Citation: D. B. Dahl (2006), Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model, in Bayesian Inference for Gene Expression and Proteomics, Kim-Anh Do, Peter Müller, Marina Vannucci (Eds.), Cambridge University Press.

10

Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model

DAVID B. DAHL

Texas A&M University

Abstract

This chapter describes a clustering procedure for microarray expression data based on a well-defined statistical model, specifically, a conjugate Dirichlet process mixture model. The clustering algorithm groups genes whose latent variables governing expression are equal, that is, genes belonging to the same mixture component. The model is fit with Markov chain Monte Carlo and the computational burden is eased by exploiting conjugacy. This chapter introduces a method to get a point estimate of the true clustering based on least-squares distances from the posterior probability that two genes are clustered. Unlike ad hoc clustering methods, the model provides measures of uncertainty about the clustering. Further, the model automatically estimates the number of clusters and quantifies uncertainty about this important parameter. The method is compared to other clustering methods in a simulation study. Finally, the method is demonstrated with actual microarray data.

10.1 Introduction

The main goal of clustering microarray data is to group genes that present highly correlated data; this correlation may reflect underlying biological factors of interest, such as regulation by a common transcription factor. A variety of heuristic clustering methods exist, including k-means clustering (MacQueen 1967) and hierarchical agglomerative clustering. These methods have had an enormous impact in genomics (Eisen et al. 1998) and are intuitively appealing. Nevertheless, the statistical properties of these heuristic clustering methods are generally not known. Model-based clustering procedures have been proposed for microarray data, including (1) the MCLUST procedure of Fraley and Raftery (2002) and Yeung et al. (2001), and (2) the Bayesian mixture model based clustering of Medvedovic and Sivaganesan (2002) and Medvedovic et al. (2004). Model-based techniques offer advantages over heuristic schemes, such as the ability to assess uncertainty about the resulting clustering and to formally estimate the number of clusters.

This chapter describes a model-based clustering procedure for microarray expression data based on a well-defined statistical model, specifically, a conjugate Dirichlet process mixture (DPM) model. In the assumed model, two genes come from the same mixture component if and only if their relevant latent variables governing expression are equal. The model itself, known as BEMMA for Bayesian Effects Model for Microarrays, was first introduced by Dahl and Newton (submitted) as a means of exploiting the clustering structure of data for increased sensitivity in a battery of correlated hypothesis tests (e.g., finding differentially expressed genes). The focus of this chapter is not finding differential expression, but rather identifying the underlying clustering structure of expression data.

Computations for Bayesian mixture models can be challenging. Unlike the finite and infinite mixture models of Medvedovic and Sivaganesan (2002) and Medvedovic et al. (2004), the proposed method is, however, conjugate. This conjugacy permits the latent variables to be integrated away, thereby simplifying to state space over which the Markov chain is run. The model is fit using Markov chain Monte Carlo (MCMC), specifically using the conjugate Gibbs sampler (MacEachern 1994; Neal 1992) and the merge–split algorithm of Dahl (2003). Each iteration of the Markov chain yields a clustering of the data.

Providing a single point estimate for clustering based on the thousands of clusterings in the Markov chain has been proven to be challenging (Medvedovic and Sivaganesan 2002). One approach is to select the observed clustering with the highest posterior probability; this is called the maximum a posteriori (MAP) clustering. Unfortunately, the MAP clustering may only be slightly more probable than the next best alternative, yet represent a very different allocation of observations. Alternatively, Medvedovic and Sivaganesan (2002) and Medvedovic et al. (2004) suggest using hierarchical agglomerative clustering based on a distance matrix formed using the observed clusterings in the Markov chain. It seems counterintuitive, however, to apply an ad hoc clustering method on top of a model which itself produces clusterings.

This chapter proposes a method to form a clustering from the many clusterings observed in the Markov chain. The method is called least-squares model-based clustering (or, simply, least-squares clustering). It selects the observed clustering from the Markov chain that minimizes the sum of squared deviations from the pairwise probability matrix that genes are clustered. The least-squares clustering has the advantage that it uses information from all the clusterings (via the pairwise probability matrix) and is intuitively appealing because it selects the "average" clustering (instead of forming a clustering via an external, ad hoc algorithm).

Section 10.2 presents the details of the proposed model, including the likelihood, prior, and how to set the hyperparameters. Section 10.2.3 describes the model fitting approach and how the conjugate nature of the model aids in its fitting. Section 10.3 details the new least-squares clustering estimator using draws from the posterior clustering distribution. Section 10.4 presents a simulation study showing that the method compares well with other clustering methods. Finally, the model is demonstrated in Section 10.5, using a microarray data set with 10,043 probe sets, 10 treatments conditions, and 3 replicates per treatment condition. This section also introduces the effects intensity plot which displays the clustering of all genes simultaneously. The chapter ends with a discussion in Section 10.6.

10.2 Model

The model-based clustering procedure presented here is based on the Bayesian Effects Model for Microarrays (BEMMA) of Dahl and Newton (submitted). The model was originally proposed as a means to gain increased sensitivity in a battery of correlated hypothesis tests by exploiting the underlying clustering structure of data. In their application, Dahl and Newton (submitted) were interested in identifying differentially expressed genes. In this chapter, we apply their model to the task of clustering highly correlated genes that may reveal underlying biological factors of interest.

