Bayesian Analyses of Longitudinal Binary Data Using Markov Regression Models of Unknown Order

A. Erkanli,* R. Soyer† and A. Angold*

December 23, 1999

Abstract

We present non-homogenous Markov regression models of unknown order as a means to assess the duration of autoregressive dependence in longitudinal binary data. We describe a subject's transition probability evolving over time using logistic regression models for his or her past outcomes and covariates. When the initial values of the binary process are unknown, they are treated as latent variables. The unknown initial values, model parameters, and the order of transitions are then estimated using a Bayesian variable selection approach, via Gibbs sampling. As a comparison with our approach, we also implement the Deviance Information Criterion (DIC) for the determination of the order of transitions. An example addresses the progression of substance use in a community sample of $n = 242$ American Indian children who were interviewed annually four times. An extension of the Markov model to account for subject-to-subject heterogeneity is also discussed.

Key Words: Alcohol and drug abuse, Bayes factors, Bayesian inference, Gibbs sampling, logistic regression, Markov models, variable selection.

*Developmental Epidemiology Program, Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Box 3454, Durham, NC 27710

†Management Science Department, George Washington University, Washington, DC 20052.

1 Introduction

We consider non-homogeneous Markov regression models of unknown order as a means of assesing duration of autoregressive dependence in longitudinal binary data. Such binary time-series data arise in repeated application of diagnostic and screening procedures on the same subjects. Conditional on the order, each subject's transition probability is modeled as an additive logistic regression of his or her past outcomes and related covariates. Thus, the parameter vector associated with the covariates is a a measure of covariate effects which is adjusted for the autoregressive dependence among the outcome variables. The unobserved initial values of the binary outcomes are treated as latent variables. Estimation of the latent variables, model parameters, and order of the transitions are undertaken within a Bayesian variable selection framework, using a Gibbs sampler. As an example, we discuss in Section 5 the progression of substance use in a sample of $n = 242$ Cherokee Indian children aged 9-16 who are participating in the Great Smoky Mountains Study (GSMS) (Costello et al., 1996), and have completed all of the four annual interviews. Our analyses reveal that effect of previous substance use behavior only extended itself over a two year period (i.e., second-order). Thus, considering a larger historical span is, for prediction purposes, may not be necessary. We also present heterogeneous Markov models to account for subject-specific variability in their transitions. This is achieved by introducing simple random-effects in the logits. When subject-specific effects are considered, however, only the previous year's substance use seems to be important, thus favoring a simpler model (i.e., first-order).

As pointed out by Liang and Zeger (1989), a major advantage of logistic transition models for longitudinal binary data is that both covariate effects and within-subject variation are addressed simultaneously on the same scale. These models are, therefore, direct extensions of the well-studied Gaussian time-series. Moreover, Markov logistic regression models may be implemented efficiently using existing software, such as `SAS PROC GENMOD` or `BUGS` (Spiegelhalter et al., 1996).

Non-Bayesian analyses of Markov models for non-normal and binary data have been discussed by many authors; see for example, Cox (1970, 1981), Bishop, Fienberg and Holland (1984, Chapter 7), Kaufmann (1987), Zeger and Qaqish (1988), Liang and Zeger (1989), and Diggle, Zeger and Liang (1994, Chapter, 10). In this previous work, statistical analyses have involved both the MLE and GEE approaches. However, one common drawback in these studies has been that the inferences are generally based on the joint distribution of the last $n(T-q)$ observations, since these models are defined conditional on the first nq observations. As Liang and Zeger (1989) have pointed out, it is undesirable in longitudinal studies with small T to ignore information in the first nq observations. The latent variable approach considered here provides a useful means of recovering information from these observations. Most of the Bayesian work involving Markov models for longitudinal binary data has concentrated on discrete state space models, see, for example, Cargnioni, Mueller and West (1997), and the references therein. In these models, the parameters are assumed to evolve over time according to a first-order Markov regression model. To the best of our knowledge, however, none of the previous involving longitudinal binary data have provided a practical means for choosing the order of the Markov regression.

The assessment of the order of a Markov model is a model selection problem encountered in many applications. Simple approaches to selecting the appropriate order include (1) fitting each model separately and comparing the "fit+complexity" type measures such as AIC or BIC, and more recently, (2) the Deviance information Criterion (DIC) developed by Spilegelhalter et al., (1998). However, both of these approaches suffer from the fact that, if the number of longitudinal observations per subject is large, then a large number of models needs to be fitted. A more sophisticated approach is to devise a Markov Chain Monte Carlo (MCMC) scheme in which the model order and parameter uncertainty are addressed simultaneously, as in the *pseudo priors* approach of Carlin and Chib (1995), and Green's (1995) *reversible jump* MCMC approach.

Alternative approaches involving Bayesian model selection have also been proposed by George and McCulloch (1993, 1996), and Kuo and Mallick (1998) for linear and generalized linear regression, and by Smith and Kohn (1996) for the non-parametric smoothing framework, using binary indicators for each predictor in the model. Using Gibbs sampling, or similar Markov Chain-Monte Carlo techniques, these model/variable selection precedures are set up to compute posterior probabilities for all the 2^p possible models having p different covariates. Although these approaches might look as computationally demanding, especially for large p, in typical applications involving generalized linear models the posterior distribution over all 2^p models usually settles in relatively small number of sub-models with higher posterior probabilities than the rest, see for example, Kuo and Mallick (1998). Moreover, these models can be easily implemented in `BUGS` software.

In this paper, we use a variable selection approach similar to those discussed in George and McCulloch and Kuo and Mallick. In our setting, however, we do not have to search over all the possible models since the number of *Markov* models is substantially smaller than the number of all possible models. For example, if there are $T = 4$ observations per subject, then there are only three Markov models to consider; the first, second and third-order.

