Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

A copula-based Markov chain model for the analysis of binary longitudinal data

Gabriel Escarela , Luis Carlos Pérez-Ruíz & Russell J. Bowater

To cite this article: Gabriel Escarela , Luis Carlos Pérez-Ruíz & Russell J. Bowater (2009) A copula-based Markov chain model for the analysis of binary longitudinal data, Journal of Applied Statistics, 36:6, 647-657, DOI: 10.1080/02664760802499287

To link to this article: https://doi.org/10.1080/02664760802499287

Journal of Applied Statistics
Vol. 36, No. 6, June 2009, 647–657

Taylor & Francis
Taylor & Francis Group

A copula-based Markov chain model for the analysis of binary longitudinal data

Gabriel Escarela[a]*, Luis Carlos Pérez-Ruíz[a] and Russell J. Bowater[b]

[a]*Departamento de Matemáticas, Universidad Autónoma Metropolitana, Unidad Iztapalapa, México D.F., Mexico;* [b]*Department of Public Health and Epidemiology, School of Medicine, University of Birmingham, Birmingham, UK*

(*Received 9 January 2008; final version received 19 September 2008*)

A fully parametric first-order autoregressive (AR(1)) model is proposed to analyse binary longitudinal data. By using a discretized version of a copula, the modelling approach allows one to construct separate models for the marginal response and for the dependence between adjacent responses. In particular, the transition model that is focused on discretizes the Gaussian copula in such a way that the marginal is a Bernoulli distribution. A probit link is used to take into account concomitant information in the behaviour of the underlying marginal distribution. Fixed and time-varying covariates can be included in the model. The method is simple and is a natural extension of the AR(1) model for Gaussian series. Since the approach put forward is likelihood-based, it allows interpretations and inferences to be made that are not possible with semi-parametric approaches such as those based on generalized estimating equations. Data from a study designed to reduce the exposure of children to the sun are used to illustrate the methods.

Keywords: copula; discrete time series; Markov regression models; maximum likelihood; probit regression model; serial correlation

1. Introduction

The use of binary data collected at various time points to examine the relationship between the probability of success and time-dependent covariates is an important issue in medical research. For instance, in order to assess the importance of maternal employment as a determinant of paediatric care utilization, Alexander and Markowitz [1] analysed daily measurements of maternal stress (present, absent) and childhood illness (present, absent) for 28 consecutive days; in the study conducted by Yu *et al.* [22], health effects of air pollution in asthmatic children were analysed using a 6-month follow-up of daily observations of wheezing (present, absent); more recently, Baker *et al.* [4] evaluated how indoor pollution from tobacco and home heating may adversely affect respiratory health in young children on the basis of a 3-year longitudinal study of the daily incidence of lower respiratory illness (present, absent).

*Corresponding author. Email: ge@xanum.uam.mx

ISSN 0266-4763 print/ISSN 1360-0532 online

DOI: 10.1080/02664760802499287
http://www.informaworld.com

Often, the most convenient way of carrying out this type of analysis is through the use of Markov chain models. In this approach, probabilities of success given the past can be determined by a set of covariates. Although, there is a well-established literature on serially correlated binary sequences [3,5,6,8,9,11,14; references therein], there are very few fully specified Markov chain models that allow for the modelling of both the marginal behaviour and the dependence between the lagged responses.

In this article, we introduce a parametric class of first-order Markov chains for binary longitudinal data. Specifically, we extend the transition model for continuous longitudinal data, in which dependence between adjacent responses follows a bivariate normal distribution, to the case which responses are binary. To do this, we use the discretized version of the Gaussian copula model to construct fully specified conditional probabilities for any given discrete margin. This gives rise to a first-order autoregressive model for analysing longitudinal binary responses, which allows for covariates.

There are several reasons why this likelihood-based methodology is an attractive way of analysing longitudinal binary data. These reasons include optimality of estimators under correct model specification, the availability of inferential procedures such as likelihood ratio tests, the flexibility to allow for unequally spaced observations, and the robustness of certain missing data structures.

2. Transition models for discrete data

To establish the context, consider a first-order discrete-time stationary Markov process, also known as AR(1), for normal responses. Let $\{Y_t, t = 1, 2, \ldots\}$ be a time series with marginal responses $Y_t \sim \mathrm{N}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_t, \sigma^2)$, for $t = 1, 2, \ldots$, where $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_t$ is the marginal mean of Y_t, $\mathbf{x}_t$ a vector of explanatory variables observed at time t, $\boldsymbol{\beta}$ the corresponding vector of unknown regression coefficients, and σ^2 the variance of the marginal response. If the joint distribution of the lagged responses Y_{t-1} and Y_t is bivariate normal with the correlation parameter ρ, the transition model has the following conditional specification

$$Y_t \mid Y_{t-1} \sim \mathrm{N}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_t + \rho[Y_{t-1} - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{t-1}], \tau^2),$$

where $\tau^2 = \sigma^2(1 - \rho^2)$ and $|\rho| < 1$.