10.2.1 Likelihood Specification

The model assumes the following sampling distribution:

$$y_{gtr} \,|\mu_g, \tau_{gt}, \lambda_g \sim N(y_{gtr} \,|\mu_g + \tau_{gt}, \lambda_g), \tag{10.1}$$

where y_{gtr} is a suitably transformed expression of replicate r $(r = 1, \ldots, R_t)$ of gene g $(g = 1, \ldots, G)$ at treatment condition t $(t = 1, \ldots, T)$ and $N(z|a, b)$ denotes the univariate normal distribution with mean a and variance $1/b$ for the random variable z. The parameter μ_g represents a gene-specific mean, the gene-specific treatment effects are $\tau_{g1}, \ldots, \tau_{gT}$, and λ_g is a gene-specific sampling precision.

The model assumes that coregulated genes have the same treatment effects and precision. That is, genes g and g' are in the same cluster if

$(\tau_{g1}, \ldots, \tau_{gT}, \lambda_g) = (\tau_{g'1}, \ldots, \tau_{g'T}, \lambda_{g'})$. The clustering can be encoded with cluster labels $c_1, \ldots, c_G$, where $c_g = c_{g'}$ if and only if genes g and g' are in the same cluster.

The gene-specific means $\mu_1, \ldots, \mu_G$ are nuisance parameters; they are not related to differential expression across treatments and they are not used to define clusters. Indeed, there can exist constant differences in expression from probe sets known to be coregulated. These constant differences may be due to the biology (e.g., mRNA degradation) or the microarray technology (e.g., hybridization differences between probes or labeling efficiency). Whatever the reason, constant differences between probe sets may naturally exist in microarray experiments, yet they are not of interest. Indeed, two genes having a constant difference across treatments is the essence of coregulation.

The nuisance parameters $\mu_1, \ldots, \mu_G$ could be handled by specifying a prior over them and integrating the likelihood implied by (10.1) over this prior. The resulting model would be nonconjugate since the prior specification (detailed in the next subsection) induces mixing with respect to both the treatment effects $\tau_{g1}, \ldots, \tau_{gT}$ and the sampling precision λ_g. (If the mixing were only with respect to the treatment effects, conjugacy would remain intact when integrating over the nuisance parameters $\mu_1, \ldots, \mu_G$.) Fitting this nonconjugate model would be computationally challenging in the presence of thousands of genes.

The following pragmatic approach is used to address the nuisance parameters $\mu_1, \ldots, \mu_G$. Select a reference treatment (taken here to be the first treatment for notational convenience). Let d_g be a vector whose elements are $y_{gtr} - \overline{y}_{g1}$ for $t \geq 2$, where $\overline{y}_{g1}$ is the mean of the reference treatment. Further, let $\tau_g = (\tau_{g2}, \ldots, \tau_{gT})$ be a treatment effect vector and $N = \sum_{t=2}^{T} R_t$ be the dimension of d_g. Simple calculations reveal that d_g is independent of the nuisance parameters $\mu_1, \ldots, \mu_G$ and distributed:

$$d_g \,|\tau_g, \lambda_g \sim N_N(d_g \,|\mathrm{X}\tau_g, \lambda_g \mathrm{M}), \tag{10.2}$$

where $N_c(z|a, b)$ is a c-dimensional multivariate normal distribution with mean vector a and covariance matrix b^{-1} for the random vector z. Also, M is an $N \times N$ matrix equal to $(\mathrm{I} + \frac{1}{R_1}\mathrm{J})^{-1}$, where I is the identify matrix and J is a matrix of ones. Finally, X is an $N \times (T - 1)$ design matrix whose rows contain all zeros except where the number 1 is needed to pick off the appropriate element of τ_g. If one prefers, the model could equivalently be written in terms of sample averages from each treatment. This would, for example, reduce the dimension of d_g from N to T.

10.2.2 Prior Specification

Clustering based on equality of τ's and λ's across genes is achieved by using a Dirichlet process prior for these model parameters, resulting in a Dirichlet process mixture (DPM) model. See Müller and Quintana (2004) and references therein for a review of the DPM model literature. The model assumes the following prior:

$$\begin{aligned} \tau_g, \lambda_g \,|F(\tau_g, \lambda_g) &\sim F(\tau_g, \lambda_g) \\ F(\tau, \lambda) &\sim DP(\eta_0 F_0(\tau, \lambda)), \end{aligned} \tag{10.3}$$

where $DP(\eta_0 F_0(\tau, \lambda))$ is the Dirichlet process having centering distribution $F_0(\tau, \lambda)$ for the random variables τ and λ and mass parameter η_0. The centering distribution $F_0(\tau, \lambda)$ is a joint distribution for τ and λ having the following conjugate density:

$$\begin{aligned} p(\tau, \lambda) &= p(\tau\,|\lambda) p(\lambda) \\ &= N_{T-1}(\tau\,|0, \lambda\Psi_0) Ga(\lambda\,|\alpha_0, \beta_0), \end{aligned} \tag{10.4}$$

where $Ga(z\,|a, b)$ is the gamma distribution with mean a/b for the random variable z and α_0, β_0, and Ψ_0 are fixed hyperparameters set based on either prior experience or current data.

