

A temporal hidden Markov regression model for the analysis of gene regulatory networks

Mayetri Gupta, Pingping Qu, and Joseph G. Ibrahim*

February 20, 2007

Abstract

We propose a novel hierarchical hidden Markov regression model for determining gene regulatory networks from genomic sequence and temporally collected gene expression microarray data. The statistical challenge is to simultaneously determine the groupings of genes and subsets of motifs involved in their regulation, when the groupings may vary over time, and a large number of potential regulators are available. We devise a hybrid Monte Carlo methodology to estimate parameters under two classes of latent structure, one arising due to the unobservable state identity of genes, and the other due to the unknown set of covariates influencing the response within a state. The effectiveness of this method is demonstrated through a simulation study and an application on an yeast cell-cycle data set.

1 Introduction

With the increased availability of large scale genomic data, discovering the roles of various components in the genome has become an important goal in the biomedical sciences. Many biological processes in complex organisms are regulated by combinations of genes forming a pathway or network (Wang et al., 2005). A fundamental question is to determine how genes interact to regulate biological pathways. Making robust statistical

*Department of Biostatistics, University of North Carolina at Chapel Hill, NC 27599, U.S.A.;
gupta@bios.unc.edu

inference from diverse types of genomic data to discover pathways of biological function presents a formidable challenge, especially as various latent correlations exist in gene behavior, and these behaviors and relationships may change over time. The first step in gene regulation constitutes the binding of certain proteins, called transcription factors (TFs) to TF binding sites (TFBSs) on the DNA sequence, which activates the genes downstream into being transcribed into messenger RNA (mRNA). The pattern common to the TFBSs for a TF is termed a *motif*, and genes that are regulated by the same TF often share a common motif. A popular strategy for studying gene regulation is a two-step procedure—searching for motifs upstream of genes after clustering the genes by similar expression patterns. However, a major problem with this approach is that if the initial clustering is inaccurate, discovery of the correct motifs may be heavily biased.

1.1 Gene regulatory processes and statistical modeling

We first describe some terminology and current approaches. A motif of length w is represented through a $4 \times w$ matrix called a position-specific weight matrix (PSWM), each column denoting the probabilities (or frequencies) of observing the 4 letters A, C, G, or T in that position (w typically ranges between 8 to 20). For example, binding sites for the yeast motif RAP1 can be represented through the PSWM (of counts):

$$\begin{array}{r}
 (A) \\
 (C) \\
 (G) \\
 (T)
 \end{array}
 \begin{array}{cccccccccccc}
 8 & 4 & 16 & 0 & 0 & 0 & 13 & 3 & 11 & 0 & 11 & 0 \\
 0 & 12 & 0 & 16 & 16 & 15 & 1 & 3 & 0 & 14 & 3 & 8 \\
 7 & 0 & 0 & 0 & 0 & 0 & 1 & 5 & 2 & 0 & 0 & 1 \\
 1 & 0 & 0 & 0 & 0 & 1 & 1 & 5 & 3 & 2 & 2 & 7
 \end{array}$$

Successful computational strategies to discover functional motifs in biological processes have mostly involved searching for conserved motifs in the *promoter* sequence adjacent to each of a set of *co-regulated* genes (Liu et al., 1995), for example, genes exhibiting similar expression patterns across a number of experimental conditions in a microarray experiment. More recently, approaches that incorporate more biological information, such as weighting promoters by expression-based scores before motif discovery, have been proposed (Liu et al., 2002). Linear model-based approaches (Conlon et al., 2003) relate

expression values to the estimated propensity of motif occurrences (summarized by a sequence motif “score”), assuming that the presence of a motif site contributes additively to the gene expression level. These approaches assume a single set of gene expression patterns, which makes it difficult to discover motifs having a positive effect on a subset of genes and a negative or neutral effect on another, a problem exacerbated in the case of temporal gene expression. Probabilistic relational models, based on the specification of multiple conditional distributions (Segal et al., 2001) have been used for modeling gene regulation; however, such models involve a degree of parametric complexity that hinders model validation approaches. One promising approach involved iterative model-based clustering and motif determination (Holmes and Bruno, 2000) but avoided a direct model-based link between expression and sequence.

1.2 A joint approach for time dependent regulatory networks

In this article we present a method that determines groups of genes and TFs regulating them, allowing for the pattern of regulation to change temporally. This approach addresses the discovery of motifs temporally involved in gene regulation, unlike methods which either ignore the time-dependence or consider each time point separately. It may not be reasonable to assume that genes can be grouped into clusters that behave similarly as a group over time. Our model allows correlations to exist between gene measurements over different time points as well as between separate genes on the same array. Several “patterns” of gene behavior may exist in a single experiment, where genes may enter or leave a particular “pattern” at a certain time point in the study.

We present a novel statistical approach that can integrate the information from expression measurements over time and genomic sequence data through a Bayesian hidden Markov model (HMM) framework, and simultaneously uncover relationships between genes and TFs that regulate them. We generalize the regression model proposed in Conlon et al. (2003) to a framework that can accommodate temporally varying motif effects.

In our model, the temporal structure of dependence in gene behavior is modeled by a latent Markov process, where at every time point, the observed expression measurement is modeled as a state-dependent function of the time and motif covariates. Using a regression framework provides interpretability of the resulting coefficients, and a systematic way to model interactions between various factors.

Unlike other two-step clustering and motif-finding approaches, our method provides a unified framework giving simultaneous estimates of gene co-regulation and time-dependent motif effects. We first describe the general Bayesian model framework (Section 2) and present a hybrid Monte Carlo procedure for model fitting, parameter estimation, and variable selection under a two-layered latent structure (Section 3). Defining the model within a hierarchical framework, we show that problems of parameter estimability, identifiability, and model validation can be addressed in a robust way. A Bayesian criterion for selecting the number of model states is discussed (Section 4), which is seen to outperform criteria such as the AIC or BIC. Finally, simulation studies and an application to a yeast cell-cycle data set demonstrates the feasibility of this model in a real biological scenario (Sections 5, 6).

1.3 A yeast cell cycle gene expression data set

Our motivating data set is from a set of yeast microarray experiments (Spellman et al., 1998), an extensive study of genes regulated in a periodic manner coincident with the cell-cycle. cDNA microarrays were used to analyze mRNA levels in cell cultures that had been synchronized by three independent methods. Gene expression was recorded as the logarithm of the ratio of expression of that gene in the sample to a baseline control. The data set was preprocessed by computing the fluorescence ratios through a local background correction (intensities of the weakest 12% of the pixels in each box) and normalized such that the average log-ratio over the course of the experiments equaled zero. Cell-cycle synchronization was inferred for genes (i) whose mRNA levels were

significantly correlated with mRNA levels of genes previously known to be cell cycle regulated, and (ii) showed a high periodicity based on a numerical score derived from Fourier analysis. A total of 800 genes were identified as being cell-cycle regulated.

For our analysis, we restrict our focus to the common group of genes between the three synchronization methods. In Section 6 we describe the analysis of the elutriation data set, which consists of measurements of 477 genes over 12 time points corresponding to approximately two cell cycles. For deriving the motif scores corresponding to each gene, position-specific weight matrices (PSWMs) and upstream promoter regions of each gene (about a 1Kb region) were downloaded from the *Saccharomyces Cerevisiae* promoter database (Zhu and Zhang, 1999). The motif scoring method is described in Section 2.2. The cell cycle data set is available from the Stanford microarray database (<http://genome-www5.stanford.edu/>).

2 A hidden Markov regression model

Now we motivate the development of the model framework. Let $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iT})$, ($i = 1, \dots, N$) denote the vector of T expression measurements made on gene i . Assume that each observation may have been generated from one of K classes, indexed by the corresponding latent vector $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{iT})$. Further, assume that we have measurements on a set of p motif covariates for each observation, the design matrix being denoted by $\mathbf{X} = (X_{il})$, ($i = 1, \dots, N$; $l = 1, \dots, p$). Let $\boldsymbol{\beta}_k = (\beta_{k1}, \dots, \beta_{kp})$ denote the regression coefficient vector for class k . First, assume Y_{ij} 's are independently generated from class k ($k = 1, \dots, K$), with the relationship between \mathbf{Y}_i and $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})$ given by

$$Y_{ij} | Z_{ij} = k \sim \mathcal{N}(\zeta_k(\mathbf{X}_i), \sigma_k^2) \equiv f_k(\cdot | \boldsymbol{\beta}_k, \sigma_k^2),$$

where $\zeta_k(\cdot)$ is shorthand for a function specifying the regression relationship, i.e. for a linear regression model, $\zeta_k(\mathbf{X}_i) = \mathbf{X}_i^T \boldsymbol{\beta}_k$. If class memberships are unknown a priori, π_k ($1 \leq k \leq K$) denoting the prior probability of being in class k , the Y_{ij} 's can be

assumed to be generated from a mixture of regression models, each component of the mixture encapsulating a separate regression. However, if the measurements are made over time, the observations Y_{i1}, \dots, Y_{iT} are intrinsically dependent. By setting the *class membership* to account for the dependence in \mathbf{Y}_i , we capture this dependence through a hidden Markov model, introducing a stochastic relationship in the latent class indicator Z_i . Assuming that the dependence structure is homogeneous over time, we model the dependence in Z_{ij} ($j = 1, \dots, T$) using a transition matrix $\boldsymbol{\tau} = ((\tau_{kl}))$, where

$$\tau_{kl} = P(Z_{ij} = l | Z_{i,j-1} = k) \quad \forall j \in \{1, \dots, T\}; \quad k, l \in \{1, \dots, K\}. \quad (2.1)$$

At the initial time point ($j = 1$), we assume the prior probability of states is given by $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$, where $P(Z_{i1} = k) = \pi_k$, ($k = 1, \dots, K$). We denote this model as a *hidden Markov regression model* (HMRM), where the conditional distribution of Y_{ij} at time j for unit i can be written in the form:

$$Y_{ij} | Z_{ij} = k \sim \mathcal{N}(\zeta_k(\mathbf{X}_i, \phi(j)), \sigma_k^2). \quad (2.2)$$

In (2.2), we have included an extra covariate $\phi(j)$ for the j -th time point, to account for time-dependence. (2.2) and (2.1) completely specify the hidden Markov regression model, up to the functional forms of $\zeta_k(\cdot)$ and $\phi(j)$. In the following section, we describe the model components in more detail.

2.1 Modeling cell-cycle effects

The two goals of interest are to: (i) determine the variables \mathbf{X} influencing the observed \mathbf{Y} and (ii) find the optimal number of classes K . We assume that the observed gene expression measurement Y_{ij} at time j is a composite of two effects, a *fixed* motif-dependent effect $\zeta_k(\mathbf{X}_i, \phi(j))$, and a *random* effect ψ_{ijk} due to gene-specific differences in expression magnitude, where $\psi_{ijk} | Z_{ij} = k \sim \mathcal{N}(\mu_{kj}, \xi_0 \sigma_k^2)$ (ξ_0 denoting a variance inflation factor). If a sufficient number of replicate measurements are available, gene-specific variances can

be modeled by using a hierarchical prior. Since such replication is often not available (as in the yeast data set), genes in a common class are currently assumed to have the same variance. Integrating out the random effect gives the marginal distribution of Y_{ij} as

$$Y_{ij}|Z_{ij} = k \sim \mathcal{N}(\mu_{kj} + \zeta_k(\mathbf{X}_i, \phi(j)), (1 + \xi_0) \sigma_k^2). \quad (2.3)$$

We next choose appropriate functional forms for $\zeta(\cdot)$ and $\phi(\cdot)$. For regulatory networks over time, the covariates of interest are motif scores corresponding to each gene. We assign each sequence promoter region a “score” covariate X_{il} ($1 \leq l \leq D$; $1 \leq i \leq N$) with respect to the set of D motif patterns, indicating its propensity to contain one or more binding sites for that motif (Section 2.2).