Our basic model is introduced in Section 2 below. In Section 3, we discuss posterior inference and estimation of the transition order using Gibbs sampling. Heterogenous extension of Markov model is discussed in Section 4. The computational details of the Gibbs sampling algorithm is given in the Appendix. In Section 5, we investigate the progression of substance use among American Indian children and adolescents using longitudinal outcomes from the *Great Smoky Mountains Study* (GSMS). Conclusions and future work are discussed in Section 6.

2 Markov Binary Regression

In this section, we describe a Markov regression model for univariate binary time-series observed in longitudinal studies. In Section 5, we consider the diagnosis of substance use based on the *Child and Adolescents Psychiatric Assessment* (CAPA) (Angold, et al. 1995) interview. The time series for each subject consists of four binary substance use statuses (1 = has used illicit substances in the past three months; 0=otherwise).

2.1 A Markov Logistic Regression Model

Consider the binary observations $(Y_{i1}, \ldots, Y_{iT})$ taken on the i-th subject at T different occasions, $t = 1, \ldots, T$, where $Y_{it} = 1$ if the i-th patient has an event at time t, and $Y_{it} = 0$, otherwise. Our main goal is to develop a parsimonious probability model for the correlated Bernoulli outcomes $\{Y_{it}\}$ in the presence of covariate information. In the most general form, one can always write

$$f(y_{i1}, \ldots, y_{iT}) = \prod_{t=1}^{T} f(y_{it}|H_{it}),$$

where $H_{it} = \{y_{i1}, \ldots, y_{it-1}\}$ is the history vector (with $H_{i1} = \{y_{i0}\}$ denoting the initial value) for the i-th subject available up to time t,

$$f(y_{it}|H_{it}) = {p_{it}}^{y_{it}}(1 - p_{it})^{1-y_{it}}, \tag{1}$$

and $p_{it} = P(Y_{it} = 1|H_{it}),\ t = 1, \ldots, T$ is the probability of a subject having an event at time t, which depends on the subject's past events through vector H_{it}. We describe p_{it} on the logistic scale as an additive function of H_{it} and a p-dimensional covariate vector x_{it}

$$\text{logit}(p_{it}) = \mu + x_{it}^{'}\delta + g(H_{it}), \tag{2}$$

where $x_{it}^{'} = (x_{i1t}, \ldots, x_{ipt})$, and $\delta^{'} = (\delta_1, \ldots, \delta_p)$. Thus, different choices of $g(\cdot)$ correspond to different beliefs about the subject's transition behaviors. In particular, the function

$$g(H_{it}) = \sum_{k=1}^{q} \gamma_k y_{it-k}, \tag{3}$$

for $q \in \{1, \ldots, q_{\max}\}$, describes a q-th order Markov logistic regression, where $1 \leq q_{\max} \leq T-1$ is the *maximal* order. With $g(\cdot)$ defined as in (3), the logistic model described in (2) has a relatively simple additive structure, though it can easily be generalized to accommodate interaction terms $y_{it-k} \times y_{i,t-l}$, $y_{i,t-k} \times x_{imt}$, or terms of the form $\sum_k \theta_k f_k(y_{it-k})$ for some known functions $f_1, \ldots, f_q$.

Motivated by the recent work of George and McCulloch (1993, 1996), Smith and Kohn (1996), and Kuo and Mallick (1998), we consider an extended version of the additive structure described by the Equations 2 and 3:

$$g(H_{it}) = \sum_{k=1}^{q_{\max}} \gamma_k \beta_k y_{it-k} \tag{4}$$

where γ_k is as defined in (2), and $\beta_k = 1$, if y_{it-k} is included in the model; $\beta_k = 0$, otherwise. Thus, $\beta_1 = 1$ and $\beta_k = 0$ for $k > 1$ corresponds to the first-order model with $g(H_{it}) = \gamma_1 y_{it-1}$;

$\beta_1 = \beta_2 = 1$ and $\beta_k = 0$ for $k > 2$ corresponds to the second order model with $g(H_{it}) = \gamma_1 y_{it-1} + \gamma_2 y_{it-2}$; and so on. In other words, the order of the transitions is completely determined by the binary vector $\beta = (\beta_1, \ldots, \beta_{q_{\max}})$. Without any loss of generality, we will set $q_{\max} = T-1$. In the example given in Section 5, $T = 4$, and thus $q_{\max} = 3$. However, in specific applications with large T we may use submodels with $q_{\max} << T - 1$ for parsimony.

Using (4) with $q_{\max} = T - 1$, the space of all Markov models can be characterized by the $T - 1$-dimensional collection of $T - 1$ tuples

$$\mathcal{M} = \{(1, 0, \ldots, 0), (1, 1, 0, \ldots, 0), \ldots, (1, 1, \ldots, 1)\},$$

where the first vector identifies the first order model, the second vector identifies the second order model, and so on. Hence, a prior distribution for $\beta \in \mathcal{M}$ can be described by a discrete distribution that puts positive mass at each point $\mathcal{M}$, and zero mass at all points that lie outside $\mathcal{M}$, e.g., $(1, 0, 1, 0, 1, \ldots, 1) \in \mathcal{M}^c$. We describe such a prior distribution for β in detail, later in Section 2.4.

In the logistic model described by Equations 2 and 4, the m-th coordinate of the parameter vector δ is a log-odds ratio associated with X_{imt}, adjusted for the subject's past outcomes, effects of remaining covariates, and transition order (model) uncertainty. In the following, we will treat μ, γ, β and δ as being mutually and independently distributed random variables having the prior distributions described in Sections 2.3 and 2.4 below. For notational clarity, we write $\theta = (\mu, \delta, \beta, \gamma)$ to denote the $2(T - 1) + p + 1$ dimensional parameter vector.

