A Bayesian autoregressive three-state hidden Markov model for identifying switching monotonic regimes in Microarray time course data

Alessio Farcomeni, Sapienza University of Rome, Italy
Serena Arima, Sapienza University of Rome, Italy

©2012 De Gruyter. All rights reserved.
A Bayesian autoregressive three-state hidden Markov model for identifying switching monotonic regimes in Microarray time course data

Alessio Farcomeni and Serena Arima

Abstract

When modeling time course microarray data special interest may reside in identifying time frames in which gene expression levels follow a monotonic (increasing or decreasing) trend. A trajectory may change its regime because of the reaction to treatment or of a natural developmental phase, as in our motivating example about identification of genes involved in embryo development of mice with the 22q11 deletion. To this aim we propose a new flexible Bayesian autoregressive hidden Markov model based on three latent states, corresponding to stationarity, to an increasing and to a decreasing trend. In order to select a list of genes, we propose decision criteria based on the posterior distribution of the parameters of interest, taking into account the uncertainty in parameter estimates. We also compare the proposed model with two simpler models based on constrained formulations of the probability transition matrix.

KEYWORDS: HMM, Bayesian statistics, MCMC, Microarray time course data

Author Notes: We would like to thank two anonymous referees whose comments strongly improved our article.
1 Introduction

The advent of microarray technology has enabled biomedical researchers to measure the expression levels of thousands of genes and a very rich statistical literature on how to analyze these data is now available. We do not attempt to review this literature, but only point the reader to the excellent books by Parmigiani et al. (2003); Speed (2003); Do et al. (2006) and to the many references therein.

We here focus on time course experiments, which involve repeated measures of the expression levels of the same genes at different time occasions. These complex experiments allow to monitor the evolution of gene expression levels over time, for instance in response to a treatment or because of natural aging processes of the cell or of the subject. There is comparatively less work on analysis of microarray time course data with respect to cross-sectional data. A general goal with time course microarrays is to cluster genes with respect to similar temporal patterns of gene expression. This is the spirit of many important works, like Eisen et al. (1998), Ramoni et al. (2002), Schliep et al. (2003), Hong and Li (2006). Ma et al. (2006) and Ma and Zhong (2008) explicitly assume an underlying function for gene expression over time. While most studies are confined to a single biological condition, some experiments may also be devoted to the identification of differentially expressed genes over multiple biological conditions; see e.g. Yuan and Kendziorski (2006) and Zhou et al. (2010).

In this work we have a rather different goal: we focus on studies with a single biological condition, and we are interested in identifying genes with an increasing or a decreasing trend at certain occasions. The underlying hypothesis is that most of the genes will show a stationary pattern of expression over time, and that local variations of expression levels correspond to the few genes that increase or decrease their activity in correspondence of a response to treatment or of the moving forward in a developmental phase of the cell or of the entire subject under study. We stress that we do not need that genes follow a monotonic trend for the entire observation period, but only in at least one occasion with respect to the preceding one.

We are motivated by an original study on genetic pathways involved in embryo development in mouse models for the 22q11 deletion syndrome. Data were collected and first analyzed in Amati et al. (2007). There are 53 genes involved in the experiment, which have been sampled at different developmental stages of the embryo, from 4.5 days post coitum (hereafter dpc
for brevity) to 14.5 dpc. All DNA samples have been hybridized against a common reference, represented by the DNA samples collected at full development of the embryo. Consequently, the recorded expression levels represent fold changes against a completed development. The objective of this analysis is to identify which genes increase or decrease their activity with respect to the previous time occasion.

In order to identify such switching monotonic regimes, we develop a Bayesian autoregressive hidden Markov model based on a three state latent Markov process. The three latent states correspond to no change, increased and decreased expression level with respect to the previous time occasion. We thus can identify local trends by using the posterior probability of being in one of the two latent states corresponding to an increasing or a decreasing trend. For a general review on the HMM refer to Zucchini and MacDonald (2009). We derive inference through an augmentation scheme, along the lines of MCMC algorithms derived for finite mixture models (Diebolt and Robert, 1994), where we compute the full conditionals for the latent indicators through efficient recursions (Baum et al., 1970; Chib, 1996).

The idea of using HMM to analyze time course microarray data is not new and has already been used for instance by Schliep et al. (2003), Yuan and Kendziorski (2006) and Zeng and Garcia-Frias (2006). The main difference with existing works is that while other works used HMM in order to remove bias due to dependence intrinsic in the repeated measurement over time (i.e., unobserved heterogeneity), in this work the latent variable is the target of the analysis. This allows us to identify local patterns, that is, when the monotonic regime of expression levels switch, rather than targeting clusters of global patterns. Furthermore, in the most general formulation of our model, we allow for a gene-specific trend. Consequently, genes can have a non-stationary pattern in a time interval, but with a completely gene-specific trend which is modeled non-parametrically through an autoregressive structure. Our model can be easily applied to single and dual channel technologies, to measurements on a single gene arising from Polymerase Chain Reaction (i.e., to time series rather than panel data); and we may or may not have replicates.

In order to identify genes we propose a criterion based on Maximum A Posteriori (MAP) selection, which minimizes the 0-1 loss; and another based on control of an estimate of the False Discovery Rate (FDR, Benjamini and Hochberg (1995)), as recommended by many authors in the context of time-course gene expression analysis; see for instance Storey et al. (2005),

The paper is organized as follows: in Section 2 we illustrate the full model. In Section 3 we describe how to estimate model parameters and to derive inference based on the estimates. In Section 4 we discuss constrained versions of the full model. In Section 5 the proposed model is applied to our motivating example and in Section 6 we conduct a small simulation study. We conclude with a brief discussion in Section 7.

2 The full model

Let $X_{igt}$ denote the (log) observed gene expression level for the $i$-th subject, the $g$-th gene and time $t$, with $t = 1, \ldots, T$, $i = 1, \ldots, n$ and $g = 1, \ldots, G$. We assume to have at least two time occasions, that is, $T > 1$. When $n = 1$ we have no replicates, and the study is therefore restricted to a single sample for each gene.

For notational simplicity we describe the full model assuming that the number of time points and replicates is equal across genes. Generalization to non-informative drop-out is straightforward.

We specify an autoregressive model of order 1 as follows:

$$X_{igt} | X_{igt-1}, \ldots, X_{igt-1} = \delta_{gt} X_{igt-1} + \varepsilon_{igt}$$

for $t = 2, \ldots, T$, $i = 1, \ldots, n$, and $g = 1, \ldots, G$; where $\varepsilon_{igt} \sim N(0, \sigma^2)$ denotes a subject and time-specific shock, and the shocks are independent; $\delta_{gt}$ is a discrete random variable to be specified better below.