Most previous work relating sequence effects with gene expression assume a linear relationship (Bussemaker et al., 2001; Conlon et al., 2003) which may be a simplification of the real biological model. However, for balancing the trade-off between parameter identifiability and model sophistication, in the face of increased model complexity, we also currently consider only linear covariate effects of the motifs. For the time covariate, the effect is cyclical, depending on which functional cluster each gene lies within and the point of the cell cycle being observed. With a total set of D motif covariates, we assume

$$\zeta_k(\mathbf{X}_i, \phi(j)) = \alpha_{1k} \sin \phi(j) + \alpha_{2k} \cos \phi(j) + \sum_{l=1}^D \beta_{kl} X_{il}, \quad (2.4)$$

where $\phi(j)$ is a suitably chosen function that maps time point j into a phase of the cell-cycle. The state-dependent sinusoidal coefficients allow the flexibility of gene clusters varying in amplitude of their effects, as well as frequency. This may help, for example, in identifying clusters having opposing time-dependent effects over the cycle. Since the data set was partially synchronized before analysis, we limited the number of harmonic components to one to avoid over-parametrization. More stringent tests could be carried out to determine an adequate number of harmonic components (see, for example, Quinn (1989)).

Next, we summarize the general data structure. For simplifying the notation, unless indicated otherwise, the subscript ranges in the following are: $i = 1, \dots, N$; $j = 1, \dots, T$; $l = 1, \dots, p$, with $p = 2 + D$. Let $\mathbf{Y}_{(k)} = \{Y_{ij} - \mu_{kj} : Z_{ij} = k\}$, and $\mathbf{x} = \{x_{ijl}\}$, where $x_{ij}^* = (x_{ij1}, \dots, x_{ijp}) = (\sin \phi(j), \cos \phi(j), X_{i1}, \dots, X_{iD})$. Further let $\mathbf{X} = ((X_{il}))$, $(1 \leq i \leq NT) \equiv \text{stack}\{x_{ij}^*, 1 \leq i \leq N; 1 \leq j \leq T\}$. \mathbf{X} is of dimension $NT \times p$, where each row corresponds to a realization of one gene (i) at one time point (j). We denote the subset of \mathbf{X} corresponding to observations in state k as $\mathbf{X}_{(k)} = \{\mathbf{X} : Z_{ij} = k\}$. Let $\mathbf{u} = (u_1, \dots, u_p)$ denote a p -dimensional vector where $u_l = 1$ (0) denotes that covariate X_l is present (absent) in the model. The subset of \mathbf{X} indexed by the *selected variables* \mathbf{u} is $\mathbf{X}^{(\mathbf{u})} = \{X_{il} \in \mathbf{X} : u_l = 1; 1 \leq i \leq NT\}$. Then, the submatrix of \mathbf{X} formed by the rows (i, j) corresponding to $Z_{ij} = k$, and $|\mathbf{u}| = \sum_{l=1}^p u_l$ columns such that $u_l = 1$, is given by $\mathbf{X}_{(k)}^{(\mathbf{u})} = \{\mathbf{X}^{(\mathbf{u})} : Z_{ij} = k\}$. Also, for the set of coefficients $\boldsymbol{\beta}_k = (\alpha_{1k}, \alpha_{2k}, \beta_{k1}, \dots, \beta_{kD})^T$ corresponding to the subset of variables in state k , indexed by \mathbf{u} , we use the notation $\boldsymbol{\beta}_k^{(\mathbf{u})} = \{\beta_{kl} : u_l = 1\}$.

2.2 Motif scoring model and covariates

Each upstream sequence is next given a *score* with respect to each position-specific weight matrix (PSWM), Θ_j , $(1 \leq j \leq D)$, of width w_j columns, to get a $N \times D$ covariate matrix of gene-sequence scores \mathbf{X} , where each row $\mathbf{X}_i = (X_{i1}, \dots, X_{iD})$ is the score vector for gene i ($1 \leq i \leq N$). As mentioned in Section 1.1, a PSWM is characterized as $\Theta = ((\theta_{jk}))$ ($1 \leq j \leq 4; 1 \leq k \leq w$), to denote a motif of w columns, where each column of the matrix $\boldsymbol{\theta}_k = (\theta_{1k}, \dots, \theta_{4k})^T$ denotes the relative frequencies of each letter in that position. Let $\boldsymbol{\theta}_0 = (\theta_{01}, \dots, \theta_{04})$ denote the relative frequencies of the four nucleotides characterizing the *background* sequence (not containing motifs) assuming that the sequence is generated from a Markov process of order zero. If the probability of motif occurrence is assumed uniform over the sequence, then the score X_{ij} for sequence i , relative to motif j , is taken to be the logarithm of the likelihood ratio between the j -th motif model and the

background model, assuming that any number of motif sites can occur in the sequence, with the constraint that there is no overlap between sites.

Let $\mathbf{x}_{[l:m]}$ denote the sequence $\{x_l, x_{l+1}, \dots, x_m\}$. For the motif covariates X_{ij} for gene i and motif j , given the position-specific PSWM Θ_j , the likelihood ratio is $\frac{P(\mathbf{x}_{[1:L_i]}|\Theta_j)}{P(\mathbf{x}_{[1:L_i]}|\boldsymbol{\theta}_0)}$, where L_i the length of the upstream sequence for gene i . Calculation of the denominator is straightforward, for both an independent and Markovian model. However, since the actual partitions of the sequence into individual sites and “background” is not known, evaluating the numerator (with Θ_j fixed) involves an exponential order sum— including all possible segmentations into possible motif sites and background sequence. Fortunately, we can use a recursive technique that has been formulated in detail in Gupta and Liu (2003) in the context of motif discovery, for more efficient computation. Let $\Phi_{k-1}(\boldsymbol{\Theta})$ be the sum of probabilities for all legitimate partitions for partial sequence $\mathbf{x}_{[1:(k-1)]}$. Then, we can recursively evaluate Φ_k as

$$\Phi_k(\boldsymbol{\Theta}) = \sum_{l \in \{1, w_j\}} \rho(\mathbf{x}_{[(k-l):k]}) \Phi_{k-l}(\boldsymbol{\Theta}), \quad (2.5)$$

where $\rho(\mathbf{x}_{[l:m]}) = P(\mathbf{x}_{[l:m]}|\Theta_j)^{1_{[m-l+1=w_j]}} P(\mathbf{x}_{[l:m]}|\boldsymbol{\theta}_0)^{1-1_{[m-l+1=w_j]}}$ evaluates the probability that the previous segment is either a motif site or a letter generated from the background, and $1_{[X=a]}$ denotes an indicator variable that takes value 1 only if $X = a$. $P(\mathbf{x}_{[l:m]}|\Theta_j)$ is calculated assuming the letter frequencies in each position are independent multinomials with parameters corresponding to columns of Θ_j . By recursively evaluating expression (2.5), the likelihood of the entire segment is calculated as $\Phi_{L_i}(\boldsymbol{\Theta})$ for sequence i ($1 \leq i \leq N$). We note that Conlon et al. (2003) use a simpler version of this score under the assumption that a sequence can contain only a single motif site.

2.3 A hierarchical prior framework

In a linear regression framework, an attractive choice of prior for the regression coefficient $\boldsymbol{\beta}$ is the conjugate g-prior (Zellner, 1986). With the error variance denoted as σ^2 , the

g-prior is of the form $g\sigma^2 I^{-1}$, where I denotes the Fisher information matrix $\mathbf{X}^T \mathbf{X}$, and g is a scalar constant. However, when the observations belong to one of K unknown classes, i.e. in a mixture or HMM framework, the density is no longer from a regular exponential family, and the information matrix cannot be derived in an analytical form. In our model, the entire set of regression coefficients for state k ($1 \leq k \leq K$) is denoted by $\boldsymbol{\beta}_k = (\alpha_{1k}, \alpha_{2k}, \beta_{k1}, \dots, \beta_{kD})^T$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K)$. Following the general form of the g-prior, we choose a conjugate prior for $\boldsymbol{\beta}$, as $\boldsymbol{\beta}_k \sim \mathcal{N}(\boldsymbol{\beta}_0, g\sigma_k^2(\mathbf{X}^T \mathbf{X})^{-1})$. This approximates the ‘‘complete data’’ g-prior with covariance $g\sigma_k^2(\sum_{i:Z_i=k} \mathbf{X}_i \mathbf{X}_i^T)^{-1}$, which cannot be directly evaluated since \mathbf{Z} is a latent variable. Our ‘‘modified’’ g-prior preserves a weakened form of the correlation structure of the likelihood, in comparison to using a simpler (e.g. independent) form, and leads to several attractive resultant properties in posterior estimates (Section 3.1.1). Also, by not assuming an a priori independence structure on $\boldsymbol{\beta}$, we can avoid imposing posterior independence among the TF effects, which appears biologically more meaningful. The choice of the scalar g , which controls the penalty for choosing models, is discussed further in Section 3.2.

For the other parameters, we use standard conjugate prior formulations. Let $\boldsymbol{\mu} = ((\mu_{kj}))$; $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_K^2)$; ($1 \leq j \leq T$; $1 \leq k \leq K$). We take $\mu_{kj} \sim \mathcal{N}(m_{k0}, v_{k0}^2)$, $\frac{1}{\sigma_k^2} \sim \text{Gamma}(w_0/2, S_0/2)$, $(\tau_{k1}, \tau_{k2}, \dots, \tau_{kK}) \sim \text{Dirichlet}(\boldsymbol{\omega}_k = \omega_{k1}, \dots, \omega_{kK})$; and $(\pi_1, \dots, \pi_K) \sim \text{Dirichlet}(\boldsymbol{\alpha}_0 = (\alpha_{01}, \dots, \alpha_{0K}))$. As a hierarchical prior for \mathbf{u} , we set $u_l | \eta \sim \text{Bernoulli}(\eta)$ ($1 \leq l \leq p$) and $\eta \sim \text{Beta}(\epsilon_1, \epsilon_2)$.

3 Parameter estimation in the HMRM

After developing the model in Section 2, we now introduce a hybrid Monte Carlo procedure for efficient model fitting and parameter estimation. Let $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\tau})$ denote the set of all parameters in the model. To start with, assume that the total number of states K , is fixed and known. The complete data posterior distribution, integrating out

the random effects $\boldsymbol{\psi}$, is given by

$$\begin{aligned}
P(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{u}) &\propto P(\mathbf{Y}|\mathbf{X}, \mathbf{Z}, \mathbf{u}, \boldsymbol{\theta})P(\mathbf{Z}|\boldsymbol{\theta})P(\mathbf{u}|\mathbf{X}, \boldsymbol{\theta})P(\boldsymbol{\theta}) \\
&\propto \prod_{k=1}^K \prod_{j=1}^T \left[\left\{ \prod_{i:Z_{ij}=k} \mathcal{N}(y_{ij}; \mu_{kj} + \zeta_k(\mathbf{X}_i, \phi(j)), (1 + \xi_0)\sigma_k^2) \times p(\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\tau}) \right\} \right. \\
&\quad \left. \times \left\{ \prod_{i=2}^N P(Z_{ij}|Z_{i,j-1}, \boldsymbol{\tau}) \right\} \right] \times IBet_{\epsilon_1, \epsilon_2}(|\mathbf{u}|, D - |\mathbf{u}|). \tag{3.1}
\end{aligned}$$

We update the model parameters through the following iterative procedure: (i) Covariate selection: update $\mathbf{u}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}$; (ii) State updating: update $\mathbf{Z}|\mathbf{Y}, \mathbf{X}, \mathbf{u}, \boldsymbol{\theta}$; (iii) Parameter updating: update $\boldsymbol{\theta}|\mathbf{Y}, \mathbf{X}, \mathbf{u}, \mathbf{Z}$. Details of each step are provided in the next section.

