Marginalized Transition Models and Likelihood Inference for Longitudinal Categorical Data

Patrick J. Heagerty
Department of Biostatistics, University of Washington, Seattle, Washington 98195, U.S.A.
e-mail: heagerty@u.washington.edu

SUMMARY. Marginal generalized linear models are now frequently used for the analysis of longitudinal data. Semiparametric inference for marginal models was introduced by Liang and Zeger (1986, Biometrics 73, 13–22). This article develops a general parametric class of serial dependence models that permits likelihood-based marginal regression analysis of binary response data. The methods naturally extend the first-order Markov models of Azzalini (1994, Biometrika 81, 767–775) and prove computationally feasible for long series.

KEY WORDS: Marginal model; Markov chain; Maximum likelihood.

1. Introduction
In a study of the natural history of schizophrenia, first-episode patients had disease symptoms recorded monthly for up to 10 years following initial hospitalization (Thara et al., 1994). In studies of the health effects of air pollution, asthmatic children recorded the presence or absence of wheezing each day for up to 6 months (Yu et al., 2000). To determine whether maternal employment correlates with pediatric care utilization and both maternal stress and childhood illness, daily measurements of maternal stress (yes, no) and childhood illness (yes, no) were recorded for 28 consecutive days (Alexander and Markowitz, 1986). In each of these examples, the primary scientific question pertains to the association between specific covariates and a high-dimensional binary response vector. Although several analysis options exist for discrete longitudinal data, there are few likelihood-based methods that can accommodate such increasingly common long categorical series. In this article, we briefly review statistical methods that are available and then introduce and motivate a new likelihood-based approach.

The seminal work of Liang and Zeger (1986) introduced estimating equations as a method of inference for the analysis of either continuous or discrete longitudinal data. The generalized estimating equations (GEE) approach adopts a semiparametric model in which only the marginal mean and the covariance of repeated measurements is specified. A complete multivariate probability model is not uniquely identified and therefore a likelihood function is not available. Semiparametric methods are attractive because estimates of regression coefficients and estimates of standard errors are valid under minimal model assumptions. However, likelihood-based methods remain attractive for several reasons, including optimality of estimators under correct model specification, the availability of inferential procedures such as likelihood ratio tests, and the robustness of likelihood-based inference to certain dropout or missing data mechanisms. GEE is sometimes criticized for its deficiency in each of these aspects: although valid inference can be obtained with any working correlation model, the resulting regression estimates may be inefficient when a poor correlation model is selected; few empirical or formal methods are available to select an appropriate correlation model; and valid estimation requires that data be missing completely at random.

In this article, we consider situations where the marginal mean regression structure is the primary target of inference. We explore the potential for flexible and computationally feasible likelihood-based analysis of serial categorical data. With continuous longitudinal response data, the multivariate Gaussian model conveniently permits likelihood-based regression inference. However, for binary longitudinal data, no simple solution exists. Generalized linear mixed models (Breslow and Clayton, 1993) have been proposed, but it is well known that these models traditionally focus on a conditional mean regression parameter rather than a marginal mean parameter. Fitzmaurice and Laird (1993) considered a reparameterization of canonical log-linear models, but their methods are computationally restricted to small and balanced longitudinal series. Molenberghs and Lesaffre (1994) studied a class of marginal models for multivariate categorical data that are also restricted in application to a small number of repeated measurements. Azzalini (1994) introduced a first-order Markov model that allowed regression analysis of equally spaced binary measurements. Although the methods of Azzalini (1994) can be applied to long measurement series, the class of models was limited to first-order Markov models. Recently, Heagerty and Zeger (2000) have classified the methods of Azzalini (1994) as a form of marginalized transition model in which a transition model that characterizes serial dependence is combined with a marginal generalized linear model that describes...
2. Marginalized Transition Models

First, we review the Markov chain models of Azzalini (1994) and describe an alternative, but equivalent, model specification in terms of the combination of a marginal regression model used to characterize the dependence of the response on covariates and a conditional regression model, or transition model (Diggle, Liang, and Zeger, 1994), used to capture the serial dependence in the response process. We then generalize the first-order marginalized transition model to allow pth-order dependence. We show that, under first-order dependence assumptions, the marginalized transition model of Azzalini (1994) is a special case of the mixed parameter model of Fitzmaurice and Laird (1993) and therefore maximum likelihood estimates have desirable robustness properties. Our focus is on binary response data \( Y_i = (Y_{i1}, \ldots, Y_{in_i}) \) observed on subjects \( i = 1, \ldots, N \) at times \( t = 1, \ldots, n_i \). We assume that associated exogenous but possibly time-varying covariates, \( X_{it} = (X_{it,1}, \ldots, X_{it,p}) \), are recorded for each subject at each occasion. Our statistical objective is to obtain estimates that allow a general pth-order dependence structure. We discuss aspects of estimation, robustness, and model checking.

## 2.1 First-Order Marginalized Transition Model

Azzalini (1994) introduced a binary Markov chain model to accommodate the serial dependence that is common in longitudinal data. A first-order Markov model assumes that the current response variable is dependent on the history only through the immediate previous response, \( E(Y_{it} \mid Y_{it-j}, j < t) = E(Y_{it} \mid Y_{it-1}) \). The transition probabilities \( p_{it,0} = E(Y_{it} \mid Y_{it-1} = 0) \) and \( p_{it,1} = E(Y_{it} \mid Y_{it-1} = 1) \) define the Markov process but do not directly parameterize the marginal mean. Azzalini (1994) parameterizes the transition probabilities through two assumptions. First, a marginal mean regression model is adopted that constrains the transition probabilities to satisfy

\[
\mu_{it}^M = p_{it,1}^M \mu_{it-1}^M + p_{it,0}^M \left( 1 - \mu_{it-1}^M \right).
\]

Second, the transition probabilities are structured through assumptions on the pairwise odds ratio,

\[
\Psi_{it} = \frac{p_{it,1}/(1-p_{it,1})}{p_{it,0}/(1-p_{it,0})},
\]

which quantifies the strength of serial correlation. The simplest dependence model assumes a time-homogeneous association, \( \Psi_{it} = \Psi_0 \); however, models that allow \( \Psi_{it} \) to depend on covariates or to depend on time are also possible.

The transition probabilities, and therefore the likelihood, can be recovered as a function of the marginal means, \( \mu_{it}^M \), and the odds ratios, \( \Psi_{it} \). Azzalini (1994) provides details on the calculations required for maximum likelihood estimation and establishes the orthogonality of the marginal mean and the odds ratio parameter in the restricted case of a time-constant (scalar) dependence model.