The specified model relies on a Markovian assumption that the expression level at time $t$ depends only on the expression level at time $t - 1$. There also is an unobserved process $\delta_{gt}$, which is independent of the observed expression level, and depends only on time and gene identity, determines whether there is no biologically significant increase or decrease of the expression level, or if there is a local trend and in what direction. More formally, the autoregressive parameter $\delta_{gt}$ is not time-fixed, but is allowed to (smoothly) change as a function of $t$ as we now outline. We let $\delta_{gt}$ be a random effect arising from a discrete distribution with three support points: $(\xi_{g1}, \xi_{g2}, \xi_{g3})$, where $\xi_{g1} \leq -1$, $\xi_{g3} \geq 1$ and $|\xi_{g2}| < 1$. Each gene can then have a time specific slope, which can either correspond to a stationary situation (second latent state), or to a non-stationary decreasing (first latent state) or increasing (third latent state) trend.
Note that we have a subject-specific time series for each gene, that is, each gene has three different possible slopes from which to choose at each time point. Subjects (i.e., replicates) only share the same gene and time specific local slope. We make an homoscedasticity assumption that the random shocks are constant over time and latent state. This assumption is reasonable as the random shocks mostly collect variability due to measurement error, which is not expected to change over time and gene after normalization and standardization of the expression levels. Nevertheless, generalization to time and/or gene-specific dispersion parameters is straightforward.

We let $\delta_{gt}$ smoothly vary over time by specifying a first order latent Markov chain as follows:

$$
\begin{align*}
\text{Pr}(\delta_{g1} = \xi_{gc}) &= \lambda_c & c = 1, \ldots, 3 \\
\text{Pr}(\delta_{gt} = \xi_{gd} | \delta_{g,t-1} = \xi_{gc}) &= \pi_{cd} & c, d = 1, \ldots, 3
\end{align*}
$$

where $\sum_c \lambda_c = 1$ collects the initial probabilities and $\pi_{cd}$ is the probability of a transition for state $\xi_{gc}$ to state $\xi_{gd}$. These latter probabilities are collected in an hidden transition matrix which we denote with $\Pi$. The underlying assumption is that the unobserved process which determines whether the expressions levels for a gene are stationary or have a trend depends only on its more recent state. This assumption is seldom found to be restrictive in the HMM literature. We finally note that the HMM we specify is slightly more complex than the basic HMM, since we relax the assumption of local independence: conditionally on the latent process, the observations are not independent, but the current observation is allowed to depend on the previous according to an autoregressive model of order 1.

There are three kinds of restrictions we impose on the random effects in order to obtain a parsimonious model. First of all, we assume the latent slopes are time-constant. This assumption translates to time-constant increases or decreases of expression levels when there is a regime switching, which is deemed to be reasonable in many biological applications. Nevertheless, this assumption can be easily relaxed in our Bayesian framework by letting $\delta_{it}$ be either $\xi_{gt1}$, $\xi_{gt2}$ or $\xi_{gt3}$; leading to a non-stationary $\delta_{gt}$. The resulting model can be compared with the model with time-constant latent slopes with the BIC, as suggested below. Note that the increase in number of parameters when relaxing this assumption may be substantial, unless $T$ is very small. A second restriction we impose is that of homogeneity of the hidden transition matrix. This assumption corresponds to a situation in which probabilities of switching regimes do not depend on time, that is, can
be assumed to be the same at the beginning, throughout, and at the end of the observation period. We note that, as argued below for our motivating example, the experiment can often be designed in order to approximately satisfy this assumption (in many cases this will correspond to measuring at equally spaced time points, in other cases it will correspond to carefully chosen unequally spaced time points). Also this assumption can be easily relaxed by replacing $\pi_{cd}$ with $\pi_{tcd}$ in our formulas above. Once again the resulting model can be compared with the one assuming homogeneous transition probabilities. Finally, we assume a Markovian structure for $\delta_{gt}$. The Markovian structure is imposed in the HMM literature in order to conveniently reduce the number of parameters. Relaxing this assumption may be much less straightforward. The likelihood may become very cumbersome to deal with if we assume a higher-order dependence, still under a Markovian structure; and assuming a general dependence for the sequence $\delta_{gt}$ would make it extremely hard to compute.

2.1 Prior distributions

Bayesian inference is carried out by completing the model proposed in the previous section with prior distributions on the corresponding unknown parameters. Of course, when prior information is available it shall be summarized in the prior distributions. In absence of prior information, we use the following independent default priors: $\lambda \sim Dir(1,1,1)$ and $\Pi_c \sim Dir(1,1,1)$, for $c = 1, 2, 3$; where $Dir(1,1,1)$ denotes a three dimensional Dirichlet distribution with parameters all equal to 1 (i.e., a Bayes-Laplace distribution), and $\Pi_c$ denotes the $c$-th row of $\Pi$. The Bayes-Laplace distribution has been recently discussed as a universal default prior for multinomial parameters by Tuyl et al. (2009).

We use truncated log-normal distributions for $\xi_{i1}$ and $\xi_{i3}$, i.e.,

$$\log(-\xi_{g1}) \sim N(0, \tau_{i1}^2)I(0, \infty),$$

and

$$\log(\xi_{g3}) \sim N(0, \tau_{i3}^2)I(0, \infty),$$

and a truncated normal for $\xi_{i2}$, i.e.,

$$\xi_{g2} \sim N(0, \tau_{i2}^2)I(-1, 1).$$
Finally for the dispersion parameter \( \sigma \) we specify an inverse Gamma distribution with location and rate parameters equal to 0.01. In our implementation we also fix \( \tau_1 = \tau_2 = \tau_3 = 1 \).

## 3 Bayesian inference

Model assumptions allow us to write down the likelihood as:

\[
L(\theta) = \prod_{i=1}^{n} \prod_{g=1}^{G} \sum_{\delta_g} p(\delta_g) \prod_{t=1}^{T} p(X_{igt}|\delta_{gt}, X_{ig,t-1}, \sigma^2),
\]

where the sum is extended to all possible configurations of the vector \( \delta_g \). This telescopic sum is infeasible for even moderate values of \( T \), and corresponds to a numerical integration with respect to the counting measure. In order to overcome the problems related to the use of the joint distribution of each vector \( \delta_g \), it is convenient to augment the problem by introducing latent indicators \( w_{gtc} \), where \( w_{gtc} = 1 \) indicates that the \( g \)-th gene is in latent state \( c \) at the \( t \)-th occasion. The augmented likelihood can be then written as:

\[
L^*(\theta) = \prod_{i=1}^{n} \prod_{g=1}^{G} \prod_{c=1}^{3} \prod_{d=1}^{3} \prod_{t=1}^{T} \lambda_{c}^{w_{gtc}} \prod_{c=1}^{3} \prod_{d=1}^{3} \prod_{t=2}^{T} \pi_{c,d}^{w_{gtc}} \prod_{c=1}^{3} \prod_{t=1}^{T} p(X_{igt}|\delta_{gt} = \xi_{gc}, X_{ig,t-1}, \sigma^2)^{w_{gtc}},
\]

where

\[
p(X_{igt}|\delta_{gt} = \xi_{gc}, X_{ig,t-1}, \sigma^2) = \phi \left( \frac{X_{igt} - \xi_{gc} - X_{ig,t-1}}{\sigma} \right),
\]

and \( \phi(\cdot) \) denotes the density of a standard normal. It can be easily seen that different assumptions on the manifest distribution (5) lead in most cases only to minor adjustments to the augmented likelihood (4), and consequently to the MCMC strategy outlined in the next section.