Similarly, if the state space of the time series is a finite or countable set, the transition distribution for the stationary AR(1) process is given by

$$F_{2|1}(y_2 \mid y_1) = \sum_{z \le y_2} \frac{f(y_1, z)}{f(y_1)}, \tag{1}$$

where $f(y_1, y_2)$ is a joint mass function with both marginal probability functions equal to $f(\cdot)$.

Note that the transition distribution of $Y_t \mid Y_{t-1}$ given by Equation (1) can be fully defined through the choice of both a copula function C and the distribution of Y_t. If $F(y_t)$ represents the marginal cdf of Y_t, the family of transition distributions of $\{Y_t\}$ can be characterized by the discrete transition density function $f_{2|1}(y_t \mid y_{t-1}) = \Pr\{Y_t = y_t \mid Y_{t-1} = y_{t-1}\}$ as follows [12]:

$$\begin{aligned} f_{2|1}(y_t \mid y_{t-1}) = \{&C[F(y_t), F(y_{t-1})] - C[F(y_t), F(y_{t-1} - 1)] - C[F(y_t - 1), F(y_{t-1})] \\ &+ C[F(y_t - 1), F(y_{t-1} - 1)]\}/f(y_{t-1}). \end{aligned}$$

Therefore, inputting choices for both the family of copulas and the family of distributions for the underlying marginals into this formula allows one to construct a model for serially correlated

data. In this study, we use, in particular, the Gaussian copula, which has the following form

$$C_\rho(v_1, v_2) = \Phi_2[\Phi^{-1}(v_1), \Phi^{-1}(v_2)], \quad (v_1, v_2)^T \in (0, 1)^2, \tag{2}$$

where $\Phi_2(\cdot, \cdot)$ is the cdf of a bivariate Gaussian distribution with mean $(0, 0)^{\mathrm{T}}$ and covariance matrix $\mathbf{R}$ equal to a 2×2 non-singular matrix with the off-diagonal elements equal to ρ, where $\rho \in (-1, 1)$, and the diagonal elements equal to 1, and $\Phi^{-1}(\cdot)$ is the inverse function of the standard Gaussian cumulative distribution. Note that when $\rho = 0$, the Gaussian copula in Equation (2) defines the independence copula $C(v_1, v_2) = v_1 v_2$. It is natural to choose the Gaussian copula since it encodes dependence in the same way that the bivariate normal distribution does using the dependence parameter ρ, with the difference that it does so for random variables with any given marginals.

The concept of a *discretized copula* was introduced by Joe [12, p. 247] in the context of binary time series. Although the copula is mainly used for continuous data, it is possible to obtain Radom–Nikodym's derivative for the copula function that yields the discretized version of the copula, which is used to construct joint distributions with given margins for discrete random vectors, as shown by Song [20].

A model for a two-state Markov chain can straightforwardly be obtained by setting the marginals F equal to Bernoulli distributions with probability of success p and the copula function equal to the Gaussian copula given by Equation (2). In the regression setting, it is possible to model the marginal probability function of Y_t as functions of covariates $\mathbf{x}_t$. Thus, a more general transition model assumes that the marginal of Y_t has the following cumulative distribution function of Y_t with covariates $\mathbf{x}_t$

$$F(y_t; \mathbf{x}_t) = q(\mathbf{x}_t)\, I_{[0,1)}(y_t) + I_{[1,\infty)}(y_t), \quad y_t \in (-\infty, \infty),$$

where $q(\mathbf{x}_t) = 1 - p(\mathbf{x}_t)$, in which $p(\mathbf{x}_t)$ is the probability of success given in terms of the covariates at time t and $I_A(y)$ is the indicator function of A, which equals 1 if $y \in A$ and equals 0 otherwise.

The transition matrix that represents a binary Markov chain of first order at time t can be expressed as

$$\mathbf{P}^{[t]} = \begin{pmatrix} 1 - p_{1|0}^{(t)} & p_{1|0}^{(t)} \\ 1 - p_{1|1}^{(t)} & p_{1|1}^{(t)} \end{pmatrix}, \tag{3}$$

where $p_{1|k}^{(t)} = \Pr\{Y_t = 1 \mid Y_{t-1} = k\}$ and $k \in \{0, 1\}$. Using the properties of the copula, the resulting one-step transition probability functions with covariate information given at time t can be represented as

$$p_{1|0}^{(t)} = \{q(\mathbf{x}_{t-1}) - C_\rho[q(\mathbf{x}_t), q(\mathbf{x}_{t-1})]\}/q(\mathbf{x}_{t-1}), \quad \text{and}$$
$$p_{1|1}^{(t)} = \{1 - q(\mathbf{x}_{t-1}) - q(\mathbf{x}_t) + C_\rho[q(\mathbf{x}_t), q(\mathbf{x}_{t-1})]\}/p(\mathbf{x}_{t-1}).$$

It follows that the corresponding conditional expectation and conditional variance are given by $\mathrm{E}[Y_t \mid Y_{t-1} = k] = p_{1|k}^{(t)}$ and $\mathrm{Var}[Y_t \mid Y_{t-1} = k] = p_{1|k}^{(t)}(1 - p_{1|k}^{(t)})$, respectively; here, $k \in \{0, 1\}$.