Heagerty and Zeger (2000) view the approach of Azzalini (1994) as combining a marginal mean model that captures systematic variation in the response as a function of covariates with a conditional mean model that describes serial dependence and identifies the joint distribution of \( Y_t \). The first-order marginalized transition model, or MTM(1), is specified by first assuming a regression structure for the marginal mean, \( E(Y_{it} \mid X_{it}) \), using a generalized linear model,

\[
g(\mu_{it}^M) = X_{it} \beta.
\]

Next, the serial dependence is specified by assuming a Markov structure or equivalently by assuming a regression model for \( \log \Psi_{it} \). Heagerty and Zeger (2000) describe the dependence model using

\[
\logit \left( \frac{\mu_{it}^C}{1 - \mu_{it}^C} \right) = E(Y_{it} \mid \beta) = \Delta_{it} + \gamma_{it,1} \Psi_{it} - 1.
\]

where \( \gamma_{it,1} = \log \Psi_{it} \). The log odds ratio \( \gamma_{it,1} \) is simply a logistic regression coefficient in the model that conditions on both \( X_{it} \) and \( Y_{it-1} \). The parameter \( \Delta_{it} \) equals \( \logit(p_{it,0}) \) and is determined implicitly by \( \beta \) and \( \gamma_{it,1} \) through equations (1) and (2). Furthermore, a general regression model can be specified for \( \gamma_{it,1} \),

\[
\gamma_{it,1} = Z_{it,1} \alpha_1,
\]

where the parameter \( \alpha_1 \) determines how the dependence of \( Y_{it} \) on \( Y_{it-1} \) varies as a function of covariates, \( Z_{it,1} \). For example, \( \gamma_{it,1} = \alpha_t \) allows serial dependence to change over time and \( \gamma_{it,1} = \alpha_0 + \alpha_1 Z_t \) allows subjects for whom \( Z_t = 1 \) to have a different serial correlation compared with subjects for whom \( Z_t = 0 \). In general, \( Z_{it} \) is a subset of \( X_{it} \) since we assume that equation (3) denotes the conditional expectation of \( Y_{it} \) given both \( X_{it} \) and \( Y_{it-1} \). Furthermore, similar to assumptions regarding \( \mu_{it}^M \), we assume that \( \mu_{it}^C \) is a conditional expectation where we condition on the full covariate series \( (X_{it,1}, \ldots, X_{it,n_i}) \) via the implicitly defined \( \Delta_{it} \).

In summary, the marginalized transition model separates the specification of the dependence of \( Y_{it} \) on \( X_{it} \) (regression) and the dependence of \( Y_{it} \) on the history \( Y_{it-1}, Y_{it-2}, \ldots, Y_{it} \) (autocorrelation) to obtain a fully specified parametric model for longitudinal binary data. A first-order model assumes that \( Y_{it} \) is conditionally independent of \( Y_{it-2}, \ldots, Y_{it} \) given \( Y_{it-1} \). The transition model intercept, \( \Delta_{it} \), is determined such that both the marginal mean structure and the Markov dependence structure are simultaneously satisfied.
2.2 General Marginalized Transition Model

Equations (3) and (4) indicate how the first-order dependence model can naturally be extended. We may assume that $Y_{it}$ depends on the history only through the previous $p$ responses, $Y_{it-1}, \ldots, Y_{it-p}$. A $p$th-order dependence model, or MTM($p$), combines the marginal mean model with

$$
\mu_{it}^C = \text{E}(Y_{it} \mid X_{it}, Y_{ij} = y_{ij} \ j < t) \quad (6)
$$

$$
\logit(\mu_{it}^C) = \Delta_{it} + \sum_{j=1}^{p} \gamma_{it,j} y_{it-j} \quad (7)
$$

$$
\gamma_{it,j} = Z_{it,j} \alpha_j \quad j = 1, \ldots, p. \quad (8)
$$

For example, a second-order (additive) marginalized transition model assumes $\logit(\mu_{it}^C) = \Delta_{it} + \gamma_{it,1} y_{it-1} + \gamma_{it,2} y_{it-2}$, and $\gamma_{it,1} = Z_{it,1} \alpha_1$, $\gamma_{it,2} = Z_{it,2} \alpha_2$. Although $\mu_{it}^C$ can also depend on the interaction $Y_{it-1} Y_{it-2}$, we assume an additive model for simplicity of presentation. Finally, we also assume that covariates $Z_{it,j}$ are a subset of $X_{it}$.

In the MTM(2), the mean parameter $\beta$ describes changes in the average response as a function of covariates without controlling for previous response variables. Figure 1 is a diagram that represents a second-order marginalized transition model. The dashed boxes indicate that the regression parameter $\beta$ models the marginal relationship between the response and covariates. Markov transition assumptions are represented by the directed edges connecting $Y_{it-j}$ to $Y_{it}$. The dependence parameters $\gamma_1$ and $\gamma_2$ describe serial dependence by quantifying how strongly the immediate past predicts the present.

The parameter space for the marginalized transition model is subject only to the constraint that the generalized linear model for $\mu_{it}^M$ must yield $\mu_{it}^M \in [0, 1]$. Using a logistic link for the transition model, $\mu_{it}^C$, implies that the parameters $\alpha_j$ are unconstrained. For a given mean model ($\beta$) and dependence model ($\alpha_j$), the intercept parameter $\Delta_{it}$ is fully constrained and must yield the proper marginal expectation $\mu_{it}^C$ when $\mu_{it}^C$ is averaged over the distribution of the history. For any finite values of $\alpha_j$, as $\Delta_{it}$ ranges from $-\infty$ to $+\infty$, the induced marginal mean monotonically increases from zero to one. Therefore, given any finite-valued dependence model and any probability distribution for the history, a unique $\Delta_{it}$ can be identified that satisfies both the transition model and the marginal mean assumptions.

The marginalized transition model for longitudinal categorical data has several desirable features similar to those of the multivariate Gaussian model for continuous response data. First, the mean model is separated from the dependence model. As a result, the interpretation of the regression parameter $\beta$ does not change as we modify assumptions regarding the dependence order. This is not true for traditional transition models, which parameterize $\mu_{it}^C$ directly as a function of covariates. For example, a direct second-order transition model may take the form $\logit(\mu_{it}^C) = X_{it} \beta^C + \gamma_1 Y_{it-1} + \gamma_2 Y_{it-2}$. Here the parameters $\gamma_j$ describe the dependence of $Y_{it}$ on the response history. However, the regression parameters $\beta^C$ contrast covariate subgroups after controlling for the previous response variables. As we change the order of the transition model, we change the conditioning variables and therefore change the interpretation of $\beta^C$. Second, the joint distribution is determined by a pair of regression assumptions. The multivariate Gaussian model, $Y_{it} \sim N(\mu_{it}, \Sigma_{it})$, can equivalently be specified through the marginal mean, $E(Y_{it} \mid X_{it}) = \mu_{it}$, and conditional assumptions including normality, conditional variances, and the coefficients $\gamma_{it,j}$ in the conditional mean $E(Y_{it} \mid X_{it}, Y_{ij} = y_{ij} \ j < t) = \Delta_{it} + \Sigma_{j=1}^{(t-1)} \gamma_{it,j} y_{it-j}$. The intercept $\Delta_{it}$ is a function of $\mu_{it}$ and $\Sigma_{it}$ (or equivalently $\gamma_j$). Finally, the marginalized transition model can be used with data where subjects have variable lengths of follow-up, permitting likelihood analysis in settings where data may be missing at random (MAR) (Laird, 1988).