### 3.1 Model estimation

Inference in the Bayesian framework is obtained through the posterior distribution, which is proportional to the prior multiplied by the likelihood. The
posterior distribution for our model, as in most cases, cannot be derived analytically and we must approximate it through a Markov Chain Monte Carlo (MCMC) algorithm specifically designed for working with the augmented likelihood.

The MCMC strategy is based on repeatedly sampling from the distribution of each parameter conditional on the observed data and the current value of all the other parameters. In order to efficiently compute the full conditionals we set up a forward-backward computation which is adapted from the hidden Markov literature (Zucchini and MacDonald, 2009). Furthermore, note that the latent indicators $w$ are new parameters which shall be efficiently sampled as described below.

Our final MCMC strategy consists in alternating sampling steps from the full conditionals and forward-backward computations in order to efficiently compute some of these full conditionals. The general step, which is summarized in Algorithm 1, is then repeated for a large number of times.

A burn-in consisting of the first $B$ iterations is discarded, and the remaining are used to approximate the posterior distribution in the usual way.

We now detail and explain each step of Algorithm 1.

First of all, we update the latent slopes. A simple Metropolis step with log-normal (for $\xi_1$ and $\xi_3$) and normal (for $\xi_2$) proposals is used. Due to the constraints on the latent slopes, there are no label switching problems.

Secondly, we use a forward-backward strategy to update the latent indicators, which we adapt from Chib (1996), see also Scott (2002). This strategy allows us sample $w_{gtc}$ from its joint full conditional distribution, dramatically speeding up convergence with respect to simple sampling from its univariate full conditional. In order to do so, we first use a forward recursion in order to compute the needed probabilities. Note that through this forward recursion we also can evaluate (3) as

$$\prod_{g=1}^{G} \sum_{c=1}^{3} \alpha_{gTc},$$

see for instance Zucchini and MacDonald (2009). Then, we sample the latent indicators backward. Refer to Chib (1996) for a detailed derivation of the rationale behind this forward-backward strategy.

We then proceed by updating the initial and transition probabilities. Dirichlet priors are conjugate with respect to the multinomial distribution,
Algorithm 1 General iteration of the MCMC strategy for fitting the full model

Update the latent slopes through a Metropolis step, where the full conditionals can be expressed as

\[
p(\xi_{gc}|X, w, \sigma^2, \Pi, \lambda) \propto p(\xi_{gc}) \prod_i \prod_t \prod_g p(X_{igt}|\delta_{gt} = \xi_{gc}, X_{ig,t-1}, \sigma^2)^{w_{gtc}}
\]

for \( g = 1, \ldots, G \) do
  \( \alpha_{g2c} = \lambda_c \prod_{i=1}^n p(X_{ig1}|\delta_{g1} = \xi_{gc}, X_{ig0}), c = 1, 2, 3. \)
  for \( t = 3, \ldots, T \) do
    \( \alpha_{gtd} = \prod_{i=1}^n p(X_{igt}|\delta_{gt} = \xi_{gd}, X_{ig,t-1}, \sigma^2) \sum_{d=1}^3 \alpha_{g,t-1,c} \pi_{cd}, d = 1, 2, 3. \)
  end for
end for

for \( g = 1, \ldots, G \) do
  \( p_{gTc} = \frac{\alpha_{gTc}}{\sum_{c=1}^3 \alpha_{gTc}}, \beta_{gTc} = 1; c = 1, 2, 3. \)
  Sample \( w_{gT} \) from a Multinomial with parameters \( (p_{gT1}, p_{gT2}, p_{gT3}). \)
  for \( t = T - 1, \ldots, 2 \) do
    \( \beta_{gtc} = \sum_{d=1}^3 \pi_{cd} \prod_{i=1}^n p(X_{i,t+1}|\delta_{g,t+1} = \xi_{gd}, X_{g,t}, \sigma^2) \beta_{g,t+1,d} \)
    \( p_{gtc} = \sum_{d=1}^3 \pi_{cd} \prod_{i=1}^n p(X_{igt}|\delta_{gt} = \xi_{gc}, X_{ig,t-1}, \sigma^2) \alpha_{g,t-1,c} \pi_{cd} \beta_{gtc} \)
    \( p_{gtc} = \sum_c p_{gtc} \)
    Sample \( w_{gt} \) from a Multinomial with parameters \( (p_{id1}, p_{id2}, p_{id3}). \)
  end for
end for

Sample the initial probabilities \( \lambda \) from \( Dir(1 + \sum_g w_{g11}, 1 + \sum_g w_{g12}, 1 + \sum_g w_{g13}). \)
for \( c = 1, 2, 3 \) do
  Sample \( \Pi_c \) from \( Dir(1 + \sum_g \sum_{t=2}^T w_{g,t-1,c} w_{gt1}, 1 + \sum_g \sum_{t=2}^T w_{g,t-1,c} w_{gt2}, 1 + \sum_g \sum_{t=2}^T w_{g,t-1,c} w_{gt3}). \)
end for

Update the dispersion parameters through a Metropolis step, where the full conditionals can be expressed as

\[
p(\sigma^2|X, w, \Pi, \lambda) \propto p(\sigma^2) \prod_{i=1}^n \prod_{g=1}^G \prod_{c=1}^3 p(X_{igt}|\delta_{gt} = \xi_{gc}, X_{ig,t-1}, \sigma^2)^{w_{gtc}}.
\]
so that we have a closed form expression for the full conditionals. The full conditionals for the initial probabilities $\lambda$ and the rows of the hidden transition matrix $\Pi_c$ are all equal to the density of a Dirichlet random variable, with parameters given in Algorithm 1.

Finally, we update the dispersion parameters with a Metropolis step.

### 3.2 Identification of local trends

Once simulations from the joint posterior distribution of all parameters have been obtained, we are ready to pursue our inferential goals. We exploit the joint posterior distribution of the parameters of interest to answer biological questions with formal inferential tools based on decision-theoretic criteria. In particular we aim at selecting those genes for which the trajectory changes its regime.