2.2 Likelihood

Assuming conditional independence across subjects $i = 1, \ldots, n$ given θ, the overall likelihood function for θ is

$$L(\theta) = \prod_{i=1}^{n} \prod_{t=1}^{T} f(y_{it} | H_{it}, \theta). \tag{5}$$

Note that the likelihood function in (5) leaves the $n(T - 1)$ initial values $Y_0 = \{Y_{i\tau}\}$, $\tau \in T_0 = \{-(T-2), -(T-3), \ldots, -1, 0\}$, and $i = 1, \ldots, n$, unspecified. Depending on the context, Y_0 can be known with certainty, observed with error, or treated as a vector of latent variables having a specified joint distribution. Thus, in essence, the likelihood function in (5) is an augmented likelihood $L(\theta, Y_0)$. In our example we assume that the initial events Y_0 of substance use are independently distributed, conditionally on a vector of initial prevalences $P_0 = \{p_\tau; \tau \in T_0\}$, as

$$[Y_{i\tau}|p_\tau] \sim \text{Bernoulli}(p_\tau), \tag{6}$$

for all $i = 1, \ldots, n$ and $\tau \in T_0$, where $p_\tau = P(Y_{i\tau} = 1)$. For the purposes of this paper, we will assume that the initial prevalences $\{p_\tau\}$ on T_0 are distributed independently and identically with a uniform distribution $U(0, a)$ where a is a fixed small number in $(0, 1)$. In the context of the substance use example of Section 5, this prior represents the belief that the probability of substance use during early childhood is likely to be very small. For this example, we set $a = 0.15$, which is slightly less than the observed prevalence of substance use (0.17) at the baseline $(t = 1)$. If the vector Y_0 is known with certainty then P_0 is degenerate at a fixed vector comprised of zero's and/or one's depending on the values of $Y_{i,\tau}$ on $\tau \in T_0$.

A commonly used model in the literature is the so-called *conditional* model where the likelihood function is specified conditional upon *only* the first nq observations $y_{i1}, \ldots y_{iq}$:

$$L_c(\theta) = \prod_{i=1}^{n} \prod_{t=q+1}^{T} f(y_{it}|y_{it-1}, \ldots, y_{it-q}, \theta),$$

where q takes a value in $\{1, \ldots, T-1\}$, and the dependency of Y_{it} on its past is described by the function in (3); see for example, Zeger and Qaqish(1988) and Diggle et al., (1994, Chapter 10). In this model θ equals to $q+p+1$-dimensional sub-vector $(\mu, \delta, \gamma_1, \ldots, \gamma_q)$ obtained by setting $\beta_k = 1$ for $k = 1, \ldots, q$, and $\beta_k = 0$ for $k = q+1, \ldots T-1$, apriori. If T is large and q is small relative to T then it would probably make little difference whether $L_c(\theta)$ or $L(\theta, Y_0)$ were used for estimating the common parameters μ and δ. For example, in applications involving continuous (cross-sectional) time-series models, typically $n = 1$, and T is very large. When T is small relative to n, however, the information loss due to the exclusion of the first nq observations could be great. When $n = 100$ and $T = 4$, for instance, the conditional model utilizes only $n(T-q) = 300, 200$ and 100 total observations, corresponding to $q = 1, 2$ and 3. In contrast, the augmented likelihood described in (5), has $nT = 400$ total observations, while also accounting for transition order uncertainty.

As pointed out by one of the referees, one could formulate the problem directly as a latent model where the initial states are latent. The parameter vector θ can then be estimated using a marginal likelihood approach. Let $L_m(\theta)$ denote the marginal likelihood. It is obtained by mixing $L(\theta, Y_0)$ over the latent initial states Y_0 and the initial prevalences P_0. The mle of θ can then be found by maximizing the marginal likelihood. Direct maximization of the marginal likelihood, however, is not trivial since it involves maximization over the discrete parameter vector β. However, maximization conditional on each possible values of the vector β can be performed, and, consequently, the corresponsindg maxima for β can be found by enumaration. This approach can be easily implemented using, for example, SAS, when the initial latent vector Y_0 is known, since, in this case, $L_m(\theta) = L(\theta, Y_0)$, for each value of β. In general, when Y_0 is unknown, the marginal likelihood can be factored as

$$L_m(\theta) = \left\{ \prod_{i=1}^{n} f(y_{i1}, \ldots, y_{iq}|\theta) \right\} L_c(\theta),$$

where the first term on the right hand side described in the brackets is the joint distribution for the first nq variables obtained by summing over the initial latent states Y_0 and integrating over P_0. For example, when $\beta = (1, 0, 0)$, i.e., the first-order model with $q = 1$, the ith term in the above product is

$$f(y_{i1}|\theta) = \sum_{k=0}^{1} w_k f(y_{i1}|y_{i,0} = k, \theta),$$

where $f(y_{i1}|y_{i,0} = k, \theta)$ is the Bernoulli density given by (1), and the $w_k \geq 0$, for $k = 0, 1$, are the mixing weights satisfying $w_0 + w_1 = 1$, and $w_1 = Pr(Y_{i,0} = 1) = E(p_0)$, the expectation being taken with respect to the prior distribution of p_0. Thus the mle approach requires maximization of $L_m(\theta)$, however, this maximization is required for each value of β, which can be computationaly inefficientc especially when T is large.

In our proposed Bayesian approach, we attempt to resolve these difficulties by embedding all possible markov models into one big meta-model, and by using Gibbs sampling. Moerover, software to do so is freely available and easy to implement. With the use of uniform priors, the Bayesian approach produces results that are similar to those which could have been obtained using a marginal mle approach, but would have required additional computational effort. Also, if there is prior information available, it too can be incorporated easily into the existing Bayesian model using informative prior distributions, which can not be achieved by maximum likelihood estimation.

2.3 Prior distributions for μ, δ,

The Markov logistic model can now be completed by specifying a joint prior distribution for θ. For the purposes of this paper, we first assume that μ, δ, γ and β are independently distributed with

- $\mu \sim N(0, \sigma)$,
- $\delta \sim N_p(0, \Omega)$,
- $\gamma \sim N_{T-1}(\gamma_0, \Sigma)$,

where σ is the known variance of μ, Ω is the known covariance matrix of δ, and γ_0 and Σ are the known mean vector and covariance matrix of γ, respectively. The prior distribution of β is described in detail in Section 2.4 below. If the heteroneous model is used, then we assume the prior independence of μ_i's and ν with other parameters with the corresponding prior distributions as dicussed in Section 4 below.