2.2.1 MTM($p$) and pr($Y_{11}, \ldots, Y_{ip}$). One disadvantage to directly using a transition model with $p$ lagged response variables is that the information in $(Y_{11}, \ldots, Y_{ip})$ is conditioned upon, and therefore does not contribute to the assessment of covariate effects. With an MTM($p$) approach, information in the initial responses regarding $\mu_{it}^M$, and thus $\beta$, is included through lower order marginalized transition models involving $E(Y_{ij} \mid X_{ij}, Y_{i1}, \ldots, Y_{ij-1})$ for $j < p$. For example, when using an MTM(2), the likelihood for the initial responses, $(Y_{i1}, Y_{i2})$, is obtained by factorization into $pr(Y_{i1} \mid X_{i1})$ (determined entirely by $\mu_{i1}^M$), and $pr(Y_{i2} \mid X_{i2}, Y_{i1})$, which involves $\mu_{i2}^M$ and the conditional model $logit(Y_{i2} \mid X_{i2}, Y_{i1}) = \Delta_{i2} + \gamma_{i2,1} Y_{i1}$. Note that $\gamma_{i2,1}$ is distinct from the first-order coefficient $\gamma_{i1,1}$, $t > 2$, in the MTM(2) dependence model, logit($\mu_{it}^C$) = $\Delta_{it} + \gamma_{it,1} Y_{it-1} + \gamma_{it,2} Y_{it-2}$, since this model involves $Y_{it-2}$ in addition to $Y_{it-1}$. Therefore, in applications, we have estimated separate lower order dependence parameters.

One practical issue when using covariates, $Z_{it,j}$, in the dependence model is whether one would assume common covariate effects for the initial state dependence model, $\gamma_{i1,1} = \alpha_{i1,0} + \alpha_{i1,1} Z_{it,1}$, and the subsequent dependence model, $\gamma_{it,1} = \alpha_{it,0} + \alpha_{it,1} Z_{it,1}$, i.e., one may consider assuming $\alpha_{i1,1} = \alpha_{i1,1}$, but since this requires connecting...
parameters from distinct models, it would need to be carefully justified.

2.3 Maximum Likelihood Estimation

The MTM(p) likelihood factors into the distribution of the first p response variables times the subsequent Bernoulli likelihood contributions with parameters \( \mu_{ij}^C \) for \( t = (p+1), \ldots, n_i \). The basic maximization algorithm starts with a model for \( pr(Y_{ij1}, \ldots, Y_{ijp}) \) using lower order marginalized transition model assumptions. The key to subsequent likelihood evaluation is that transition probabilities, \( \mu_{ij}^C \), can be sequentially recovered as a function of the parameters \( \beta \) and \( \alpha_1, \ldots, \alpha_p \). The dependence parameters enter into \( \mu_{ij}^C \) directly, but the intercept \( \Delta_{it} \) is an implicit function of \( \beta \) and \( \alpha_1, \ldots, \alpha_p \). We obtain \( \Delta_{it} \) by solving the marginal constraint equation,

\[
\mu_{il}^M = \sum_{y_{it-1}, \ldots, y_{it-p}} pr(Y_{it} = 1 | Y_{it-1} = y_{it-1}, \ldots, Y_{it-p} = y_{it-p}) \times pr(Y_{it-1} = y_{it-1}, \ldots, Y_{it-p} = y_{it-p})
\]

In order to obtain a solution, we require the initial state probability \( pr(Y_{i1} = y_{i1}, \ldots, Y_{ip} = y_{ip}) \) from which all subsequent p-dimensional probabilities can be obtained by multiplying \( pr(Y_{ij} = y_{ij}, \ldots, Y_{ij+(p-1)} = y_{ij+(p-1)}) \) times \( \mu_{ij+p} \) and then summing over \( y_{ij} \). Details for MTM(2) are provided in Appendix B.

The computational complexity of MTM(p) likelihood evaluation for subject i is \( O(n_i 2^p) \). Calculations required to compute and update the p-dimensional history are \( O(2^p) \), and each observation requires such calculations. Therefore, with a fixed dependence order, the computational burden only increases linearly with the length of the observation series, \( n_i \). The order of alternative likelihood methods is generally much greater with the iterative proportional fitting algorithm of Fitzmaurice and Laird (1993), requiring exponentially increasing computational time, \( O(2^{n_i}) \).

Azzalini (1994) suggested that the MTM(1) has general robustness properties but only established a consistency result for a restricted scenario. Viewing the MTM(1) as adopting a logistic regression for \( \mu_{ij}^C \) allows us to show that the MTM(1) is a special case of the mixed-parameter model of Fitzmaurice and Laird (1993), with \( \gamma = vec(\gamma_{it,1}) \) the canonical log-linear interaction parameter. Appendix A provides details that show \( \beta \) and \( \alpha_1 \) are orthogonal. The implication of orthogonality is that the maximum likelihood estimate, \( \hat{\beta} \), remains consistent for \( \beta \) even if the dependence model is incorrectly specified. Use of the MTM(1) maximum likelihood estimate \( \hat{\beta} \) and a sandwich variance estimator (Huber, 1967; White, 1982) provides a likelihood-motivated version of GEE appropriate in serial data situations since the point estimate \( \hat{\beta} \) will be consistent and valid standard errors can be obtained without requiring correct Markov order or correct modeling of \( \gamma_{it,1} \).

For general \( p \)-th order models, \( \beta \) and \( (\alpha_1, \ldots, \alpha_p) \) may not be orthogonal, and consistent estimation of mean regression parameters requires appropriate dependence modeling.

One practical advantage of marginalized transition modeling is that several simple procedures can be used to assess the dependence model assumptions. To establish the appropriate order, we can use direct transition models and regress \( Y_{it} \) on \( X_{it} \) and prior responses. Approximate score tests for an additional lagged response (increased order from \( p \) to \( p+1 \)) take simple forms. For example, using a MTM(p), we can test the assumption \( \alpha_{p+1} = \gamma_{it,p+1} = 0 \) with the statistic

\[
U_{p+1} = \sum_{i=1}^{N} \sum_{t=1}^{n_i} Y_{it-(p+1)} (Y_{it} - \mu_{it}^C)
\]

where \( \mu_{it}^C \) is the fitted conditional mean. This statistic only approximates the likelihood score statistic because it ignores \( \partial \Delta_{it}/\partial \alpha_{p+1} \). The approximate score statistic is intuitive since it simply evaluates the correlation between the \( (p+1) \)st lagged response and the conditional residual obtained by fitting a marginalized transition model with the first \( p \) lagged responses.

2.4 Maximum Likelihood Versus GEE

Since GEE is a widely available method that can be used to estimate marginal regression parameters, the additional effort required for likelihood estimation warrants justification. In this section, we consider the efficiency of GEE methods relative to maximum likelihood and consider the bias that obtains in the use of GEE when data are missing at random (MAR) rather than missing completely at random (MCAR).