First of all, we must outline a gene selection strategy. We are interested in the event

$$\bigcup_{t=2}^T \{ \delta_{gt} = \xi_g^1 \cup \delta_{gt} = \xi_g^3 \},$$

for $g = 1, \ldots, G$; that is, we are interested in sorting out those genes for which an increasing or a decreasing trend has been observed at least in one time occasion. All other genes can be deemed to have had a stationary trajectory over the observation period, that is, to have substantially maintained the expression levels observed at baseline at time $t = 1$.

The posterior probability of event (7), which we denote with $\eta_g$, can be estimated from the MCMC samples and corresponds, for each gene, to the proportion of iterations of the sampler in which at least one latent indicator $w_{gt} = 1$ or $w_{gt} = 3$.

We propose to selects genes whose posterior probability of event (7) is larger than a threshold $\delta_{cut}$, i.e.,

$$\eta_g > \delta_{cut}. \quad (8)$$

In order to select $\delta_{cut}$ we could minimize the posterior expected loss. The classical 0-1 loss leads to set $\delta_{cut} = 0.5$, i.e., a maximum a posteriori selection. On the other hand, minimization of the posterior expected loss does not allow us to formally control the expected number of false positives. To do so, we first of all estimate the FDR. The FDR can be loosely defined as the number of falsely selected genes over the number of selected genes, if
any. The FDR for any $0 \leq \delta_{cut} \leq 1$ can be consequently estimated as

$$FDR(\delta_{cut}) = \frac{\sum_{g=1}^{G}(1 - \eta_g)I(\eta_g > \delta_{cut})}{\sum_{g=1}^{G}I(\eta_g > \delta_{cut})},$$

(9)

where $I(C)$ denotes the indicator function of event $C$. In words, for any cut off $\delta_{cut}$ we have at the numerator the sum of the posterior probabilities of being stationary for the entire observation period for the selected genes, and the number of selected genes at the denominator. We use the convention that $0/0=0$. A level is a-priori set for FDR, e.g., $q = 0.05$ or $q = 0.1$, and consequently we can set $\delta_{cut}$ as

$$\delta_{cut} = \sup\{0 \leq \delta \leq 1 : FDR(\delta) < q\}.$$  

(10)

A useful note for solving (10) is that $FDR(\delta)$ is easily seen to be a step function with jumps at the sorted unique values of $\eta_g$.

It is important here to underline that we only are using an estimate of the FDR, and we can make no claims of FDR control in a frequentist sense. Control of an estimate of the FDR is convenient in practice, and is often performed since the estimates are expected to be close to the realized FDR when the model assumptions hold well. Consequently, in many cases the FDR is often controlled or out-of-control only by a small fraction when the model assumptions are not violated severely.

In our simulations below we find that the FDR is in fact controlled when the data generating model is fit. This is confirmed also, for instance, by Alfo et al. (2011) in a broader similar simulation study in a similar setting.

As usual with tests of hypotheses, for the selected genes it can be only concluded that the effect is not null: for at least one time occasion an increasing or a decreasing trend has been observed. After genes have been selected we must estimate an effect size (Kirk, 2007), which can be used to identify the most interesting genes.

In order to do so, we propose the following effect measure for ranking genes:

$$t_g = \sum_{t=2}^{T} \hat{\xi}_{g1}^2 \Pr(\delta_{gt} = \xi_{g1}|X) + \hat{\xi}_{g3}^2 \Pr(\delta_{gt} = \xi_{g3}|X),$$

(11)

for $g = 1, \ldots, G$; where $\hat{\xi}_{ge}$ is the median of the posterior distribution of $\xi_{ge}$. Effect size measure (11) summarizes, for each gene, the magnitude of the slope at time $t$ in non-stationary cases, weighted by the posterior probability
of the corresponding latent states. Selected genes can then be sorted accord-
ing to (11), and we can further investigate the most interesting (i.e., largest
in absolute value) latent slopes and posterior probabilities.

4 Constrained and generalized formulations

We can generalize (1) in three respects. First of all, we can use a time or
gene specific variance for the error term $\varepsilon_{igt}$. This would lead to minor ad-
justments to Algorithm 1, in practice we would only have to repeat the last
step for each dispersion parameter separately. We prefer the homoschedas-
ticity assumption in this paper as this would increase sensibly the number of
parameters, and recommend doing so only when data are not standardized
for some reason and the number of replicates is large. Secondly, we could use
an $AR(p)$ rather than an $AR(1)$ model. We have found the $AR(1)$ model
rather satisfactory, and recommend increasing the autoregressive parameter
only when the number of time occasions is large and the first $p-1$ occasions
are not of interest for detection of trends. The $AR(1)$ structure is the only,
in this formulation, that allows to detect a trend immediately after base-
line. Finally, we could estimate a common time-specific expectation for gene
expression levels. This would lead to a full model as follows:

$$X_{igt}|X_{igt-1}, \ldots, X_{ig1} = \delta g t \mu g,t-1 + \varepsilon_{igt}. \quad (12)$$

Since the actual expression level is not the main interest of the analysis, we
prefer to decrease the variability of the estimates by using the actual observed
values. Note that when $n = 1$, as in our motivating example, (1) and (12)
coincide.

We can also propose some more parsimonious representations of (1)
in order to (i) reduce the number of parameters and (ii) provide formulations
which can lead to speculations about the biological mechanisms behind the
data. Several alternative formulations can be obtained by constraining the
parameters involved in (1) and (2).

There are basically two constraints which are important on the hidden
transition matrix $\Pi$: the first constraint specifies independence for latent
transitions, and corresponds to identical rows for $\Pi$. This corresponds to
a situation in which trends depend only on the specific time occasion and
gene, but not on the state of the gene at the previous occasion. The second
constraint imposes global monotonicity, i.e., that the outcome can only be
monotonically increasing, monotonically decreasing, or constant over time, and corresponds to a diagonal transition matrix (i.e., a latent class model in which genes choose a latent state at baseline and transitions are not allowed).

We also could impose constraints on the latent slopes, for instance requiring that $|\xi_{g1}| = \xi_{g3}$, thereby forcing increasing and decreasing trends to be of the same magnitude. Slopes can also be assumed to be equal across genes, i.e., $\xi_{gc} = \xi_c$ for $c = 1, 2, 3$.

In order to approximate the posterior of each possible constrained model, we can still use an MCMC strategy along the lines of Algorithm 1, where we appropriately modify the steps in which constrained parameters are involved. Specifically, when identical rows are assumed for $\Pi$, we update the rows sampling only once from a Dirichlet distribution with parameter vector $(\gamma_1, \gamma_2, \gamma_3)$, where $\gamma_j = 1 + \sum_{g=1}^{G} \sum_{t=3}^{T} \sum_{c=1}^{3} w_{g,t-1,c} w_{gtj}, \ j = 1, 2, 3$.

