates were 0's, only eight genes had their cluster-specific mean estimates as 0's across all seven clusters, and hence were regarded as non-informative; previous studies demonstrated that there were indeed a large number of the genes differentially expressed between ALL and AML samples (Thomas et al., 2001; Pan, 2002).

3.3 Gene Function Discovery Using Gene Expression Profiles

Because many genes still have unknown functions, a biologically important subject is to computationally predict gene functions using, for example, gene expression profiles (Brown et al., 2000). The premise is that co-expressed genes are likely to share the same biological function. Due to incomplete knowledge, it is equally important to discover new gene functions; there is functional heterogeneity within most of functional categories, and there are probably many more uncharacterized gene functions. Clustering gene expression profiles has become a popular approach for both gene function prediction and discovery (Eisen et al., 1998; Wu et al., 2002; Zhou et al., 2002; Xiao and Pan, 2005). We stress that gene function discovery is more of a clustering or unsupervised

learning problem, as opposed to supervised learning: first, we do not restrict the genes of the same function to be in the same cluster/class; each of the multiple clusters of the genes coming from the same functional category may suggest some novel subcategory, a refinement of the original functional category. Second, we allow the existence of some unknown and novel classes: some genes of unknown function do not have to be predicted to have any one of the known functions because they may have some unknown new functions. Here, we considered gene function discovery using gene expression data for yeast *S cerevisiae*. Specifically, we used a gene expression data set containing 300 microarray experiments with gene deletions and drug treatments (Hughes et al., 2000). Variable selection was highly relevant here: first, as shown in simulation, incorrectly using noise attributes might degrade the performance of clustering, obscuring some interesting clustering structures; second, it was also biologically important to identify which microarray experiments (i.e., attributes) were informative in clustering, linking putative functions of gene clusters to biological perturbations underlying the microarray experiments.

One difficulty in evaluating the performance of a clustering algorithm for real data is how to choose an appropriate criterion. Although our interest was in clustering for gene function discovery, for the purpose of evaluations, we treated the problem as supervised learning: each gene had its response variable as its known function; gene functions were downloaded from the MIPS database (Mewes et al., 2004). For illustration, we only considered two gene functions, *cytoplasm* and *mitochondrion*, with 100 genes in each class as training data; we used other 406 and 212 genes in the two functional classes as test data.

We first used the training data without their class labels: we clustered the 200 gene expression profiles into, say K, clusters. Then, for each cluster, based on the class labels of the training observations assigned to the cluster, we assigned a class label or class probability to the cluster. There were two ways to do so, namely, hard classification and soft classification. For hard classification, we assign each cluster a class label that was possessed by the majority of the observations in the cluster. For a test observation that was assigned to a cluster, we predicted its class label as that of the cluster. For soft classification, for each cluster, we first calculated the proportion of the training observations in each class. Suppose P_k^c was such a proportion for class c in cluster k. For a test observation, if it was assigned to cluster k with posterior probability τ_k, it was classified to class c with probability $\sum_{k=1}^{K} P_k^c \tau_k$; summing these probabilities all over, we obtained an expected number of test observations assigned to each class.

3.3.1 USING THE ORIGINAL DATA

Both standard clustering and penalized clustering selected $K = 7$ by BIC (Table 7), with their predictive performances for the test data given in Table 8. The results were close. For penalized clustering, some MPLEs of the cluster-specific mean parameters were exactly zero; their numbers ranged from 36 to 126 in the seven clusters. However, there was no single attribute for which the mean parameter MPLEs were all zero in the seven clusters, hence all the 300 attributes were used in final clustering. This example showed that our penalized clustering performed as well as the standard clustering method for data with none or few non-informative attributes.

As a comparison, we applied the nearest shrunken centroids (NSC) (Tibshirani et al., 2003), random forests (RF) (Breiman, 2001), and support vector machines (SVM) (Vapnik, 1998), and thus treating the problem as supervised learning; the NSC was specifically developed for classification with gene expression data, while the RF and SVM are two state-of-the-art machine learning tools.