2.4.1 MCAR with a group-by-time design. To compare the efficiency of GEE and maximum likelihood estimators, we generated longitudinal binary data with a group-by-time regression structure,

\[
\logit E(Y_{it} | X_{it}) = \beta_0 + \beta_1 time_{it} + \beta_2 group_i + \beta_3 group_i time_{it},
\]

where \( t = 1, 2, \ldots, 20 \), \( time_{it} = (t - 1)/19 \), and \( group_i = 0 \) or 1 for all times \( t \). We used 100 subjects with \( group = 0 \) and 100 subjects with \( group = 1 \). Missing data was created by randomly generating the number of observations for each subject as uniform between 5 and 20. A second-order dependence model with \( \alpha_1 = 2.5 \) and \( \alpha_2 = 1.0 \) was used to induce serial correlation. We compare the efficiency of three GEE estimators and a MTM(1) ML estimator relative to the correctly specified MTM(2) ML estimator for \( \beta = (-1.50, 0.50, 0.50, -0.25) \). Table 2 shows that the variance of the MTM(2) ML estimator was between 86.7 and 89.7% of the variance obtained using GEE with working independence. A similar modest amount of inefficiency is observed for GEE with an exchangeable correlation model, while GEE with an AR(1) correlation results in efficiencies greater than 98%. The MTM(1) ML estimator is also highly efficient relative to the correct MTM(2) estimator.

2.4.2 MCAR with a cross-over design. A second simulation was conducted based on a cross-over design. In this model, the group indicator is time dependent and takes the value one for observations \( t = 1, 2, \ldots, 10 \) and then switches to the value zero for observations \( t = 11, 12, \ldots, 20 \). A linear trend in time is also assumed: \( time_{it} = (t - 1)/19 \). The mean model was \( \logit E(Y_{it} | X_{it}) = \beta_0 + \beta_1 time_{it} + \beta_2 group_i \) and a second-order dependence model was used with \( \alpha_1 = 2.5 \) and \( \alpha_2 = 1.0 \). A total of 200 subjects was simulated with \( n_i \) uniform in [5, 20] and with an equal number of subjects for whom \( group_i \) switches from zero to one and subjects for whom \( group_i \) switches from one to zero. Table 2 shows that the GEE estimators with working independence and working exchangeable correlation models yield poor relative efficiencies of 64.9 and 60.6%, respectively. Again, we find that
AR(1) GEE and MTM(1) estimators retain high efficiency. Therefore, GEE estimators that adopt a working correlation structure that approximates the true correlation structure obtain high efficiency while poor correlation choices may be inefficient.

2.4.3 MAR with a group-by-time design. Finally, we consider the performance of GEE and ML estimators when the data are missing at random (MAR). Specifically, we consider a monotone missing data mechanism where all observations are missing after a dropout time and where the probability of dropout depends on the past response values. Specifically, we generated response vectors according to the dropout model logit\(P(\text{dropout} = t | \text{dropout} \geq t) = -2.5 + 0.75Y_{t-1} + 0.5Y_{t-2}\). Such a process is termed MAR for longitudinal data (Laird, 1988), and it is well known that correctly specified ML estimators provide consistent parameter estimates, while available case moment estimators such as GEE may be biased. An alternative to likelihood inference in MAR situations is inverse probability weighted GEE (Robins, Rotnitzky, and Zhao, 1995), which requires specification and estimation of a selection model for the missing data mechanism. We used the same group-by-time mean model and second-order dependence model that was used to evaluate efficiency under MCAR as described in Section 2.4.1. Table 3 presents the average regression estimates and the percent relative bias. We find substantial bias in the GEE estimators. Bias in the coefficient of time ranges from −52.6 to +90.0%, depending on the choice of the working correlation model. Similarly, we find biases in the interaction coefficient estimators that range from −52.6 to +90.0%. Bias also obtains for the incorrect MTM(1) ML estimator, while the correctly specified ML estimator is approximately unbiased. Therefore, when data are incomplete due to dropout that is associated with observed outcomes (i.e., is MAR), GEE estimates may be biased, while a proper likelihood analysis can yield valid estimates.

### 3. Example

In this section, we analyze data from the Madras Longitudinal Schizophrenia Study, which investigated the course of positive and negative psychiatric symptoms over the first year after initial hospitalization for disease (Thara et al., 1994). Scientific interest is in factors that correlate with the course of illness. Our analysis addresses whether the rate of decline in symptom prevalence differs across patient subgroups defined by age-at-onset and gender. To statistically address the scientific questions, we use marginalized transition models to estimate the interaction between time and age-at-onset and the interaction between time and gender in a marginal logistic regression model. We compare parameter estimates obtained using maximum likelihood under different dependence assumptions and estimates obtained using GEE.

The outcome \(Y_{it}\) is the presence or absence of the symptom thought disorders recorded each month, \(t = 0, \ldots, 11\), during the first year following hospitalization for schizophrenia. Not surprisingly, a large fraction of the \(N = 86\) subjects have symptoms at the time of hospitalization (56/86 = 65% at month = 0). The crude prevalence of thought disorders decreases during the first year, with only 6/69 = 9% presenting symptoms at month = 11. To evaluate if the recovery differs for subjects with older age-at-onset or for women, we fit a marginal logistic regression model with main effects for month (\(t = 0, \ldots, 11\)), age (0 = age-at-onset ≥ 20, 1 = age-at-onset < 20), gender (0 = male, 1 = female), and the interaction between time and the two subject-level covariates. These data contain 42/86 female subjects and 30/86 subjects with early age-at-onset. Evaluation of the interaction terms addresses whether the course of recovery appears to differ by subgroup.

#### Table 1

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>GEE independence</th>
<th>GEE exchangeable</th>
<th>GEE AR(1) MTM(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.897</td>
<td>0.894</td>
<td>0.986</td>
</tr>
<tr>
<td>Time</td>
<td>0.867</td>
<td>0.826</td>
<td>0.992</td>
</tr>
<tr>
<td>Group</td>
<td>0.885</td>
<td>0.879</td>
<td>0.980</td>
</tr>
<tr>
<td>Group × time</td>
<td>0.885</td>
<td>0.840</td>
<td>0.990</td>
</tr>
</tbody>
</table>

#### Table 2

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>GEE independence</th>
<th>GEE exchangeable</th>
<th>GEE AR(1) MTM(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.802</td>
<td>0.751</td>
<td>0.973</td>
</tr>
<tr>
<td>Time</td>
<td>0.835</td>
<td>0.835</td>
<td>0.973</td>
</tr>
<tr>
<td>Group</td>
<td>0.649</td>
<td>0.606</td>
<td>0.953</td>
</tr>
</tbody>
</table>
Bias of GEE estimators and MTM maximum likelihood estimators when data are missing at random (MAR) for a group-by-time analysis. For 1000 simulated data sets, response vectors for 100 subjects with group = 0 and 100 subjects with group = 1 were generated using a second-order marginalized transition model and then subject to monotone missingness using the model logit[\(P(\text{dropout} = t \mid \text{dropout} \geq t)\) = \(-2.5 + 0.75Y_{it-1} + 0.5Y_{it-2}\). Estimates were obtained using MTM(2), the correctly specified ML estimator; MTM(1); and GEE using independence, exchangeable, or AR(1) correlation models. Displayed is the average regression coefficient estimate and relative bias, 100 \times (\hat{\beta} - \beta)/\beta for \(\beta = (-1.5, 0.5, 0.5, -0.25)\).