10.2.3 Sampling from the Posterior Distribution

Quintana and Newton (2000) and Neal (2000) have good reviews and comparisons of methods for fitting DPM models. We suggest fitting the proposed model using MCMC. The centering distribution $F_0(\tau, \lambda)$ in (10.4) is conjugate to the likelihood for τ_g and λ_g in (10.2). Thus, the model parameters may be integrated away, leaving only the clustering of the G genes. As a result, the stationary distribution of a Markov chain for the model is $p(c_1, \ldots, c_G|d_1, \ldots, d_g)$, the posterior distribution of the clustering configurations. This technique was shown by MacEachern (1994) and MacEachern et al. (1999) to greatly improve the efficiency of Gibbs sampling and sequential importance sampling, respectively. Efficiency is very important if the model is to be useful in practice.

It should be noted that the technique of integrating away the model parameters is merely a device used for model fitting. Inference on the model parameters $\tau_1, \ldots, \tau_G$ and $\lambda_1, \ldots, \lambda_G$ can still be made by sampling from posterior distribution of the model parameters (i.e., (10.6) in next subsection) after having obtained samples from the posterior clustering distribution.

The Gibbs sampler can be used to sample from the posterior clustering distribution of conjugate DPM models (MacEachern 1994; Neal 1992). The Gibbs sampler repeatedly takes a gene out of the clustering and draws a new cluster label from the full conditional distribution. Because the Gibbs sampler only moves one gene at a time, it may explore the posterior clustering distribution rather slowly. Jain and Neal (2004) and Dahl (2003) present merge–split algorithms that attempt to update more than one cluster label at a time. The Gibbs sampler and both of these merge–split samplers require the evaluation of the posterior predictive distribution. The next subsection gives the full conditional distribution and the posterior predictive distribution for the proposed model.

10.2.4 Full Conditional and Posterior Predictive Distributions

The full conditional distribution is essential for fitting the model using the Gibbs sampler. Let c_{-i} denote the collection of all cluster labels except that corresponding to gene i. For notational convenience, let the cluster labels in c_{-i} be numbered from 1 to k and let $k+1$ be the label of an empty cluster. Finally, let n_c be the number of cluster labels equal to c (not counting c_i), unless cluster c is empty, in which case, n_c is set to the mass parameter η_0. The full conditional distribution is a multinomial distribution given by

$$p(c_i = c \,|\, c_{-i}, d_1, \ldots, d_G) \propto n_c \int B(d_i|\tau, \lambda) p(\tau, \lambda|D_c)\, d\tau\, d\phi, \tag{10.5}$$

for $c = 1, \ldots, k+1$, where $B(d_i|\tau, \lambda)$ is the normal distribution in (10.2), and $p(\tau, \lambda|D_c)$ is the density of the posterior distribution of τ and λ based on the prior $F_0(\tau, \lambda)$ in (10.4) and all differences d_j for which $j \neq i$ and $c_j = c$. In the case of an empty cluster, $p(\tau, \lambda|D_c)$ is just the density of the prior $F_0(\tau, \lambda)$ and n_c is set to the mass parameter η_0 instead of 0; otherwise, it is rather straightforward to show that

$$\begin{aligned} p(\tau, \lambda \,|D_c) &\propto p(\tau \,|\lambda, D_c) p(\lambda \,|D_c) \\ &= N_{T-1}(\tau \,|\Psi_{n_c}^{-1} S_1, \lambda \Psi_{n_c}) Ga(\lambda \,|\alpha_{n_c}, \beta_1), \end{aligned} \tag{10.6}$$

where

$$\begin{aligned} \Psi_{n_c} &= \Psi_0 + n_c \mathrm{X}'\mathrm{M}\mathrm{X}, \\ \alpha_{n_c} &= \alpha_0 + \frac{n_c N}{2}, \\ \beta_1 &= \beta_0 + \frac{1}{2} S_2 - \frac{1}{2} S_1' \Psi_{n_c}^{-1} S_1, \\ S_1 &= \sum_{d \in D_c} \mathrm{X}'\mathrm{M}d, \quad \text{and} \\ S_2 &= \sum_{d \in D_c} d'\mathrm{M}d. \end{aligned} \tag{10.7}$$

The integral in (10.5) refers to the posterior predictive distribution of d belonging to cluster c. For conjugate DPM models, this distribution can usually be found in closed form. In the present model, the posterior predictive distribution for a new difference vector d^* evaluated at d (when its cluster label c^* is c and given the data D_c having cluster label c) has the following density:

$$p(d^* = d \,|c^* = c, D_c) = c_n \frac{\beta_1^{\alpha_{n_c}}}{\beta_2^{\alpha_{n_c+1}}}, \tag{10.8}$$

where

$$\begin{aligned} \beta_2 &= \beta_0 + \frac{1}{2}S_2 + \frac{1}{2}d'\mathrm{M}d - \frac{1}{2}(\mathrm{X}'\mathrm{M}d + S_1)'\Psi_{n_c+1}^{-1}(\mathrm{X}'\mathrm{M}d + S_1) \\ c_n &= \frac{\Gamma(\alpha_{n_c+1})}{\Gamma(\alpha_{n_c})}\sqrt{\frac{\mid \Psi_n \mid\mid \mathrm{M} \mid}{\mid \Psi_{n_c+1} \mid (2\pi)^N}}. \end{aligned} \tag{10.9}$$

It is interesting to note that (10.8) is not the usual multivariate Student t-distribution.

