Joint mean-correlation model and its application to NorStOP binary data

Cheng Peng¹, Yihe Yang², Jie Zhou² and Jianxin Pan³

¹College of Mathematics, Sichuan University, Chengdu, Sichuan, 610065, China
²Department of Mathematics, University of Manchester, Manchester M13 9PL, UK
³Email: jianxin.pan@manchester.ac.uk

Abstract. In this paper we propose a generalized estimating equations method to joint model the mean and correlation structures for longitudinal binary data based on Gaussian copula, and apply it to a large cohort study for the UK’s North Staffordshire Osteoarthritis Project (NorStOP), where the responses can be pre-processed to binary variables. The resulting estimators for the mean and correlation parameters are proven to be consistent and asymptotically normally distributed. Since the theory and simulation results were studied in our previous manuscript [1], we give a brief introduction of our approach but mainly focus on its application to the NorStOP data.

1. Introduction
The UK North Staffordshire Osteoarthritis Project (NorStOP) [2-3] for the local population whose age is over 50 years was a clinical research aiming to investigate physical and psychological health for older people. The NorStOP study was a health questionnaire survey which lasted for 6 years and included the subject information like age, gender, alcohol consumption, cost of living, body mass index (BMI), pain interference degree (PI) as well as anxiety and depression status. We switch the depression status to the binary responses and explore its dependence on other explanatory variables. Meanwhile, characterizing the correlation relationship between pairs of binary responses is another important objective. Many methods have been studied for modeling mean and correlation for binary data. Standard methods including logistic regression, GEE methods and random intercept models were used by Kuchibhatla and Fillenbaum (2003) [4] to model correlated binary outcomes, but these methods cannot deal with the issue on how the correlations between pairs of binary outcomes depend on certain covariates. A class of marginal models which separately model the mean and correlations of binary responses with explanatory covariates were proposed by Dale (1986) [5], Prentice (1988) [6] and Liang et al. (1992) [7]. Lipsitz and Laird (1991) [8] proposed generalized estimating equation to estimate the odds ratio which is used to model the association between binary outcomes, and Carey et al. (1993) [9] also introduced alternating logistic regressions based on pairwise odds ratios for modeling correlated binary data. Cessie and Houwelingen (1994) [10] discussed and compared different association measures for the dependence among correlated binary responses. More recently, Chaganty and Joe [11] studied the efficiency of GEE for binary responses when choosing various working correlation matrices.

Nevertheless, since binary data contains much less information than any other kinds of data and many parameters in covariance matrices need to be estimated, modeling correctly the covariance structures is still a big challenge. In order to resolve this issue, on the basis of latent Gaussian copula...
model proposed by Fan et al. (2016) [12], we introduce a joint mean-correlation model for analyzing longitudinal binary data. Assume that the binary observations are generated through dichotomizing a latent Gaussian variable at an unknown cutoff, the association between each pair of outcomes can be characterized by a function of the correlation between their latent variables. It is worth mentioning that, the above assumption is a natural mechanism for producing correlated binary variables, which is also easy to implement in terms of computation. In this paper, we consider a semiparametric regression model for estimating the mean parameters while the two commonly-used correlation structures Compound Symmetry (CS) and AR(1) for latent correlation matrices are considered. Furthermore, the technique of generalized estimating equations (GEE, Liang and Zeger 1986, Prentice 1988, Prentice and Zhao 1991) [6,13,14] is used to estimate all the parameters in the models, and misspecification of the working correlation matrix does not lead to inconsistency of the parameter estimators obtained by GEE (Wang and Carey 2003) [15].

The rest of the article is organized as follows. In section 2, we simply introduce the models and the algorithm used for estimating the unknown parameters. In section 3, we employ the proposed method to analyze the NorStOP data. Section 4 gives some concluding remarks.

2. Models and main algorithm

Let $Y_i = (y_{ij1}, y_{ij2}, \ldots, y_{ijm})'$ be $m$ repeated binary observations at time points $T_i = (t_{i1}, t_{i2}, \ldots, t_{im})'$ of the $i$-th subject ($i = 1, \ldots, n$), $W_i = (w_{i1}, w_{i2}, \ldots, w_{im})'$ be observed values of a covariate such as BMI. Here we suppose that $\{t_{ij}\}$ and $\{w_{ij}\}$ are scaled into the interval $(0,1)$. Assume that the first two moments of the binary response $y_{ij}$ are $E(y_{ij}|x_{ij}, w_{ij}) = u_{ij}$ and $\text{Cov}(y_{ij}, y_{ik}|x_{ij}, x_{ik}, t_{ij}, t_{ik}) = \sigma_{ijk}$, where $x_{ij}$ is a $p_i$-vector covariate. Without loss of generality, the matrices $\Sigma_i = (\sigma_{ijk})$ are supposed to be positive definite. Based on the definition of the latent Gaussian copula model (Fan et al. 2016) [12], we use $R_i$ to denote the latent correlation matrix of $Y_i$, and the bridge function which links the latent correlation coefficient $R_{ijk}$ to the covariance $\sigma_{ijk}$ for binary observations is given by