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>GEE independence</th>
<th>GEE exchangeable</th>
<th>GEE AR(1)</th>
<th>ML MTM(1)</th>
<th>ML MTM(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.519 (+1.3)</td>
<td>-1.537 (+2.5)</td>
<td>-1.503 (-0.2)</td>
<td>-1.504 (-0.3)</td>
<td>-1.502 (-0.1)</td>
</tr>
<tr>
<td>Time</td>
<td>0.237 (-52.6)</td>
<td>0.950 (+90.0)</td>
<td>0.373 (-25.4)</td>
<td>0.364 (-27.2)</td>
<td>0.487 (-2.6)</td>
</tr>
<tr>
<td>Group</td>
<td>0.493 (-1.4)</td>
<td>0.485 (-3.0)</td>
<td>0.494 (-1.2)</td>
<td>0.496 (-0.8)</td>
<td>0.495 (-1.0)</td>
</tr>
<tr>
<td>Group \times time</td>
<td>-0.295 (+18.0)</td>
<td>-0.186 (-25.8)</td>
<td>-0.280 (-12.0)</td>
<td>-0.267 (+6.8)</td>
<td>-0.244 +2.4</td>
</tr>
</tbody>
</table>

The majority of subjects have complete data, but 17/86 subjects have only partial follow-up that ranges from 1 to 11 months. Regression analysis of subject discontinuation (dropout) suggests that subjects who currently have symptoms, \(Y_{it} = 1\), are at increased risk to discontinue at time \(t + 1\) (odds ratio 1.716, \(p\text{-value} = 0.092\)). However, if an equally valid working independence model is used, we obtain \(\hat{\beta}_0 = -0.113\), with \(Z = -0.113/0.096 = 1.18\), \(p\text{-value} = 0.239\). Unfortunately, the choice of working dependence model can impact point estimates and significance levels, and without objective criteria, we cannot formally choose among various asymptotically valid estimators.

By using a likelihood-based method rather than a semiparametric method, we are able to compare alternative dependence models using likelihood ratios. Table 5 presents maximum likelihood estimates, adopting both first-order and second-order marginalized transition models. The simplest marginalized transition model is the first-order time homogeneous dependence model, \(\gamma_{it,1} = \alpha_{1,0}\). Point estimates and standard errors for \(\beta\) are quite close to those obtained using GEE. The estimated first-order coefficient \(\hat{\alpha}_{1,0} = 3.166\) indicates that the odds of symptoms at \(\text{month} = t\) are \(\exp(3.166)\) times greater among subjects who previously had symptoms, \(Y_{t-1} = 1\), compared with subjects who previously did not have symptoms, \(Y_{t-1} = 0\). Model 2 allows the first-order coefficient to depend on covariates, \(\gamma_{it,1} = \alpha_{1,0} + \alpha_{1,1}\text{month} + \alpha_{1,2}\text{age} + \alpha_{1,3}\text{gender}\), and suggests that serial dependence is increasing over time, \(\hat{\alpha}_{1,1} = 0.173\), with \(Z = 0.173/0.081 = 2.135\), \(p\text{-value} = 0.032\). The observed time trend in serial dependence is expected in situations where patients stabilize (either with or without symptoms). Comparing the deviances,

Table 3

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coef.</th>
<th>Mod. SE</th>
<th>Emp. SE</th>
<th>Coef.</th>
<th>Mod. SE</th>
<th>Emp. SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marginal mean, (\beta)</td>
<td>0.643</td>
<td>0.202</td>
<td>0.305</td>
<td>0.553</td>
<td>0.296</td>
<td>0.291</td>
</tr>
<tr>
<td>Month</td>
<td>-0.254</td>
<td>0.038</td>
<td>0.059</td>
<td>-0.235</td>
<td>0.053</td>
<td>0.055</td>
</tr>
<tr>
<td>Age</td>
<td>0.811</td>
<td>0.305</td>
<td>0.493</td>
<td>0.638</td>
<td>0.440</td>
<td>0.461</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.388</td>
<td>0.286</td>
<td>0.449</td>
<td>-0.161</td>
<td>0.412</td>
<td>0.420</td>
</tr>
<tr>
<td>Month \times age</td>
<td>-0.137</td>
<td>0.064</td>
<td>0.094</td>
<td>-0.101</td>
<td>0.089</td>
<td>0.085</td>
</tr>
<tr>
<td>Month \times gender</td>
<td>-0.113</td>
<td>0.063</td>
<td>0.096</td>
<td>-0.150</td>
<td>0.090</td>
<td>0.089</td>
</tr>
</tbody>
</table>

\(\alpha\) The estimated lag-1 correlation is \(\hat{\rho} = 0.590\).
Table 5

Maximum likelihood estimates using marginalized transition models for schizophrenia symptoms

<table>
<thead>
<tr>
<th>Variable</th>
<th>MTM(1), model 1</th>
<th>MTM(1), model 2</th>
<th>MTM(2), model 3</th>
<th>MTM(2), model 4a</th>
<th>MTM(2), model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
<td>SE</td>
<td>Coef.</td>
<td>SE</td>
<td>Coef.</td>
</tr>
<tr>
<td>Marginal mean, ( \beta )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.534</td>
<td>0.300</td>
<td>0.576</td>
<td>0.285</td>
<td>0.576</td>
</tr>
<tr>
<td>Month</td>
<td>-0.236</td>
<td>0.054</td>
<td>-0.231</td>
<td>0.103</td>
<td>-0.238</td>
</tr>
<tr>
<td>Age</td>
<td>0.650</td>
<td>0.422</td>
<td>0.626</td>
<td>0.432</td>
<td>0.588</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.142</td>
<td>0.413</td>
<td>-0.197</td>
<td>0.404</td>
<td>-0.150</td>
</tr>
<tr>
<td>Month ( \times ) age</td>
<td>-0.112</td>
<td>0.086</td>
<td>-0.100</td>
<td>0.092</td>
<td>-0.101</td>
</tr>
<tr>
<td>Month ( \times ) gender</td>
<td>-0.144</td>
<td>0.083</td>
<td>-0.144</td>
<td>0.089</td>
<td>-0.140</td>
</tr>
<tr>
<td>First-order coefficient, ( \alpha_1 )</td>
<td></td>
<td></td>
<td>2.161</td>
<td>0.500</td>
<td>2.911</td>
</tr>
<tr>
<td>Intercept</td>
<td>3.166</td>
<td>0.228</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial</td>
<td>0.173</td>
<td>0.081</td>
<td>0.222</td>
<td>0.117</td>
<td>0.156</td>
</tr>
<tr>
<td>Month</td>
<td>0.168</td>
<td>0.480</td>
<td>0.560</td>
<td>0.600</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.274</td>
<td>0.481</td>
<td>0.369</td>
<td>0.555</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second-order coefficient, ( \alpha_2 )</td>
<td></td>
<td></td>
<td>0.650</td>
<td>0.295</td>
<td>1.303</td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial</td>
<td></td>
<td></td>
<td>-0.081</td>
<td>0.119</td>
<td></td>
</tr>
<tr>
<td>Month</td>
<td></td>
<td></td>
<td>-0.438</td>
<td>0.663</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td>-0.447</td>
<td>0.630</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximized log likelihood</td>
<td>-337.19</td>
<td>-334.43</td>
<td>-334.44</td>
<td>-331.66</td>
<td>-332.93</td>
</tr>
</tbody>
</table>