When a diagonal transition matrix is used, we update the parameters $\lambda$ from a Dirichlet distribution with parameters $(1 + \sum_{g} \sum_{t} w_{gt1}, 1 + \sum_{g} \sum_{t} w_{gt2}, 1 + \sum_{g} \sum_{t} w_{gt3})$. When $\xi_{g1} = -\xi_{g3}$, we sample $\xi_{g3}$ from $p(\xi_{g3}) \Pi_{i} \Pi_{t} p(X_{igt}|\delta_{gt} = -\xi_{g3}, X_{ig,t-1}, \sigma^2)^{w_{gt1}} p(X_{igt}|\delta_{gt} = \xi_{g3}, X_{ig,t-1}, \sigma^2)^{w_{gt3}}$.

4.1 Model choice

Once we are willing to consider more than one model (e.g., model (1) and few of its constrained versions), in the Bayesian framework we should set up computation of Bayes factors, that is, generalized likelihood ratios. See for instance Kass and Raftery (1995). On the other hand, approximation of Bayes factors would be very cumbersome in the proposed framework since the marginal likelihood of competing models is not analytically available.

We therefore propose to choose the model with the best goodness of fit as assessed by an approximation of the Bayesian Information Criterion (BIC). Note that Schwarz (1978) shows that BIC is itself an approximation of the Bayes factor when objective priors are used.

We propose to approximate the BIC of model $\mathcal{M}$ with

$$BIC_{\mathcal{M}} = -2 \log(L(\hat{\theta})) + \log(G)d, \quad (13)$$

where $d$ is the number of free parameters involved in model $\mathcal{M}$; $L(\hat{\theta})$ is the likelihood; and $\hat{\theta}$ represents the posterior mode for the parameters under the chosen model $\mathcal{M}$. The difference with the usual BIC is that the likelihood is computed at the maximum likelihood estimates, while we use here the
Bayesian estimates. We prefer to do so for two reasons: first of all, these are already available from the MCMC output; secondly, these are the estimates that are used for inference and therefore we believe the model should be evaluated according to those values. Note that due to the well known convergence of the Bayesian and frequentist estimates, and since the likelihood is a continuous function of the parameters, (13) will converge to the BIC of Schwarz (1978) as $G$ grows. We can finally select the model corresponding to the smallest BIC value. When comparing two competing models, $M_1$ and $M_2$, we prefer model $M_1$ ($M_2$) when

$$
\Delta BIC_{12} = BIC_{M_1} - BIC_{M_2}
$$

is negative (positive).

We also consider a pure Bayesian model choice tool, the deviance information criteria (DIC). From a Bayesian perspective, the DIC criterion, introduced by Spiegelhalter et al. (2002) as a measure of model comparison and adequacy, is a generalization of the Akaike Information Criterion for hierarchical Bayesian models. The model with the smallest DIC shows the best fit to the data and offers the best predictions for a replicated dataset with the same structure which has generated the observed data.

5 Data analysis

The proposed model is applied to data coming from a study about the embryo development of mice with the 22q11 deletion (Amati et al., 2007). Deletion of the 22q11.2 chromosomal region is responsible of many malformations such as congenital heart defects, hypoplasia of the thymus gland and craniofacial dysmorphism. Although it has been demonstrated that many such phenotypic traits are due to changes in gene regulation of a subset of genes mapping within the critical area, little is known concerning how this microdeletion acts on deregulating the expression of 22q11 genes. Identification of key genes involved in specific developmental processes requires an understanding of the patterns of gene expression in a specific tissue at a specific time. We consider the expression profile of the mouse orthologous genes in embryos at 7 developmental stages (from 4.5 dpc to 14.5 dpc), corresponding to the pharyngeal development. We considered 53 genes that constitute the 22q11DS and spotted them on a low-density DNA microarray. Observed signals have been
normalized through dye-swap normalization. We specify a model through the following equation:

$$X_{gt}|X_{g,t-1}, \ldots, X_{g1} = \delta_{gt}X_{g,t-1} + \varepsilon_{gt}$$

(14)

for \(t = 2, \ldots, 7\), and \(g = 1, \ldots, 53\); where \(\varepsilon_{gt} \sim N(0, \sigma^2)\) Normalized gene expression levels are plotted in Figure 1. It is clear from the plot that not all gene expression levels are measured at each time. In particular, only the expression levels of 25 genes have been measured at each time and for the remaining genes at least one measurement is missing (the percentage of missing entries is about 17%). This is not an issue in the Bayesian approach, as in our implementation missing data can be embedded in the model as unknowns and estimated jointly with the other parameters. Indeed we can divide the vector of \(X\) in two components: \(X^{\text{obs}}\), that contains the data what we do observe and \(X^{\text{miss}}\), that contains the data that we do not observe. Algorithm in Section 3.1 is augmented with data imputation at each iteration by generating \(X^{\text{miss}}\) from its full conditional

$$p(X_{igt}^{\text{miss}}|\delta_{gt} = \xi_{gc}, X_{ig,t-1}, \sigma^2) = \phi \left( \frac{X_{igt}^{\text{miss}} - \xi_{gc}X_{ig,t-1}}{\sigma} \right).$$

The proposed model has been estimated using the algorithm described in Subsection 3.1, augmented for missing data as described above. Two chains have been initialized with different starting points. We allow 10000 iterations for the sampler to converge and another 5000 for sampling from the joint posterior. The posterior samples are finally thinned by 10 for estimating the posterior distribution. The model is implemented using R software (R Development Core Team, 2011) and the output is obtained in about 40 seconds on a dual processor 2.4 GHz machine. The computational time does not dramatically increase when the number of genes is large. For 50 genes measured at 50 occasions, 10000 iterations of the proposed algorithm take about 60 seconds. For 20000 genes measured at 10 occasions, it takes about 5 minutes. Chain convergence has been ascertained by visual inspection using standard convergence diagnostic tools, such as trace plots and autocorrelation plots.

Consider for instance the first two columns of Figure 2, where we show trace plots and autocorrelations for the parameters \(\xi_{g1}\), \(\xi_{g2}\) and \(\xi_{g3}\) corresponding to gene Crabp1. The two chains (red and black) show a satisfactory mixing and no significant autocorrelation can be detected. In the
Figure 1: *Normalized expression measurements of 53 genes that constitute 22q11DS (Amati et al., 2007).*
Figure 2: Convergence diagnostics for parameter $\xi$ of gene Crabp1. The first column shows the trace plots of the two Markov chains and the second column shows the autocorrelation for chain 1. Trace plots show a good mixing between the two chains (red and black lines) and the ACFs do not exceed the significant threshold. In the last column, the marginal posterior distributions of $\xi_1, \xi_2, \xi_3$ are shown for the same gene.
last column of Figure 2 we show the histogram of the marginal posterior distribution of the same parameters. Similar plots have been done for all parameters and do not give evidence of convergence warnings.