$$\sigma_{ijk} = E(Y_{ij}Y_{ik}) - E(Y_{ij})E(Y_{ik}) = P(Z_{ij} > C_{ij}, Z_{ik} > C_{ik}) - P(Z_{ij} > C_{ij})P(Z_{ik} > C_{ik})$$

$$= P(Z_{ij} \leq C_{ij}, Z_{ik} \leq C_{ik}) - P(Z_{ij} \leq C_{ij})P(Z_{ik} \leq C_{ik})$$

$$= \Phi_2(C_{ij}, C_{ik}; R_{ijk}) - \Phi(C_{ij})\Phi(C_{ik}) = F(C_i; R_i), \quad (1)$$

where $\Phi_2(u, v; \tau)$ is the cumulative distribution function of the standard bivariate normal distribution with correlation $\tau$; $\{Z_{ij}, i = 1, \ldots, n, j = 1, \ldots, m\}$ are the latent Gaussian variables and $\{C_{ij}, i = 1, \ldots, n, j = 1, \ldots, m\}$ are the cutoff points which can be estimated through $C_{ij} = \Phi^{-1}(1-\mu_{ij})$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard univariate normal distribution.

2.1. Generalized linear models and GEE

By means of the semiparametric regression [16,17,18], we construct two models for the parameters of mean and latent correlation as follows,

$$\ln \left( \frac{u_{ij}}{1-u_{ij}} \right) = x_{ij}' \beta + f(w_{ij}), \quad R_{ijk} = h(t_{ij} - t_{ik}; \alpha), \quad (2)$$

where $\beta$ and $\alpha$ are the associated unknown parameters; $R_{ijk}$ is $(j, k)$-th element of correlation matrix $R_i; x_{ij}, w_{ij}$ and $t_{ij}$ are $p_i \times 1$ vector of covariates, value of covariate in nonparametric model $f(\cdot)$ and time point of $j$-th observation of $i$-th individual, respectively. In this article, for a fixed value of $\alpha$, the function $h(\cdot; \alpha)$ is assumed to be monotone and differentiable. Furthermore, we consider the stationary case, which means that the correlation only depends on the time lag. For example, the AR(1) correlation $h(t_{ij} - t_{ik}; \alpha) = \alpha^{|j-k|}$, where $\alpha$ is a parameter in $(-1,1)$.

The nonparametric function $f$ can be parametrized by some basis functions. For instance, we can use the following trigonometric basis functions

$$s_i(x) = \cos((i\pi x), i = 1, \ldots, k,$$
where \(k\) is the number of trigonometric basis functions. The BIC criterion can be used to select the number of basis functions. We treat the nonparametric function \(f\) as a linear regression function with those basis function as covariates. That is, \(f(w)\) can be approximated by \(\varphi(w)' \rho\). Hence, the regression model can be written as

\[
\ln \left( \frac{u_{ij}}{1 - u_{ij}} \right) \approx x_i' \beta + \varphi(w_{ij})' \rho = b_i' \theta,
\]

where \(b_i' = (x_i', \varphi(w_{ij}))\) and \(\theta = (\beta', \rho')'\). Based on the aforementioned models, we propose the generalized estimating equations below:

\[
S_1(\theta) = \sum_{i=1}^{n} \left( \frac{\partial \mu_i}{\partial \theta} \right) \Sigma_i^{-1} (y_i - \mu_i) = 0,
\]

\[
S_2(\alpha)|_{\theta = \hat{\theta}} = \sum_{i=1}^{n} \left( \frac{\partial \eta_i}{\partial \alpha} \right) M_i^{-1} (H_i - \eta_i) = 0,
\]

where \(H_i\) is the column vector obtained by vectorizing the lower triangular part of estimated covariance matrix (not including the diagonal elements) through column by column and \(\eta_i\) is the corresponding part of the matrix \(F(C_i; R_i(\alpha)) = (F(C_{ij}, C_{ik}; R_{ijk}(\alpha)))_{1 \leq j, k \leq m}\). The solutions of the above two estimating equations, \(\hat{\theta}\) and \(\hat{\alpha}\) say, are termed the GEE estimators of \(\theta\) and \(\alpha\). Note that, \(S_2(\alpha)\) is conditional on the estimator \(\hat{\theta}\) so that the consistency of estimator \(\hat{\alpha}\) can be guaranteed.