\( \Delta D = 2 \times (337.19 - 334.43) = 5.52 \), indicates that the additional flexibility of model 2 (three additional degrees of freedom) is marginally warranted (\( p = 0.137 \)). Changes in the first-order dependence assumptions have a minor impact on the point estimates, \( \hat{\beta} \), and standard errors. Recall that, since the MTM(1) has \( \beta \) orthogonal to \( \alpha \), the resulting maximum likelihood estimates, \( \hat{\beta} \), are consistent even if the dependence model is incorrectly specified.

To evaluate the specification of the dependence order, we can compute score tests using only the MTM(1) fit or use likelihood ratio tests comparing second-order with first-order models. The score test of \( \gamma_{it,2} = \alpha_{2,0} = 0 \) obtained from model 1 is \( U_2 = 0.926 \), \( p \)-value=0.336. Model 3 is a second-order marginalized transition model with scalar first- and second-order coefficients. The first-order coefficient model includes the variable initial, an indicator variable for month = 1, allowing \( \alpha_1 \) to be used for both the second-order model, \( Y_{it} = Y_{it-1}, Y_{it-2} \), and the initial state, \( Y_{it} = Y_{i0} \), which is a purely first-order distribution. The second-order coefficient, \( \hat{\alpha}_{2,0} = 0.638 \), is significant based on the Wald statistic, \( Z = 0.638/0.296 = 2.155 \), \( p \)-value = 0.031. Comparison of deviances yields \( \Delta D = 2 \times (337.19 - 334.44) = 5.50 \), \( p \)-value = 0.064 on 2 d.f. Model 4 allows both \( \gamma_{it,1} \) and \( \gamma_{it,2} \) to depend on covariates and yields a modest decrease in the deviance (\( \Delta D = 5.56 \) on 6 d.f.). A simplified second-order model, \( \gamma_{it,1} = \alpha_{1,0} + \alpha_{1,0,initial} + \alpha_{1,0,month} \) and \( \gamma_{it,2} = \alpha_{2,0} \), evaluates whether both second-order effects and first-order time trends are supported. Model 5 yields a maximized log likelihood of -332.93 and \( \hat{\alpha}_{1,3} = 0.156 \), \( Z = 0.156/0.096 = 1.625 \), and reduction in deviance of \( \Delta D = 2 \times (334.44 - 332.93) = 3.02 \) on 1 d.f. (\( p \)-value = 0.082) relative to model 3. Model 5 also achieves the smallest Akaike Information Criterion value among the first-order and second-order models considered. A score test for third-order effects, \( U_3 = 0.072 \), indicates the adequacy of a second-order model.

The marginalized transition model fitted using maximum likelihood allows a thorough model-based analysis of schizophrenia symptoms. We use the maximized log likelihood to establish an appropriate dependence model and then evaluate the evidence regarding differences in the disease course across subgroups defined by age-at-onset and gender. We find that a second-order time inhomogeneous dependence model is appropriate. The estimated rate of decline in the log odds of symptoms for the reference group (males with later age-at-onset) is -0.234 per month and is statistically significant. Women appear to recover at a faster rate, with a 95% confidence interval for the interaction between time and gender equal to -0.149 ± 0.174. Subjects with early age-at-onset appear to also recover more quickly, with a 95% confidence interval for the interaction between time and age equal to -0.100 ± 0.178. However, neither interaction is significant at the nominal 0.05 level.

4. Discussion

This article introduces a flexible class of models for serial categorical response data that permit computationally feasible likelihood-based marginal regression analysis. Our methods are a natural extension of Azzalini (1994), and specific models have direct connections to Fitzmaurice and Laird (1993). The marginalized transition model constructs separate mod-
els for the average response and for the correlation among longitudinal observations. Details regarding estimation and model checking are presented for equally spaced binary response data, and a worked example illustrates application with a moderately long response series (n_t = 12).

Several aspects warrant further research. First, investigation of application to data that are not equally spaced is important. The methods that we develop can be modified to permit general time spacing if Markov dependence assumptions remain plausible. In these situations, the transition coefficients (e.g., \( \gamma_{1,1} \)) may be specified to depend on the time lag between observations. Characterizing the relationship between marginalized Markov models with general time spacing and discrete time models with intermittent missing data or continuous time Markov models would be useful. Second, a detailed investigation should study use of MTM(p) maximum likelihood estimators with data subject to MAR and nonignorable dropout processes. Third, for time-dependent covariates, we have assumed that the full covariate conditional mean structure, \( E(Y_t | X_{it}) = E(Y_t | X_{i1}, \ldots, X_{im}) \), and that this represents the target of inference. In many applications, we may be interested in partly conditional means, \( E(Y_{it} | X_{ij}, j \leq t) \), which will be different than \( E(Y_{it} | X_{ij} \forall j) \) whenever \( Y_t \) influences \( X_{it+k} \) for some \( k \geq 1 \). Extension of the MTM(p) likelihood approach to estimation of partly conditional means in the stochastic covariate scenario is possible if a model for the covariate process is adopted and incorporated into the likelihood maximization. Finally, although we have focused on binary response data, the MTM(p) can easily be extended to other categorical response types. For example, with ordinal data, we may adopt global odds ratios (Dale, 1986) to characterize serial dependence within a marginalized transition framework.

RÉSUMÉ
Les modèles linéaires généralisés marginaux sont maintenant souvent utilisés pour analyser des données longitudinales. L'influence semi-paramétrique pour les modèles marginaux a été introduite par Liang et Zeger (1986). Cette article présente une classe paramétrique générale de dépendance sérielle qui permet une analyse de régression marginale de données de réponse binaire, basée sur la vraisemblance. La méthode est une extension naturelle des modèles markoviens du premier ordre d’Azzalini (1994), et ne pose pas de difficultés de calcul pour les séries longues.