Figure 3 shows the posterior probability of each latent status for gene *Crabp1* at each time, that is the posterior probability $Pr(\delta_{gt} = \xi_{gc}|X)$ for $c = 1, 2, 3$. We denote with 1 the non-stationarity decreasing situation ($\xi_{gc} \leq -1$), with 2 the stationarity situation ($|\xi_{gc}| < 1$) and with 3 the non-stationarity increasing situation ($\xi_{gc} > 1$). Latent trajectory of gene *Crabp1* is stationary at time 1, 2. At time 3, the posterior probability of state 3 is slightly larger than the posterior probability of latent state 2: after time 3 the latent trajectory changes its slope such that the probability $Pr(\delta_{gt} = \xi_{g3}|X)$ is larger than the posterior probability of the stationary situation.

In Figure 4 we report the posterior distribution of the probability transition matrix $\Pi$ and the 95% credibility intervals. This can be used to obtain a general idea of transitions within the data. As it seems clear, the hypothesis of diagonal transition matrix is not supported since credibility intervals of $\Pi_{11}$, $\Pi_{22}$ and $\Pi_{33}$ do not lie around 1. Hence there are transitions during the observation period. Similarly, the posterior distribution for $\Pi$ seems to deliver very few evidence of identical rows; hence these transitions are not independent of the current state. We will perform a goodness of fit analysis in the next section. We can also interpret the posterior mean for the transition probabilities. From state 1 these means are (0.15, 0.41, 0.44); from state 2 (0.03, 0.75, 0.22) and from state 3 (0.04, 0.72, 0.22). Consequently, we can conclude that stationarity is highly persistent (a stationary gene leaves this state only with probability 0.25), and such genes more often switch to an increasing trend, which is also persistent. On the other hand, decreasing trends are not very persistent (only 15% probability of staying in the state), and can easily switch to increasing (with 44% probability as estimated by the posterior mean).

Using criterion (8) with $\delta_{cut} = 0.5$ or with $\delta_{cut} = 0.4$ arising from (10) with $q = 0.1$ we obtain the same list of 15 genes, which are shown in Table 1. Similar results can be obtained using values of $\delta_{cut}$ ranging from 0.3 to 0.5. We also report the posterior probability $Pr(\delta_{gt} = \xi_{g2}|X)$, while genes are sorted as suggested in (11).

Note for instance that the gene with largest estimated effect, *Crabp1* is likely non-stationary from dpc 9.5 to the end of embryo development. The estimated slopes are large as reported earlier. We do not report the other
Figure 3: Posterior probability of latent status for gene Crabp1 at each time. We denote 1 the non-stationarity decreasing situation ($\xi_{gc} \leq -1$), 2 the stationarity situation ($|\xi_{gc}| < 1$) and 3 the non-stationarity increasing situation ($\xi_{gc} > 1$).
Figure 4: Probability transition matrix $\Pi$: posterior distribution and 95% credibility intervals.
Table 1: Posterior probability of $Pr(\delta_{gt} = \xi_{g2}|X)$ for genes selected by criterion in Equation (8). Genes are sorted according to (11).

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>6.5</th>
<th>7.5</th>
<th>8.5</th>
<th>9.5</th>
<th>11.5</th>
<th>14.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crabp1</td>
<td>0.64</td>
<td>0.43</td>
<td>0.42</td>
<td>0.18</td>
<td>0.21</td>
<td>0.32</td>
</tr>
<tr>
<td>Hand2</td>
<td>0.01</td>
<td>0.25</td>
<td>0.59</td>
<td>0.88</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td>Foxc2</td>
<td>0.98</td>
<td>0.86</td>
<td>0.72</td>
<td>0.38</td>
<td>0.36</td>
<td>0.91</td>
</tr>
<tr>
<td>Hcf2</td>
<td>0.01</td>
<td>0.10</td>
<td>0.89</td>
<td>0.84</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Aldh1a2</td>
<td>0.98</td>
<td>0.86</td>
<td>0.72</td>
<td>0.32</td>
<td>0.36</td>
<td>1.00</td>
</tr>
<tr>
<td>Usp18</td>
<td>0.37</td>
<td>0.31</td>
<td>0.02</td>
<td>0.98</td>
<td>0.97</td>
<td>0.71</td>
</tr>
<tr>
<td>Arvcf</td>
<td>0.80</td>
<td>0.26</td>
<td>0.27</td>
<td>0.56</td>
<td>0.98</td>
<td>0.78</td>
</tr>
<tr>
<td>Comt</td>
<td>0.35</td>
<td>0.36</td>
<td>0.72</td>
<td>0.32</td>
<td>0.75</td>
<td>0.82</td>
</tr>
<tr>
<td>Cldn5</td>
<td>0.77</td>
<td>0.72</td>
<td>0.66</td>
<td>0.62</td>
<td>0.38</td>
<td>0.99</td>
</tr>
<tr>
<td>Prodh</td>
<td>0.73</td>
<td>0.67</td>
<td>0.31</td>
<td>0.94</td>
<td>0.88</td>
<td>0.67</td>
</tr>
<tr>
<td>Prodh</td>
<td>0.74</td>
<td>0.79</td>
<td>0.75</td>
<td>0.23</td>
<td>0.22</td>
<td>0.56</td>
</tr>
<tr>
<td>Clc3</td>
<td>0.79</td>
<td>0.78</td>
<td>0.76</td>
<td>0.29</td>
<td>0.68</td>
<td>0.69</td>
</tr>
<tr>
<td>Sema3c</td>
<td>0.83</td>
<td>0.85</td>
<td>0.57</td>
<td>0.18</td>
<td>0.83</td>
<td>0.90</td>
</tr>
<tr>
<td>Snap29</td>
<td>0.81</td>
<td>0.81</td>
<td>0.86</td>
<td>0.90</td>
<td>0.22</td>
<td>0.48</td>
</tr>
<tr>
<td>Stk22a</td>
<td>0.22</td>
<td>0.37</td>
<td>0.64</td>
<td>0.85</td>
<td>0.81</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Estimated slopes for reasons of space. We can anyway give further comments. The second gene in the list, Hand2, instead is non stationary at the beginning of embryo development and then becomes stationary. More precisely, this gene seems to be active at baseline and then its activity decreases steadily until 7.5 dpc. The third gene in the list, Foxc2 seems to show local trends at 9.5 and 11.5 dpc. It seems like this gene is active until 8.5 dpc, it then decreases its activity until 11.5 dpc, after which no other local trends are observed. We give more details below.