2.4 Prior Distributions on $\mathcal{M}$

There are several ways to describe a prior distribution for the vector $\beta \in \mathcal{M}$. One simple way is to assign a discrete probability distribution $P = \{p_1, \ldots, p_{T-1}\}$, where $p_k = P(\beta_1 = 1, \ldots, \beta_k = 1, \beta_j = 0, k < j \leq T-1)$, for $k = 1, \ldots, T-2$, subject to $\sum_{k=1}^{T-1} p_k = 1$. For example, $p_k = 1/(T-1)$ describes a uniform distribution on $\mathcal{M}$. To simulate such a prior, we can generate a multinomial variate M with $P(M = k) = p_k$ for $k = 1, \ldots, T-1$, and then set $\beta_j = 1$ for $j \leq k$ and $\beta_j = 0$ for $j > k$, when $M = k$ is observed.

One can also use a simple Markov prior for the vector β. Consider the case $T = 4$ and $q_{\max} = 3$. Notice that β_1 is always 1 in any vector $\beta \in \mathcal{M} = \{(1,0,0), (1,1,0), (1,1,1)\}$, so we can set $\alpha_1 = P(\beta_1 = 1) = 1$. Now, β_2 can be equal to either 0 or 1. Let $\alpha_2 = P(\beta_2 = 1|\beta_1 = 1) = P(\beta_2 = 1)$; $0 < \alpha_2 < 1$. Conditional on the value of β_2, β_3 can be equal to either 0 or 1. If $\beta_2 = 0$, then $P(\beta_3 = 0|\beta_2 = 0) = 1$, since the transition $\beta_2 = 0 \rightarrow \beta_3 = 1$ is not allowed in a Markov Model. If $\beta_2 = 1$, then both transitions $\beta_2 = 1 \rightarrow \beta_3 = 0$ and $\beta_2 = 1 \rightarrow \beta_3 = 1$ are admissible, so we can set $P(\beta_3 = 1|\beta_2 = 1) = \alpha_3$ for some $0 < \alpha_3 < 1$. Thus, we obtain the following joint prior distribution for β:

$$p(\beta_1, \beta_2, \beta_3) = \begin{cases} 1 - \alpha_2 & \text{if } \beta_2 = \beta_3 = 0 \\ \alpha_2(1 - \alpha_3) & \text{if } \beta_2 = 1, \beta_3 = 0 \\ \alpha_2\alpha_3 & \text{if } \beta_2 = 1, \beta_3 = 1 \\ 0 & \text{otherwise} \end{cases}$$

For example, the uniform prior $p_k = 1/3$ corresponds to the choice $\alpha_2 = 2/3$ and $\alpha_3 = 1/2$. If $\alpha_2 = 1/2$ and $\alpha_3 = 1/2$, then $p_1 = 1/2$, $p_2 = p_3 = 1/4$, i.e., P puts half of its mass on the first order model, and distributes the remaining half equally between the second and third order models.

In the general case of $T > 4$, a Markov prior for β can be described by setting $\beta_1 = 1$, generating $\beta_2 \sim \text{Bernoulli}(\alpha_2)$, and setting sequentially

$$\beta_k|\beta_{k-1} = \begin{cases} 0 & \text{if } \beta_{k-1} = 0 \\ \sim \text{Bernoulli}(\alpha_k) & \text{if } \beta_{k-1} = 1, \end{cases} \tag{7}$$

for $k = 3, ..., T-1$. For example, choosing $\alpha_k = (T-k)/(T-k+1)$, for $k = 2, \ldots, T-1$, corresponds to the uniform prior $p_k = 1/(T-1)$ on $\mathcal{M}$.

Using the augmented likelihood function in (5) and the priors described above, the posterior distributions of μ, γ, β, δ, latent variables Y_0 and initial probabilities P_0 can now be simulated by using the Gibbs sampler dicussed in the next section.

3 Posterior Distributions and Model Order Estimation

The joint posterior distribution of θ, Y_0 and P_0 is obtained as proportional to

$$L(\theta, Y_0)p(\theta)f(Y_0|P_0)p(P_0) \tag{8}$$

where $L(\theta, Y_0)$ is the augmented likelihood described in Equation 5, $p(\theta)$ is the joint prior density of θ, $f(Y_0|P_0)$ is the joint density for Y_0 conditional on P_0, as described in (6), and $p(P_0) = \prod_{\tau \in T_0} U(0, a) = a^{-(T-1)}$. Unfortunately, the joint posterior distribution described in (8) is not available in closed analytical form. We therefore implement a Gibbs sampling approach (Gelfand and Smith, 1990) which utilizes subsequent draws from the set of conditional posterior distributions induced by the joint posterior in (8). The details of the Gibbs sampling approach is given in the Appendix to save space.

The posterior analysis described above produces a Monte Carlo sequence $\theta^{(r)} = (\mu^{(r)}, \delta^{(r)}, \beta^{(r)}, \gamma^{(r)})$, $r = 1, \ldots, R$ from the joint posterior distribution of θ. Having obtained such a sequence, there are several ways of summarizing information in the data about model order and parameter uncertainty. Since the interpretation of the basic parameters μ and δ is the same (log-odds ratios across all Markov models), statistical inferences about them can be made by utilizing their marginal posterior distributions. As such, these distributions account for both the model order uncertainty and within-subject correlations. In addition to these parameters, we may also be interested in the posterior distribution of $\eta = (\gamma_1\beta_1, \ldots, \gamma_{T-1}\beta_{T-1})$. Here, the marginal posterior distribution of η_k describes the "strength" of the association between Y_t and Y_{t-k}, for $k = 1, \ldots, T-1$.

