

Modelling Gene Expression Data over Time: Curve Clustering with Informative Prior Distributions.

JON C. WAKEFIELD

Department of Statistics and Biostatistics, University of Washington, Seattle, USA
jonno@u.washington.edu

CHUAN ZHOU

Department of Biostatistics, University of Washington, Seattle, USA
czhou@u.washington.edu

STEVE G. SELF

Fred Hutchinson Cancer Research Center, Seattle, USA
sgs@hivnet.fhcrc.org

SUMMARY

The aim of many microarray experiments is to discover genes that exhibit similar behaviour, that is, co-express. A common approach to analysis is to apply generic clustering algorithms that produce a single cluster allocation for each gene. Such a strategy does not account for the experimental context, and provides no measure of the uncertainty in this classification. The model introduced in this paper is specifically tailored to the application in hand, so allowing the incorporation of prior information, and quantifies cluster membership probabilistically. In this paper we are interested in the situation in which the experiments are indexed by a variable that has a natural ordering such as time, temperature or dose level, and propose a four-stage hierarchical model for the analysis of such data. This model assumes that each gene follows one of a number of underlying trajectories, where the number may be assumed unknown, and the specific form of the trajectory depends on the experimental context. An initial filter based on a Bayes factor reduces the dimensionality of the data. The model is illustrated using two experiments carried out on yeast, one during sporulation, and another during cell-cycle. A summary measure that we emphasize is the posterior probability of co-expression which, in the context of our model, corresponds to the probability that two or more genes fall in the same group.

Keywords: BAYES FACTORS; BIRTH-DEATH MARKOV CHAIN MONTE CARLO; CLUSTERING;
CO-EXPRESSION; MICROARRAY EXPERIMENTS; MIXTURE MODELS.

1. INTRODUCTION

Recently there has been a huge interest in the analysis of gene expression data from DNA microarray experiments (e.g. Collins, 1999). This technology allows the simultaneous recording of activity, as measured by messenger RNA (mRNA) levels that reflect the level of transcription from DNA to RNA, in a large number of genes. The resultant data allow the possibility for gaining a greater understanding of the transcription dynamics on a genome-wide scale, under varying experimental conditions.

The data we analyse in Sections 4 and 5 arise from cDNA microarrays in which a fluorescently labelled cDNA sample is obtained for each time point and, to control for the unknown amount of DNA in the sample, is added to another fluorescently labelled sample from a reference experiment with another dye. The combined sample is hybridised to arrays containing templates on each gene in the genome. The amount of signal recorded reflects the extent of mRNA expression for each gene and at each time point, relative to the reference. While acknowledging their importance, we do not consider many of the generic issues that are pertinent to the analysis of data from microarray experiments such as signal extraction, array and dye effects, normalization, and background subtraction.

There are various scientific questions that may be assessed via microarray experiments, in this paper we are interested in the situation in which the experiments may be indexed by an ordered variable such as time, temperature, or the dose level of a toxin. In our scenario, the aim is to gain insight into those genes that behave similarly over the course of the experiment. By comparing genes of unknown function with profiles that are similar to genes of known function, clues to function may be obtained. Hence, *co-expression* of genes is of interest. A variety of approaches to this problem have been proposed. By far the most common is to apply generic supervised or unsupervised clustering algorithms to the data. For example, Eisen *et al.* (1998) use hierarchical clustering with distance measured via correlation, Chu *et al.* (1998) cluster to *known* profiles again using correlation, and Tamayo *et al.* (1999) use self-organising maps. These approaches, though useful as a starting point are deficient in a number of respects. First, the data are clustered on the basis of the raw measurements and so classifications can be sensitive to outlying observations. To overcome this, often the raw data are “filtered” to remove aberrant observations, though this procedure has an ad hoc flavour (specific recipes are given in Section 3). Second, no measure of the certainty of the classification is given. A remedy to this latter problem has been proposed by Kerr and Churchill (2000), who describe a method for accessing the uncertainty by bootstrapping the residuals from an analysis of variance model that includes the gene-time effects of interest; the proportion of bootstrap samples that cluster to the original classification then gives a measure of the reliability. Finally, clustering algorithms are generic and are in no way tuned to the application in hand (except perhaps for the measure of dissimilarity used), and in particular do not allow the incorporation of prior information.

In this paper we propose a model-based approach to this problem in which, heuristically speaking, we explicitly model the trajectory as a function of the ordering variable (e.g. time) and a gene-specific set of parameters. We then cluster on the basis of the latter, with our probabilistic framework providing quantitative measures of classification. The structure of this paper is as follows. In Section 2 we describe our modelling framework, and in Section 3 an initial screen using Bayes factors that is applied to reduce the dimensionality of the data. In Sections 4 and 5 we apply our model to two data sets, both collected in yeast, one during sporulation, and one during the cell-cycle. These analyses are not intended to represent substantive contributions to yeast genetics but have been selected to highlight posterior probabilities of co-expression (the sporulation data), and the use of prior information of a specific (periodic) form (cell-cycle data). Section 6 contains a concluding discussion.