	Standard	Penalized	
K	BIC	BIC	λ
1	51420	49830	> 0
2	46993	46409	5
3	44252	44170	2
4	42674	42352	5
5	41944	41382	5
6	41891	40716	2
7	*41797*	*38368*	5
8	42138	39448	5

Table 7: BIC values for various numbers (K) of clusters in standard and penalized clustering for Hughes' gene expression data.

	Hard classification				Soft classification			
	Standard		Penalized		Standard		Penalized	
Truth	Pred=1	Pred=2	1	2	1	2	1	2
1	375	31	377	29	265.5	140.5	271.9	134.1
2	111	101	107	105	83.5	128.5	81.4	130.6
Accuracy	0.770		0.780		0.638		0.651	

Table 8: Predictions of standard clustering and penalized clustering over a separate test data set for Hughes' gene expression data. BIC selected $K = 7$ for both the methods.

	NC ($P = 300$)		NSC ($P = 6$)		RF		SVM	
Truth	Pred=1	Pred=2	1	2	1	2	1	2
1	313	93	304	102	316	90	362	44
2	61	151	68	144	48	164	81	131
Accuracy	0.751		0.725		0.777		0.798	

Table 9: Predictions over a separate test data set based on the nearest shrunken centroid without shrinkage (called NC) and with shrinkage (NSC), random forests (RF) and support vector machine (SVM) for Hughes' gene expression data.

Note that it is in general unfair to compare the predictive performance of a clustering method against that of a classification or supervised learning method; our purpose here was to use the modern classifiers as benchmarks. We used the default setting of R function `randomForest()` for the RF, and used `pamr()` and `svm()` for the NSC and SVM respectively; for the latter two, a 5-fold cross-validation was used to choose tuning parameters, such as the shrinkage parameter Δ in the NSC.

For the NSC, with the selected Δ, only six attributes remained in the final model; however, using all the attributes gave a slightly higher accuracy (Table 9). It was interesting to note that the NSC failed to perform better than either clustering with hard classification. There was some similarity between these two: if we regarded each cluster as a single class, then clustering with hard classification worked in a similar manner as the NSC. Nevertheless, a difference between the two was that, the NSC assumed only one cluster for each class, whereas clustering with hard classification allowed observations of the same class to go to different clusters. Unsurprisingly, the random forests and SVM also performed well for the data.

3.3.2 USING THE DATA WITH ADDED NOISE

It seemed that there were none or few non-informative attributes in the gene expression data for gene function prediction. To mimic other real applications, where a large number of microarray experiments were available, of which however only a fraction are informative, we added 700 noise attributes to the gene expression data. Each noise variable was generated from a standard Normal distribution independent of each other.

Table 10 summarizes the results of model fitting. By BIC, standard clustering selected $K = 2$ whereas penalized clustering chose $K = 6$. With $K = 2$, standard clustering gave quite bad results for the test data with an accuracy only at about 50% (Table 11), while it performed much better with $K = 6$ (results not shown); in contrast, penalized clustering gave much higher accuracy rates. Note that with many noise attributes, in agreement with that in the previous simulation study, standard clustering probably under-estimated the true number of the clusters of interest at $K = 2$. In addition, a distinct advantage of penalized clustering was that it could correctly identify most non-informative attributes: among the added 700 noise attributes, penalized clustering correctly identified 508 such attributes; in total, 521 attributes whose cluster-specific means were estimated to be 0 for all the clusters in penalized clustering.

By comparison, the NSC performed poorly (Table 12). By 5-fold cross-validation, the method selected a model with only one attribute. Using all attributes (or some subsets of the attributes) in the NSC did not help either. In contrast, both the RF and SVM performed well. It was not clear

	Standard	Penalized	
K	BIC	BIC	λ
1	173221	167960	> 0
2	*172209*	164909	20
3	173663	163148	15
4	175812	161830	15
5	177712	160906	15
6	181799	*159805*	10
7	185537	159917	15

Table 10: BIC values for various numbers (K) of clusters in standard and penalized clustering for Hughes' gene expression data with added noise.