3.1 Covariate selection

The initial number p of covariates included in the model may potentially be very large. Covariates that are not significantly correlated with the response may introduce noise and result in inaccurate fitting, especially as the clusters are determined by the state-specific regression relationship between the response and covariate. We thus include a variable selection step using the evolutionary Monte Carlo (EMC) procedure, which has been shown to be highly efficient in high-dimensional problems (Liang and Wong, 2000).

3.1.1 Evolutionary Monte Carlo procedure

EMC is a *population-based* Monte Carlo procedure that involves (i) sampling simultaneously from parallel Monte Carlo chains using *tempered* versions of the target distribution, ranging from lowest (target distribution) to highest (“flattest” distribution), maximizing mixing and (ii) using local Metropolis-Hastings moves of *mutation* and *crossover*.

In the variable selection step, we need to compare models of dimensions $|\mathbf{u}|$ and $|\mathbf{v}|$ ($|\mathbf{u}| = \sum_{l=1}^p u_l$). We first calculate a marginalized likelihood, integrating out the

parameters whose dimensions vary with \mathbf{u} . With conjugate priors for β_k and σ_k^2 , we get

$$\begin{aligned}
\mathcal{H}(\mathbf{u}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\mu}) &= P(\mathbf{Y}|\mathbf{X}^{(u)}, \mathbf{Z}, \boldsymbol{\mu}, \mathbf{u}) \tag{3.2} \\
&= \int P(\mathbf{Y}|\mathbf{X}^{(u)}, \mathbf{Z}, \boldsymbol{\mu}, \mathbf{u}, \boldsymbol{\beta}^{(u)}, \boldsymbol{\sigma}^2, \boldsymbol{\tau}) P(\mathbf{u}|\boldsymbol{\eta}) p(\boldsymbol{\beta}^{(u)}, \boldsymbol{\sigma}^2, \boldsymbol{\eta}, \boldsymbol{\tau}) d\boldsymbol{\beta}^{(u)} d\boldsymbol{\sigma}^2 d\boldsymbol{\eta} d\boldsymbol{\tau} \\
&= IBet_{\epsilon_1, \epsilon_2}(|u|, D - |u|) IDir_{\boldsymbol{\alpha}_0}(\mathbf{m}_1) \times \prod_{k=1}^K \left[IDir_{\boldsymbol{\omega}_k}(\mathbf{t}_k) \times \right. \\
&\quad \left. IIGamma_{\frac{w_0}{2}, \frac{s_0}{2}}\left(\frac{n_k}{2}, \frac{R_k^{(u)}}{2}\right) \frac{[2\pi(1 + \xi_0)]^{-n_k/2} \left| \frac{\mathbf{X}^{(u)T} \mathbf{X}^{(u)}}{g} \right|^{\frac{1}{2}}}{\left| \frac{\mathbf{X}^{(u)T} \mathbf{X}^{(u)}}{1 + \xi_0} + \frac{\mathbf{X}^{(u)T} \mathbf{X}^{(u)}}{g} \right|^{\frac{1}{2}}} \right], \tag{3.3}
\end{aligned}$$

where $IBet_{a,b}(c, d)$ denotes the inverse ratio of the normalizing constants for the Beta distributions $\text{Beta}(a + c, b + d)$ and $\text{Beta}(a, b)$, and $IIGamma_{a,b}(c, d) = \frac{(b+d)^{a+c}\Gamma(a)}{b^a\Gamma(a+c)}$ denotes the ratio of normalizing constants for the Inverse-Gamma distributions $IG(a + c, b + d)$ and $IG(a, b)$. $IDir_{\boldsymbol{\omega}_k}(\mathbf{t}_k)$ denotes the ratio of normalizing constants for the Dirichlet distributions $Dir(\mathbf{t}_k + \boldsymbol{\omega}_k)$ and $Dir(\boldsymbol{\omega}_k)$, where $\mathbf{t}_k = (t_{k1}, \dots, t_{kK})$; $t_{kl} = \sum_{i=1}^N \sum_{j=2}^T 1_{[Z_{i,j-1}=k, Z_{ij}=l]}$ is the number of observed transitions between states k and l . $\mathbf{m}_1 = (\mathbf{m}_{11}, \dots, \mathbf{m}_{1K})$; $\mathbf{m}_{1k} = \sum_{i=1}^N 1_{[Z_{i,1}=k]}$, and $n_k = \sum_{i=1}^N \sum_{j=1}^T 1_{[Z_{ij}=k]}$. Also,

$$R_k^{(u)} = \frac{1}{1 + \xi_0} \mathbf{Y}_{(k)}^T \mathbf{Y}_{(k)} + \frac{1}{g} \boldsymbol{\beta}_0^{(u)T} \mathbf{X}^{(u)T} \mathbf{X}^{(u)} \boldsymbol{\beta}_0^{(u)} - \tilde{\boldsymbol{\beta}}_k^{(u)T} \tilde{\boldsymbol{\Sigma}}_{\beta k(u)}^{-1} \tilde{\boldsymbol{\beta}}_k^{(u)}, \tag{3.4}$$

where the posterior estimates for the covariance and mean of β_k are

$$\tilde{\boldsymbol{\Sigma}}_{\beta k(u)} = \left[\frac{\mathbf{X}_{(k)}^{(u)T} \mathbf{X}_{(k)}^{(u)}}{1 + \xi_0} + \frac{\mathbf{X}^{(u)T} \mathbf{X}^{(u)}}{g} \right]^{-1} \text{ and } \tilde{\boldsymbol{\beta}}_k^{(u)} = \tilde{\boldsymbol{\Sigma}}_{\beta k(u)} \left[\frac{\mathbf{X}_{(k)}^{(u)T} \mathbf{Y}_{(k)}}{1 + \xi_0} + \frac{\mathbf{X}^{(u)T} \mathbf{X}^{(u)}}{g} \boldsymbol{\beta}_0^{(u)} \right]. \tag{3.5}$$

R_k can be interpreted as a cluster-specific *residual* (see Appendix A). Also, for a large value of g , the R_k term has the attractive property of reducing to a scaled version of the frequentist residual sum of squares in the regression framework (Appendix B). Conversely, for a small value of g , implying a strongly informative prior for $\boldsymbol{\beta}$, R_k represents a sum of squares scaled by the *involutory* matrix $(I - 2H_k)$, where $(I - 2H_k)^T(I - 2H_k) = I$ (Harville, 1997). It is interesting to note that while $\mathbf{Y}_{(k)}^T(I - H_k)\mathbf{Y}_{(k)} = \mathbf{Y}_{(k)}^T(I - H_k)(I - H_k)\mathbf{Y}_{(k)}$ represents the residual sum of squares, $\mathbf{Y}_{(k)}^T(I - 2H_k)(I - 2H_k)\mathbf{Y}_{(k)} = \mathbf{Y}_{(k)}^T\mathbf{Y}_{(k)}$,

the total sum of squares. Since $(I - 2H_k)$ is not positive definite (though it is non-singular), taking a very small value of g may lead to instability or inestimability of parameter estimates in case $\mathbf{Y}_{(k)}^T(I - 2H_k)\mathbf{Y}_{(k)}$ is negative. Thus it is desirable to perform a few pilot runs before selecting an appropriate g for the analysis. Further details of the EMC procedure are given in Appendix C.

A remaining task is to sample the states and parameters. A Gibbs sampling procedure may be used to sample states, one gene at a time, successively from the conditional distributions $P(Z_{i,t}|Z_{i,1}, \dots, Z_{i,t-1}, Z_{i,t+1}, \dots, Z_{i,T}, \mathbf{X}, \mathbf{u}, \boldsymbol{\theta})$. However, a more efficient method in this scenario is a recursive data augmentation (DA) step, based on *forward sum-backward sampling* techniques used in computation of the likelihood in hidden Markov-type models (Gupta and Liu, 2003). This represents a *grouped* sampling step rather than a conditional update, and has been shown to have better convergence properties than the Gibbs sampler (Liu et al., 1994). Since conjugate priors are used, each of the updating steps is of closed analytical form and straightforward. The state and parameter updating steps are detailed in Appendices D and E.

3.2 Sensitivity of covariate selection to choice of g

A large g in the prior for β_k results in robust posterior inference for β , but over-penalizes larger models. As $g \rightarrow \infty$, the Bayes factor for comparing any choice of predictors to the null model tends to zero, making the model selection procedure inconsistent (see Appendix F and Bartlett (1957)). For regular families, the choice of $g = n$ represents a “unit information prior”, leading to Bayes factors that behave as the BIC (Kass and Wasserman, 1995). In the case of HMMs this result does not hold as the observations are correlated. Our empirical studies suggest that the BIC tends to overestimate the number of components.

One alternative is to take a hyperprior on g . By assuming $g \sim \text{Inv-}\chi^2(\nu)$, $(\beta_k - \beta_0) \sim t_\nu(\cdot; \sigma_k^2(\mathbf{X}^T \mathbf{X})^{-1})$, a scaled t-distribution with ν degrees of freedom. The robustness of

the t distribution makes it an attractive alternative to using a fixed g -prior. In the case of the marginal distribution (3.2), integrating out g is analytically intractable; however, one can sample g from its posterior distribution during the MCMC procedure. This involves adding a sampling step for g , using:

$$\frac{1}{g} | \mathbf{Y}, \mathbf{X}, \mathbf{u}, \mathbf{Z}, \boldsymbol{\beta} \sim \text{Gamma} \left(\frac{\nu + K|u|}{2}, \frac{1 + \sum_{k=1}^K \sigma_k^{-2} (\boldsymbol{\beta}_k^{(u)} - \boldsymbol{\beta}_0^{(u)})^T \mathbf{X}^{(u)T} \mathbf{X}^{(u)} (\boldsymbol{\beta}_k^{(u)} - \boldsymbol{\beta}_0^{(u)})}{2} \right).$$

In applications, a numerically stable procedure appears to be initiating the algorithm with a large g value, and sampling g with other parameters once the algorithm stabilizes. In general, one must check that the diagonal elements of $g(\mathbf{X}^T \mathbf{X})^{-1}$ are not too small.

4 Bayesian criterion-based model selection

To estimate the number of states K in the HMRM, a usual approach would be calculating the Bayes factor for models with different K . The HMRM involves the latent variables \mathbf{u} and \mathbf{Z} , relating to the included covariates and hidden states. To evaluate the Bayes factor, we need to compute the marginal probabilities $P(\mathbf{Y} | \mathcal{M}, \mathbf{X}) = \int_{\boldsymbol{\theta}} \sum_{\mathbf{u}} \sum_{\mathbf{Z}} P(\mathbf{Y} | \mathcal{M}, \mathbf{X}, \boldsymbol{\theta}, \mathbf{Z}, \mathbf{u}) P(\boldsymbol{\theta} | \mathcal{M}) P(\mathbf{Z}) P(\mathbf{u}) d\boldsymbol{\theta}$, for each model \mathcal{M} , integrating out the latent variables. These sums are analytically intractable, and computational methods for calculating the Bayes factor (Meng and Wong, 1996; Chen and Shao, 1997; Green, 1995) are typically either computationally expensive, or unstable due to the irregular nature of the likelihood function.