2. MODEL DESCRIPTION

2.1 Clustering Model

Let Y_{it} denote, for gene i , the log-ratio of mRNA expression level measured during experiment t , relative to a reference experiment, $i = 1, \dots, N$, $t = 1, \dots, T$. The experiments may be indexed by any ordered variable, but for descriptive purposes we suppose that the variable is time and denote the actual sampling times by X_t , where for simplicity we have assumed that the design is the same for each gene. We then model these data via the following four-stage hierarchical model.

Stage One: For the observed data we have

$$y_{it} = f(\boldsymbol{\theta}_i, X_t) + e_{it},$$

where $e_{it} \sim_{iid} N(0, \sigma_e^2)$ and $f(\boldsymbol{\theta}_i, X_t)$ denotes the form of the trajectory and depends on a gene-specific set of parameters, $\boldsymbol{\theta}_i$, and the experiment time, X_t .

Stage Two: Conditional on C known trajectories, we introduce trajectory membership indicators Z_i that reflect the underlying trajectory that gene i follows so that $\boldsymbol{\theta}_i = \boldsymbol{\theta}^c$ if $Z_i = c$, $c = 1, \dots, C$. We model $\boldsymbol{\theta}^c | \boldsymbol{\phi}, C$, $c = 1, \dots, C$, as arising from a distribution that depends on unknown parameters $\boldsymbol{\phi}$. This is a prior for the collection of trajectories.

Stage Three: We assume that

$$\Pr(Z_1 = z_1, \dots, Z_N = z_N | \boldsymbol{\pi}, C) = \prod_{i=1}^N \Pr(Z_i = z_i | \boldsymbol{\pi}, C),$$

where $\Pr(Z_i = z_i | \boldsymbol{\pi}, C) = \pi_{z_i}$, $c = 1, \dots, C$, and $\boldsymbol{\pi} = (\pi_1, \dots, \pi_C)$. In the examples of this paper we take this prior to be a product of Dirichlet distributions, $\text{Di}(1, \dots, 1)$.

Stage Four: Finally, we place prior distributions on σ_e^2 , $\boldsymbol{\pi} | C$ and $\boldsymbol{\phi}$, and possibly C (if the latter is assumed unknown).

We now provide an interpretation of the random variables Z_i that represent the gene expression curve that gene i is following. The change in expression levels describing transcription from DNA to RNA that occurs within the cell nucleus is determined by, amongst other things, the regulatory proteins that are responsible for gene i and on the timings at which these proteins are acting (either to activate or suppress activity). The curve membership indicators can therefore be viewed as summarizing the proteins that are relevant for gene i , and if two genes lie in the same cluster it is evidence of shared transcription factors.

Our approach to assume a mixture model is crucially different to the ‘‘model-based’’ clustering approach of Yeung *et al.* (2001) who analyze similar data but simply assume that the data arise from a mixture of T -dimensional normal distributions and hence do not acknowledge the time-ordering of the data (the analysis would be unchanged if the time ordering were permuted). In particular it would be desirable to allow serial dependence, within such an approach, but the MCLUST software (Fraley and Raftery (1998)) that is used by Yeung *et al.* (2001) does not allow for this possibility. Medvedovic and Sivaganesan (2002) also describe a Bayesian hierarchical model for microarray data, but again do not consider specific curve forms.

2.2 Computation

For fixed C , samples may be generated from the model described in the previous section using a Metropolis-Hastings Markov chain Monte Carlo (MCMC) algorithm. For unknown C either reversible jump MCMC (Green, 1995), or birth-death MCMC (Stephens, 2000a) may be used. We use the latter since it is relatively straightforward to implement in our context. The algorithm obtains samples from the posterior $p(C, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi}, \sigma_e^2 | \mathbf{y})$ by simulating a continuous time marked point process to sample points $\boldsymbol{\theta}^c$ with associated ‘‘marks’’ π_c , with a Gibbs sampler for $\boldsymbol{\phi}, \sigma_e^2$ (and possibly $\boldsymbol{\theta}$). In our examples convergence was diagnosed informally by examination of trace plots of parameters and comparison of results from multiple chains initiated at different starting points. We also ran fixed C analyses and compared the results to the BDMCMC algorithm.

We follow Stephens (2000a) quite closely and take the birth rate to be mean of the Poisson prior that we specify for C . For a birth we simulate $\pi \sim \text{Be}(1, C)$, and the new trajectory vector from the second stage prior, which in our two examples is a normal distribution. The death rate for component c , d_c , is the likelihood for the collection $\boldsymbol{\pi}, \boldsymbol{\theta}$ with $\pi_c, \boldsymbol{\theta}^c$ removed, divided by the likelihood for $\boldsymbol{\pi}, \boldsymbol{\theta}$. Component c is selected to die with probability $d_c / \sum_{c'} d_{c'}$.

As discussed in detail in Richardson and Green (1997) (and the accompanying discussion), and Stephens (2000b) there is a fundamental non-identifiability associated with mixture problems in that the posterior contains $C!$ modes of equal height that are indistinguishable from the data alone. To identify parameters uniquely some form of ‘‘labelling’’ must be carried out. In Sections 4 and 5 we describe specific labelling methods, but the fundamental question of co-expression of genes i and i' may be answered without resolving this problem since the quantity $\Pr(Z_i = c, Z_{i'} = c | \mathbf{y})$ is invariant to re-labelling.