	Hard classification				Soft classification			
	Standard		Penalized		Standard		Penalized	
Truth	Pred=1	Pred=2	1	2	1	2	1	2
1	263	143	307	99	203.0	203.0	265.4	140.6
2	149	63	73	139	106.0	106.0	82.2	129.8
Accuracy	0.527		0.722		0.500		0.639	

Table 11: Predictions of standard clustering and penalized clustering over a separate test data set for Hughes' gene expression data with added noise.

	NC ($P = 1000$)		NSC ($P = 1$)		LDA		RF		SVM	
Truth	Pred=1	Pred=2	1	2	1	2	1	2	1	2
1	213	193	200	206	275	131	312	94	322	84
2	105	107	111	101	65	147	50	162	57	155
Accuracy	0.518		0.487		0.683		0.767		0.772	

Table 12: Predictions for a separate test data set based on several classifiers for Hughes' gene expression data with added noise.

why the NSC did not work in this example. One possible explanation was that there were multiple centroids for each class, contrary to the assumption of the NSC that there was only a single one for each class. Hence, we applied linear discriminant analysis (LDA), which imposed an assumption similar to that of the NSC. Although there was a warning message from the R function `lda()` (due to $P > n$), the LDA performed much better than the NSC.

4. Discussion

Penalized likelihood has been widely used in model regularization, particularly for variable selection. A general theory has been laid out, see, for example, Fan and Li (2001), but mainly in the context of regression and classification. We are not aware of any other penalized likelihood approaches to multivariate model-based clustering, which we have studied in this article. In particular, it is confirmed that with the chosen L_1 penalty function, it yields a simple thresholding, enabling automatic variable selection. Our numerical examples demonstrate the usefulness of our proposal, especially for "high dimension, low sample size" settings. In particular, our numerical studies suggest the following two points. First, clustering without variable selection may fail to uncover interesting structures underlying the data. Second, best subset selection not only is computationally infeasible for clustering high-dimensional data, but also may fail in small problems. In addition to high computational demand, a key issue with best subset selection is the lack of an appropriate model selection criterion: if a conventional criterion is adopted based on the correctness of a model, because of the existence of many correct models, the criterion will not be useful; for example, any model containing one cluster based on any noise variable or their combinations is correct, but of no interest, in clustering analysis.

The basic idea proposed here is generalizable to semi-supervised learning where some, but not all, observations have class labels. An approach to semi-supervised learning is to conduct clustering analysis (i.e., class discovery) simultaneously with supervised learning (i.e., classification) with a mixture model (McLachlan and Peel, 2002). Alexandridis et al. (2004) proposed such a semi-supervised learning approach with an application to tumor classification and class discovery. A drawback of their approach was that variable selection had to be taken prior to clustering/classification: they conducted variable selection using either supervised learning or other heuristics, then used the selected variables in the subsequent clustering/classification. Pan et al. (2006) extended the penalized likelihood approach discussed here to semi-supervised learning so that variable selection is accomplished simultaneously along with model fitting (i.e., clustering/classification). In particular, their simulation results clearly demonstrated the advantage of simultaneous variable selection and model fitting over that of separating variable selection from model fitting.

We have used a common diagonal covariance matrix for all clusters. There are several practical reasons. First, in "high dimension, low sample size" settings, which are of particular interest here, an unrestricted covariance matrix is infeasible for $P > n$; some modeling is necessary. Second, in the context of linear discriminant analysis with "high dimension, low sample size" data, it has been found, both empirically and theoretically, that a diagonal covariance matrix may work better than non-diagonal ones (Tibshirani et al., 2003; Bickel and Levina, 2004); in fact, naive Bayes classifiers are well known to work well in these settings. These results suggest a possible advantage of using a diagonal covariance matrix in clustering. Finally, a careful examination reveals that other more flexible choices of the within-cluster/class covariance matrix, for example, allowing different diagonal covariance matrices for different clusters/classes, destroy the mechanism of variable selection