Instead, we use an alternative approach through a Bayesian goodness-of-fit statistic constructed from the posterior predictive distribution of the data. The L-measure (Ibrahim et al., 2001), and its calibration distribution allows the formal comparison of two competing models. Let \mathbf{W}_i , ($i = 1, \dots, N$) denote the future values from an imagined replicate experiment, with the same sampling distribution as \mathbf{Y}_i , ($i = 1, \dots, N$) as in Section 2. The generalized L-measure (Ibrahim et al., 2001) is

$$L(\mathbf{y}, \nu) = E[(\mathbf{W} - E(\mathbf{W} | \mathbf{y}))(\mathbf{W} - E(\mathbf{W} | \mathbf{y}))]^T + \nu(E(\mathbf{W} | \mathbf{y}) - \mathbf{y})(E(\mathbf{W} | \mathbf{y}) - \mathbf{y})^T, \quad (4.1)$$

where the expectations are taken with respect to the posterior predictive distribution $P(\mathbf{W}|\mathbf{y})$, and $0 \leq \nu \leq 1$ (see Appendix G for details). The two terms in (4.1) can be interpreted as corresponding to a variance and bias term, and hence choosing $\nu = 1$ would reduce (4.1) to the matrix of Euclidean distance between \mathbf{y} and \mathbf{W} . An attractive feature of the L-measure approach, compared to Bayes factors, is that this criterion gives an absolute judgment of whether the selected model provides adequate fit to the data, rather than a comparison between models. To calibrate the L-measure, we compute the empirical distribution of $D(\mathbf{y}, \nu) = \hat{L}_c(\mathbf{y}, \nu) - \hat{L}_t(\mathbf{y}, \nu)$, where \hat{L}_c and \hat{L}_t denote the estimated trace of the L-measure for the candidate and “true” models. In practice, as the true model is unknown, we follow Ibrahim et al. (2001) in replacing the true model by the “criterion-minimizing” model, and generating samples $\tilde{\mathbf{y}}$ from the prior predictive distribution of \mathbf{y} .

5 Simulation studies

As a proof-of-principle test for the new approach, simulation studies were conducted to test the (i) consistency of the method in estimating parameters in presence of noise and increasing dimensionality of the data, and (ii) robustness of the method to misspecification of hyperparameters.

Data was generated from a two-cluster and five-cluster model with 400 response variables (genes) and 15 “true” covariates (motifs), with measurements over 18 time points. The motif scores were simulated from a set of real yeast PSWMs through the scoring function described in Section 2.2. We also generated a set of 35 dummy covariates corresponding to each gene, some of which resembled true motif scores not significantly correlated with the response, and the remaining as noise. Given the motif scores, and cluster identity Z_{ij} at time point j , for gene i , the distribution of gene expression scores

was given by

$$Y_{ij}|Z_{ij} = k \sim \mathcal{N}\left(\mu_{kj} + \alpha_{1k} \sin \phi(j) + \alpha_{2k} \cos \phi(j) + \sum_{l=1}^{15} \beta_{kl} X_{il}, (1 + \xi_0) \sigma_k^2\right).$$

We tested the algorithm both by fixing g such that $g(\mathbf{X}^T \mathbf{X})^{-1}$ ranged between 1 and 100 and also allowing g to vary once the algorithm had achieved some stability- either way, we observed no significant variability in the posterior estimates. The next tests for consistency and robustness were done assuming the total number of true states K was known. ACFs for all parameters for different specifications of ξ_0 were seen to have a maximum autocorrelation of 0.05 at lag 5 (Supplementary material) so the algorithm was assumed to have converged to the stationary distribution over the range of iterations being considered (with a burn-in of about 5000). A typical run of 50,000 iterations on a data set of the size mentioned took a median CPU time of 23.43 hours on a IBM BladeCenter cluster with dual Intel Xeon 2.4GHz nodes running RedHat Linux 7.3.

Tests for consistency. We tested the performance of the algorithm under increasing levels of noise, by introducing sets of “noise” variables into the algorithm, 15, 25, and 35 variables in turn. Results are presented for the $K = 5$ data set; the $K = 2$ data set gives even more accurate results due to the decreased complexity of the model, and are not shown here. Table 1 shows that increasing the number of covariates has almost no effect on the selection of the correct ones. This observation suggests that the algorithm is robust to mis-inclusion of covariates which have no effect on the response, and thus should be able to efficiently select covariates even when the total dimensionality is high, as long as the true covariates have a significant effect on the response. To estimate the performance of the algorithm when unknown or unmodeled motifs interact with the pathway, we also ran a similar analysis excluding one and two “true” motifs from the model fitting step (Table 1). This leads to a very slight increase in the errors of variable selection. The parameter estimates are consistent, with the exception of σ^2 , for which the informative prior appears to have a biasing effect for clusters of very small sizes.

[Table 1 about here]

Tests for robustness to hyperparameter specification. We tested the performance of the algorithm under varying amounts of noise from the random effects component ξ_0 , and under varying degrees of misspecification of ξ_0 . Results (Figure 1 in Appendix) show that the algorithm performs consistently well for a range of values of ξ_0 (given in the log scale), but as expected, at extremely high values of ξ_0 the performance declines as more noise is introduced into the data. However, the performance is extremely robust to parameter settings- for lower true values of ξ_0 , the magnitude of set values of ξ_0 appears to have no effect on either the MSEs of parameter estimates, misclassification rates, or correct selection of covariates.

Model selection using the L-measure. For the simulated data set, we next applied the model selection method based on the L-measure to test whether the method could recover the true value of K . The results here are presented for the $K = 2$ data set. We used the posterior estimates to construct the L-measure for values of ν varying between 0 and 1. It can be seen that the true model with $K = 2$ outperforms all the other models (Table 2). To check whether the fitted model was not significantly different from the true, we constructed the calibration distribution for the difference $D(\mathbf{y}, \nu) = \hat{L}_{K=2}(\mathbf{y}, \nu) - \hat{L}_t^*(\mathbf{y}, \nu)$, where $\hat{L}_t^*(\mathbf{y}, \nu)$ denotes the L-measure for the true model with the true parameter values. For almost all values of ν (except for very small ones), it appeared that the model fits quite well. As a comparison, we also computed the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) model selection criteria based on the likelihood calculated through recursion (Appendix D) at the modal parameter estimates for $K = 2, 3, 4$. As can be seen in Table 2, while AIC and BIC tend to slightly overfit the data, the L-measure clearly picks out the correct number of states. Since the effective sample size (ESS) is an issue with the correct calculation of the BIC for the dependent data, we computed the BIC for both extremes with sample

size equaling N (under-estimate, assuming each gene contributes an ESS of 1) and NT (over-estimate, assuming measurements for a gene over all time points are independent); the “true” ESS would be between these two extremes. However, since over-fitting occurs for both extremes, it appears that the L-measure is probably a more accurate model selection criterion in this case.

[Table 2 about here]

6 Case study: Analysis of yeast cell-cycle data

In this section we describe the analysis of the yeast cell cycle data set (Spellman et al., 1998) applying the new methodology. The HMRM model was fitted to the yeast cell cycle data over a range of values for the total number of states K , with a starting set of 24 motif covariates, and the L-measure used to determine the optimal number of states. We chose the hyperparameter settings as suggested by our simulation studies, ξ_0 in a range of 0.1 – 5, g in the range 1 to 100, and the results were virtually identical within this range. We report here the results for the settings of $\xi_0 = 0.1$, $g = 100$, $\epsilon_1 = \epsilon_2 = 2$, $\omega_{kk} = 1$, $\omega_{k,-k} = 9$ ($k = 1, \dots, K$), and data-dependent priors for $S_0 = \text{Var}(\mathbf{Y})$ and $w_0 = \sqrt{N}$. Table 3 indicates that $K = 3$ or 4 are almost identical to the model with the lowest value of the L-criterion ($K = 5$).

[Table 3 about here]

6.1 Data analysis

Posterior convergence diagnostics were computed using the R CODA package. Auto-correlation plots for all parameters and trace plots of the number of selected variables (Figure 3 in Supplementary Material) showed adequate convergence on taking about 50,000 iterations of the sampler. Posterior samples were used to make inference regarding the classification of genes and to judge whether there were significant effects of certain

transcription factors on groups of genes. Figure 1 gives the motifs that showed significant effects (posterior probability of selection greater than 50%) and the 95% confidence intervals for the regression coefficient for each state, with $K = 3$ and 5. For almost all selected motifs, the posterior intervals do not overlap between states, and in at least one state shows a highly significant positive or negative effect on gene expression. For $K = 5$, significant positive effects (details in Supplementary Table 2) are shown by the TFs GAL4, GCR1, MATalpha2, MIG1, XBP1 (state 1); CSRE, GCN4, MCM1, PHO2, UASPHR (state 2); MCB, PHO4, RAP1, STE12, SWI5 (state 4); while significant negative effects are shown by CAR1 repressor, CSRE, MCM1, PHO4, RLM1, ROX1, STE12, UASPHR (state 1); GCR1, MCB, MIG1, RAP1, SWI5, XBP1 (state 2); CSRE, GAL4, GCN4, GCR1, MATalpha2, UASPHR (state 4); and SMP1 (state 5). State 3 overall seems to show weaker motif effects. The significant transcription factors picked out by $K = 3$ states include many of the same motifs, suggesting that the extra states are formed by subdividing some of the previous states (e.g. state 1 is similar for the two sets, while state 2 for the $K = 5$ model is similar to state 3 in the $K = 3$ model). By comparing to the motifs listed in Spellman et al. (1998), we see that a number of factors known to have strong effects on the cell cycle, such as MCB, MCM1 and SWI5, are detected by the method, as well as motifs that are active at specific time points (Table 4). By jointly modeling temporal expression with motif sequence scores, we thus can get a simultaneous picture of groups of genes regulated by certain transcription factors, that may have opposing effects in different groups. Since the groups are allowed to change over time, this implies that we can uncover the pattern of regulation of transcription factors over the cell-cycle, not being limited by a fixed grouping of genes.

[Figure 1 about here]

6.2 Biological validation

We compared our results with previous inference on the same data set (Spellman et al., 1998) and also re-analyzed the data using a stepwise multiple regression approach on the lines of Conlon et al. (2003). The stepwise method uses each time point separately, and assumes all genes in a single group without clusters. It uses the AIC and a forward-backward procedure to select significant covariates. In comparison, the HMRM detects the overall influence of motifs over time, and allows genes to belong to different states at different times. Table 4 indicates that the HMRM succeeds in uncovering more motif effects, that are also observed to be cluster-specific. Motifs that are not found significant by HMRM are neither found by stepwise regression, with the exception of TBP. The stepwise method misses known cell-cycle regulators MCB and SWI5, while MCM1 signals are weak (picked up at only one time point). One reason why they might not appear significant is that they have opposing effects in sub-groups of genes at different time points (e.g. SWI5 has a negative effect in group 1 and a positive effect in group 3 for the $K = 3$ model), opposing effects nullifying the overall one. Also, although a few transcription factors show continual effects over a phase of the cell cycle (e.g. GCN4, MIG1), in many cases most motifs only show significant effects sparsely, as no information is borrowed over neighboring time points to judge whether the overall effect over a period of time is significant, which is biologically more meaningful.

[Table 4 about here]

We compared how the clusters based jointly on motif effects and gene expression correlate to groups of genes categorized by their involvement in a functional pathway over certain points in the cell cycle, as shown in Figure 7 of Spellman et al. (1998). Our results indicate that a number of functional groupings at particular phases of the cell cycle are shown to have high over-representation in certain clusters (Table 5): for example, genes involved in DNA repair (G1), DNA synthesis (G1), cell-cycle control

(G1), budding (G1 and M/G1), mitosis (M), nutrition (M) and mating (M/G1). Even more promising is the fact that the analysis for gene clusters are remarkably consistent for the $K = 3$ and $K = 5$ models; the pathway-related genes are found to be grouped together at the relevant phase irrespective of the model chosen, which is important for robust inference. The only significant difference is for the **budding:fatty acids** pathway for which the $K = 3$ and $K = 5$ models pick out two different groups, which are active at different points of the cell-cycle, hence are still consistent. It can be noted here that a single-cluster based analysis, such as the stepwise regression approach, does not provide us a mechanism for attempting this kind of pathway inference which can generate useful biological hypotheses for further testing.

[Table 5 about here]

7 Discussion

Treating gene expression clustering as a temporal variable may help in discovering relationships between functional sequence motifs, and groups of genes they regulate. Differing groups of genes may behave as a *cluster* at different time points, influenced by different groups of transcription factors. If two groups of genes are *differentially regulated* by a TF, e.g. in one group the TF acts by inducing the response, while in the other it has a repressive effect, grouping the genes together will lead to losing the effect of the TF for the entire set. In order to uncover the relationships between TFs and genes that are involved in biological pathways of interest, it is thus desirable to determine how these TF effects may vary between groups and over time.