3. INITIAL FILTERING

The method we have proposed may be computationally prohibitive for a large number of genes (depending on the functional form of the trajectories), and so we describe an initial screen to reduce the dimensionality of the data. In the microarrays literature, similar procedures are carried out but in a relatively ad hoc manner. For example, Eisen *et al.* (1998) only include genes for study if their expression levels deviate from those at time zero by at least a factor of three in at least two time points, while Tamayo *et al.* (1999) exclude yeast genes that do not show a relative change of two units and an absolute change of 35 units; values of 3 and 100 are used for human genes, though no justification for these values was given. We define genes as ‘‘un-interesting’’ if their mRNA levels remain at a constant level over the time course of the experiment.

We propose a very simple approach in which we compare the models $M_0 : \mu_1 = \mu_2 = \dots = \mu_T = \mu$ and $M_1 : \text{not } M_0$, via the Bayes factor $p(\mathbf{y}_i | M_1) / p(\mathbf{y}_i | M_0)$, where

$$I_1 = p(\mathbf{y}_i | M_1) = \prod_{t=1}^T \int_{\mu_t} \int_{\sigma_e^2} p(y_{it} | \mu_t, \sigma_e^2) \times \pi(\mu_t) \pi(\sigma_e^2) d\mu_t d\sigma_e^2, \quad (1)$$

and

$$I_0 = p(\mathbf{y}_i | M_0) = \prod_{t=1}^T \int_{\sigma_e^2} p(y_{it} | \mu_t = \mu, \sigma_e^2) \times \pi(\mu) \pi(\sigma_e^2) d\mu d\sigma_e^2. \quad (2)$$

We assume that $Y_{it} | \mu_t, \sigma_e^2 \sim_{iid} N(\mu_t, \sigma_e^2)$, with priors $\pi(\mu)$ and $\pi(\sigma_e^{-2})$ given, respectively, by $\mu_t \sim_{iid} N(m_0, v_0)$, $t = 1, \dots, T$, and $\sigma_e^{-2} \sim Ga(a_0, b_0)$. Due to the normalization we choose $m_0 = 0$ and take $v_0 = 2^2$ to reflect the range of variation observed in similar experiments (e.g. Eisen *et al.*, 1998). The data of Chu *et al.* (1998) includes data at $t = 0$, relative to another $t = 0$ experiment that therefore act as a “reference” and reflect measurement error, σ_e^2 , only. Hence we place a highly informative prior on σ_e^2 for the calculation of Bayes factors. The reference data were analysed under the model $\hat{Y}_0 | \mu_0, \sigma_e^2 \sim N(\mu_0, \sigma_e^2/N)$ with the improper prior $\pi(\mu_t, \sigma_e^2) \propto \sigma_e^{-2}$ which leads to the posterior $\sigma_e^{-2} | \mathbf{y}_0 \sim Ga\{(N-1)/2, (N-1)s^2/2\}$, where s^2 is the unbiased estimator of σ^2 . Since N is large the posterior distribution is highly concentrated.

Importance sampling based on sampling from the prior is relatively efficient given our informative priors, and results in (1) and (2) being estimated, respectively, by $\hat{I}_1 = S^{-1} \sum_{s=1}^S \prod_{t=1}^T p(y_{it} | \mu_t^{(s)}, \sigma_e^{2(s)})$, and $\hat{I}_0 = S^{-1} \sum_{s=1}^S \prod_{t=1}^T p(y_{it} | \mu^{(s)}, \sigma_e^{2(s)})$, where $\mu_t^{(s)}, \mu^{(s)} \sim_{iid} \pi(\mu)$, $\sigma_e^{2(s)} \sim_{iid} \pi(\sigma_e^2)$. It would be desirable to analyse all of the data together with the clustering model described in Section 4.2, with a “zero” cluster for the un-interesting genes, but the increase in computational burden is unlikely to be worthwhile in terms of identifying interesting clusters.

4. EXAMPLE 1: SPORULATION DATA

4.1 Data Description

We first describe the experiment of Chu *et al.* (1998) which contained data on seven microarrays with a reference at time $t = 0$, and activity during sporulation measured at times $X_t = \{0.5, 2, 5, 7, 9, 11.5\}$ so that $T = 6$. Chu *et al.* (1998) were interested in genes that exhibited similar profiles, and to this end they created seven “characteristic curves” by averaging, via a visual inspection, genes contained in each profile. Figure 1 gives the sets of genes in each profile, along with the mean trajectories. After an initial screen in which 80% of genes were eliminated they clustered the remaining genes to each profile (using correlation as the distance measure). This approach provides a list of genes that appear to conform to each profile, but does not give a measure of the uncertainty of this classification, in common with other distance-based clustering procedures (e.g. Eisen *et al.* 1998; Tamayo *et al.* 1999).