10.2.5 Setting the Hyperparameters

Lacking strong prior belief about the hyperparameters η_0, α_0, β_0, and Ψ_0, an empirical Bayes procedure can be used. Notice that (10.7) implies that $\Psi_{n+1} = \Psi_n + \mathrm{X}'\mathrm{M}\mathrm{X}$ and $\alpha_{n+1} = \alpha_n + \frac{N}{2}$. That is, for each additional observation, Ψ_n and α_n are incremented by $\mathrm{X}'\mathrm{M}\mathrm{X}$ and $\frac{N}{2}$, respectively. It is natural, therefore, to set the hyperparameter Ψ_0 to $n_0\mathrm{X}'\mathrm{M}\mathrm{X}$ and the hyperparameter α_0 to $n_0\frac{N}{2}$, for $n_0 > 0$ representing the number of observations that prior experience is worth. By default, we recommend $n_0 = 1$.

As shown in (10.1) and (10.4), the hyperparameters α_0 and β_0 are, respectively, the shape and rate parameters of the gamma prior distribution for the precision of an observation in a given cluster. We recommend setting α_0 and β_0 such that the mean of this distribution, α_0/β_0, matches a data-driven estimate of the expected precision for a cluster. Equivalently, in terms of the standard deviation, choose α_0 and β_0 so that $\sqrt{\beta_0/\alpha_0}$ matches the estimated standard deviation for a cluster. The software implementation of BEMMA uses the median standard deviation across all probe sets if no value is specified by the user. Since $\alpha_0 = n_0\frac{N}{2}$ (from the previous paragraph), specifying the expected standard deviation implies a value for β_0.

The final hyperparameter to consider is the mass parameter η_0, which affects the distribution on the number of clusters. The mass parameter in DPM models has been well studied (Escobar 1994; Escobar and West 1995; Liu 1996;

Medvedovic and Sivaganesan 2002). From Antoniak (1974), the prior expected number of clusters is

$$K(G) = \sum_{g=1}^{G} \frac{\eta_0}{\eta_0 + g - 1}.$$

In some DPM model applications, the mass parameter is set to 1.0. This seems overly optimistic for microarray experiments since, for example, it implies a prior belief that there are less than 12 clusters in data set with 50,000 genes. We use an empirical Bayes approach which sets η_0 such that the posterior expected number of clusters equals the prior expected number of clusters. The software implementation of BEMMA provides this option.

10.3 Inference

Draws $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_B$ from the posterior clustering distribution can be obtained using MCMC, where B is a number of sampled clusterings. Several methods have been used to arrive at a point estimate of the clustering using draws from the posterior clustering distribution. Perhaps the simplest method is to select the observed clustering that maximizes the density of the posterior clustering distribution. This is known as the maximum a posteriori (MAP) clustering. Unfortunately, the MAP clustering may only be slightly more probable than the next best alternative, yet represent a very different allocation of observations.

For each clustering $\boldsymbol{c}$ in $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_B$, an association matrix $\delta(\boldsymbol{c})$ of dimension $G \times G$ can be formed whose (i, j) element is $\delta_{i,j}(\boldsymbol{c})$, an indicator of whether gene i is clustered with gene j. Element-wise averaging of these association matrices yields the pairwise probability matrix of clustering, denoted $\widehat{\pi}$. Medvedovic and Sivaganesan (2002) and Medvedovic et al. (2004) suggest forming a clustering estimate by using the pairwise probability matrix $\widehat{\pi}$ as a distance matrix in hierarchical agglomerative clustering. It seems counterintuitive, however, to apply an ad hoc clustering method on top of a model which itself produces clusterings.

We introduce the least-squares model-based clustering (or, simply, least-squares clustering), a new method for estimating the clustering of observations using draws from a posterior clustering distribution. As with the method of Medvedovic and Sivaganesan (2002), the method is based on the pairwise probability matrix $\widehat{\pi}$ that genes are clustered together. The method differs, however, in that it selects one of the observed clusterings in the Markov chain as the point estimate. Specifically, the least-squares clustering $\boldsymbol{c}_{\text{LS}}$ is the observed clustering $\boldsymbol{c}$ which minimizes the sum of squared deviations of its association matrix $\delta(\boldsymbol{c})$ from the pairwise probability matrix $\widehat{\pi}$:

$$\boldsymbol{c}_{\mathrm{LS}} = \underset{c \in \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_B\}}{\arg\min} \sum_{i=1}^{G} \sum_{j=1}^{G} (\delta_{i,j}(\boldsymbol{c}) - \widehat{\pi}_{i,j})^2. \tag{10.10}$$

The least-squares clustering has the advantage that it uses information from all the clusterings (via the pairwise probability matrix) and is intuitively appealing because it selects the "average" clustering (instead of forming a clustering via an external, ad hoc clustering algorithm).

Uncertainty about a particular clustering estimate can be accessed from the posterior clustering distribution. For example, one can readily estimate the probability that two genes are clustered together by computing the relative frequency of this event among the clusterings in the Markov chain. Also, the posterior distribution of the number of clusters is easily obtained.