If the focus is on a single model for future use and prediction, we may then use the posterior distribution of $P(\beta|\mathbf{y})$, since it describes the posterior probability of the Markov models under consideration. Letting $M = k$ to denote the event $\{\beta_j = 1;\ j \leq k, \beta_j = 0;\ j > k\}$, and $p_k = P(M = k)$ to denote the prior probability reflecting our prior beliefs about the likely values of the transition order for a given transition process, we may seek the Markov model having the highest posterior probability. In a particular application, however, the observed data may not be sufficiently informative to identify a single model, so we may need to calibrate the models by using the Bayes factors (Kass and Raftery, 1995).

3.1 Bayes Factors

In our application, the Bayes factor for the hypothesis $H_k : M = k$ against $H_q : M = q$ is ratio of the posterior and prior odds

$$\begin{aligned} BF &= \{P(H_k|\mathbf{y})/P(H_q|\mathbf{y})\}\{P(H_k)/P(H_q)\} \\ &= f(\mathbf{y}|M = k)/f(\mathbf{y}|M = q), \end{aligned} \tag{9}$$

where $P(H_j) = P(M = j) = p_j$ and $P(H_j|\mathbf{y}) = P(M = j|y)$. Note that, the predictive densities $f(\mathbf{y}|M = k)$ and $f(\mathbf{y}|M = q)$ are not directly estimated using Gibbs sampling, but their ratio is, as a result of Bayes' Theorem, since $P(M = j|\mathbf{y}) = f(\mathbf{y}|M = j)p_j/f(\mathbf{y})$, where $f(\mathbf{y}) = \sum_{k=1}^{T-1} f(\mathbf{y}|M = k)p_k$. We do not need to compute $f(\mathbf{y})$ either, because it cancels out in the Bayes factor. Larger values of BF provide evidence in favor of H_k, while smaller values of BF support H_q.

3.2 Deviance Information Criterion

An alternative way of comparing Markov models of different orders is to use a global model selection criterion, such as the *Deviance Information Criterion* (DIC), recently described in Speilgelhalter et al. (1998). For a generic parameter vector θ, DIC is defined as $\overline{D} + p_D$, where $D = -2\log l(\theta)$, $\overline{D} = E_{\theta|y}[D]$, and $p_D = \overline{D} - D(\widehat{\theta})$, with $\widehat{\theta}$ denoting the posterior mean. The DIC has the general "fit+complexity" form shared by many other model choice criterion, where the fit is measured by the posterior mean of the -2(log-likelihood) and complexity is measured by the *effective* number of parameters, p_D. Speilgelhalter et al. (1998) show that for non-hierarchical models e.g., fixed-effects models where the prior distribution of θ is completely specified with no unknown parameters, DIC $\approx$ AIC (the Akaike Information Criterion) with $p_D \approx p = \dim(\theta)$. In the case of the Markov models considered in this paper, $D(\theta, Y_0) = -2\log L(\theta, Y_0)$, where θ now equals to sub-vector $(\mu, \delta, \gamma_1, \ldots, \gamma_q)$. If Y_0 is known, then D is a function of θ alone. The posterior mean $\overline{D}$ is estimated from the sampled values for D, corresponding to the sampled values of θ (and of Y_0 when needed).

Although the DIC cannot be reduced to a probability statement such as $P(M = k|y)$, as used in the the Bayes factor above, we think of it as a useful measure to consider for a quick comparison of the Markov models. Computing DIC, however, may require several runs of the Gibbs sampler depending on the transition order. For example, if $T = 4$, we then set apriori the vector $\beta = (\beta_1, \beta_2, \beta_3)$ to $(1, 0, 0)$ for the first-order model, to $(1, 1, 0)$ for the second-order model, and to $(1, 1, 1)$ for the third-order model, respectively. Then, running the Gibbs sampler seperately produces posterior estimates for the sub-vectors of θ and DIC for each of the three Markov models. The model with a minimum DIC can be used for future use. For comparison with our full probability model, we also include the DIC in the substance use example in Section 5.

4 Extension: Heterogeneous Model

Our approach so far relies on the basic assumption that all the subject-to-subject heterogeneity is accounted for by the introduction of subject and time varying covariates in the logistic model. However, in some real applications there might be unknown sources of heteregoneity not presented in the model. In such circumstances, the logistic model should be modified in a number of ways to accomodate such heterogeneity. At a first glance, a trivial modification that seems approriate is to replace the parameter vectors β and γ in Equation 3 with subject-specific variants β_i and γ_i, such that, a priori, $E(\beta_i) = \beta$, and $E(\gamma_i) = \gamma$. However, such approach now requires the identification and estimation of $2n$ additional parameters which would be computationally demanding even for a moderate number of subjects n and time points T. A much simpler and more familiar approach to accomodate heterogeneity is to introduce a random-effect intercept in Equation 3:

$$\text{logit}(p_{it}) = \mu + \mu_i + x_{it}^{'}\delta + g(H_{it}), \tag{10}$$

where μ_i is a random-effect parameter distributed as $\mu_i \sim N(0, \nu)$, and $\nu \sim \text{Gamma}(a, b)$; $a > 0$ and $b > 0$ are being known. Assuming prior independence of $\mu_1, \ldots, \mu_n$ and ν and other parameters in the model, and augmenting the parameter vector θ with them, the likelihood function in Equation 5 can now be used to make simultaneous inference for μ_i's and ν, along with other parameters. Their posterior distributions can be obtained, for example, by slightly modifiying the Gibbs sampling algorithm discussed in Section 3, although we may have a different dependence structure among the observables Y_{it} since the effects δ, β and γ are now being adjusted for the random-effects μ_i.