4.2 Model Description

In this experiment Y_{it} represents the \log_2 -ratio of expression for gene i at time t , relative to time $t = 0$, $i = 1, \dots, 6118$, $t = 1, \dots, 6$. We re-iterate that the aim is to discover structure over time, and in particular to determine genes that follow common trajectories. In the absence of further information, we would expect the trajectories to be smooth after the onset of sporulation (that is following time $X_1 = 0.5$) which occurs rapidly. To this end we assume a first-order random walk model so that at stage one we have

$$y_{it} | \theta_{it}, \sigma_e^2 \sim_{iid} N(\theta_{it}, \sigma_e^2),$$

with at stage two $\theta_{it} = \theta_t^c$ if $Z_i = c$ and

$$\theta_t^c = \theta_{t-1}^c + u_t,$$

for $t = 2, \dots, T$, and $u_t \sim_{iid} N(0, \Delta_t \tau^2)$, where $\Delta_t = X_t - X_{t-1}$ so that observations closer in time are more likely to be similar. For the first time point we have $\theta_1^c \sim_{iid} N(0, \tau_1^2)$.

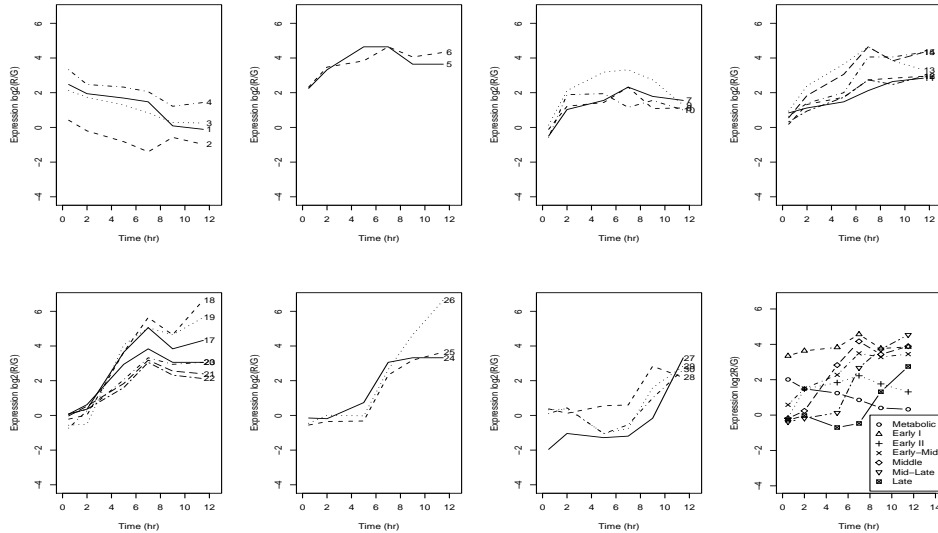


Figure 1. Hand-picked genes in each of seven groups (from Chu et al. 1998), along with mean trajectories (panel 8).

With reference to the model described in Section 4.2, $\phi = (\tau_1^2, \tau^2)$ and, conditional on C , the data are modelled as arising from the C trajectories $\theta^1 = (\theta_1^1, \dots, \theta_T^1), \dots, \theta^C = (\theta_1^C, \dots, \theta_T^C)$. Collapsing over stages two and three shows that we are modelling the data as arising from a mixture of C underlying smooth trajectories, i.e. $\theta_i = \sum_{c=1}^C \pi_c \theta^c$.

We now discuss the fourth stage prior which here requires specification for $\sigma_e^2, \tau_1^2, \tau^2$ and C . The prior for σ_e^2 is identical to that described in Section 3. For specification of priors for each of the variances τ_1^2 and τ^2 we pick a “most likely” and an “upper value” for the standard deviation. These values are then converted to the inverse variance scale, and we pick the parameters of the gamma distribution to line up the mode with the most likely point, and the 95% point of the distribution with the upper value (which requires a numerical search). We choose the modal value for τ_1^2 to be 2^2 (which is consistent with $\mu_0 \sim N(0, 2^2)$), and for τ^2 (which is a conditional variance) we take the most likely value to be 1.5^2 (so that in one unit time interval we expect the trajectory to be within ± 3.0 with probability 0.95); we take 3^2 and 2^2 as upper values for τ_1^2 and τ^2 , respectively. Figure 2 shows four simulations from this prior with fixed $C = 9$ and 20 genes within each cluster. The prior for C was Poisson with mean 15.

4.3 Analysis

We applied the initial filter described in Section 3 to the $N = 6118$ genes. Figure 3 displays the 100 “top” and 100 “bottom” genes in terms of $\Pr(M_1 | \mathbf{y})$. We see the same qualitative behaviour of the trajectories as that observed in the prior simulations of Figure 2, though in the latter the groups are less distinct. We chose to analyse the 1104 genes for which $\Pr(M_1 | \mathbf{y}) > 0.999$. We varied the size of the variance on the normal prior for μ_t and found that though the resultant Bayes factors showed sensitivity to this choice, the ranking on the genes with highest $\Pr(M_1 | \mathbf{y})$ was robust.