10.4 Simulation Study

This section provides a simulation study comparing the proposed clustering method to several standard methods. To assess the robustness of the clustering methods, four degrees of clustering are considered:

Heavy clustering: Data with 12 clusters of 100 genes per cluster.
Moderate clustering: Data with 60 clusters of 20 genes per cluster.
Weak clustering: Data with 240 clusters of 5 genes per cluster.
No clustering: Data with no clustering of the genes.

Each data set has 1,200 genes. The simulated experimental design is a time-course experiment (with three time points) and two groups, making in all $T = 6$ treatments.

Each cluster may be classified as either containing genes that are differentially expressed or equivalently expressed. Clusters that are equivalently expressed have equal treatment effects for the two treatments within a time point. Clusters that are differentially expressed have independently sampled treatment effects at one or more of the time points. In all cases, the precision λ for a cluster is a draw from a gamma distribution with mean 1 and variance $1/10$, the treatment effects $\tau_1, \ldots, \tau_6$ for a cluster are drawn independently from a normal distribution with mean 0 and variance $(9\lambda)^{-1}$, and the gene-specific shift μ is drawn from a normal distribution with mean 7 and variance 1.

Regardless of the degree of clustering, each data set contains 300 genes that are differentially expressed. A third of the differentially expressed clusters have unequal treatment effects at only one time point, a third have unequal treatment effects at two time points, and the remaining third have unequal treatment effects at all three time points. Finally, the observed data is drawn as specified in (10.1), with the first time point having five replicates per treatment and the other time points having three replicates.

10.4.1 Simulation Results

The MAP and least-squares clusterings based on the BEMMA model (as described in Section 10.3) were computed for each simulated data set and are labeled "BEMMA(map)" and "BEMMA(least-squares)," respectively. To compare the performance of BEMMA, the MCLUST procedure (Fraley and Raftery 1999, 2002) and hierarchical clustering (Hartigan 1975; Ihaka and Gentleman 1996) were applied to the simulated data. Specifically, the following methods were used:

MCLUST: The Mclust( ) function of the mclust package of R (Ihaka and Gentleman 1996)

HCLUST(correlation,average): Hierarchical clustering where the distance between genes was one minus the square of the Pearson correlation of the sample treatment means and using the "average" agglomeration method

HCLUST(correlation,complete): Hierarchical clustering using correlation distance and using the "complete" agglomeration method

HCLUST(effects,average): Hierarchical clustering where the distance between genes was the Euclidean distance between the sample treatment effects and using the "average" agglomeration method

HCLUST(effects,complete): Hierarchical clustering using effects distance and using the "complete" agglomeration method

Hierarchical clustering is a heuristic clustering procedure, while BEMMA and MCLUST are model-based clustering procedures. The number of clusters in the data is unspecified in the proposed model. For simplicity, the number of clusters for the other clustering methods was set to the true number of clusters.

There are many indices for measuring the agreement between two clusterings. In a comprehensive comparison, Milligan and Cooper (1986) recommend

Table 10.1. *Adjusted Rand Index for BEMMA and Other Methods*

Degree of clustering	Clustering method	Adjusted Rand index w/95% C.I.
Heavy	MCLUST	0.413 (0.380, 0.447)
	BEMMA(least-squares)	0.402 (0.373, 0.431)
	BEMMA(map)	0.390 (0.362, 0.419)
	HCLUST(effects,average)	0.277 (0.247, 0.308)
	HCLUST(effects,complete)	0.260 (0.242, 0.279)
	HCLUST(correlation,complete)	0.162 (0.144, 0.180)
	HCLUST(correlation,average)	0.156 (0.141, 0.172)
Moderate	BEMMA(least-squares)	0.154 (0.146, 0.163)
	MCLUST	0.144 (0.136, 0.152)
	BEMMA(map)	0.127 (0.119, 0.135)
	HCLUST(effects,complete)	0.117 (0.111, 0.123)
	HCLUST(effects,average)	0.101 (0.095, 0.107)
	HCLUST(correlation,average)	0.079 (0.075, 0.083)
	HCLUST(correlation,complete)	0.073 (0.068, 0.078)
Weak	MCLUST	0.050 (0.048, 0.052)
	HCLUST(effects,complete)	0.045 (0.043, 0.048)
	BEMMA(least-squares)	0.042 (0.040, 0.043)
	HCLUST(effects,average)	0.037 (0.035, 0.038)
	BEMMA(map)	0.031 (0.030, 0.033)
	HCLUST(correlation,average)	0.029 (0.027, 0.030)
	HCLUST(correlation,complete)	0.027 (0.025, 0.029)

Note: Large values of the adjusted Rand index indicate better agreement between the estimated and true clustering.

the adjusted Rand index (Hubert and Arabie 1985; Rand 1971) as the preferred measure of agreement between two clusterings. Large values for the adjusted Rand index mean better agreement. That is, an estimated clustering that closely matches the true clustering has a relatively large adjusted Rand index.

Table 10.1 shows the adjusted Rand index for BEMMA and the other clustering methods. Under heavy, moderate, and weak clustering, the MCLUST does very well. BEMMA too performs well. Notice that the newly proposed least-squares clustering method of Section 10.3 performs better than the MAP clustering method. The hierarchical clustering procedures generally do not perform very well, especially those based on the correlation distance matrix.