5 Progression of substance use

In this section, we apply the models developed in Sections 2 and 4 to a problem relating to substance use. Previous studies involving the first three annual waves of data from the *Great Smoky Mountains Study* (Costello et al., 1996; Federman et al., 1997) shown evidence of increasing prevalence rates with age, of alcohol, marijuana and tobacco use (both cigarettes and chewing tobacco) among American Indian and other youths living in the same rural area in Southern Appalachia. Other studies based on samples of young adolescents (Gfellner and Hundleby; 1995) and a recent Bayesian model (Erkanli et al., 1999) also give similar rates of substance use among American Indians and other racial ethnic groups. Our goal here is to establish whether there are similar relationships between the *transitional rate* of substance use and age of the subjects. The measure used for substance use is the *Child and Adolescent Psychiatric Assessment* (CAPA) (Angold et al., 1995), an interview that elicits information about symptoms that contribute to a wide range of psychiatric diagnoses. For the purpose of this paper, we use the binary variable Y_{it}, where $Y_{it} = 1$, if I-th subject had used alcohol, tobacco or marijuana in the three months preceding time t, and $Y_{it} = 0$ otherwise. Thus, the covariate vector x_{it} consists of a single variable; the age (ranging from 9 to 16 year-olds) of the subject at the time of the interview. The parameter δ is the unknown log-odds ratio associated with age, adjusted for the the temporal correlations among the substance use binary scores $\{Y_{it}\}$. Here, the parameter vectors $\gamma = (\gamma_1, \gamma_2, \gamma_3)$, and $\beta = (\beta_1, \beta_2, \beta_3)$ together describe the dependency among the Y_{it}'s. For sensitivity analyses, we consider the two cases for the initial events Y_0 and prevalences P_0:

- **Case 1.** $p_\tau = 0$, $Y_{i,\tau} = 0$, $\tau \in T_0$.
- **Case 2.** $p_\tau \sim U(0, 0.15)$, $Y_{i,\tau} \sim \text{Bernoulli}(p_\tau)$ $\tau \in T_0$.

Here, Case 1 represents the belief that there are no cases of substance use prior to the first interview, while Case 2 allows some cases of substance use prior to the first interview.

5.1 Priors

In describing prior uncertainty, we consider relatively uninformative but proper priors for μ and δ: $\mu \sim N(0, 100)$, and $\delta \sim N(0, 100)$. As the prior distribution of γ we chose a 3-dimensional multivariate normal with the mean vector $\gamma_0 = (0, 0, 0)$ and variance-covariance matrix $\Sigma = \text{diag}(100, 100, 100)$. As described in Section 2.3, we chose a uniform prior for β by setting $\alpha_1 = 1$, $\alpha_2 = 2/3$ and $\alpha_3 = 1/2$.

5.2 Implementation

The computations were performed using `BUGS` software on a Pentium PC running under Windows NT. To improve the convergence behavior of the sampler, we used the transformations $\mu' = \mu + \delta\bar{x}$, and $x'_{it} = x_{it} - \bar{x}$, where $\bar{x} = \sum_{i,t} x_{it}/nT$. We first ran the sampler for the first, second and third order models seperately (by setting apriori, β to (1,0,0), (1,1,0) and (1,1,1), respectively) and computed the DIC for each of them. We then ran the full model. For each case, we used 5000 iterations for burn in and 20000 for computations, corresponding to five different starting values of the sampler that are generated automatically in `BUGS` (generated randomly using the piors described above). No convergence problems were encountered; all the sampling trajectories seemed to settle after 1500-2000 iterations. A typical run with 1000 iterations took about 2 minutes for Case 1, and 6 minutes for Case 2, respectively. To save space, we do not display any convergence plots or diagnoses, however, all BUGS code can be obtained from the first author upon request.

5.3 Results

Tables 1 and 2 show that the second order model fits better than the first and third order models. It is interesting to note that under Case 1 the effective number of parameters is close to the dimension of θ, confirming that asymptotically the DIC is equal to AIC. In the second case,

Table 1: Comparison of models using DIC.

	$\overline{D}$		p_D		DIC	
Order	**Case 1**	**Case 2**	**Case 1**	**Case 2**	**Case 1**	**Case 2**
1	785.88	733.16	3.04	59.73	788.92	792.89
2	779.71	713.11	3.92	59.77	783.63	772.88
3	779.79	670.69	4.03	181.09	783.82	851.79

however, the effective number of parameters is substantially larger, reflecting the contribution of Y_0 and P_0 to the total number of parameters. The posterior model probabilities, and Bayes factors under the uniform 1/3 prior on $\mathcal{M}$ are given in Table 2. Table 3 exhibits the posterior summaries under the full and the second-order models for μ, δ, γ_k and β_k. The posterior means of the second-order transition probability, along with the 95% credible intervals, for the event $Y_t = 1$ are displayed in Figures 1.a and 1.b as a function of age and the past (y_{t-1}, y_{t-2}). The agreement of our full probability model with the model chosen by DIC, the second-order model, is quite interesting. The DIC approach is likely to be more time consuming than our full model approach when T is large, say 20, because one would need to fit 19 different models to pick the "best" model.

Table 3, and Figures 1a and 1b together show that each subject's substance use transitional behavior can be attributed to two components; a relatively small effect of age itself, and a much larger effect of previous substance use behavior. These are effects that could quite well have been demonstrated using other well-known methods, i.e., using random-effects logistic regression models. However, we can also see that the effect of previous substance use behavior only extended itself over a two year period. Considering a larger historical span may, therefore, not be necesseray. When conducting longitudinal analyses of developmental process it is often difficult to know in advance how long a historical period to take into account for prediction purposes. Indeed, different theories may predict different durations for the effects of earlier behavior or other factors. The method presented here provides a useful means of addressing

Table 2: Posterior model probabilites $P(M = m|y)$ and Bayes Factors B_{ij}.

Order	**Case 1**	**Case 2**
1	0.166	0.003
2	0.736	0.844
3	0.098	0.153
B_{12}	0.225	0.004
B_{13}	1.694	0.019
B_{23}	7.510	5.163

these issues directly.