We found that the posterior for C was highly sensitive to the prior choices, adding to the need for informative prior distributions. Figure 4 shows the trace plot of C

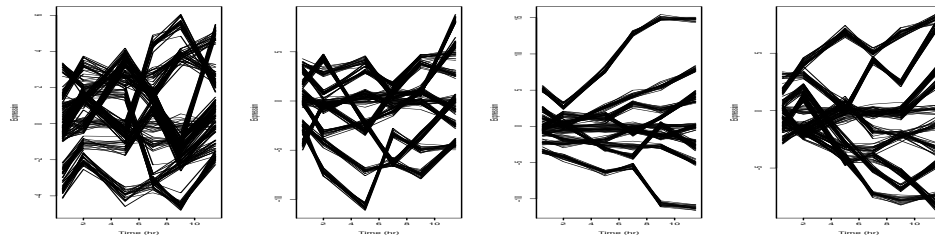


Figure 2. Four simulations from the random walk prior, with measurement error added. There are $C = 10$ groups, each of which contain 20 genes.

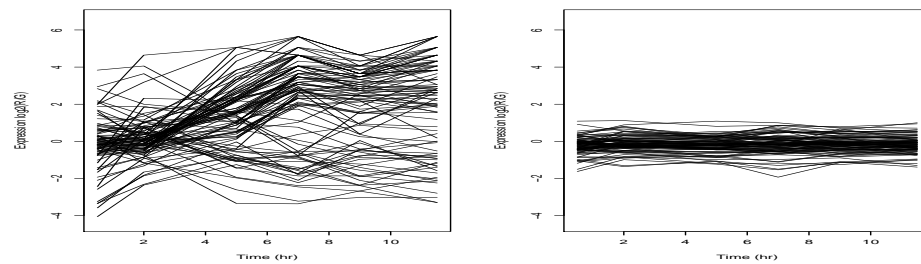


Figure 3. Expression levels versus time for 100 genes with the highest (left panel), and the lowest (right panel) values of $\Pr(M_1 | \mathbf{y})$ where M_1 is the model of non-constant level.

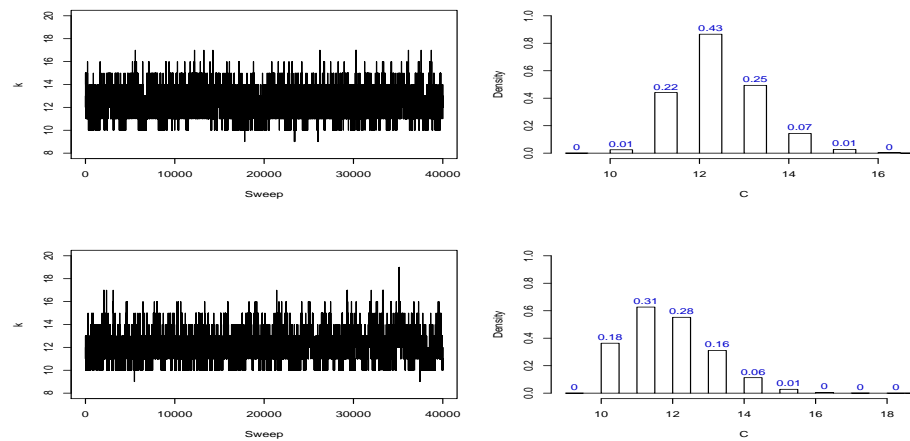


Figure 4. Trace plot of the number of clusters C and the posterior distribution of C (after a burn-in of 40,000 iterations), from BDMCMC analysis of the sporulation data (top row), and the cell-cycle data (bottom row).

versus iteration number, and the marginal posterior of C , following a burn-in of 40,000 iterations. For 80,000 iterations, the computer time was approximately 4 hours on a Unix cluster.

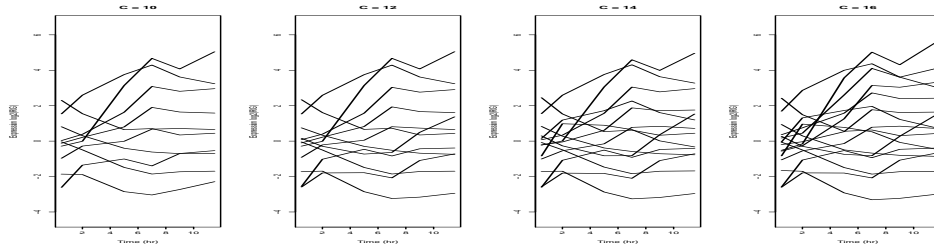


Figure 5. Mean expression levels as a function of time, for numbers of cluster $C = 10, 12, 14, 16$.

For illustration, Figure 5 shows the curves that result from values of $C = 10, 12, 14, 16$ (these may be compared with Figure 4B of Chu *et al.* 1998). Re-labelling was carried out on the basis of the mean at the second time-point (since these were relatively distinct) and showed good agreement with the decision theoretic approach of Stephens (2000b). As C increases the extreme trajectories remain relatively constant. The posterior medians of σ_e were 0.50, 0.47, 0.45, 0.44 for $C = 10, 12, 14, 16$, respectively. The prior median of the informative prior on σ_e was 0.25 which suggests there is some model misspecification (or that the measurement error increases after the zero time point). For the $C = 11$ case we plot, in Figure 6, the mean curves along with the genes clustered to these curves.