In summary, the simulation study suggests that the least-squares clustering is able to estimate the true clustering relatively well. It does about as well as MCLUST and much better than hierarchical clustering, even though

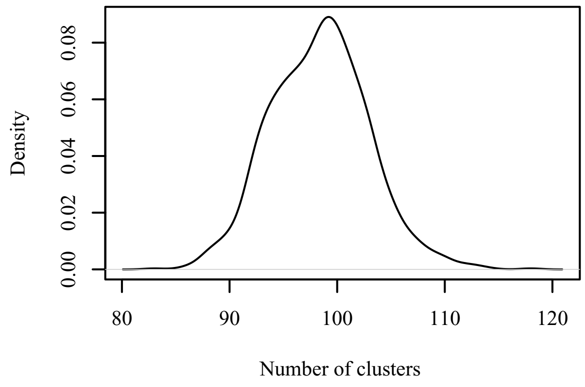

Fig. 10.1. Posterior distribution of the number of clusters.

BEMMA does not have the benefit of knowing the true number of clusters. Further, the model-based nature of BEMMA allows one to readily assess the variability in the estimated clustering. Finally, when information on differential expression is desired, BEMMA is also shown by Dahl and Newton (submitted) to be a very sensitive method for detecting differentially expressed genes.

10.5 Example

The proposed method was implemented on a replicated, multiple treatment microarray experiment. Researchers were interested in the transcriptional response to oxidative stress in mouse skeletal muscle and how that response changes with age. Young (5-month-old) and old (25-month-old) mice were treated with an injection of paraquat (50 mg/kg). Mice were sacrificed at 1, 3, 5, and 7 hours after paraquat treatment or were sacrificed having not received paraquat (constituting a baseline). Thus, $T = 10$ experimental conditions were under consideration. Edwards et al. (2003) discuss the experimental details. All treatments were replicated three times. Gene expression was measured on $G = 10{,}043$ probe sets using high-density oligonucleotide microarrays manufactured by Affymetrix (MG-U74A arrays). The data was background-corrected and normalized using the Robust Multichip Averaging (RMA) method of Irizarry et al. (2003) as implemented in the affy package of BioConductor (Gentleman et al. 2004). For a review of the issues and procedures for background-correction and normalization, see Irizarry et al. (2003) and Dudoit et al. (2002)

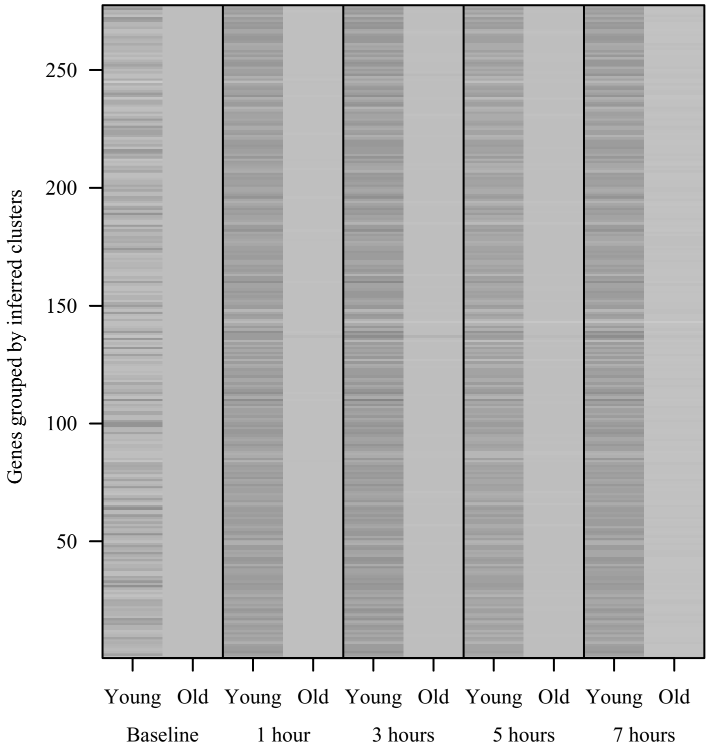

Fig. 10.2. Effects intensity plot for genes clustered with the probe set of interest. This effects intensity plot shows the estimated treatment effects for the other probe sets that were clustered with the probe set of interest in the least-squares clustering. Rows correspond to the genes in this cluster. The reference treatment is old at baseline and is shaded gray. Lighter shades indicate underexpression relative to the reference, whereas darker shades indicate overexpression.

10.5.1 Burn-in and Posterior Simulation

The model was fit using MCMC. The hyperparameters α_0, β_0, Ψ_0, and η_0 were set according to Section 10.2.5, resulting in the prior and posterior expected number of clusters being 98 (i.e., mass parameter $\eta_0 = 15$).