The hidden Markov regression model framework allows for (i) determining covariates (motifs and phase of cell-cycle) that have significant effects on the response (gene expression) (ii) covariate effects varying between states (iii) gene clusters varying over time. The hierarchical model framework also leads to nicely interpretable properties of the poste-

rior estimates and asymptotic comparability to the frequentist framework. In addition, although the model as discussed here does not directly find novel motifs, a de-novo motif discovery method may be formulated that uses the model predictions to generate groups of genes for motif discovery. Our approach also seems to induce a “tighter” clustering of genes, by grouping noisy genes which are not significantly affected by the covariates as a separate cluster, as seen in the yeast cell-cycle application. However, a principled way to induce tight clustering under a model based framework, still needs to be studied in detail. Our approach allows a novel, and to our belief the first, model-based method for determining groups of genes being influenced by separate sets of covariates over time. Application to yeast cell cycle data succeeds in detecting “regulatory modules”- groups of genes regulated by sets of TFs, that match previous biological knowledge. The joint modeling approach also succeeds in discovering more known functional TFs than the stepwise method in the yeast data (e.g. MCB, SWI5)– one likely reason for this is the separation of effects of groups of TFs that have differential effects on groups of genes.

Our approach to determining the number of clusters K is based on a Bayesian model choice criterion, the L-measure. The use of this approach is supported by the fact that (i) the number of biologically “interesting” clusters are within an approximately known, small range of values, and (ii) inference for the significant covariates appears consistent for a small range of K around the “optimal” value. For instance, for the yeast cell-cycle data, if we choose five states instead of three, the gene-cluster allocation, and cluster-specific covariates remain essentially the same, with a few new genes (having weaker covariate effects) being sub-stratified. Another promising direction for avoiding the model choice issue is by considering extensions to models such as the infinite mixture model based on the Dirichlet process, which we intend to explore further in future work.

Acknowledgments

This research is supported by funding from the National Institutes of Health and the Environmental Protection Agency. The authors are grateful to two anonymous referees and the editor, whose comments significantly improved the content and presentation in this article.

References

- Bartlett, M. S. (1957). A comment on D.V. Lindley’s statistical paradox. *Biometrika*, 44:533–4.
- Bussemaker, H. J., Li, H., and Siggia, E. D. (2001). Regulatory detection using correlation with expression. *Nature Genetics*, 27:167–174.
- Chen, M.-H. and Shao, Q.-M. (1997). On Monte Carlo methods for estimating ratios of normalizing constants. *The Annals of Statistics*, 25:1563–1594.
- Conlon, E. M., Liu, X. S., Lieb, J. D., and Liu, J. S. (2003). Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl Acad. Sci. USA*, 100(6):3339–3344.
- Green, P. J. (1995). Reversible jump MCMC and Bayesian model determination. *Biometrika*, 82:711–732.
- Gupta, M. and Liu, J. S. (2003). Discovery of conserved sequence patterns using a stochastic dictionary model. *J. Am. Stat. Assoc.*, 98(461):55–66.
- Harville, D. A. (1997). *Matrix Algebra from a Statistician’s Perspective*. Springer-Verlag.
- Holmes, I. and Bruno, W. (2000). Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Proc Int Conf Intell Syst Mol Biol.*, 8:202–10.

- Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). Criterion-based methods for Bayesian model assessment. *Statistica Sinica*, 11(2):419–443.
- Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Stat. Assoc.*, 90:928–934.
- Liang, F. and Wong, W. H. (2000). Evolutionary Monte Carlo: applications to c_p model sampling and change point problem. *Statistica Sinica*, 10:317–342.
- Liu, J. S., Neuwald, A. F., and Lawrence, C. E. (1995). Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.*, 90:1156–1170.
- Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81:27–40.
- Liu, X., Brutlag, D. L., and Liu, J. S. (2002). An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotech.*, 20(8):835–9.
- Meng, X. L. and Wong, W. (1996). Simulating ratios of normalising constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 6:831–860.
- Quinn, B. G. (1989). Estimating the number of terms in a sinusoidal regression. *J. Time Ser. Anal.*, 10(1):71–75.
- Segal, E., Taskar, B., Gasch, A., Friedman, N., and Koller, D. (2001). Rich probabilistic models for gene expression. *Bioinformatics*, 17:S243–52.
- 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 the yeast *saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell.*, 9(12):3273–97.

- Wang, W., Cherry, J. M., Nochomovitz, Y., Jolly, E., Botstein, D., and Li, H. (2005). Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. *Proc Natl Acad Sci U S A*, 102(6):1998–2003.
- Zellner, A. (1986). *On assessing prior distributions and Bayesian regression analysis with g-prior distributions*, volume Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti of Goel, P.K. and Zellner, A. (Eds.), page 233. North-Holland, Amsterdam.
- Zhu, J. and Zhang, M. (1999). SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, 15(7):607–611.

Table 1: Comparing misclassification and variable selection rates for simulation studies, for number of states in the HMRM equaling 5. “Var sel” denotes the results of variable selection: “Sens” denotes the sensitivity, or proportion of correct variables found; “Spec” denotes specificity, the proportion of selected variables that are correctly selected. “1–Misc(%)” is 1 minus the misclassification rate, based on state allocation by the sampler.

Parameter	noise 15	noise 25	noise 35	missing motif 12	missing motif 15	missing motifs 12 & 15
Var sel						
Sens(%)	1	1	1	0.93	1	0.92
Spec(%)	1	1	1	0.92	0.96	0.88
1–Misc(%)	0.8	0.8	0.8	0.77	0.69	0.67
β Bias ²	.0364	.0372	.0363	.0554	.091	.109
(var)	(.0141)	(.0143)	(.014)	(.0309)	(.0402)	(.0642)
σ^2 Bias ²	24.7142	24.7622	24.6217	152.714	126.4307	279.9485
(var)	(.1299)	(0.1289)	(.1289)	(.7261)	(.7873)	(1.7794)
τ Bias ²	.0117	.0117	.0117	.0118	.0122	.0122
(var)	(.0004)	(.0004)	(.0004)	(.0004)	(.0005)	(.0005)
μ Bias ²	.0002	.0002	.0002	.0005	.0008	.001
(var)	(.0078)	(.0079)	(.0078)	(.0152)	(.0166)	(.0235)

Table 2: Comparison of model selection criteria for simulation study. The columns give the number of states K ; the mode of the log-likelihood; the AIC; the BIC with sample size N ($N \times T$); and the L -measure evaluated for $\nu = 0.5$. The optimal choice of K for each criterion is highlighted, showing that only the L -measure gives the correct selection of K , while AIC and BIC overfit by choosing 3 states.

K	$L(\hat{\theta})$	AIC	BIC	$\log L_K(\mathbf{y}, 0.5)$
2	-3516.75	-3570.75	-3670.751(-3748.79)	6.97
3	-2157.95	-2247.95	-2414.62(-2544.69)	11.25
4	-2176.41	-2304.41	-2541.45(-2726.43)	11.26

Table 3: Model selection based on L -criterion for yeast cell-cycle data. Based solely on the L -measure, the optimal model choice is for $K^* = 5$. Using the calibration distribution for $D(\mathbf{y}, \nu)$ shows that models with $K = 3, 4$ or 5 are essentially indistinguishable (95% posterior intervals of the difference almost symmetric about zero), hence we may choose either the most parsimonious model, with $K = 3$, or $K = 5$ for inference.

K	$\log L_K(\mathbf{y}, 0.5) - \log L_{K^*}(\mathbf{y}, 0.5)$	95% CI for $D(\mathbf{y}, \nu)$
2	0.2109	(0.0426, 0.1321)
3	0.0447	(-0.0322, 0.0373)
4	0.0283	(-0.0208, 0.04)

Table 4: *Transcription factors selected by (i) stepwise regression separately over each time point and (ii) the HMRM. The “+” and “-” signs denote positive and negative motif effects, based on the 95% confidence (posterior) interval for the regression coefficient. The confidence (posterior) intervals are given in Tables 1-3 in the Supplementary material.*

TF	Stepwise Regression		HMRM	
	Effect	Time	Effect (Cluster)	% Selected
ABF1				
CSRE			- (1), + (2), - (3)	0.95
GAL4			+ (1), - (4)	1
GCN4	+, +	180, 210	+ (2), - (4)	1
GCR1	-	330	+ (1), - (2), - (4)	1
MATalpha2			+ (1), - (4)	0.99
MCB			- (2), + (4)	1
MCM1	-	120	- (1), + (2)	0.8
MIG1	-, +, +, +	30, 270, 300, 330	+ (1), - (2)	1
PHO2			+ (2)	1
PDR1/PDR3				
PHO4	-	300	- (1), + (4)	0.98
REB1				
ROX1	+	0	- (1), - (3)	0.62
RAP1	-	240	- (2), + (4)	1
RLM1			- (1), + (5)	0.98
CAR1 Repressor			- (1)	0.98
SMP1			- (5)	0.88
SWI5			- (2), + (4)	1
STE12	+	330	- (1), + (4)	0.58
TBP	-	390		
UASPHR	+	330	- (1), + (2), - (4)	1
XBP1	+, -	60, 240	+ (1), - (2)	0.91

Table 5: *Functional group enrichment of genes under the models with $K = 3$ and $K = 5$. Column 1 denotes the functional pathways discovered by Spellman et al. (1998). Column 2 denotes the proportion of genes involved in a function, that were predicted by the HMRM as being in the same state at a time point corresponding to the phase they are known to be active in the cell-cycle (Column 3). The correspondence between phase and time point is determined from Spellman et al. (1998). Column 4 gives the gene names and Column 5 the state label in the HMRM.*

Function	Gene proportion in state		Cell cycle phase		Gene names		State	
	$K = 3$	$K = 5$	$K = 3$	$K = 5$	$K = 3$	$K = 5$	$K = 3$	$K = 5$
DNA repair	6/8	6/8	early G1	early G1	DHS1 DUN1 MSH2 MSH6 PMS1 RAD53	DHS1 DUN1 MSH2 MSH6 PMS1 RAD53	3	2
DNA synthesis	8/12	7/12	early G1	mid G1	DPB2 POL1 POL2 POL32 RFA1 RFA3 RFC5 TOP3	CDC2 CTF4 DPB2 POL2 POL32 RFA3 RFC5	1	2
DNA chromatin	4/6	4/6	S	S	HHF2 HHO1 HHT2 HTA2	HHF2 HHO1 HHT2 HTA2	2	5
Budding: Glycosylation	5/5	4/5	end G1	mid G1	OCH1 PMT1 PMT3 PSA1 QRI1	OCH1 PMT1 PMT3 PSA1	3	2
Fatty acids	5/5	4/5	M/G1	G1	ELO1 FAA1 FAA3 FAA4 FAS1	LPP1 PSD1 SUR1 SUR2	2	2
Mitosis	4/4	3/4	M	M	APC1 CDC5 CDC20 TEM1	APC1 CDC20 TEM1	1	1
Mating	7/7	5/7	M/G1	M/G1	AFR1 AGA2 ASH1 FUS1 GPA1 SAG1 SST2	AFR1 FUS1 GPA1 SAG1 SST2	2	2
Cell cycle control	4/5	–	early G1	–	CLN2 HSL1 PCL1 PCL2	–	1	
Nutrition	–	6/9	–	early M	–	DIP5 FET3 PHO3 RGT2 VCX1 ZRT1		2