FIGURE 1 HERE

For comparison, we also include results obtained from using the random-effects model in Table 4. These results are based on the additional prior specifications with $a = 1$, $b = 0.1$ as the hyperparameters of ν to reflect relatively weak prior information. The posterior model probabilities for the heterogeneous model are $P(M = 1|y) = 0.636$, $P(M = 2|y) = 0.206$, and $P(m = 3|y) = 0.158$ for Case 1, and $P(M = 1|y) = 0.643$, $P(M = 2|y) = 0.162$, and $P(M = 3|y) = 0.114$ for Case 2, respectively. These results suggest a heterogeneous first-order model. So, care must be taken when interpreting these results.

6 Conclusions

The Markov regression models discussed in this paper provide a general framework for longitudinal binary data analyses where both model order and parameter uncertainty are of special concern. We have also shown how these models can easily be implemented using the Gibbs sampling. The tools developed here can also be applied, with some modifications, to multivariate longitudinal binary data analyses.

Table 3: Posterior means and standard deviations.

	Model-averaged		Second-order	
Parameter	**Case 1**	**Case 2**	**Case 1**	**Case 2**
μ'	-1.95 (0.11)	-2.28 (0.14)	-1.96 (0.11)	-2.28 (0.14)
δ	0.23 (0.05)	0.29 (0.06)	0.23 (0.05)	0.28 (0.06)
γ_1	2.23 (0.25)	2.24 (0.24)	2.19 (0.23)	2.24 (0.23)
γ_2	0.68 (1.34)	1.23 (0.38)	0.82 (0.30)	1.25 (0.32)
γ_3	-0.01 (3.02)	0.06 (2.91)	–	–
β_1	1	1	1	1
β_2	0.83 (0.37)	0.99 (0.05)	1	1
β_3	0.10 (0.30)	0.15 (0.36)	0	0

Acknowledgments

This paper was supported by a grant from the National Institute of Mental Health.

References

Angold, A., Prendergast, M., Cox, A., Harrington, R., Simonoff, E., and Rutter, M. (1995). The Child and Adolescents Psychiatric Assessment (CAPA). *Psychological Medicine*, **25**, 739-753.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P.W. (1975), *Discrete Multivariate Analysis*, Cambridge, MA: MIT Press.

Cargnoni, C., Mueller, P., and West, M. (1997). Bayesian forecasting of multinomial time series through conditionally Gaussian dynamic models. *Journal of the American Statistical Association* **92**, 640-647.

Carlin, B. P., and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods For Data Analysis.* London: Chapman and Hall.

Table 4: Posterior means and standard deviations under the heterogeneous model.

Parameter	**Case 1**	**Case 2**
μ'	-1.47 (1.07)	-1.51 (1.05)
δ	0.44 (0.11)	0.44 (0.11)
γ_1	0.59 (0.42)	0.67 (0.46)
γ_2	-0.19 (2.58)	-0.13 (2.71)
γ_3	-0.18 (2.96)	-0.14 (2.93)
β_1	1	1
β_2	0.35 (0.47)	0.31 (0.45)
β_3	0.17 (0.37)	0.14 (0.35)
τ	0.19 (0.08)	0.20 (0.09)

Carlin, B. P, and Chib, S. (1995). Bayesian model choice via Markov chain monte carlo. *Journal of Royal Statistical Society*, **57**, 473-485.

Costello, E. J., Angold, A., Burns, B., Stangl, D., Tweed, D., Erkanli, A., and Worthman, C. (1996). The Great Smoky Mountains Study of Youth: Prevalence and correlates of DSM-III-R disorders. *Archives of General Psychiatry*, 53, 1129-1136.

Cox, D. R. (1961), *The Analysis of Binary Data*, London: Methuen.

Diggle, J. P., Liang K-Y., and Zeger, S. L. (1994). *Analysis of Longitudinal data.* New York: Oxford university Press.

Erkanli, A., Soyer, R., and Stangl, D. (1997). Bayesian inference in two-phase prevalence studies. *Statistics in Medicine* **16**, 1121-1133.

Erkanli, A., Soyer, R., and Costello, E. J. (1999). Bayesian inference in longitudinal two-phase prevalence studies. *Biometrics.* In press.

Federman, E. B., Costello, E. J., Angold, A., Farmer, E. M. Z., and Erkanli, A. (1997). Development of substance use and psychiatric comorbidity in an epidemiologic study of white and American Indian young adolescents: The Great Smoky Mountains Study. *Drug and Alcohol Dependence*, **44**, 69-78.

Gelfand, A. E., and Smith, A. M. F. (1990). Sampling-based approaches to calculating marginal densities.*Journal of the American Statistical Association*, 85, 398-409.

George, E. I., and McCulloch, R. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881-889.

George, E. I., and McCulloch, R. (1998). Approaches to Bayesian variable selection. *Statistica Sinica*, **7**, 339-374.

Gfellner, B. M., and Hundleby, J. D. (1995). Patterns of drug use among native and white adolescents: 1990-1993. *Canadian Journal of Public Health*, **86**, 95-97.

Gilks W. R., and Wild, P., (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41, 337-348.

Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics*, **53**,, 455-472.

Green, R. (1995). Reversible Markov chain monte carlo computations and Bayesian model determination, *Biometrika*, **82**, 711-732.

Liang, K-Y., and Zeger, S. L. (1989). A class of logistic regression models for multivariate binary time series. *Journal of the American Statistical Association*, **84**, 447-451

Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *Journal of American Statistical Association*, **90**, 773-795.

Kaufmann, H. (1987). Regression models for non-stationary categorical time series: asymptotic estimation theory. *Annals of Statistics*, **15**, 863-871.

Kuo, L., and Mallick, B. K. (1995). Bayesian variable selection. *Sanhkya.* In press.

Smith, M., and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, **75**, 317-367.

Spiegelhalter, D., Thomas, D., Best, N., and Gilks, W. R. (1996). *BUGS Manual 0.5.* Cambridge: MRC Biostatistics Unit.

Spiegelhalter, D., Thomas, D., Best, N., and Carlin, B. P. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. *Journal of Royal Statistical Society, Series B.* In press.