REFERENCES


APPENDIX A

Log-Linear Parameterization for MTM(1)

We show that the first-order marginalized transition model has mean regression parameters, \( \beta \), that are orthogonal to the association parameters, \( \alpha_1 \). Azzalini (1994) showed orthogonality in the restricted situation where the pairwise odds ratio is constant over time. By showing that the MTM(1) model corresponds to a mixed parameter log-linear model (Fitzmaurice and Laird, 1993), we are able to apply the results of Barndorff-Nielsen and Cox (1994) and establish the orthogonality of the parameterization. Specifically, we show that the
MTM(1) is a quadratic exponential family model with $\alpha_1$ structuring the canonical interaction parameters.

Without loss of generality, we assume that a logistic model is used for the marginal mean regression. Using the fact that the likelihood contribution from each subject can be factored into the product of the probability for the first observation times, the subsequent one-step transition probabilities allow a sequential calculation of the likelihood. Consider a vector of responses from a single subject, $Y = (Y_1, Y_2, \ldots, Y_n)$. The logistic MTM(1) assumes

$$
\mu_t^M = \Pr(Y_t = 1 | X_t), \\
\logit (\mu_t^M) = X_t \beta,
$$

$$
\mu_t^C = \Pr(Y_t = 1 | X_t, Y_j = y_j < t), \\
\logit (\mu_t^C) = \Delta_t + \gamma_{t,1} Y_{t-1}, \\
\gamma_{t,1} = Z_{t,1} \alpha_1.
$$

To obtain the likelihood function, we compute the sequential products of transition probabilities,

$$
\Pr(Y_1 = y_1) = \exp(y_1 \eta_1) /[1 + \exp(\eta_1)] = \exp(\theta_{0,1} + y_1 \eta_1),
$$

with $\theta_{0,1} = -\log[1 + \exp(\eta_1)]$ and $\eta_1 = \logit(\mu_1^M)$;

$$
\Pr(Y_1 = y_1, Y_2 = y_2) = \Pr(Y_2 = y_2 | Y_1 = y_1) \Pr(Y_1 = y_1) = \exp(\theta_{0,2} + \theta_1 y_1 + \Delta_2 y_2 + \gamma_{2,1} y_2 y_1),
$$

with $\theta_{0,2} = \theta_{0,1} - \log[1 + \exp(\Delta_2)]$ and $\theta_1 = \eta_1 - \log[1 + \exp(\Delta_2 + \gamma_{2,1} y_1 y_2)];$

$$
\Pr(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) = \Pr(Y_2 = y_2 | Y_1 = y_1) \Pr(Y_1 = y_1, Y_2 = y_2) = \exp \left( \theta_{0,3} + \sum_{j=1}^{2} \theta_j y_j + \Delta_3 y_3 + \sum_{j=2}^{3} \gamma_{j,1} y_j y_{j-1} \right),
$$

with $\theta_{0,3} = \theta_{0,2} - \log[1 + \exp(\Delta_3)]$ and $\theta_2 = \Delta_2 - \log[1 + \exp(\Delta_3 + \gamma_{3,1})] + \log[1 + \exp(\Delta_3)].$

Continuing the product yields

$$
\Pr(Y_1 = y_1, \ldots, Y_n = y_n) = \exp \left( \theta_{0,n} + \sum_{j=1}^{n} \theta_j y_j + \sum_{j=2}^{n} \gamma_{j,1} y_j y_{j-1} \right),
$$

with $\theta_{0,n} = \Sigma_{j=1}^{n} - \log[1 + \exp(\Delta_j)], \theta_j = \Delta_j - \log[1 + \exp(\Delta_{j+1} + \gamma_{j+1,1})] + \log[1 + \exp(\Delta_{j+1})]$ for $j < n$, and $\theta_n = \Delta_n$, where, for simplicity, we adopt $\Delta_1 = \eta_1$.

Therefore, the MTM(1) model is a reparameterization of the canonical log-linear model $(\theta^{(1)}, \gamma^{(1)}) \rightarrow (\mu_1^M, \gamma^{(1)})$, where $\theta^{(1)} = (\theta_1, \ldots, \theta_n)$ and $\gamma^{(1)} = (\gamma_{1,1}, \ldots, \gamma_{n,1})$. This implies that $\beta$ and $\alpha_1$ are orthogonal (Barndorff-Nielsen and Cox, 1994). The results of Fitzmaurice and Laird (1993) imply the consistency of the MLE $\hat{\beta}$ regardless of whether the association model is correctly specified (i.e., consistency robust to either incorrect lag assumptions or incorrect regression structure for $\gamma_{j,1}$).

**APPENDIX B**

**Maximum Likelihood Estimation for MTM(2)**

In this section, we provide details on the calculation of score equations for a single subject with a series of $n$ binary measurements. Our notation suppresses dependence on covariates $X_t$. Define the marginal probability as $\mu_t^M = \Pr(Y_t = 1)$. Let $h(x) = \log^{-1}(x)$. Define the transition probability for time $t$ as

$$
\mu_t^C = \Pr(Y_t = 1 | Y_{t-1} = y_{t-1}, Y_{t-2} = y_{t-2}) = h(\Delta_t + \gamma_{t,1} y_{t-1} + \gamma_{t,2} y_{t-2}).
$$

**Model**

We assume a logistic model for the marginal mean, although other links can easily be accommodated. Define the three regression models that comprise the MTM(2),

$$
\logit (\mu_t^M) = X_t \beta, \\
\gamma_{t,1} = Z_{t,1} \alpha_1, \\
\gamma_{t,2} = Z_{t,2} \alpha_2.
$$

**Likelihood**

The likelihood contribution from a single subject takes the form

$$
\mathcal{L} = \prod_{t=3}^{n} \left( \mu_t^C \right)^{y_t} (1 - \mu_t^C)^{(1-y_t)} \times \Pr(Y_2 = y_2, Y_1 = y_1)
$$

$$
\mathcal{L} = \mathcal{L}^{(2)} \times \mathcal{L}^{(1)}.
$$

To evaluate the likelihood, we need to evaluate both the contribution from the initial state, $\mathcal{L}^{(1)}$, and the subsequent contribution from each transition probability, $\mathcal{L}^{(2)}$. We first discuss calculations required for evaluation of $\mathcal{L}^{(2)}$ and then discuss $\mathcal{L}^{(1)}$.

**Transitions**

Score functions from $\mathcal{L}^{(2)}$ take the familiar generalized linear model form

$$
\frac{\partial}{\partial \theta} \log \mathcal{L}^{(2)} = \sum_{t>2} \frac{\partial \mu_t^C}{\partial \theta} (y_t - \mu_t^C) / \left\{ \mu_t^C (1 - \mu_t^C) \right\}.
$$

In order to compute $\mu_t^C$ and $\partial \mu_t^C / \partial \theta$, we require the value $\Delta_t$ that satisfies the marginal mean constraints, i.e.,

$$
\mu_t^M = \sum_{j,k} \Pr(Y_t = 1 | Y_{t-1} = y_{t-1}, Y_{t-2} = y_{t-2}) \pi_{j,k}^{(t)},
$$

where

$$
\pi_{j,k}^{(t)} = \Pr(Y_{t-1} = j, Y_{t-2} = k).
$$

First, we describe a sequential algorithm that computes $\mu_t^C$ and required derivatives based on available estimates of $\pi_{j,k}^{(t)}$ (and derivatives). Next, we discuss how $\mu_t^C$ and $\pi_{j,k}^{(t)}$ can be carried forward to obtain $\pi_{j,k}^{(t+1)}$ (and derivatives). Finally, we discuss initialization based on calculations for $\mathcal{L}^{(1)}$.