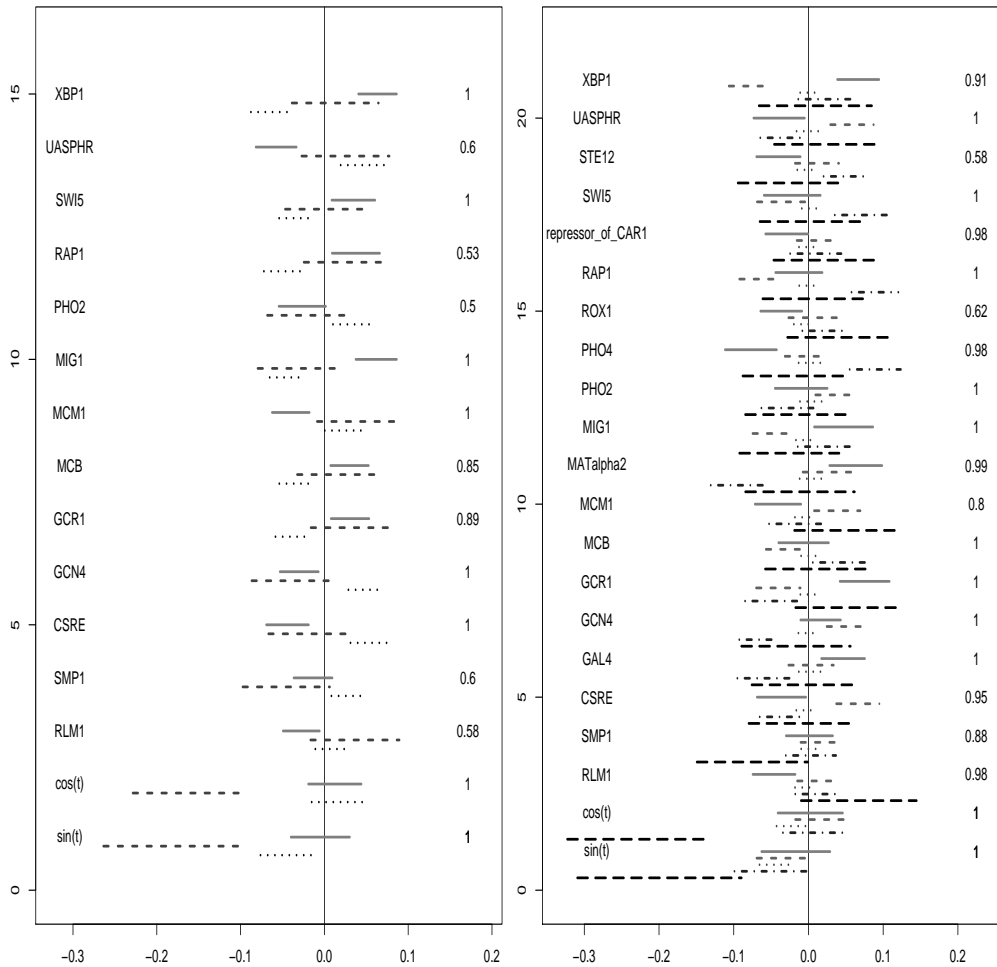


Figure 1: 95% posterior intervals for the motif and time coefficients for each state in the HMRM with (a) $K = 3$ and (b) $K = 5$. On each panel, the left column gives the names of the transcription factor motifs, and the number on the right gives the posterior marginal frequency of selection of each covariate in the model. The horizontal lines give the posterior confidence intervals, with each state being denoted by a different line type (The solid, dashed and dotted lines in panel 1 give the intervals for the regression coefficient in each of the three states; while the five line types in panel 2 represent the intervals for $K = 5$).

Supplementary Materials for “A temporal hidden Markov
regression model for the analysis of gene regulatory
networks”

Mayetri Gupta, Pingping Qu, and Joseph G. Ibrahim*

February 20, 2007

*Department of Biostatistics, University of North Carolina at Chapel Hill, NC 27599, U.S.A.;
gupta@bios.unc.edu

APPENDIX A: Interpretation of R_k as a residual term

The posterior estimate of β_k , $\tilde{\beta}_k$ can be interpreted as a weighted mean of the maximum likelihood estimate of the regression coefficient for state k , and the prior coefficient β_0 . For notational simplicity, considering the number of variables $|\mathbf{u}|$ fixed at present, we drop the indicator for the selected variables \mathbf{u} from the notation and write,

$$\tilde{\beta}_k = \tilde{\Sigma}_{\beta k} \frac{V_k}{1 + \tau_0} \hat{\beta}_k + \tilde{\Sigma}_{\beta k} \frac{V}{g} \beta_0,$$

where $V_k = \mathbf{X}_{(k)}^T \mathbf{X}_{(k)}$, $V = \mathbf{X}^T \mathbf{X}$, $\hat{\beta}_k = (\mathbf{X}_{(k)}^T \mathbf{X}_{(k)})^{-1} \mathbf{X}_{(k)}^T \mathbf{Y}_{(k)}$. It can be seen that $\tilde{\Sigma}_{\beta k} = \frac{V_k}{1 + \tau_0} + \frac{V}{g}$. Now, let $R_k = \frac{1}{1 + \tau_0} \mathbf{Y}_{(k)}^T \mathbf{Y}_{(k)} + \frac{1}{g} \beta_0^T \mathbf{X}^T \mathbf{X} \beta_0 - \tilde{\beta}_k^T \tilde{\Sigma}_{\beta k}^{-1} \tilde{\beta}_k$. Then, using the definition of R_k from Eqn (3.4), we have the following result:

Proposition 1.1.

$$\begin{aligned} R_k &= \frac{1}{1 + \tau_0} \left[\mathbf{Y}_{(k)}^T \mathbf{Y}_{(k)} - \hat{\beta}_k^T V_k \hat{\beta}_k \right] - (\hat{\beta}_k - \beta_0)^T M_k (\hat{\beta}_k - \beta_0) \\ &= \frac{1}{1 + \tau_0} \mathbf{Y}_{(k)}^T (I - H_k) \mathbf{Y}_{(k)} - (\hat{\beta}_k - \beta_0)^T M_k (\hat{\beta}_k - \beta_0), \end{aligned}$$

$$\text{where } M_k = \frac{V \tilde{\Sigma}_{\beta k} V_k}{g(1 + \tau_0)} = \frac{\mathbf{X}_{(k)}^T \mathbf{X}_{(k)}}{1 + \tau_0} \left[\frac{\mathbf{X}_{(k)}^T \mathbf{X}_{(k)}}{1 + \tau_0} + \frac{\mathbf{X}^T \mathbf{X}}{g} \right]^{-1} \frac{\mathbf{X}^T \mathbf{X}}{g},$$

and H_k is the projection matrix for regression within state k , i.e. $H_k = \mathbf{X}_{(k)} (\mathbf{X}_{(k)}^T \mathbf{X}_{(k)})^{-1} \mathbf{X}_{(k)}^T$.

Proposition 1.1 implies that under the HMRM, R_k can be treated as a state-specific residual term, similar to a residual term in the single state model.

Proof of Proposition 1.1.

By definition,

$$R_k = \frac{1}{1 + \tau_0} \mathbf{Y}_{(k)}^T \mathbf{Y}_{(k)} + \frac{1}{g} \beta_0^T \mathbf{X}^T \mathbf{X} \beta_0 - \tilde{\beta}_k^T \tilde{\Sigma}_{\beta k}^{-1} \tilde{\beta}_k.$$

Using the definition of $\tilde{\beta}_k = \tilde{\Sigma}_{\beta k} \frac{V_k}{1 + \tau_0} \hat{\beta}_k + \tilde{\Sigma}_{\beta k} \frac{V}{g} \beta_0$, where $V_k = \mathbf{X}_{(k)}^T \mathbf{X}_{(k)}$, $V =$

$\mathbf{X}^T \mathbf{X}$, $\hat{\boldsymbol{\beta}}_k = (\mathbf{X}_{(k)}^T \mathbf{X}_{(k)})^{-1} \mathbf{X}_{(k)}^T \mathbf{Y}_{(k)}$, we have

$$\begin{aligned} R_k &= \frac{1}{1+\tau_0} \mathbf{Y}_{(k)}^T \mathbf{Y}_{(k)} + \frac{1}{g} \boldsymbol{\beta}_0^T V \boldsymbol{\beta}_0 - \left[\frac{\boldsymbol{\beta}_0^T V \tilde{\Sigma}_{\beta k} V \boldsymbol{\beta}_0}{g^2} + 2 \frac{\boldsymbol{\beta}_0^T V \tilde{\Sigma}_{\beta k} V_k \hat{\boldsymbol{\beta}}_k}{g(1+\tau_0)} + \frac{\hat{\boldsymbol{\beta}}_k^T V_k \tilde{\Sigma}_{\beta k} V_k \hat{\boldsymbol{\beta}}_k}{(1+\tau_0)^2} \right] \\ &= \frac{1}{1+\tau_0} \mathbf{Y}_{(k)}^T \mathbf{Y}_{(k)} + \boldsymbol{\beta}_0^T M_k \boldsymbol{\beta}_0 - 2 \boldsymbol{\beta}_0^T M_k \hat{\boldsymbol{\beta}}_k - \frac{\hat{\boldsymbol{\beta}}_k^T V_k \tilde{\Sigma}_{\beta k} V_k \hat{\boldsymbol{\beta}}_k}{(1+\tau_0)^2}, \end{aligned}$$

with $M_k = \frac{V \tilde{\Sigma}_{\beta k} V_k}{g(1+\tau_0)}$, since

$$\frac{V}{g} - \frac{V \tilde{\Sigma}_{\beta k} V}{g^2} = \frac{V \tilde{\Sigma}_{\beta k}}{g} \left[\tilde{\Sigma}_{\beta k}^{-1} - \frac{V}{g} \right] = \frac{V \tilde{\Sigma}_{\beta k}}{g} \frac{V_k}{1+\tau_0} = M_k.$$

Then,

$$\begin{aligned} R_k &= \frac{1}{1+\tau_0} \mathbf{Y}_{(k)}^T \mathbf{Y}_{(k)} - (\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_0)^T M_k (\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_0) - \hat{\boldsymbol{\beta}}_k^T \left[\frac{V_k \tilde{\Sigma}_{\beta k} V_k}{(1+\tau_0)^2} + M_k \right] \hat{\boldsymbol{\beta}}_k \\ &= \frac{1}{1+\tau_0} [\mathbf{Y}_{(k)}^T \mathbf{Y}_{(k)} - \hat{\boldsymbol{\beta}}_k^T V_k \hat{\boldsymbol{\beta}}_k], \end{aligned}$$

$$\text{since } \frac{V_k \tilde{\Sigma}_{\beta k} V_k}{(1+\tau_0)^2} + \frac{V \tilde{\Sigma}_{\beta k} V_k}{g(1+\tau_0)} = \left(\frac{V_k}{1+\tau_0} + \frac{V}{g} \right) \frac{\tilde{\Sigma}_{\beta k} V_k}{1+\tau_0} = \tilde{\Sigma}_{\beta k}^{-1} \frac{\tilde{\Sigma}_{\beta k} V_k}{1+\tau_0} = \frac{V_k}{1+\tau_0}.$$

Hence the result.

APPENDIX B: Frequentist evaluation of R_k

Proposition 1.2. *For any fixed set of covariates indexed by u ,*

$$\lim_{g \rightarrow \infty} R_k = \frac{1}{1+\tau_0} \mathbf{Y}_{(k)}^T (I - H_k) \mathbf{Y}_{(k)}.$$

Also, if $\beta_0 = 0$,

$$\lim_{g \rightarrow 0} R_k = \frac{1}{1+\tau_0} \mathbf{Y}_{(k)}^T (I - 2H_k) \mathbf{Y}_{(k)}.$$

Proof of Proposition 1.2.