Six of the selected genes were also identified in Amati et al. (2007) using testing procedures for genes differentially expressed with respect to full embryo development. This latter strategy of course gives us much less information, and is aimed at identifying different effects in the data, so that a comparison between our list and the list in Amati et al. (2007) is not completely meaningful. Nevertheless, genes Foxc2 and Crabp1 were already selected and have been also confirmed as differentially expressed with respect to completed embryo development by PCR. These genes play a key role in regulating the embryo development of mice with the 22q11 deletion.
In fact, *Foxc2* gene encodes for a transcription factor expressed when forming somites, in head mesoderm and in endothelial and mesenchymal cells of the developing heart and blood vessels. Mutations in *Foxc2* cause the human lymphedema-distichiasis syndrome, an autosomal dominant disorder characterized principally by lymphedema of the limbs and double rows of eyelashes. One of the complications of this disorder may include cardiac defects, suggesting that *Foxc2* is a gene with pleiotropic effects acting during development. Amati et al. (2007) argued that *Foxc2* is expressed at all the developmental stages analyzed (from 4.5 to 14.5 dpc) and that its expression levels remain constant except in the stages from 9.5 dpc to 11.5 dpc when they decrease. This is partly confirmed by our analysis since at time 9.5 and 11.5 a significant change in the trajectory slope is detected.

### 5.1 Model comparison

In this subsection we compare our full model with some of its constrained or generalized formulations, in order to gain more insights into the time-course experiment we are analyzing. In particular, we focus on different constrained parameterizations of the hidden transition matrix Π, which is at the core of the HMM. We compare the proposed model, labeled $M^∗$, with three alternative models which differ from the proposed model in the following aspects:

- **$M_1$**: the transition matrix $\Pi_{M_1}$ has identical rows;
- **$M_2$**: the transition matrix $\Pi_{M_2}$ is a diagonal matrix;
- **$M_3$**: autoregressive time-varying coefficient model;
- **$M_4$**: autoregressive time-varying coefficient model with regression coefficient modeled as a nonparametric smooth function (TVAR model).

In other words, $M_1$ differs from our $M^∗$ since it allows independent transitions among the latent states; on the other hand, in $M_2$ transitions are not allowed once a latent state has been chosen at baseline. Following the suggestions of one of the referees, we compare the proposed model with an autoregressive time-varying coefficient models $M_3$ and $M_4$. In model $M_3$ the parameter $\delta_{gt}$ is regarded as an unconstrained time-varying coefficient for gene $g$:

$$X_{igt} | X_{igt-1}, \ldots, X_{ig1} = \delta_{gt} X_{igt-1} + \varepsilon_{igt} \quad (15)$$
where $\delta_{gt} \sim N(0, \tau^2_g)$ and $\tau^2_g \sim IG(0.01, 0.01)$.

The main difference with the proposed model is that $\delta_{gt}$ is modeled as a continuous distribution while in the proposed model $\delta_{gt}$ is modeled as a random effect arising from a discrete distribution with three support points defining the stationary status of the process. Model $M_4$ differs from $M_3$ since the regression coefficient is modeled as a nonparametric smooth function. In particular we consider a time-varying coefficient model in which the regression coefficient is modeled as the following second order polynomial spline:

$$
\delta_{gt} = b_{gt0}^{(0)} + b_{gt0}^{(1)}(X_{ig,t-1} - a_0) + b_{gt0}^{(2)}(X_{ig,t-1} - a_0)^2 + \sum_{k=1}^{K} b_{gtk}^{(2)}(X_{ig,t-1} - a_k)^2
$$

where the polynomial coefficients $b_{itk}^{(p)}$ and knots $a_k$ are unknown parameters, which are estimated as part of the model. Since the number of knots does not affect significantly the model fit, we have fixed $K = 3$. Following Huang and Shen (2004) and Rajan (1997), we have used $t-1$ as threshold lag that minimizes the BIC. For what concerns priors, we assume the knots $a_k$ are uniformly distributed on $(a_0, a_{K+1})$ where $a_0$ and $a_{K+1}$ are fixed respectively equal to the 10% and 90% quantile of the data; while polynomial coefficients have independent normal distributions with 0 mean and $10^3$ variance. In order to select the degree of the polynomial spline, we have compared a linear, a quadratic and a cubic polynomial spline. A quadratic polynomial spline has been chosen according to the BIC criterion.

We believe TVAR models make a good comparison thanks to complete flexibility, but may be too complex on real data. Furthermore, they lack the general advantage of our hidden state assumption which directly clusters genes.

Table 2 summarizes the differences in terms of overall fit indices for the proposed model $M^*$, and models $M_1$, $M_2$, $M_3$ and $M_4$: although the proposed model involves a larger number of parameters, all information criteria agree on the fact that our $M^*$ is to be preferred over each of the other models. With respect to model $M_3$, sensitivity about the prior distributions has been investigated: we have tried using an “informative” prior for $\tau^2_g$, centering the $IG$ on the posterior estimate obtained running the model with the first (flat) prior. The DIC and BIC still favor our model since $\text{DIC}=1204.321$ and $\Delta \text{BIC}=17.864$. We have also tried other priors for $\tau^2_g$, which are not reported for reasons of space, and the BIC still favored our model ($\tau^2_g$ modeled as lognormal, we have $\text{DIC}=1214.32$).
Table 2: Overall fit indices for models $M^*$, $M_1$, $M_2$ and $M_3$. $\Delta BIC$ denotes the differences of BIC of models $M_1$, $M_2$ and $M_3$ with respect to the proposed model. The larger the differences, the more $M^*$ is favored.

<table>
<thead>
<tr>
<th>Model</th>
<th>DIC</th>
<th>$\Delta BIC$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M^*$</td>
<td>1179.950</td>
<td>–</td>
</tr>
<tr>
<td>$M_1$</td>
<td>1205.519</td>
<td>4.779</td>
</tr>
<tr>
<td>$M_2$</td>
<td>1224.715</td>
<td>46.942</td>
</tr>
<tr>
<td>$M_3$</td>
<td>1207.831</td>
<td>18.342</td>
</tr>
<tr>
<td>$M_4$</td>
<td>1263.721</td>
<td>51.725</td>
</tr>
</tbody>
</table>

More specifically, we conclude that transitions are not independent, and consequently change in expression levels during development are determined not only by the dpc, but also by the presence of absence of a trend at the previous dpc. Further, the transition matrix is not diagonal: a latent class model would lead to biased estimates on these data, as genes switch states during embryo development, that is, there are only local but not necessarily global trends.

6 Simulation study

In order to evaluate the performance of the proposed model in a controlled setting, we perform a small simulation study. We compare the proposed model, labeled $M^*$ with its constrained formulations, $M_1$ and $M_2$ and with time varying model in Equation (15).