Zeger, S. L. and Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics*, **44**, 1019-1031.

APPENDIX: POSTERIOR COMPUTATIONS

For notational convenience let $\mathbf{y}$ denote the nT dimensional observed data vector, and $D = (\mathbf{y}, Y_0)$, the complete data vector. For a generic parameter θ_m, write $p(\theta_m|\theta_{-m})$ to denote the conditional prior density of θ_m given all the other parameters, θ_{-m}, except θ_m; write $p(\theta_m|\ldots)$ for the conditional posterior density for θ_m given D and all the other parameters except θ_m.

For simulating μ and δ the full conditional posterior densities are:

$$\begin{aligned} p(\mu|\ldots) &\propto L(\theta, Y_0)p(\mu) && (11) \\ p(\delta|\ldots) &\propto L(\theta, Y_0)p(\delta). && (12) \end{aligned}$$

For simulating $\{\gamma_k\}$, for $k = 1, \ldots, T-1$, the full conditionals are:

$$p(\gamma_k|\ldots) \propto \{I(\beta_k = 1)L(\theta, Y_0) + I(\beta_k = 0)\}p(\gamma_k|\gamma_{-k}), \tag{13}$$

where $I(A)$ is the indicator function for the event A. In (13), the likelihood function $L(\theta, Y_0)$ is used for simulating γ_k only if the term y_{it-k} is included in the model, i.e., $\beta_k = 1$. Otherwise, i.e., if $\beta_k = 0$, γ_k is simulated from the conditional prior $p(\gamma_k|\gamma_{-k})$ obtained from the multivariate normal prior $N_{T-1}(\gamma_0, \Sigma)$ given in Section 2.3. Hence, strictly speaking, an improper prior cannot be used for γ.

For simulating β, first note that $p(\beta_1 = 1|\ldots) = 1$, because $p(\beta_1 = 1) = 1$. For $k \geq 2$:

$$\begin{aligned} p(\beta_k = 0|\ldots) &\propto L(\theta_{-k}, \beta_k = 0, Y_0)p(\beta_k = 0|\beta_{-k}) && (14) \\ p(\beta_k = 1|\ldots) &\propto L(\theta_{-k}, \beta_k = 1, Y_0)p(\beta_k = 1|\beta_{-k}). && (15) \end{aligned}$$

The influence of the prior distribution for β introduced in Section 2.4 can be seen in the Bernoulli distributions given in (14) and (15) above.

The conditional posterior distributions of μ, δ and γ (when $\beta_k = 1$) in (11), (12) and (13) above are not available in closed form, while the conditional posterior distribution of β_k is a Bernoulli distribution.

The simulation of the initial prevalences in P_0 is performed using the full conditionals, for each $\tau \in T_0$,

$$p(p_\tau|\ldots) \propto p_\tau^{y_\tau}(1 - p_\tau)^{n - y_\tau} I\{p_\tau \in (0, a)\} \tag{16}$$

where $y_\tau = \sum_{i=1}^{n} y_{i,\tau}$, and $I\{B\}$ is the indicator function of the event B. This is a truncated $\mathrm{Beta}(y_\tau+1, n-y_\tau+1)$ distribution on the interval $(0,a)$, $a<1$, which can be simulated exactly.

To simulate the latent vector Y_0, we make use of the full conditional discrete distribution

$$f(y_{i,\tau}|\ldots) \propto \left\{ \prod_{t=1}^{T-1+\tau} f(y_{it}|y_{it-1},\ldots,y_{it-(T-1)},\theta) \right\} f(y_{i,\tau}|p_\tau), \tag{17}$$

for each $\tau \in T_0$ and $i=1,\ldots,n$, where $f(y_{i,\tau}|p_\tau)$ is the Bernoulli prior described in (6), and $f(y_{i,\tau}|\ldots)$ denotes the conditional posterior density of $y_{i,\tau}$. If P_0 is given, but Y_0 is latent, then only (16) is needed. If Y_0 is observed with error, so that P_0 is unknown, then only (17) is needed, although estimation of θ is independent of that of P_0 in this case. If Y_0 is known with certainty, neither (16), nor (17) are needed. If both Y_0 and P_0 are unknown, then the simulation of θ is augmented using both conditionals in (16) and (17) above. Also note that, in Equation 17, only the conditionals in the bracket corresponding to the "active" β_k's are relevant.

Simulations from (11), (12), and (13) (when $\beta_k = 1$) can be performed in many ways, for example, using the Metropolis-Hastings method (Carlin and Louis, 1996). The Adaptive Rejection Sampling method (ARS) (Gilks et al., 1992) can also be used since these conditionals are also log-concave in μ, δ, and γ_k, respectively. Sampling from γ_k is trivial when $\beta_k = 0$, since (13) is a univariate normal distribution. Simulations from (14), (15), and (16) are also trivial, as they are available in closed analytical forms. In the substance use example below, we used `BUGS` (Spiegelhalter, et al., 1996) software which currently implements both the Adaptive Metropolis (Gilks et al., 1995) and ARS Rejection Sampling methods for the simulations.

Also, if the markov prior in (6) is used, it can be re-expressed for implementing in BUGS as $\beta_k = I(\beta_{k-1}=1)z_k$, where $z_k \sim \mathrm{Bern}(\alpha_k)$ is a latent variable, independent of β_k, and β_{k-1} .

Note that for the heterogeneous model, the Gibbs sampler in Section 3 above will be augmented with conditional distributions $p(\mu_i|\ldots) \propto L(\theta, Y_0)N(0,\nu)$ for $i=1,\ldots,n$, and $p(\nu|\ldots) \propto L(\theta, Y_0)\mathrm{Gamma}(a,b)$.

Figure 1: Posterior mean of the second-order transition probability $P(Y_t^* = 1|y_{t-1}^*, y_{t-2}^*)$ as a function of age and past $H_{it}^* = (y_{t-1}^*, y_{t-2}^*)$. Panel a: **Case 1**, and Panel b: **Case 2**.