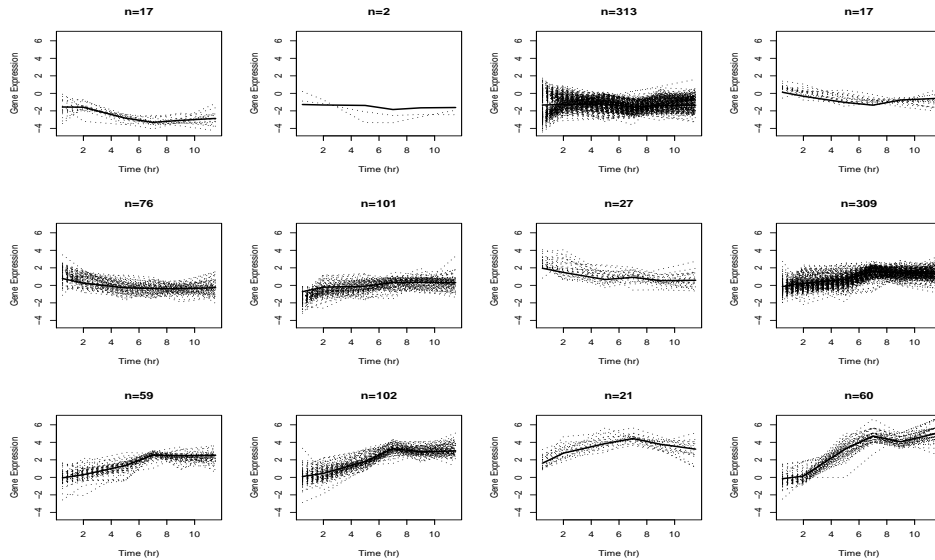


Figure 6. Posterior profiles and genes classified to each of these profiles (using MAP classification), conditional on $C = 12$ clusters.

In Figure 7 we summarize the pairwise probabilities based on the BDMCMC analysis, averaging across all C . For illustration, the genes we examine are the 30 hand-picked genes highlighted by Chu *et al.* (1998), and reproduced in Figure 1. If the cluster membership was consistent with this figure then we would see blocks of shaded areas for genes in the same group (close to the diagonal), and white in the other areas. In fact we see that although there is a greater probability of co-expression close to the diagonal, there are both genes within the same hand-picked collection which do not appear

to co-express, and genes in other groups that co-express. Concentrating on the sixth group in Figure 1, gene 26 would not appear to co-express with any of the other 29 genes, while genes 24 and 25 co-express with each other and also with genes 11, 12, 16, 20, 21 and 22. Hence we see that our model offers new insights into co-expression.

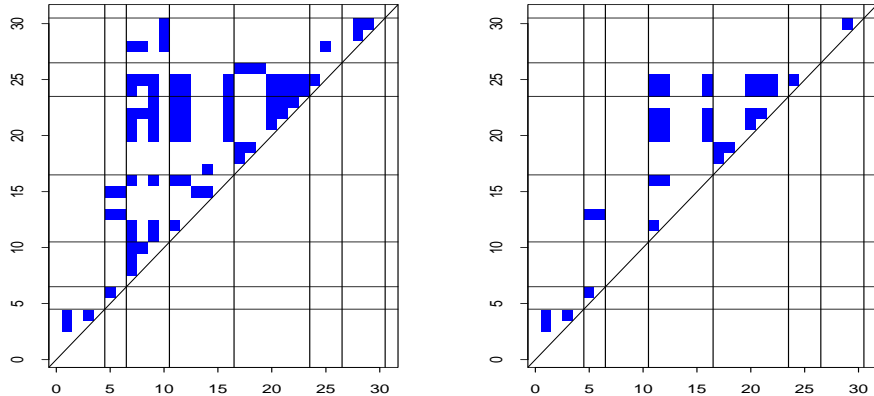


Figure 7. Heat map showing pairwise probabilities of common cluster membership of the 30 genes in Figure 1. The solid lines separate the different groups. On the left the shaded squares denote those pairwise probabilities greater than 0.5, while on the right the cut-off probability is 0.8.

5. EXAMPLE 2: CELL-CYCLE DATA

5.1 Data Description

Spellman *et al.* (1998) describe a number of microarray experiments that were carried out to create a comprehensive list of yeast genes whose expression levels vary periodically within the cell cycle. For illustration we analyse one of the data sets that measured levels on $N = 4489$ genes that were synchronised via α factor, and randomly select 800 genes. Expression levels were measured every 7 minutes for 140 minutes (so that $T = 20$ measurements were recorded in total). The expression levels of the 800 genes are shown in Figure 9, after application of our model. We take as our objective the identification of genes that display similar periodicity.