A convenient way to account for concomitant information in this model is to use the probit link in the marginal probability model. This link implies that $p(\mathbf{x}_t) = \Phi(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_t)$, where Φ is the standard normal cumulative distribution function and $\boldsymbol{\beta}$ is the vector of coefficients. Due to the symmetry of the normal distribution, we find that $q(\mathbf{x}_t) = \Phi(-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_t)$, and thus, for the normal copula, we obtain $C_\rho[q(\mathbf{x}_t), q(\mathbf{x}_{t-1})] = \Phi_2(-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_t, -\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{t-1})$; here, Φ_2 has dependence parameter equal to ρ.

2.1 *Unequally spaced data*

Although the methods presented here are focused on equally spaced data, the discretized Gaussian copula model lends itself to formulating growth curve analysis in the form of a continuous-time autoregressive structure [13] to study longitudinal data when each subject is observed at different unequally spaced time points. If the Markov dependence assumption remains plausible, the methodology can be modified to permit general time spacing, which can lead to a continuous time model as discussed by Cox and Snell [8].

Two commonly used forms for the correlation function for Gaussian data are given by

$$\rho(u) = \rho^u, \quad |\rho| < 1,$$

so that it decreases as the distance in time, u, increases, and

$$\rho(u) = \exp\{-\kappa u^\nu\}, \quad \kappa > 0.$$

These two formulations can be straightforwardly adapted to the discretized AR(1) model described earlier by substituting ρ with $\rho(t_u - t_l)$, where t_u and t_l ($t_u > t_l$) represent two successive observation time points of any given subject. Thus, the resulting model will embody the dependence between successive observations on the same subject through one of these decay-of-correlation structures.

3. Maximum likelihood estimation

Let $y_{i\,1}, \ldots, y_{i\,m_i}$ denote equally spaced responses of the ith subject at time points t_{ij}, $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$. The contribution to the likelihood for the ith subject can be written as

$$L_i = f(y_{i\,1};\, \mathbf{x}_{i\,1}) \prod_{k=2}^{m_i} f(y_{i\,k} \mid y_{i\,k-1}; \mathbf{x}_{i\,k}, \mathbf{x}_{i\,k-1}). \tag{4}$$

In the presence of missing data, the likelihood in Equation (4) has to be modified. When the missing value mechanism is at random, the likelihood only takes into account the indexes that correspond to the observed data; for instance, if there are three observation time points and only the second observation is missing for the ith individual, the likelihood replaces the one-step probabilities by two-step probabilities $\Pr\{Y_{i\,3} = y_{i\,3} \mid Y_{i\,1} = y_{i\,1}\}$, which are obtained from the product $\mathbf{P}_i^{[2]}\mathbf{P}_i^{[3]}$, where $\mathbf{P}_i^{[t]}$ denotes the transition probability matrix for individual i as defined in Equation (3).

The copula-based transition model is not linear, so numerical techniques need to be used in order to determine the maximum likelihood estimators (MLEs). For convenience, we estimate the parameter $\theta = \operatorname{arctanh}(\rho)$, which takes values in $(-\infty, \infty)$ and ensures that $-1 < \rho < 1$. In this study, the covariance matrix will be calculated from the inverse of the observed information matrices of the MLEs $\hat{\boldsymbol{\beta}}$ and $\operatorname{arctanh}(\hat{\rho})$.

An important characteristic of fitting the transition model in the manner described is that the process of finding a parsimonious model can follow standard ideas based on the likelihood ratio; for example, the type of approach used in fitting generalized linear models [16]. However, p-values and the tests based on them tend to give unsatisfactory results when the sample size is large [18]. With large data sets, the introduction or removal of almost any parameter from the model turns out to be significant, despite the fact that the change does not reveal a genuine insight.

For this reason, in order to identify the appropriate covariates to be included in the model, we use the Bayesian information criterion (BIC) as our main model choice criterion. The value of

BIC for a model, in particular, is given by [19]

$$\text{BIC} = -2 \log \prod_{i=1}^{n} L_i + n_p \log(N),$$

where n_p denotes the number of independent parameters in the model and N is the total number of non-missing observations; when the data are complete, $N = m \times n$ [2]. The criterion used here chooses the model for which BIC is the smallest.