Two Markov chains were run from one of two extreme starting configurations: (1) all genes belonging to a single cluster, or (2) each gene belonging to its own cluster. One iteration of the Markov chain consisted of a Gibbs scan (accounting for more than 97% of the CPU time) and five sequentially allocated merge–split proposals of Dahl (2003). The moving average (of size 50) of the number of clusters was monitored. When these averages crossed, the chains were declared to be burned-in. Trace plots of various univariate

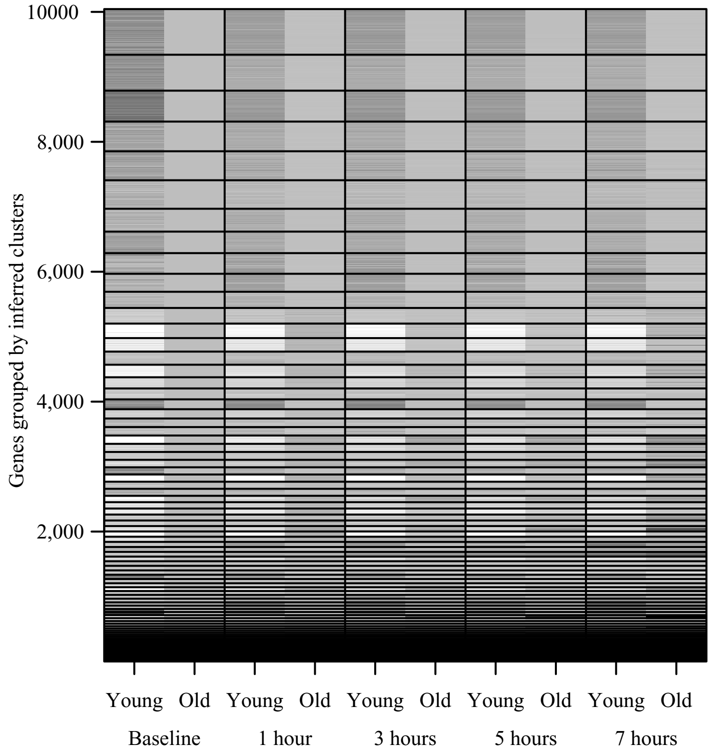

Fig. 10.3. Effects intensity plot for all probe sets. This figure, based on the least-squares clustering, shows the estimated treatment effects of all the clusters simultaneously and sorts the clusters based on size.

summaries of the chains support this burn-in procedure. Two desktop computers independently implemented this burn-in procedure and then sampled from the posterior for less than four days. To reduce disk storage requirements, the sample was thinned by saving only one in 100 states, leaving a total of 1,230 nearly independent draws from the posterior distribution.

10.5.2 Inference

The least-squares clustering (described in Section 10.3) of the expression data had 105 clusters, ranging in size from 1 to 700 probe sets. Pairwise probabilities of coregulation can readily be obtained by examining the relative frequency that genes are clustered together in states of the Markov chain. The posterior distribution of the number of clusters is given in Figure 10.1.

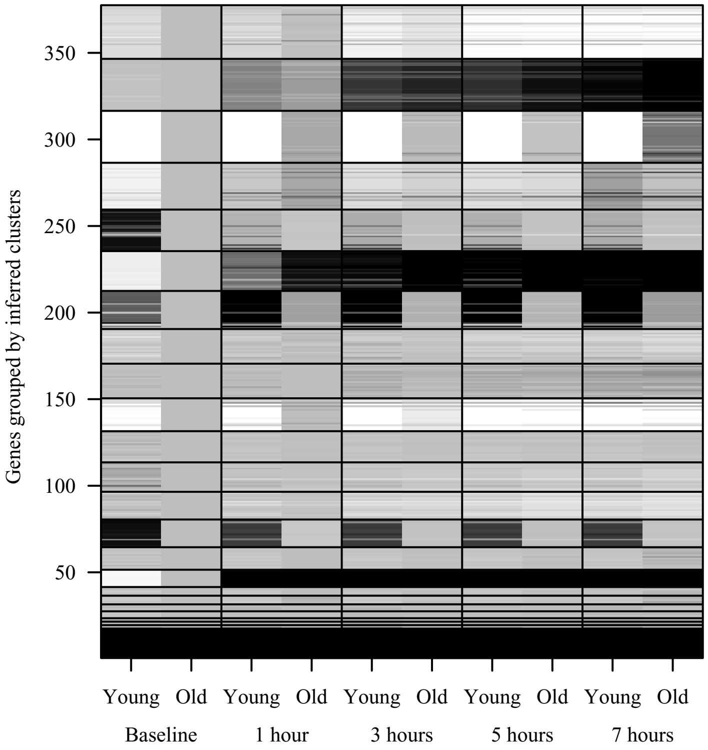

Fig. 10.4. Effects intensity plot for the 4% of probe sets in the smallest clusters.

Probe set 92885_at was identified as scientifically interesting based on another analysis. Biologists may be interested in the other probe sets that are clustered with probe set 92885_at. The proposed clustering procedure provides this information. Figure 10.2 graphically shows the treatment effects for the other probe sets that were clustered with probe set 92885_at in the least-squares clustering. The columns represent the 10 different treatment conditions and the rows correspond to the probe sets in this cluster. The reference treatment is old at baseline and is shaded gray. At other treatments, lighter shades are used to indicate underexpression relative to the reference and the darker shades indicate overexpression.