in model-based clustering, as the use of a common diagonal covariance matrix in the NSC for the same purpose of variable selection for classification. For example, consider an attribute x_p with distribution $N(0,1)$ or $N(0,2)$ for the two clusters respectively: although its means are equal, because of its different variances in the two clusters, it is still informative to discriminating between the two clusters. It is unclear how to realize automatic variable selection for other more general covariance matrices in penalized model-based clustering. Nevertheless, we acknowledge that it may be desirable to use more flexible covariance structures, for example, a non-diagonal covariance matrix, in some applications (McLachlan et al., 2003), and more work is needed to explore how to realize variable selection with such a choice in penalized model-based clustering.

In penalized/regularized methods, an important issue is the choice of the penalization parameter. Although cross-validation and other data-resampling methods can be adopted, due to their high computational cost and possibly sub-optimal performance (Efron, 2004), we have proposed a modified BIC as a model selection criterion. Based on the new results on degrees of freedom in the context of L_1-penalized regression (Efron et al., 2004; Zou et al., 2004), we propose counting only non-zero components of the maximum penalized likelihood estimate when calculating the effective number of parameters in BIC. Although it seemed to work well in our numerical examples, theoretical justifications and further evaluations are needed.

Acknowledgments

WP was supported by NIH grant HL65462 and a UM AHC Development grant, XS by NSF grants IIS-0328802 and DMS-0604394. WP thanks Benhuai Xie, Guanghua Xiao and Peng Wei for assistance with the gene expression data. The authors thank the two reviewers and the Action Editor for many helpful and constructive comments.

References

R. Alexandridis, S. Lin, and M. Irwin. Class discovery and classification of tumor samples using mixture modeling of gene expression data. *Bioinformatics*, 20:2546-2552, 2004.

P. J. Bickel, and E. Levina. Some theory for Fisher's linear discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations. *Bernoulli*, 10:989-1010, 2004.

L. Breiman. Random forests. *Machine Learning* 45:5-32, 2001.

M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussle. Knowledge-based analysis of microarray gene expression data using support vector machines. *Proc Natl Acad Sci USA*, 97:262-267, 2000.

W. C. Chang. On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, 32:267-275, 1983.

G. Ciuperca, A. Ridolfi, and J. Idier. Penalized maximum likelihood estimator for normal mixtures. *Scandinavian Journal of Statistics*, 30:45-59, 2003.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *JRSS-B*, 39:1-38, 1977.

B. Efron. The estimation of prediction error: covariance penalties and cross-validation. *JASA*, 99:619-632, 2004.

B. Efron, T. Hastie T, I. Johnstone I, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407-499, 2004.

M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95:14863-14868, 1998.

J. Fan, and R. Li. Variable selection via nonconcave penalized likelihood and its Oracle properties. *JASA*, 96:1348-1360, 2001.

C. Fraley, and A. E. Raftery. How many clusters? Which clustering methods? - Answers via model-based cluster analysis. *The Computer Journal*, 41:578-588, 1998.

C. Fraley, and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611-631, 2002.

C. Fraley, and A. E. Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. Technical report 486, Dept. of Statistics, University of Washington, 2005.

J. H. Friedman, and J. J. Meulman. Clustering objects on subsets of attributes (with discussion). *J. R. Stat. Soc. Ser. B*, 66:815-849, 2004.

D. Ghosh D, and A. M. Chinnaiyan. (2002). Mixture modeling of gene expression data from microarray experiments. *Bioinformatics*, 18:275-286, 2002.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531-537, 1999.

P. J. Green. On use of the EM for penalized likelihood estimation. *J. R. Stat. Soc. Ser. B*, 52:443-452, 1990.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* Springer, 2001.

P. D. Hoff. Discussion of 'Clustering objects on subsets of attributes' by Friedman and Meulman. *J. R. Stat. Soc. Ser. B*, 66:845-846, 2004.