The assumptions made about a particular transitional model may be checked by plotting a modification of the conditional randomized quantile residuals proposed by Dunn and Smyth [10]. Let $a_{it} = F(y_{it} - 1 \mid H_{it})$ and $b_{it} = F(y_{it} \mid H_{it})$, where $H_{i1} = \{\mathbf{x}_{i1}\}$ represents the present history of the ith subject at time $t = 1$ and $H_{it} = \{y_{it-1}; \mathbf{x}_{it}, \mathbf{x}_{it-1}\}$ represents the past and present history of the ith subject at time t, for $t \geq 2$. Thus, $F(y_{it} \mid H_{it})$ equals the marginal distribution F when $t = 1$ and equals the transition distribution $F_{2|1}$ when $t \geq 2$. The randomized quantile residual for y_{it} is then defined by

$$r_{it} = \Phi^{-1}(u_{it}), \tag{5}$$

where u_{it} is a random value from the uniform distribution in the interval $(a_{it}, b_{it}]$.

Under the assumed model, the residuals defined in Equation (5) should be independent and follow a standard normal distribution. Thus, in order to assess the quality of fit of the model, standard residual plots can be used to assess how closely these residuals conform to this hypothesis.

4. The sun exposure data

We analyse data from a clinical trial designed to assess the reduction of sun exposure in children. This data set is from the research of Lori A. Crane, PhD, University of Colorado Health Sciences Center (NIH CA 74592). Infants were enrolled prior to 6 months of age and were randomized into an intervention or a control group. Parents of the intervention group received sun protection kits and advice at well-child visits between 2 and 36 months of age. Annual parent surveys at 1, 2, and 3 years of age were used to follow the intervention up. In this study, we categorize the response 'avoids mid-day sun exposure' as never or seldom (coded 0) and frequently or always (coded 1). The two explanatory variables were the factor `TR`, the treatment, and the `time` of the observation. There were 680 children in the follow-up study from which 1398 had three observations, 286 had two observations and 66 had one observation. Therefore, the total number of non-missing observations is $N = 1750$.

By using the probit link in the model for the marginals, we fit the discrete copula transition model described in Section 2 to the sun exposure data using the method of maximum likelihood. For simplicity, we make the assumption that missing observations are missing at random. In analysing the same data set, this assumption is also made by Jones *et al.* [14].

Note that the generalized estimating equation (GEE) approach could be used to estimate the marginal regression parameters by regarding the correlation between the responses as a nuisance parameter [21]. Therefore, since the GEE approach is a widely available method, we proceed to compare the results of using the copula transition model described earlier with those of the semi-parametric GEE. Although it should be noted that unlike the approach presented here, the GEE approach is not based on an exact formulation of the likelihood. We used the function `nlm` of the R language to minimize $-2 \times \log \prod_{i=1}^{n} L_i$, the *deviance*. Also, we used the R package `geepack` to estimate the parameters in the GEE method.

By maximizing the likelihood specified in Equation (4), we also fitted the copula transition model under the assumption that all observations are independent, so that comparisons could be

Table 1. Deviances and BICs of various models taking into account combinations between time and TR.

	Independence		AR(1)		
Formula	Deviance	BIC	Deviance	BIC	$\hat{\rho}$ with 95% C.B.
`time` $*$ TR	1768.57	1798.44	1660.45	1697.79	0.566 (0.473, 0.648)
`time` $+$ TR	1768.68	1791.09	1660.51	1690.38	0.566 (0.473, 0.648)
`time`	1776.47	1791.40	1665.59	1688.00	0.571 (0.478, 0.651)
`TR`	1837.06	1852.00	1732.30	1754.70	0.544 (0.451, 0.625)
Null	1844.18	1851.65	1736.72	1751.66	0.548 (0.456, 0.629)

made of deviances and coefficients included in the model when different dependence structures are assumed. The deviances of the maximized log-likelihoods for the various combinations of the covariates included in each model along with BIC measures are displayed in Table 1. It can be seen that, using the model selection criterion based on deviance, which consists of keeping a covariate if $\Delta\text{Deviance} > \chi^2_{0.05,\text{df}=1} = 3.84$ when removing the covariate, the best-fitting model for either dependence structure is that which includes the main effects of `time` and `TR` but not the interaction. When it comes to using the BIC criterion, the outcome differs; whilst the best-fitting specification for the independence model includes `time` and `TR`, the AR(1) model just includes `time`, suggesting that `TR` fails to be significant after accounting for `time` effects.