For model $M_3$, gene selection strategy is based on the marginal posterior distribution of $\delta_{gt}$. In particular, we are interested in the event

$\bigcup_{t=2}^{T} \{ \delta_{gt} \in (-\infty; -1) \cup \delta_{gt} \in (1, +\infty) \}$

As with decision criteria (8) we select genes whose posterior probability for this event is larger than a threshold $\delta_{cut}$ calibrated controlling for the FDR as in Equation (10), with $q = 0.10$.

We simulate 100 datasets from model $M^*$ and fit $M^*$, $M_1$, $M_2$ and $M_3$. For each dataset, we select the list of genes for which the trajectory changes its regime according to criterion in Equation (10), with $q = 0.10$. Similarly, we simulate 100 datasets from model $M_1$ and fit $M^*$, $M_1$, $M_2$ and $M_3$ and
Table 3: Models $M^*$, $M_1$, $M_2$ and $M_3$ are compared in terms of average gene detection performance with respect to 100 data sets have been simulated from the proposed model $M^*$ and from $M_1$. The observed averaged True Discovery Rate (TDR), False Discovery Rate (FDR) and False Non Discovery Rate (FNR) are reported in the last three columns.

<table>
<thead>
<tr>
<th>True Model</th>
<th>Model</th>
<th>TDR</th>
<th>FDR</th>
<th>FNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M^*$</td>
<td>$M^*$</td>
<td>0.90</td>
<td>0.10</td>
<td>0.03</td>
</tr>
<tr>
<td>$M_1$</td>
<td>$M^*$</td>
<td>0.82</td>
<td>0.17</td>
<td>0.04</td>
</tr>
<tr>
<td>$M_2$</td>
<td></td>
<td>0.55</td>
<td>0.68</td>
<td>0.18</td>
</tr>
<tr>
<td>$M_3$</td>
<td></td>
<td>0.80</td>
<td>0.22</td>
<td>0.04</td>
</tr>
<tr>
<td>$M_1$</td>
<td>$M^*$</td>
<td>0.82</td>
<td>0.12</td>
<td>0.05</td>
</tr>
<tr>
<td>$M_1$</td>
<td></td>
<td>0.87</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td>$M_2$</td>
<td></td>
<td>0.51</td>
<td>0.71</td>
<td>0.22</td>
</tr>
<tr>
<td>$M_3$</td>
<td></td>
<td>0.81</td>
<td>0.21</td>
<td>0.04</td>
</tr>
</tbody>
</table>

select the list of genes of interest. When simulating from $M^*$ we generate data assuming a hidden transition matrix $\Pi$ equal to the posterior mean estimate obtained on the real data in the previous section. For model $M_1$, we use a transition probability matrix whose rows are all equal to the first row of matrix $\Pi$ used for model $M^*$. Table 3 shows the average percentage, out of 100 simulations, of the truly identified genes (TDR), the average actual False Discovery Rate (FDR) and the actual False Non Discovery Rate (FNR) for each model. The FNR is defined as the number of incorrectly non-identified genes over the number of non-identified genes. When the data are simulated from $M^*$, the true model outperforms the other models in terms of all rates. All models except $M^*$ lead to an unacceptable FDR, large FNR and low TDR. When data are simulated from model $M_1$, the true model is favored but it performs only slightly better than $M^*$. $M_2$ leads to a much lower TDR, and is hence conservative; while $M_3$ leads to an unacceptable FDR. In synthesis, our model $M^*$ seems to perform well and to be robust to the data generating process to some extent. On the other hand, $M_1$ and $M_2$ could be too restrictive for general data generating processes like $M^*$ and $M_3$ could be too general and lead to many false discoveries when the data are actually generated from simpler models like $M^*$ and $M_1$. 

Submission to Statistical Applications in Genetics and Molecular Biology
7 Conclusion

We have proposed a three-state autoregressive latent Markov model which allows us to smoothly model switching monotonic regimes for microarray time-course data. We have assumed an $AR(1)$ structure, which can be easily generalized to $AR(p)$, with $p \geq 1$, or to other assumptions for the manifest distribution. Note that when $p > 1$, it is not straightforward to define a trend. In that case, one latent state should always correspond to a situation in which the vector of $p$ slopes is contained within the unit ball, in order to identify stationarity in the $AR(p)$ model. One or more latent states should then be specified corresponding to trends of interest (e.g., a latent state could correspond to a vector outside the unit ball with positive components, or whose sum is positive, and so on).

Our model can be directly applied when the probability of latent transitions can be assumed to be constant from one time point to another (i.e., we have assumed homogeneity of the latent transition matrix). Note that this does not necessarily correspond to having equally spaced time points. In our application, time points are not equally spaced. They have been designed a priori in this way since rat embryology tells us that events are expected at those time points. Hence, the probability of transition between 4.5 and 6.5 dpc should be approximately the same as that between 6.5 and 7.5 dpc even if these are not equally spaced, and similarly for the other time points. In other studies one may have unequally spaced time points, but a fixed probability of transition may be expected only between equally spaced time points. In those cases one should regard the time points which should have been measured (e.g., 5.5, 10.5, 12.5 and 13.5 dpc in our example) as missing data, which can be treated as described. See also (Zucchini and MacDonald, 2009, Chapter 2) on this issue.

We have proposed a Bayesian approach to inference. We note that one could also obtain the maximum likelihood estimator by using an ad-hoc EM-type algorithm, but that inference on certain quantities of interest would be much more cumbersome.

We have focused on the case of a single biological condition, where the main interest is in identifying genes with local trends. A possibility for further work in this area resides in generalization to multiple biological conditions, like in the works of Yuan and Kendziorski (2006) and Yoneya and Mamitsuka (2007); or in general to include covariates.

Our model can be also seen as a time-course version of the latent class
model specified for cross-sectional microarray data in Alfo et al. (2011), with
the difference that we do not rely on discretizing the observed expression
levels. Discretization is also a possibility for further elaboration of the pro-
posed model, and could lead to robust inference and to identification of genes
with a pre-specified effect size. One could finally allow the latent slopes to
be time-specific, which would lead to latent trajectory models (Bollen and
Curran, 2006). A similar result would be achieved by increasing the number
of regimes to an unknown number \( k + 1 \). The number of regimes could then
be chosen with a transdimensional MCMC.

References

class model for robust semiparametric gene discovery,” *Statistical Applica-
tions in Genetics and Molecular Biology*, 10.

Amati, F., M. Biancolella, A. Farcomeni, S. Giallonardi, D. Bueno,
S.and Minella, L. Vecchione, G. Chillemi, A. Desideri, and G. Novelli

technique occurring in the statistical analysis of probabilistic functions of

Benjamini, Y. and Y. Hochberg (1995): “Controlling the false discovery rate:
A practical and powerful approach to multiple testing,” *Journal of the


Chib, S. (1996): “Calculating posterior distributions and modal estimates in

through Bayesian sampling,” *Journal of the Royal Statistical Society (Ser.
B)*, 56, 363–375.