5.2 Model Description

In this example our model formulation is strongly driven by prior information. Specifically we assume that the data arise from a mixture of functions with periodic structure. Specifically, we assume that

$$Y_{it} = R_i \cos(\omega t + \phi_i) + e_{it} = A_i X_{1t} + B_i X_{2t} + e_{it},$$

where R_i is the amplitude and ϕ_i the phase of gene i , $X_{1t} = \sin(\omega t)$ and $X_{2t} = \cos(\omega t)$, and $\omega = 2\pi/p$ with $p = 66$ minutes as the known period (obtained from Spellman *et al.* 1998). It may seem more natural to model in terms of amplitude and phase, but

because of the irregular constraints on the collection (R_i, ϕ_i) we prefer to formulate our mixture model in terms of $\theta_i = (A_i, B_i)$. Figure 8 gives the least squares estimates for (A_i, B_i) (left panel), and $(\log R_i, \log[(\pi/2 - \phi_i)/(\pi/2 + \phi_i)])$ (right panel) and clearly shows the irregularity of the joint distribution of the latter which does not allow us to simply parameterize in terms of functions of the phase and amplitude.

At the second stage of the model we assume that θ_i arise from bivariate normal distributions, i.e. $\theta_i^c | \mathbf{m}, \mathbf{V} \sim_{iid} N_2(\mathbf{m}, \mathbf{V})$, where we assume that \mathbf{m}, \mathbf{V} are known. We estimate these from the data (along the lines followed by Richardson and Green (1997) and Stephens, 2000a). In particular we take $\mathbf{m} = (-0.215, -0.005)$, which correspond to the means of the least squares estimates, and \mathbf{V} with elements $(4.45 \ 0 \ 0 \ 2.13)$ which are the ranges of the least squares estimates squared. Given that our model here is exploratory we are not troubled by the mild dependence of the prior on the data. The priors for σ_e^2 and C were as for the sporulation data. Figure 9 shows simulations from our prior distribution. We are modeling the A_i, B_i pairs as arising from a bivariate normal distribution and so, in this example we would not want to filter out the constant genes since this would leave a ‘‘hole’’ close to zero. For these data (although we acknowledge the random error around each curve in our first stage distribution) we have effectively reduced the dimensionality of the data from 20 to 2.

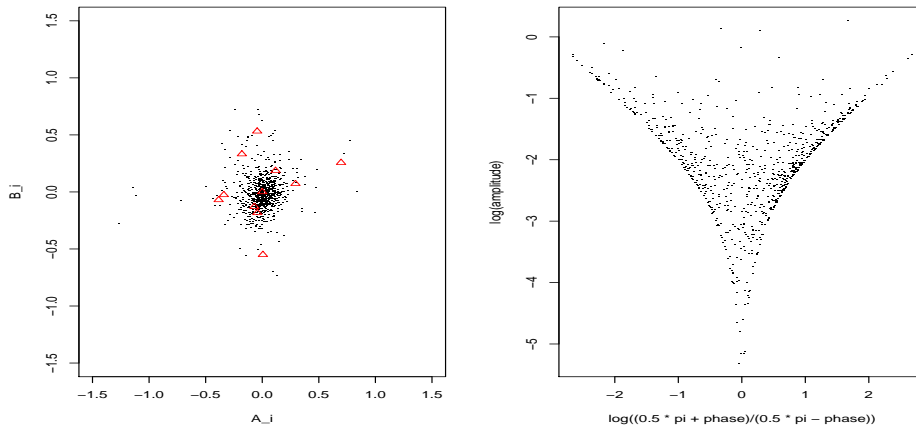


Figure 8. Least squares estimates \hat{A}_i, \hat{B}_i from the model $E[Y_{it} | A_i, B_i] = A_i \sin(\omega t) + B_i \cos(\omega t)$, for gene i (left panel), and $\log(\hat{R}_i), \log\{(\pi/2 + \hat{\phi}_i)/(\pi/2 - \hat{\phi}_i)\}$ (right panel), $i = 1, \dots, 800$.

5.3 Analysis

The bottom row of Figure 4 shows the behavior of C as a function of iteration number, and the posterior distribution of C . We see that the posterior for the latter is concentrated between 10 and 15. The computational overhead was a little less in this example, due to the use of a linear model that reduces a number of the required calculations. For the $C = 11$ analysis we relabelled on the basis of B^c (which were relatively distinct); again we found good agreement with the decision theoretic approach of Stephens (2000b). The means $\theta^c = (A^c, B^c)$, $c = 1, \dots, 11$, are displayed in the left hand panel of Figure 8. In Figure 10 we show the collection of profiles, conditional on $C = 11$. We classified each gene to a profile based on $\Pr(Z_i | \mathbf{y})$ and these classifications are

displayed in Figure 10. The largest cluster is in the third panel of the second row and corresponds to the “zero” cluster which has a flat profile. We see that the classifications look reasonable, though there is clearly some model misspecification. In particular a number of the trajectories look to have attenuated profiles as time increases.