Table 1 also shows the estimated ρ for the copula transition model and 95% confidence intervals for this estimate based on the approximation $(\hat{\theta} - \theta)/\widehat{\text{se}}_{\hat{\theta}} \sim \text{N}(0, 1)$, so that the confidence limits for ρ are calculated as $\tanh(\hat{\theta} \pm 1.96 \times \widehat{\text{se}}_{\hat{\theta}})$. It is interesting to see that the estimator of the dependence parameter ρ remains within the same range without being influenced by the set of covariates included in the model. To test for the significance of the dependence parameter ρ, it is possible to perform the null hypothesis test $H_0 : \rho = 0$, which corresponds to the independence model versus the alternative hypothesis $H_1 : \rho \neq 0$. Given that the confidence intervals for ρ do not include zero, then assuming the large sample properties mentioned here, this null hypothesis would in fact be rejected. This decision is consistent when one compares the corresponding BICs. Thus, the dependence between adjacent observations is significant, and the parameter ρ must be taken into account rather than consideration being given to a model that assumes independence, i.e. the usual probit regression model.

Table 2 shows regression estimates and their standard errors corresponding to the marginal model time $+$ TR when employing the probit specification under the assumption of independence and for the GEE and AR(1) approach when dependence is allowed. Note that all parameters are statistically significant, including TR ($p < 0.05$), and that the coefficients for the GEE and AR(1) approaches are similar.

However, unlike the GEE approach, the AR(1) model adopted here allows for the estimation of the transition matrices. Using the marginal model time $+$ TR in the AR(1) specification, the estimated transition probabilities and 95% confidence bands for the control group are given by

Table 2. Parameter estimation for the marginal model for time $+$ TR using the independence, GEE and AR(1) approaches.

	Independence		GEE		AR(1)	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	1.394	0.100	1.375	0.093	1.376	0.101
`time`	-0.348	0.043	-0.333	0.036	-0.334	0.040
`TR`	0.190	0.068	0.189	0.085	0.184	0.082

the following transition matrices

$$\hat{\mathbf{P}}^{[2]} = \begin{pmatrix} 0.579 & 0.421 \\ (0.475, 0.679) & (0.321, 0.525) \\ 0.180 & 0.820 \\ (0.110, 0.272) & (0.728, 0.890) \end{pmatrix} \quad \text{and} \quad \hat{\mathbf{P}}^{[3]} = \begin{pmatrix} 0.659 & 0.341 \\ (0.538, 0.768) & (0.232, 0.462) \\ 0.258 & 0.742 \\ (0.155, 0.385) & (0.615, 0.845) \end{pmatrix},$$

whereas for the treatment group, the estimated transition matrices are given by

$$\hat{\mathbf{P}}^{[2]} = \begin{pmatrix} 0.534 & 0.466 \\ (0.392, 0.673) & (0.327, 0.608) \\ 0.143 & 0.857 \\ (0.064, 0.265) & (0.735, 0.936) \end{pmatrix} \quad \text{and} \quad \hat{\mathbf{P}}^{[3]} = \begin{pmatrix} 0.615 & 0.385 \\ (0.453, 0.763) & (0.237, 0.547) \\ 0.214 & 0.786 \\ (0.096, 0.378) & (0.622, 0.904) \end{pmatrix}.$$

It can be seen that both the probability of moving from failure to success and the probability of staying in the state of success are roughly 0.08 lower for the second transition when compared with the first transition. However, this difference of 0.08 lies within the tolerance of the 95% confidence bands for the estimated transition probabilities. The difference in the transition probabilities between the control and treatment groups is very small and lies well within the tolerance of the 95% confidence bands for these estimated probabilities. This suggests that the transitions between not avoiding the midday sun and avoiding the midday sun do not differ significantly between the intervention and control groups. Therefore, from observation time one onwards, the intervention would appear to be ineffective in encouraging parents to reduce their child's sun exposure.

We thus, remove `TR` from the marginal model as indicated by the BIC criterion. Table 3 shows regression estimates and their standard errors corresponding to the marginal model, which only takes into account time for the three approaches. The corresponding fitted transition matrices of the AR(1) model are as follows

$$\hat{\mathbf{P}}^{[2]} = \begin{pmatrix} 0.560 & 0.440 \\ (0.460, 0.658) & (0.342, 0.540) \\ 0.161 & 0.838 \\ (0.098, 0.244) & (0.756, 0.902) \end{pmatrix} \quad \text{and} \quad \hat{\mathbf{P}}^{[3]} = \begin{pmatrix} 0.640 & 0.360 \\ (0.522, 0.748) & (0.252, 0.478) \\ 0.235 & 0.765 \\ (0.140, 0.354) & (0.646, 0.860) \end{pmatrix}.$$

It can be seen that removing TR as a covariate in the marginal model does not greatly affect the difference in the estimated transition probabilities between the first and second transitions.

When the assumption that the transition model follows a specific dependence structure is questionable, it is possible to carry out a graphical analysis to assess that such a dependence

Table 3. Parameter estimation for the marginal model with time using the independence, GEE and AR(1) approaches.