This chapter introduces the effects intensity plot which displays the entire clustering in one plot. An effects intensity plot is produced by making a plot like Figure 10.2 for each cluster and then stacking them in order of size. Figure 10.3 shows an effects intensity plot for the least-squares clustering. Since some of the clusters are very small, the smaller clusters are difficult to see. To better see the small clusters, Figure 10.4 shows only 4% of the probe sets corresponding to the smallest clusters. Notice that the smaller clusters exhibit more variation from the reference treatment than do the larger clusters. The effects intensity plot can help researchers visualize a clustering and identify clusters for additional study.

10.6 Conclusion

This chapter describes a model-based clustering procedure for microarray expression data based on a conjugate Dirichlet process mixture model. The model was first proposed by Dahl and Newton (submitted) to exploit clustering for increased sensitivity in a battery of correlated hypothesis tests. This chapter shows how the model can also be used as a clustering procedure. The model is fit with MCMC and the computational burden of the DPM model is eased by exploiting conjugacy. This chapter also introduced least-squares model-based clustering in which a point estimate of the true clustering is based on squared distances for the pairwise probability matrix. Unlike ad hoc clustering methods, the model provides measures of uncertainty about the clustering. Further, the model automatically estimates the number of clusters and quantifies uncertainty about this parameter. The method compares well to other clustering methods in a simulation study and the demonstration shows its feasibility using a large microarray data set.

Bibliography

Antoniak, C. E. (1974), "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *The Annals of Statistics*, 2, 1152–1174.

Dahl, D. B. (2003), "An improved merge-split sampler for conjugate Dirichlet Process mixture models," Technical Report 1086, Department of Statistics, University of Wisconsin – Madison.

Dahl, D. B. and Newton, M. A. (submitted), "Using clustering to enhance hypothesis testing."

Dudoit, S., Yang, Y. H., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002), "Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation," *Nucleic Acids Research*, 30, e15.

Edwards, M., Sarkar, D., Klopp, R., Morrow, J., Weindruch, R., and Prolla, T. (2003), "Age-related impairment of the transcriptional responses to oxidative stress in the mouse heart," *Physiological Genomics*, 13, 119–127.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998), "Cluster analysis and display of genomic-wide expression patterns," *Proceedings of the National Academy of Sciences (USA)*, 95, 14863–14868.

Escobar, M. D. (1994), "Estimating normal means with a Dirichlet process prior," *Journal of the American Statistical Association*, 89, 268–277.

Escobar, M. D. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Fraley, C. and Raftery, A. E. (1999), "MCLUST: Software for model-based cluster analysis," *Journal of Classification*, 16, 297–306.

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

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Li, F. L. C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004), "Bioconductor: Open software development for computational biology and bioinformatics," *Genome Biology*, 5, R80.

Hartigan, J. A. (1975), *Clustering Algorithms*, John Wiley & Sons, New York.

Hubert, L. and Arabie, P. (1985), "Comparing partitions," *Journal of Classification*, 2, 193–218.

Ihaka, R. and Gentleman, R. (1996), "R: A language for data analysis and graphics," *Journal of Computational and Graphical Statistics*, 5, 299–314.

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., and Speed, T. (2003), "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, 4, 249–264.

Jain, S. and Neal, R. M. (2004), "A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model," *Journal of Computational and Graphical Statistics*, 13, 158–182.

Liu, J. S. (1996), "Nonparametric hierarchical Bayes via sequential imputations," *The Annals of Statistics*, 24, 911–930.

MacEachern, S. N. (1994), "Estimating normal means with a conjugate style Dirichlet process prior," *Communications in Statistics, Part B – Simulation and Computation*, 23, 727–741.

MacEachern, S. N., Clyde, M., and Liu, J. S. (1999), "Sequential importance sampling for nonparametric Bayes models: The next generation," *The Canadian Journal of Statistics*, 27, 251–267.

MacQueen, J. (1967), "Some methods for classification and analysis of multivariate observations," in *The 5th Berkeley Symposium on Mathematical Statistics and Probability,* Vol. 1, eds. Cam, L. M. L. and Neyman, J., University of California Press, Barkeley, pp. 281–297.

Medvedovic, M. and Sivaganesan, S. (2002), "Bayesian infinite mixture model based clustering of gene expression profiles," *Bioinformatrics*, 18, 1194–1206.

Medvedovic, M., Yeung, K., and Bumgarner, R. (2004), "Bayesian mixture model based clustering of replicated microarray data," *Bioinformatrics*, 20, 1222–1232.

Milligan, G. W. and Cooper, M. C. (1986), "A study of the comparability of external criteria for hierarchical cluster analysis," *Multivariate Behavioral Research*, 21, 441–458.

Müller, P. and Quintana, F. (2004), "Nonparametric Bayesian data analysis," *Statistical Science*, 19, 95–110.

Neal, R. M. (1992), "Bayesian mixture modeling," in *Maximum Entropy and Bayesian Methods: Proceedings of the 11th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, eds. Smith, C. R., Erickson, G. J., and Neudorfer, P. O., Kluwer Academic Publishers, Dordrecht, pp. 197–211.

Neal, R. M. (2000), "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, 9, 249–265.

Quintana, F. A. and Newton, M. A. (2000), “Computational aspects of nonparametric Bayesian analysis with applications to the modeling of multiple binary sequences,” *Journal of Computational and Graphical Statistics*, 9, 711–737.

Rand, W. M. (1971), “Objective criteria for the evaluation of clustering methods,” *Journal of the American Statistical Association*, 66, 846–850.

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