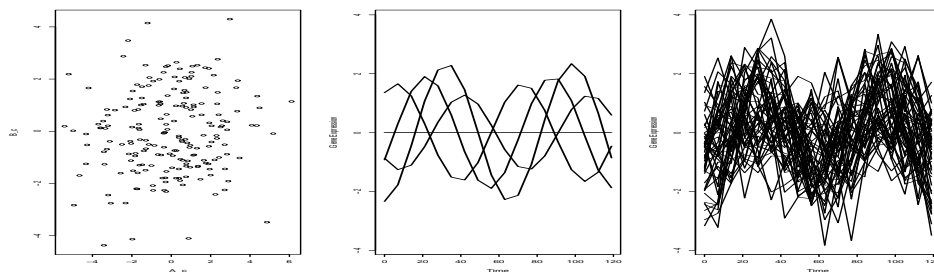


Figure 9. 200 simulations of (A_c, B_c) for the cell-cycle parameters (left); five trajectories without random error (centre); 55 simulations from 5 groups, 4 simulated plus zero group with random error (right).

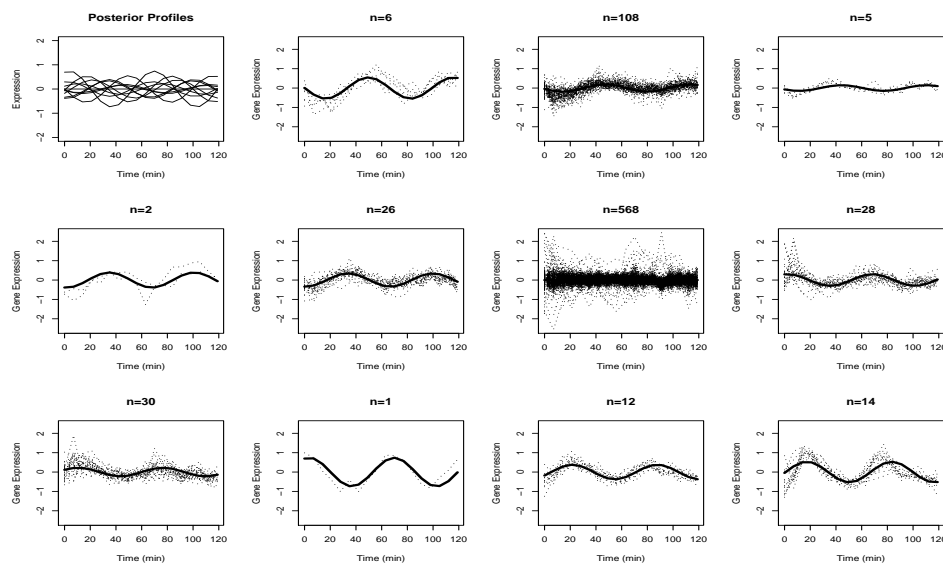


Figure 10. Posterior profiles and genes classified to each of these profiles (using MAP classification), conditional on $C = 11$ clusters.

6. DISCUSSION

In this paper we have introduced a model that allows a quantitative description of gene co-expression, in contrast to clustering techniques that are currently used. A systematic comparison with these more traditional techniques is unfortunately beyond the scope of this paper. In investigations not reported here, we have found that the number of clusters, and sometime the co-expression probabilities, may be highly sensitive to the

prior distributions and we would have far less faith in our quantitative conclusions if the priors we had used were not based on biological and experiment-specific information.

Although we have emphasized that posterior probabilities of co-expression are invariant to component re-labelling, for other summaries such as the reporting of mean trajectories, the problem remains. An alternative to the procedure followed for the examples, and is closer to the context is to label on the basis of known marker genes.

In our model formulation of Section 2 we assumed a priori that the trajectory indicator variables Z_i were independent. In practice there will often be substantial information available to place collections of genes in the same cluster with high probability. We are currently working on how to extend our independence model, using ideas from Markov random field modelling.

More substantively, our eventual aim is to combine expression and sequence data. Models for the latter have been extensively developed, see for example Liu (1999). A Bayesian framework is ideally suited to such an endeavor, since it allows a natural combination of multiple data sources, and the incorporation of prior information, which is likely to be essential in complex problems such as these.

ACKNOWLEDGEMENTS

The authors would like to thank Matthew Stephens for useful discussions, and in particular for advice on birth-death MCMC and label-switching.

REFERENCES

- Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P., and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science* **282**, 699–705.
- Collins, F. (1999). The chipping forecast. *Nature Genetics* **21**, 1–60 .
- Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868.
- Fraley, C. and Raftery, A. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal* **41**, 578–588.
- Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.
- Kerr, K. and Churchill, G. (2000). Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sci. USA* **98**, 8961–8965.
- Liu, J., Neuwald, A., and Lawrence, C. (1999). Markov structures in biological sequence alignments. *J. Amer. Statist. Assoc.* **94**, 1–15.
- Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model-based clustering of gene expression profiles. *Bioinformatics* **18**, 1194–1206.
- Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Roy. Statist. Soc. B* **59**, 731–792.
- Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell* **9**, 3273–3297.
- Stephens, M. (2000a). Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *Ann. Statist.* **28**, 40–74.
- Stephens, M. (2000b). Dealing with label-switching in mixture models. *J. Roy. Statist. Soc. B* **62** , 795–809.
- Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., and Dmitrovsky, E. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* **96**, 2907–2912.
- Yeung, K., Fraley, C., Murua, A., Raftery, A., and Ruzzo, W. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17**, 977–987.