	Independence		GEE		AR(1)	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	1.479	0.095	1.461	0.085	1.459	0.094
`time`	−0.346	0.043	−0.331	0.035	−0.331	0.040

structure is present within the data. For this, we define the profile log-likelihood as

$$\ell^*(\rho) = \max_{\boldsymbol{\beta}|\rho}\{\log L(\boldsymbol{\beta}, \rho)\} = \log L[\hat{\boldsymbol{\beta}}(\rho), \rho]. \tag{6}$$

We are particularly interested in the local shape of ℓ^* at $\rho = \hat{\rho}$; if the curve is very flat, then the dependence structure is not well represented within the data, as discussed in Copas and Li [7 and references therein].

Figure 1 displays the profile log-likelihoods ℓ^* for the transition copula model plotted against ρ. We can note that the curve is 'well behaved' in the sense that its shape is fairly close to quadratic for values close to the maximum, which gives support to the use of the asymptotic properties that underlie the fitting method. Also, the height of the log-likelihood curve around the maximum appears to be fairly sensitive to the value of ρ. This is evidenced by the fact that drawing a line at a height of $0.5 \times \chi^2_{1:0.05}$ below the maximum of this graph and reading off the points of intersection produces a 95% confidence interval for ρ that is reasonably narrow. This confidence interval of (0.478, 0.651) is in fact similar in both limits and width to the 95% confidence interval for ρ given earlier. We thus conclude that the data appear to give sufficient information about local departures of ρ from its maximum likelihood estimate.

For the example being considered, a histogram, a kernel density plot, and a normal Q–Q plot of the standardized residuals r_{it}, as defined in Section 3, are given in Figure 2. By examining these plots, it would appear that the proposed model fits the data reasonably well. Also, standard diagnostic plots did not demonstrate the existence of dependence between these residuals.

Figure 3 shows bar charts corresponding to observed and expected frequencies of the possible response patterns in the sun exposure data. It can be seen from the similarities between the probabilities that the model predicts well what was actually observed.

Jones *et al.* [14] use a formulation that specifies the transition probabilities in terms of the steady-state probabilities, which account for covariates through a logit link, and the conditional transition intensities, which account for covariates through a log link, thus, two parameters are included in the model for each covariate. However, when analysing the effect of the treatment,

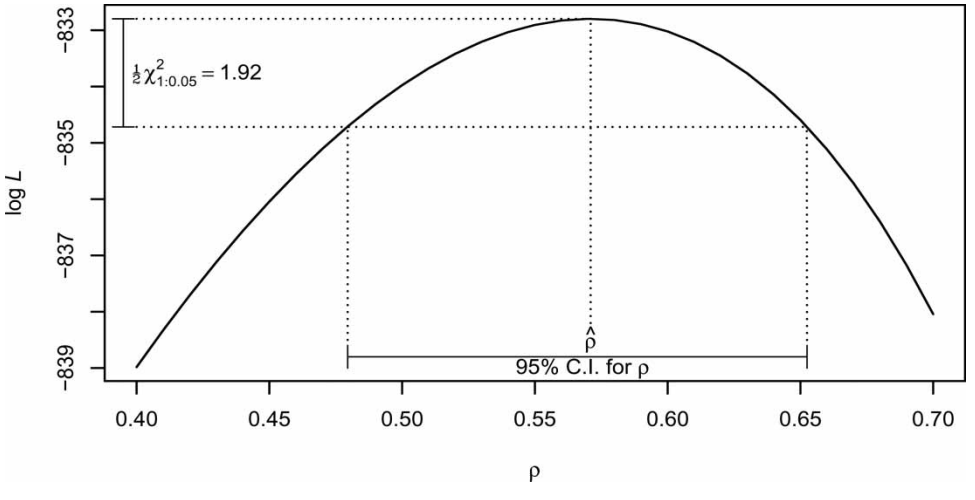

Figure 1. Profile log-likelihood for ρ as estimated by using the copula transition model, together with a 95% confidence interval.