By Proposition 1.1, $R_k = \frac{1}{1+\tau_0} \mathbf{Y}_{(k)}^T (I - H_k) \mathbf{Y}_{(k)} - (\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_0)^T M_k (\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_0)$. The first term is independent of g , and is equivalent to the ordinary least squares residual sum of

squares multiplied by a factor of $\frac{1}{1+\tau_0}$. In the second term, the only part dependent on g is the matrix M_k . Now assuming \mathbf{u} is fixed, and dropping \mathbf{u} from the superscript,

$$M_k = \frac{V_k}{1+\tau_0} \left[gV^{-1} \left(\frac{V_k}{1+\tau_0} + \frac{V}{g} \right) \right]^{-1} = \frac{V_k}{1+\tau_0} \left[g \frac{V^{-1}V_k}{1+\tau_0} + I \right]^{-1} \quad (1.1)$$

Taking limits of (1.1) as $g \rightarrow \infty$, we get $\lim_{g \rightarrow \infty} M_k = \frac{V_k}{1+\tau_0} \times 0 = 0$, so $\lim_{g \rightarrow \infty} R_k = \frac{1}{1+\tau_0} \mathbf{Y}_{(k)}^T (I - H_k) \mathbf{Y}_{(k)}$. Also, $\lim_{g \rightarrow 0} M_k = \frac{V_k}{1+\tau_0}$. When $\beta_0 = 0$, it can be seen that $\lim_{g \rightarrow 0} R_k = \frac{1}{1+\tau_0} \mathbf{Y}_{(k)}^T (I - 2H_k) \mathbf{Y}_{(k)}$. In the case that the data are normalized, taking $\beta_0 = 0$ is a reasonable assumption in practice.

APPENDIX C: Evolutionary Monte Carlo (EMC) algorithm

To conduct EMC, we first prescribe a set of temperatures, $t_1 > t_2 > \dots > t_M = 1$, one for each member in the population. Then, using the marginalized conditional probability for a configuration of covariates \mathbf{u} (Eqn. 3.3), we set

$$\pi_i(u_i) \propto \exp[-\log \mathcal{H}(\mathbf{u} | \mathbf{Z}, \mathbf{X}, \boldsymbol{\mu}, g) / t_i],$$

and let $\pi(\mathbf{u}) \propto \prod_{i=1}^M \pi_i(u_i)$. The population $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_M)$ is updated iteratively using two types of moves: *mutation* and *crossover*. In the mutation operation, a unit \mathbf{u}_k is randomly selected from the current population and mutated to a new vector \mathbf{v}_k by changing the values of some of its bits chosen at random. The new member \mathbf{v}_k is accepted into the population with probability $\min(1, r_m)$, where $r_m = \pi_k(\mathbf{v}_k) / \pi_k(\mathbf{u}_k)$. In the crossover, two individuals, \mathbf{u}_j and \mathbf{u}_k , say, are chosen at random from the population. Next, a crossover point x is chosen randomly over the positions 1 to D , and two new units \mathbf{v}_j and \mathbf{v}_k are formed by switching between the two individuals the segments on the right side of the crossover point. The two ‘‘children’’ are accepted into the population to replace their parents \mathbf{u}_j and \mathbf{u}_k with probability $\min(1, r_c)$, where $r_c = \frac{\pi_j(\mathbf{v}_j)\pi_k(\mathbf{v}_k)}{\pi_j(\mathbf{u}_j)\pi_k(\mathbf{u}_k)}$. If rejected, the two parents are kept unchanged. After convergence, the samples of \mathbf{u}_M , corresponding to temperature $t_M = 1$, follow the target distribution (Eqn 3.3).

APPENDIX D: State updating in the HMRM

To sample states of the HMRM, we use the identity

$$P(Z_{i1}, \dots, Z_{iT} | \mathbf{Y}, \boldsymbol{\theta}) = P(Z_{iT} | \mathbf{Y}, \boldsymbol{\theta}) P(Z_{i,T-1} | \mathbf{Y}, \boldsymbol{\theta}, Z_{iT}) \cdots P(Z_{i1} | \mathbf{Y}, \boldsymbol{\theta}, Z_{iT}, \dots, Z_{i2}).$$

First, the *forward-sum* expression $f_t(k)$ representing the partial likelihood of unit i upto time j ($1 \leq j \leq T$), is evaluated recursively as follows:

$$\begin{aligned} f_t(k) &= P(Y_{i1}, \dots, Y_{ij}, Z_{ij} = k | \mathbf{X}, \mathbf{u}, \boldsymbol{\theta}) \\ &= P(Y_{ij} | Z_{ij} = k, \mathbf{X}, \mathbf{u}, \boldsymbol{\theta}) \sum_{l=1}^K f_{j-1}(l) \tau_{lk}, \end{aligned} \quad (1.2)$$

where $Y_{ij} | Z_{ij} = k, \mathbf{X}, \mathbf{u}, \boldsymbol{\theta} \sim \mathcal{N}(\mu_{kj} + \beta_k^{(u)}(\mathbf{X}_i, \phi(j)), (1 + \tau_0)\sigma_k^2)$ from Eqn. (2.3). The initializing condition for the recursion is given by $f_1(k) = P(Y_{i1} | Z_{i1} = k, \mathbf{X}, \mathbf{u}, \boldsymbol{\theta})$.

Next, the sampling step for unit i can be carried out as follows:

$$Z_{iT} \sim P(Z_{iT} = k | Y_{i1}, \dots, Y_{iT}, \mathbf{X}, \mathbf{u}, \boldsymbol{\theta}) = \frac{f_T(k)}{\sum_{l=1}^K f_T(l)} \quad (1.3)$$

$$Z_{i,j-1} \sim P(Z_{i,j-1} = k | Z_{ij}, Y_{i1}, \dots, Y_{ij}, \mathbf{X}, \mathbf{u}, \boldsymbol{\theta}) = \frac{f_{j-1}(k) \tau_{k,Z_{ij}}}{\sum_{l=1}^K f_{j-1}(l) \tau_{l,Z_{ij}}} \quad (1.4)$$

for $j = T-1, T-2, \dots, 1$. All the expressions required for Eqns. (1.4) have been calculated in the forward sum step (1.2).

APPENDIX E: Conditional posterior distributions

Parameter updating is through the following steps:

- (i) Update $\boldsymbol{\beta}_k$ from its posterior distribution. $\boldsymbol{\beta}_k^{(u)} | \mathbf{Y}, \mathbf{X}, \mathbf{u}, \mathbf{Z}, \boldsymbol{\mu}, \sigma_k^2 \sim \mathcal{N}(\tilde{\boldsymbol{\beta}}_k^{(u)}, \sigma_k^2 \tilde{\Sigma}_{\boldsymbol{\beta}_k}^{(u)})$, where $\tilde{\Sigma}_{\boldsymbol{\beta}_k}$ and $\tilde{\boldsymbol{\beta}}_k$ are defined in Eqn (3.5).

- (ii) Update $\boldsymbol{\mu}$. $\mu_{kj} | \mathbf{Y}, \mathbf{X}, \mathbf{u}, \mathbf{Z}, \boldsymbol{\beta}, \sigma_k^2 \sim \mathcal{N}(\tilde{\mu}_{kj}, \tilde{\Sigma}_{\mu k})$, where

$$\tilde{\Sigma}_{\mu k} = [v_{k0}^{-2} + n_k \sigma_k^{-2}]^{-1}, \quad \tilde{\mu}_{kj} = \tilde{\Sigma}_{\mu k} [v_{k0}^{-2} m_{k0} + N_{kj} \bar{y}_{kj} / \sigma_k^2].$$

Here $\bar{y}_{kj} = \frac{1}{N_{kj}} \sum_{i: Z_{ij}=k} [Y_{ij} - \beta_k^{(u)}(\mathbf{X}_i, \phi(j))]$, and $N_{kj} = \sum_{i=1}^N 1_{[Z_{ij}=k]}$.

(iii) Update σ_k^2 . For more efficient computation we *collapse* the posterior distribution by integrating out β , and updating σ_k^2 from

$$\frac{1}{\sigma_k^2} \Big| \mathbf{Y}, \mathbf{X}, \mathbf{u}, \mathbf{Z}, \boldsymbol{\mu} \sim \text{Gamma}\left(\frac{n_k + w_0}{2}, \frac{S_0 + R_k^{(u)}}{2}\right),$$

where $n_k = \sum_{i=1}^N \sum_{j=1}^T 1_{[Z_{ij}=k]}$, $\mathbf{Y}_{(k)}$ and $\mathbf{X}_{(k)}^{(u)}$ are as defined before, and the residual term $R_k^{(u)}$ is as defined in Eqn (3.4).

(iv) Update $\boldsymbol{\tau}$. For $k = 1, 2, \dots, K$,

$$(\boldsymbol{\tau}_{k1}, \dots, \boldsymbol{\tau}_{kK}) \Big| \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{u} \sim \text{Dirichlet}(\mathbf{t}_k + \boldsymbol{\omega}_k),$$

where $\mathbf{t}_k = (t_{k1}, \dots, t_{kK})$, and $t_{kl} = \sum_{i=1}^N \sum_{j=2}^T 1_{[Z_{i,j-1}=k, Z_{ij}=l]}$ (for $1 \leq k, l \leq K$).

(v) Update $\boldsymbol{\pi}$. $\boldsymbol{\pi} \Big| \mathbf{Y}, \mathbf{X}, \mathbf{u}, \mathbf{Z} \sim \text{Dirichlet}(\mathbf{m}_1 + \boldsymbol{\alpha}_0)$.

APPENDIX F: Consistency of covariate selection

The value of g in the prior distribution for β_k determines the sensitivity of variable selection. The larger the value of g , the less is the influence of the prior. From expression (3.3), it can be shown that the null model is favored over all other models irrespective of the true underlying model if g is large, making the model selection procedure inconsistent.

Proposition 1.3. *For a non-empty covariate set \mathbf{u} , $\lim_{g \rightarrow \infty} \mathcal{H}(\mathbf{u} | \mathbf{X}^{(u)}, \mathbf{Z}, \boldsymbol{\mu}, g) = 0$.*

Proof of Proposition 1.3.

For a fixed \mathbf{u} , ignoring all terms independent of g in Eqn (3.3), and dropping the superscript \mathbf{u} from the notation, we have

$$\lim_{g \rightarrow \infty} \mathcal{H}(\mathbf{u} | \mathbf{X}^{(u)}, \mathbf{Z}, \boldsymbol{\mu}, g) = C \times \prod_{k=1}^K \left[\frac{|V/g|^{\frac{1}{2}}}{|V_k/(1 + \tau_0) + V/g|^{\frac{1}{2}}} \times \text{IGamma}_{\frac{w_0}{2}, \frac{S_0}{2}}\left(\frac{n_k}{2}, \frac{R_k^{(u)}}{2}\right) \right], \quad (1.5)$$

where C is a constant independent of g . By proposition 1.2, the second term in (1.5) tends to a constant independent of g as $g \rightarrow \infty$. So, we have

$$\lim_{g \rightarrow \infty} \frac{|V/g|}{|V_k/(1 + \tau_0) + V/g|} = \lim_{g \rightarrow \infty} \frac{|V|}{|gV_k/(1 + \tau_0) + V|}$$

Hence for any $|\mathbf{u}| > 0$, $\lim_{g \rightarrow \infty} \mathcal{H}(\mathbf{u} | \mathbf{X}^{(u)}, \mathbf{Z}, \boldsymbol{\mu}, g) = 0$.

APPENDIX G: Model selection using the L-measure

We used the trace of the L-measure, computed in Eqn. (4.1), as a summary for model selection purposes- other summaries, such as the determinant, may also be used. Under the HMRM, we then have $L(\mathbf{y}, \nu) = \sum_{i=1}^N \text{tr} \text{Cov}(\mathbf{W}_i | \mathbf{y}) + \nu \sum_{i=1}^N (\boldsymbol{\lambda}_i - \mathbf{y}_i)^T (\boldsymbol{\lambda}_i - \mathbf{y}_i)$, where $\boldsymbol{\lambda}_i = E(\mathbf{W}_i | \mathbf{y})$, $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})$, ($i = 1, \dots, N$), and the expectations and covariances are computed under the posterior predictive distribution of $\mathbf{W}_i | \mathbf{y}$. Now let $\theta = (\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\tau}, \boldsymbol{\pi}, \mathbf{Z}, \mathbf{u})$ denote the generic set of parameters and latent variables. Since the posterior predictive distribution is analytically intractable, we instead use the MC estimate for $L(\mathbf{y}, \nu)$,

$$\hat{L}(\mathbf{y}, \nu) = \sum_{i=1}^N \sum_{j=1}^T \{E_{\theta | \mathbf{Y}}[E(Z_{ij}^2 | \theta)] - \hat{\lambda}_{ij}^2\} + \nu \sum_{i=1}^N \sum_{j=1}^T (\hat{\lambda}_{ij} - y_{ij})^2, \quad (1.6)$$

where the inner expectations, over $P(\mathbf{W}_i | \theta)$ have simple closed forms, and $\hat{\lambda}_{ij} = E_{\theta | \mathbf{Y}}[E(Z_{ij} | \theta)]$.