With current values for $\pi_{j,k}^{(t)}$, we can obtain $\Delta_t$ using Newton–Raphson to solve the equation that links the marginal mean to the conditional expectation given by the transition
probability
\[ \mu_t^M = \sum_{j,k} h(\Delta_t + \gamma_{t,1} + \gamma_{t,2}k) \pi_{j,k}^{(t)} \]

\[ \frac{\partial \mu_t^M}{\partial \Delta_t} = \sum_{j,k} h_{j,k}^{(t)} (1 - h_{j,k}^{(t)}) \pi_{j,k}^{(t)} = A_t, \]

where

\[ h_{j,k}^{(t)} = h(\Delta_t + \gamma_{t,1} + \gamma_{t,2}k). \]

For score and information computations, we require additional derivatives of \( \Delta_t \),

\[ \frac{\partial \mu_t^M}{\partial \beta} = \sum_{j,k} h_{j,k}^{(t)} (1 - h_{j,k}^{(t)}) \frac{\partial \Delta_t}{\partial \beta} \pi_{j,k}^{(t)} + \sum_{j,k} h_{j,k}^{(t)} \frac{\partial \pi_{j,k}^{(t)}}{\partial \beta} \]

\[ \frac{\partial \Delta_t}{\partial \beta} = \frac{\partial \Delta_t}{\partial \gamma_{t,1}} = 0 = \sum_{j,k} h_{j,k}^{(t)} (1 - h_{j,k}^{(t)}) \left[ \frac{\partial \Delta_t}{\partial \gamma_{t,1}} + j \right] \pi_{j,k}^{(t)} \]

\[ + \sum_{j,k} h_{j,k}^{(t)} \frac{\partial \pi_{j,k}^{(t)}}{\partial \gamma_{t,1}} \]

\[ = A_t \frac{\partial \Delta_t}{\partial \gamma_{t,1}} + \sum_{j,k} h_{j,k}^{(t)} (1 - h_{j,k}^{(t)}) \left[ j \right] \pi_{j,k}^{(t)} \]

\[ + \sum_{j,k} h_{j,k}^{(t)} \frac{\partial \pi_{j,k}^{(t)}}{\partial \gamma_{t,1}} \]

\[ \frac{\partial \Delta_t}{\partial \gamma_{t,1}} = \frac{\partial \Delta_t}{\partial \gamma_{t,2}} = -\left[A_t - B_t + C_t \right] / A_t \]

\[ = -[D_t + E_t] / A_t. \]

After obtaining computations for time \( t \), we make calculations for time \( t+1 \) that require an update \( (\pi_{j,k}^{(t+1)}, \mu_{j,k}^{(t)}) \rightarrow (\pi_{j,k}^{(t+1)}, \mu_{j,k}^{(t)}) \) as well as updates for derivatives,

\[ \pi_{j,k}^{(t+1)} = \text{pr}(Y_t = i, Y_{t-1} = j) \]

\[ = \sum_k \text{pr}(Y_t = i | Y_{t-1} = j, Y_{t-2} = k) \pi_{j,k}^{(t)} \]

\[ \pi_{j,k}^{(t+1)} = \sum_k h(\Delta_t + \gamma_{t,1} + \gamma_{t,2}k) \pi_{j,k}^{(t)} \]

\[ \frac{\partial \pi_{j,k}^{(t+1)}}{\partial \beta} = \sum_k h_{j,k}^{(t)} (1 - h_{j,k}^{(t)}) \frac{\partial \Delta_t}{\partial \beta} \pi_{j,k}^{(t)} + h_{j,k}^{(t)} \frac{\partial \pi_{j,k}^{(t)}}{\partial \beta} \]

\[ \frac{\partial \pi_{j,k}^{(t+1)}}{\partial \gamma_{t,1}} = \sum_k h_{j,k}^{(t)} (1 - h_{j,k}^{(t)}) \left[ \frac{\partial \Delta_t}{\partial \gamma_{t,1}} + j \right] \pi_{j,k}^{(t)} + h_{j,k}^{(t)} \frac{\partial \pi_{j,k}^{(t)}}{\partial \gamma_{t,1}} \]

\[ \frac{\partial \pi_{j,k}^{(t+1)}}{\partial \gamma_{t,2}} = \sum_k h_{j,k}^{(t)} (1 - h_{j,k}^{(t)}) \left[ \frac{\partial \Delta_t}{\partial \gamma_{t,2}} + k \right] \pi_{j,k}^{(t)} + h_{j,k}^{(t)} \frac{\partial \pi_{j,k}^{(t)}}{\partial \gamma_{t,2}} \]

and

\[ \pi_{0,j}^{(t+1)} = \sum_k [1 - h(\Delta_t + \gamma_{t,1} + \gamma_{t,2}k)] \pi_{j,k}^{(t)} \]

\[ \frac{\partial \pi_{0,j}^{(t+1)}}{\partial \beta} = -\sum_k h_{j,k}^{(t)} (1 - h_{j,k}^{(t)}) \frac{\partial \Delta_t}{\partial \beta} \pi_{j,k}^{(t)} + (1 - h_{j,k}^{(t)}) \frac{\partial \pi_{j,k}^{(t)}}{\partial \beta} \]

\[ \frac{\partial \pi_{0,j}^{(t+1)}}{\partial \gamma_{t,1}} = -\sum_k h_{j,k}^{(t)} (1 - h_{j,k}^{(t)}) \left[ \frac{\partial \Delta_t}{\partial \gamma_{t,1}} + j \right] \pi_{j,k}^{(t)} + \left(1 - h_{j,k}^{(t)} \right) \frac{\partial \pi_{j,k}^{(t)}}{\partial \gamma_{t,1}} \]

\[ \frac{\partial \pi_{0,j}^{(t+1)}}{\partial \gamma_{t,2}} = -\sum_k h_{j,k}^{(t)} (1 - h_{j,k}^{(t)}) \left[ \frac{\partial \Delta_t}{\partial \gamma_{t,2}} + k \right] \pi_{j,k}^{(t)} + \left(1 - h_{j,k}^{(t)} \right) \frac{\partial \pi_{j,k}^{(t)}}{\partial \gamma_{t,2}}. \]

Initial State

Evaluation of \( L^{(1)} \) = \( \text{pr}(Y_1 = y_1, Y_2 = y_2) \) requires the bivariate distribution. The marginal means \( \mu_M^1 \) and \( \mu_2^M \) combined with the pairwise odds ratio specifies \( L^{(1)} \). We specify the odds ratio for the initial pair using the \( Z_1 \) matrix. This allows us to set the log odds ratio equal to \( \gamma_{t,1} \) if we choose or to allow a separate estimate for the association between the first pair of responses.