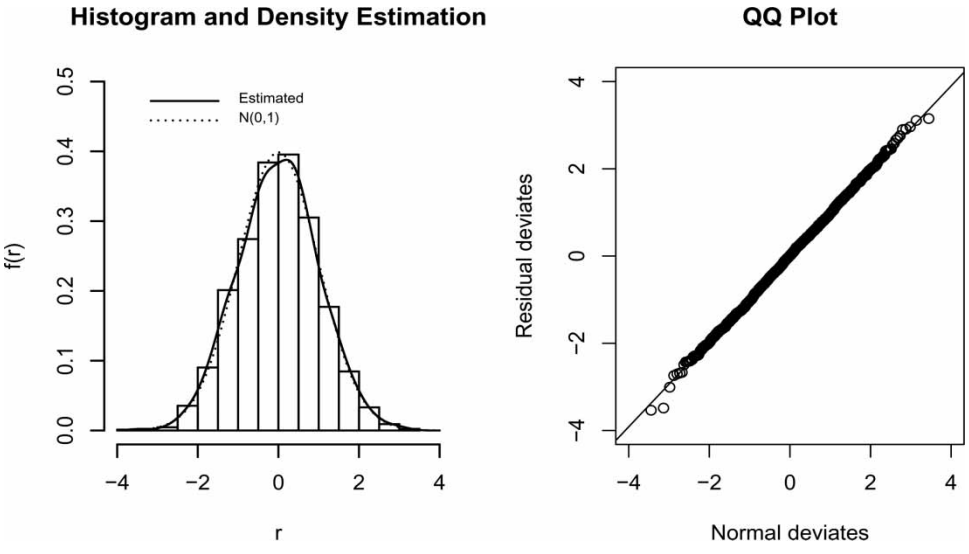

Figure 2. Randomized quantile residuals for the transitional mixture model of the sun exposure data with `time` as the sole covariate.

it can be seen from the changes in deviances that result from introducing and removing parameters from the model in Jones *et al.* [14] that if the BIC criterion had been used to choose the most parsimonious model, then the final model would not be dependent on the treatment group in the same way that the copula model under consideration does not possess this kind of dependence.

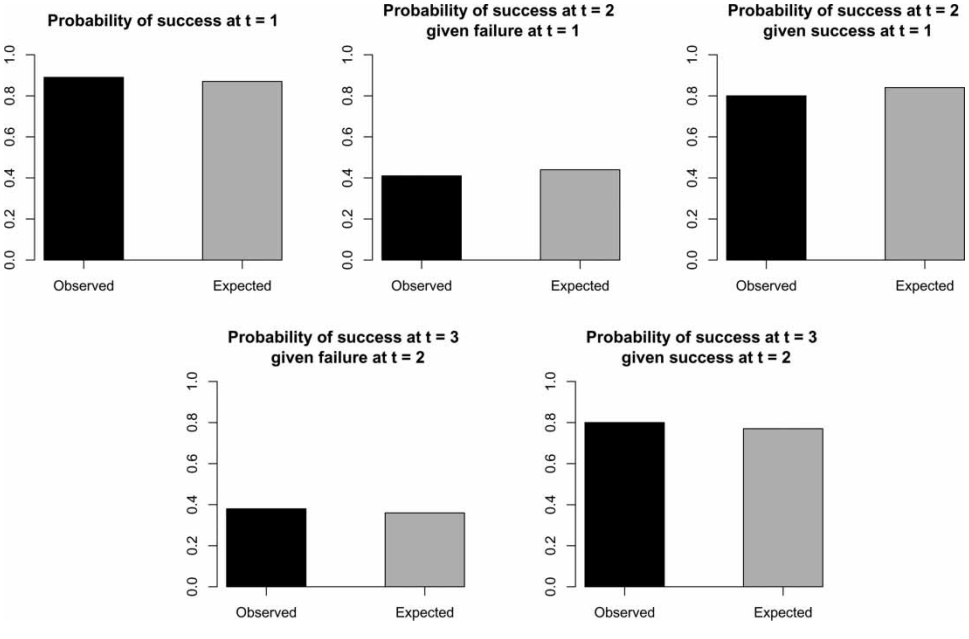

Figure 3. Observed and expected frequencies of the sun exposure data.

5. Discussion

This article introduces a flexible transition (Markov) model for serial binary response data that permits feasible likelihood-based regression analysis. The method presented here is a practical alternative to, and is a natural extension of, the AR(1) model for Gaussian responses and supplements other approaches for analysing binary longitudinal data.

Use of the discretized-copula model described here allows one to construct separate models for the marginal response and for the dependence among longitudinal observations and also allows one to draw inferences and interpretations that are not possible with other approaches such as the semi-parametric GEE. In particular, if the purpose of the longitudinal study is to estimate probabilities of event occurrences, then the copula model presented here represents a convenient method to obtain these kinds of predictions. Also, as, unlike the GEE approach, this copula approach is based on an exact formulation of the likelihood, covariate selection can be performed by using the standard criteria based on the likelihood such as BIC.

Regarding the sun exposure study, use of the transition model provided a straightforward method of comparing treatments over time through easy to interpret summaries, such as transition probabilities and marginal regression coefficients. The modelling is very similar to the type of analysis that is associated with the use of the usual probit model.

The approach undertaken in this study can be generalized further by considering other classes of copulas and a different link function such as the logit. How to perform an exploratory analysis of the data and then choose an appropriate parametric family of transition models remains an open question. Therefore, the chosen model should, as always, be validated by carrying out diagnostic checks and goodness-of-fit tests.