Using a posterior sample of size M , it can be shown that we can estimate $E_{\theta | \mathbf{Y}}[E(Z_{ij} | \theta)]$

by

$$\bar{\lambda}_{ij} = \frac{1}{M} \sum_{r=1}^M \hat{\lambda}_{ij}^{(r)},$$

where $\hat{\lambda}_{ij}^{(r)} = \mu_{Z_{ij}^{(r)}, j} + \mathbf{X}_i^T \boldsymbol{\beta}_{Z_{ij}^{(r)}}$, and estimate $E_{\theta | \mathbf{Y}}[E(Z_{ij}^2 | \theta)]$ by

$$(1 + \tau_0) \frac{1}{M} \sum_{r=1}^M \sum_{i=1}^N \sum_{j=1}^T \sigma_{Z_{ij}^{(r)}}^2 + \hat{V}_{ij},$$

where $\hat{V}_{ij} = \frac{1}{M} \sum_{r=1}^M \sum_{i=1}^N \sum_{j=1}^T (\hat{\lambda}_{ij}^{(r)} - \bar{\lambda}_{ij})^2$.

Supplementary figures and tables

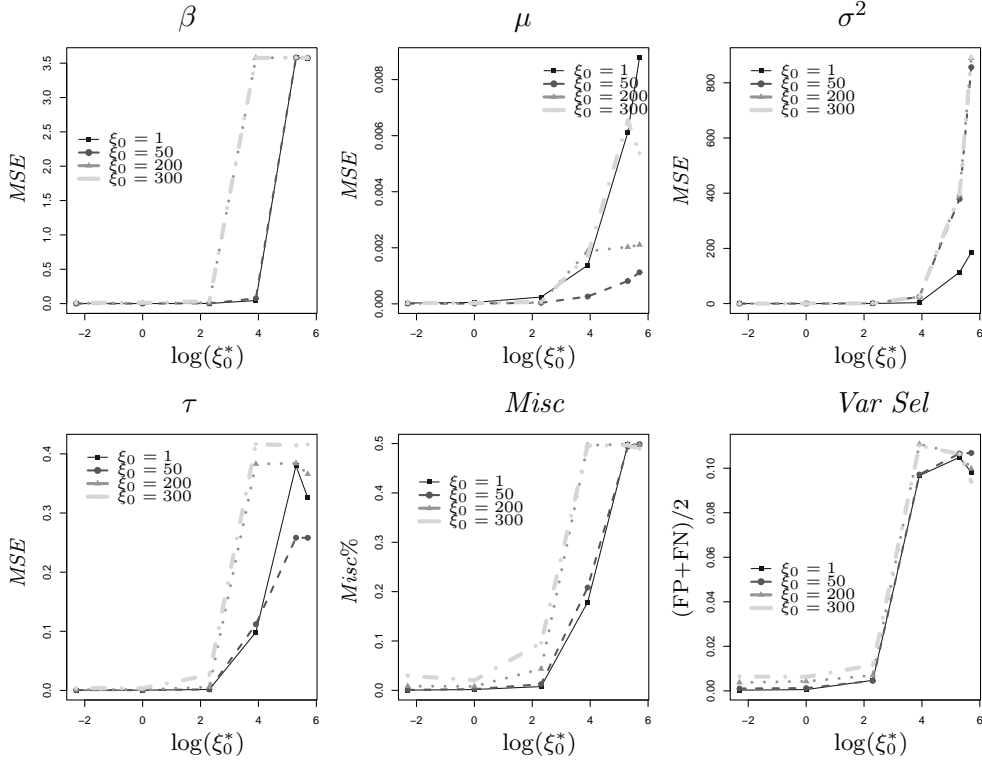


Figure 1: *Robustness of model fit varying settings and specification of ξ_0 , for the $K = 2$ data set. The first four of the six panels (from top left to bottom right) represent the MSEs of parameters β, μ, σ^2 and τ under different true values ξ_0^* (horizontal axes) and different values of ξ_0 set during the run of the algorithm (shaded curves, darker for higher values of ξ_0). The fifth panel represents the misclassification rate (“Misc%”) of the algorithm under similar settings, while the last panel represents the average error in variable selection rates, given by $\frac{FP+FN}{2}$, where FP represents false positives or proportion of extra variables selected, and FN represents false negatives, or proportion of true variables missed. Incidentally, none of the runs under any settings missed any of the true variables, i.e. $FN = 0$.*

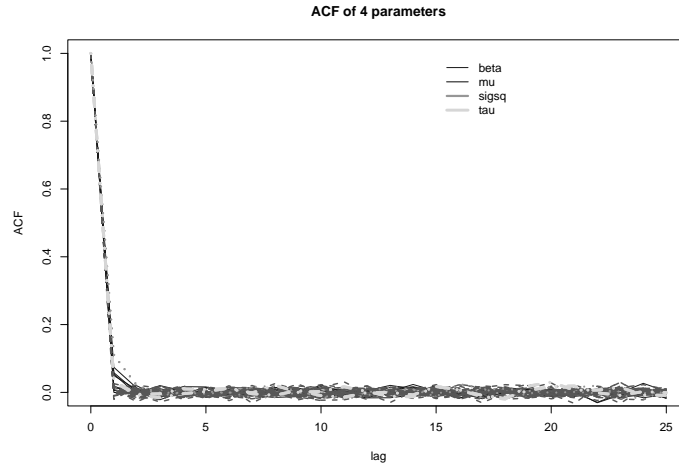


Figure 2: Autocorrelation functions for all parameters for the simulated data set with $\tau_0 = 1$. Each shaded line represents a different parameter, the darkest line representing the 2×4 components of β , and the lightest representing the 4 components of the transition probability matrix τ .

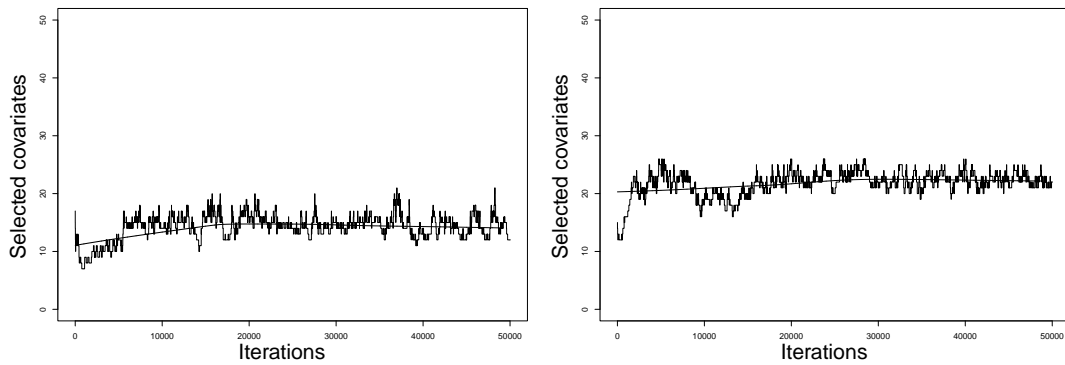


Figure 3: Trace plot for the number of selected variables $|\mathbf{u}|$ in the model over iterations of the sampler, with $K = 3$ and $K = 5$.

Table 1: *Confidence intervals for significant transcription factors by cluster for the HMRM with $K = 3$. Column 3 and 4 give the 95% posterior intervals for the regression coefficient; column 5 indicates whether the TF up-regulates (+) genes in the cluster, or down-regulates (-) them.*

cluster	TF	2.5%	97.5%	up/down
1	RLM1	-0.0496	-0.0055	-
1	CSRE	-0.0695	-0.0186	-
1	GCN4	-0.0533	-0.0068	-
1	GCR1	0.0077	0.0536	+
1	MCB	0.0071	0.0530	+
1	MCM1	-0.0626	-0.0177	-
1	MIG1	0.0372	0.0865	+
1	RAP1	0.0090	0.0663	+
1	SWI5	0.0085	0.0607	+
1	UASPHR	-0.0820	-0.0331	-
1	XBP1	0.0404	0.0865	+
3	SMP1	0.0084	0.0505	+
3	CSRE	0.0311	0.0756	+
3	GCN4	0.0287	0.0637	+
3	GCR1	-0.0584	-0.0188	-
3	MCB	-0.0538	-0.0113	-
3	MCM1	0.0005	0.0456	+
3	MIG1	-0.0653	-0.0250	-
3	PHO2	0.0103	0.0557	+
3	RAP1	-0.0722	-0.0259	-
3	SWI5	-0.0540	-0.0110	-
3	UASPHR	0.0194	0.0721	+
3	XBP1	-0.0877	-0.0443	-

Table 2: *Confidence intervals for significant transcription factors by cluster for the HMRM with $K = 5$.*

cluster	TF	2.50%	97.50%	up/down
1	CAR1 repressor	-0.0578	-0.0002	-
1	CSRE	-0.0690	-0.0031	-
1	GAL4	0.0170	0.0757	+
1	GCR1	0.0418	0.1086	+
1	MATalpha2	0.0280	0.0987	+
1	MCM1	-0.0718	-0.0097	-
1	MIG1	0.0075	0.0866	+
1	PHO4	-0.1118	-0.0423	-
1	RLM1	-0.0748	-0.0174	-
1	ROX1	-0.0644	-0.0084	-
1	STE12	-0.0698	-0.0105	-
1	UASPHR	-0.0737	-0.0049	-
1	XBP1	0.0384	0.0947	+
2	CSRE	0.0371	0.0950	+
2	GCN4	0.0243	0.0708	+
2	GCR1	-0.0701	-0.0105	-
2	MCB	-0.0571	-0.0027	-
2	MCM1	0.0074	0.0696	+
2	MIG1	-0.0749	-0.0230	-
2	PHO2	0.0092	0.0633	+
2	RAP1	-0.0926	-0.0396	-
2	SWI5	-0.0690	-0.0044	-
2	UASPHR	0.0291	0.0874	+
2	XBP1	-0.1067	-0.0577	-
4	CSRE	-0.0652	-0.0033	-
4	GAL4	-0.0952	-0.0236	-
4	GCN4	-0.0926	-0.0432	-
4	GCR1	-0.0848	-0.0097	-
4	MATalpha2	-0.1308	-0.0584	-
4	MCB	0.0058	0.0794	+
4	PHO4	0.0551	0.1238	+
4	RAP1	0.0576	0.1203	+
4	STE12	0.0202	0.0789	+
4	SWI5	0.0352	0.1077	+
4	UASPHR	-0.0645	-0.0049	-
5	SMP1	-0.1495	-0.0007	-

Table 3: *Confidence intervals for significant transcription factors found by stepwise regression.*

time point	TF	2.50%	97.50%
2	MIG1	-0.1309	-0.0236
7	GCN4	0.0084	0.0797
8	GCN4	0.0067	0.0746
9	RAP1	-0.0614	-0.0009
	XBP1	-0.0676	-0.0071
10	MIG1	0.0185	0.0821
	STE12	0.0113	0.0747
11	MIG1	0.0165	0.0751
	PHO4	-0.0615	-0.0028
12	GCR1	-0.0747	-0.0012
	MIG1	0.0031	0.0761
	UASPHR	0.0051	0.0783
14	TBP	-0.0659	-0.0027