P. D. Hoff. Subset clustering of binary sequences, with an application to genomic abnormality data. *Biometrics*, 61:1027-1036, 2005.

P. D. Hoff. Model-based subspace clustering. *Bayesian Analysis*, 1:321-344, 2006.

T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. Functional Discovery via a Compendium of Expression Profiles. *Cell*, 102:109-126, 2000.

A. Jasra, C. C. Holmes, and D. A. Stephens. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20:50-67, 2005.

S. Kim, M. G. Tadesse, and M. Vannucci. Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, 93:877-893, 2006.

H. Li, and F. Hong. Cluster-Rasch models for microarray gene expression data. *Genome Biology*, 2: research0031.1-0031.13, 2001.

J. S. Liu, J. L. Zhang, M. J. Palumbo, C. E. Lawrence. Bayesian clustering with variable and transformation selection (with discussion). *Bayesian Statistics*, 7:249-275, 2003.

O. L. Mangasarian, and E. W. Wild. Feature selection in k-median clustering. *Proceedings of SIAM International Conference on Data Mining, Workshop on Clustering High Dimensional Data and its Applications*, April 24, 2004, La Buena Vista, FL, pages 23-28.

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

G. J. McLachlan, and D. Peel. *Finite Mixture Model.* New York, John Wiley & Sons, Inc, 2002.

G. J. McLachlan, D. Peel, and R. W. Bean. Modeling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis*, 41:379-388, 2003.

H. W. Mewes, C. Amid, R. Arnold, D. Frishman, U. Guldener, G. Mannhaupt, M. Munsterkotter, P. Pagel, N. Strack, V. Stumpflen, J. Warfsmann, and A. Ruepp. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, 32:D41-D44, 2004.

W. Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 12:546-554, 2002.

W. Pan, X. Shen, A. Jiang, and R. P. Hebbel. Semi-supervised learning via penalized mixture model with application to microarray sample classification. *Bioinformatics*, 22:2388-2395, 2006.

A. E. Raftery. Discussion of "Bayesian clustering with variable and transformation selection" by Liu et al. *Bayesian Statistics*, 7:266-271, 2003.

A. E. Raftery, and N. Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101:168-178, 2006.

S. Richardson, and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. *JRSS-B*, 59:731-758, 1997.

G. Schwarz. Estimating the dimensions of a model. *Annals of Statistics*, 6:461-464, 1978.

X. Shen, and J. Ye. Adaptive model selection. *Journal of the American Statistical Association*, 97:210-221, 2002.

M. G. Tadesse, N. Sha, and M. Vannucci. Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100:602-617, 2005.

J. G. Thomas, J. M. Olson, S. J. Tapscott, and L. P. Zhao. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research*, 11:1227-1236, 2001.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *JRSS-B*, 58:267-288, 1996.

R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Class prediction by nearest shrunken centroids, with application to DNA microarrays. *Statistical Science*, 18:104-117, 2003.

V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

L. F. Wu, T. R. Hughes, A. P. Davierwala, M. D. Robinson, R. Stoughton, and S. J. Altschuler. Large-scale prediction of saccharomyces cerevisiae gene function using overlapping transcriptional clusters. *Nature Genetics*, 31:255-265, 2002.

G. Xiao, and W. Pan. Gene function prediction by a combined analysis of gene expression data and protein-protein interaction data. *Journal of Bioinformatics and Computational Biology*, 3:1371-1389, 2005.

K. Y. Yeung, and W. L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17:763-774, 2001.

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*, 17:977-987, 2001.

X. Zhou, M. C. Kao, and W. H. Wong. Transitive functional annotation by shortest-path analysis of gene expression data. *Proc Natl Acad Sci USA*, 99:12783-12788, 2002.

H. Zou, T. Hastie, and R. Tibshirani. On the "Degrees of Freedom" of the Lasso. Technical report, Dept. of Statistics, Stanford University, 2004. Available at http://stat.stanford.edu/∼hastie/pub.htm.