In principle, it is possible to extend the transition copula models described earlier to higher-order representations by using a discretized version of a multivariate copula. In practice, however, this can be quite cumbersome and computationally expensive. A more parsimonious approach than the discretization of a multivariate copula could be to use the *mixture transition distribution* model of order p introduced by Raftery [17] which is characterized by the following transition density

$$f(y_t \mid y^{[t,p]}) = \sum_{k=1}^{p} \omega_k \, f_k(y_t \mid y_{t-k}), \tag{7}$$

where $y^{[t,p]} = (y_{t-1}, \ldots, y_{t-p})$, $\sum_{k=1}^{p} \omega_k = 1$, $\omega_k \geq 0$, $k = 1, \ldots, p$, and $f_k(y_t \mid y_{t-k})$ denotes the one-step ahead transition density corresponding to lag y_{t-k}. In the context of continuous time series data, Le *et al.* [15] have modelled Raftery's AR(p) structure using mixtures of Gaussian transitions, which is equivalent to using mixtures of Gaussian copula conditional models with Gaussian margins. However, for the modelling of serially correlated binary data, the one-step ahead transition probabilities in Equation (7) could in fact be modelled effectively with the conditional discretized-copula model proposed in the present study. Such an extension of the model warrants further research.

Acknowledgements

The authors thank two referees for helpful comments and suggestions. We also thank Dr Richard Jones for providing us with the sun exposure data.

References

[1] C.A. Alexander and R. Markowitz, *Maternal employment and use of pediatric clinic services*, Med. Care 24 (1986), pp. 134–147.

[2] R. Azari, L. Li, and C.-L. Tsai, *Longitudinal data model selection*, Comput. Stat. Data Anal. 50 (2006), pp. 3053–3066.
[3] A. Azzalini, *Logistic regression for autocorrelated data with application to repeated measures*, Biometrika 81 (1994), pp. 767–775; correction: 84 (1997), p. 989.
[4] J.B. Baker, I. Hertz-Picciotto, M. Dostál, J.A. Keller, J. Nozicka, F. Kotesovec, J. Dejmek, D. Loomis, and R.J. Stram, *Coal home heating and environmental tobacco smoke in relation to lower respiratory illness in Czech children, from birth to 3 years of age*, Environ. Health Perspect. 114 (2006), pp. 1126–1132.
[5] P. Billingsley, *Statistical methods in Markov chains*, Ann. Math. Stat. 32 (1961), pp. 12–40.
[6] N.E. Breslow and D.G. Clayton, *Approximate inference in generalized linear mixed models*, J. Am. Stat. Assoc. 88 (1993), pp. 9–25.
[7] J.B. Copas and H.G. Li, *Inference for non-random samples*, J. R. Stat. Soc. B, 59 (1997), pp. 55–95.
[8] D.R. Cox and E.J. Snell, *Analysis of binary data*, 2nd ed., Chapman & Hall, London, 1989.
[9] P. Diggle, P. Heagerty, K.-Y. Liang, and S. Zeger, *Analysis of Longitudinal Data*, 2nd ed., Oxford University Press, New York, 2002.
[10] P.K. Dunn and G.K. Smyth, *Randomized quantile residuals*, J. Comput. Graph. Stat. 5 (1996), pp. 236–244.
[11] P.J. Heagerty, *Marginalized transition models and likelihood inference for longitudinal categorical data*, Biometrics 58 (2002), pp. 342–351.
[12] H. Joe, *Multivariate Models and Dependence Concepts*, Chapman & Hall, New York, 1997.
[13] R.H. Jones and F. Boadi-Boateng, *Unequally spaced longitudinal data with AR(1) serial correlation* Biometrics 47 (1991), pp. 161–175.
[14] R.H. Jones, X. Su, and G.K. Grunwald, *Continuous time Markov models for binary longitudinal data* Biomet. J. 48 (2006), pp. 411–419.
[15] N.D. Le, R.D. Martin, and A.E. Raftery, *Modeling flat stretches, burst and outliers in time series using mixture transition distribution models*, J. Am. Stat. Assoc. 91 (1996), pp. 1504–1515.
[16] P. McCullagh and J.A. Nelder, *Generalized Linear Models*, 2nd ed., Chapman & Hall, London, 1989.
[17] A.E. Raftery, *A model for high-order Markov chains*, J. R. Stat. Soc. Ser. B 47 (1985), pp. 528–539.
[18] A.E. Raftery, *Bayesian model selection in social research*, Sociol. Methodol. 25 (1995), pp. 111–163.
[19] G. Schwarz, *Estimating the dimension of a model*, Ann. Stat. 8 (1978), pp. 461–464.
[20] P.X.-K. Song, *Multivariate dispersion models generated from Gaussian copula*, Scand. J. Stat. 27 (2000), pp. 305–330.
[21] J. Yan and J.P. Fine, *Estimating equations for association structures*, Stat. Med. 23 (2004), pp. 859–880.
[22] O. Yu, L. Sheppard, T. Lumley, J.Q. Koenig, and G.G. Shapiro, *Effects of ambient air pollution on symptoms of asthma in Seattle-area children enrolled in the CAMP study*, Environ. health Perspect. 108 (2006), pp. 1209–1214.