2.2. Main algorithm

The quasi-Fisher scoring algorithm is applied to obtain the numerical solutions for all the parameters in our models. The parameter estimators \(\hat{\theta}\) and \(\hat{\alpha}\) can be iteratively updated using weighted generalized least squares. Specifically, given \(\Sigma_i\), the mean parameter estimates \(\{\theta^{(k+1)}, k = 0,1,2, \cdots\}\) can be updated through

\[
\theta^{(k+1)} = \theta^{(k)} + \left[ \sum_{i=1}^{n} \left( \frac{\partial \mu_i}{\partial \theta} \right) \Sigma_i^{-1} \left( \frac{\partial \mu_i}{\partial \theta} \right)' \right]^{-1} \left[ \sum_{i=1}^{n} \left( \frac{\partial \mu_i}{\partial \theta} \right) \Sigma_i^{-1} (y_i - \mu_i) \right]|_{\theta = \theta^{(k)}}.
\]

On the other side, given \(\theta\) and \(M_i\), the latent correlation parameter \(\{\alpha^{(k+1)}, k = 0,1,2, \cdots\}\) can be calculated by

\[
\alpha^{(k+1)} = \alpha^{(k)} + \left[ \sum_{i=1}^{n} \left( \frac{\partial \eta_i}{\partial \alpha} \right) M_i^{-1} \left( \frac{\partial \eta_i}{\partial \alpha} \right)' \right]^{-1} \left[ \sum_{i=1}^{n} \left( \frac{\partial \eta_i}{\partial \alpha} \right) M_i^{-1} (H_i - \eta_i) \right]|_{\alpha = \alpha^{(k)}}.
\]

3. Application

We apply the proposed estimation method to the NorStOP study, which is a large cohort study of middle-aged and elderly people in the North Staffordshire of the UK between 2002 and 2008. For more information about the dataset, we refer to Thomas et al. (2007) [2]. This dataset comprises the information collected from patients aged 50 and over through questionnaires. And the information includes age, gender, alcohol consumption, income status, body mass index (BMI), pain interference (PI) degree of osteoarthritis, and depression status. In our study, age is a numeric covariate; gender is a binary covariate which is categorized as 1 if the participant is male and 0 otherwise; alcohol consumption is a binary covariate which is assigned as 1 if the participant drinks almost every day and 0 otherwise; income status and PI are ordinal covariates whose levels range from 1 to 4 and from 1 to 5, respectively; BMI is calculated from self-reported height and weight, and here we truncate its values to the interval (15,40); depression status is the binary response which might depends on the aforementioned covariates. This dataset is longitudinal and balanced. Although it has some missing values, for convenience we exclude those individuals whose have missing values. The sample size of the pre-processed dataset is 2008.
The objective of this research is to find the dependence of depression status on other explanatory covariates and the association between depression status at different time points. Aiming to model jointly the mean and covariance structures for the NorStOP data, we use the following generalized semiparametric mean model:

$$\ln \left[ \frac{E(depres_{ij})}{1 - E(depres_{ij})} \right] = \beta_0 + f_0(BMI_{ij}) + \beta_1 age_i + \beta_2 gender_i + \beta_3 alcohol_i + \beta_4 income_{1i}$$

$$+ \beta_5 income_{2i} + \beta_6 income_{3i} + \beta_7 PI_{1i} + \beta_8 PI_{2i} + \beta_9 PI_{3i} + \beta_{10} PI_{4i}, \quad (8)$$

where the income status and the PI are dummy variables; $f_0$ is an unknown nonparametric function and we use the aforementioned trigonometric basis functions to estimate $f_0$. The BIC suggests that the optimal number of basis functions is 2 and we present the fitting results in Figure 1 below. For simplicity, we only present the case for independence structure when modelling $M_i$ in estimating equation $S_2$. And also, the correlation structures adopted in modelling the latent correlation matrices are as follows, for $j, k = 1, \ldots, m$,

- **Compound Symmetry (CS):** $R_{ijk}(\alpha) = \begin{cases} \alpha_1, & j \neq k \\ 1, & j = k \end{cases}$

- **AR(1):** $R_{ijk}(\alpha) = \alpha_{1}^{\lfloor j-k \rfloor}$

Figure 1. The fitted curves for nonparametric function $f_0$ and estimated confidence limits for two different latent correlation structures. (a) Compound Symmetry, (b) AR(1).

Table 1 lists the parameter estimates and standard errors for $\beta$ and $\alpha$ by using our method. All the parameters are proven to be significant via the test of significance. According to Table 1, it is clear that a participant can be relieved from depression when his/her age grows or economic status improves. Moreover, female or intemperate person are likely to be depressive, which is in accordance with our ‘stereotype’. Table 1 also shows that the estimates of $\beta$ are very similar in the two cases, which implies that the resulting estimators of mean parameters are robust against misspecification of $\Sigma_i$ or $R_i$, so do their standard errors. Besides, the estimated values of the parameter $\alpha$ in the two cases indicate that there exists a significant latent correlation among depression status at different time points. Aiming to choose the optimal structure, we compute QIC (Pan 2001) [19] for the parameter estimator.
\( \hat{\beta} \). And regarding the values of QIC in the two cases, AR(1) correlation structure is slightly better than CS.

**Table 1.** The NorStOP data. Generalized estimating equation estimates of all parameters in the mean and latent correlation, with standard errors in parentheses. The QIC values are calculated to select the appropriate latent correlation structure for the data.

| Compound Symmetry | AR(1) |
|-------------------|-------|
| \( \beta_0 \)     | -2.1489(0.0593) -2.1559(0.0594) |
| \( \beta_1 \)     | -0.0089(0.0068) -0.0090(0.0068) |
| \( \beta_2 \)     | -0.0957(0.1054) -0.1068(0.1053) |
| \( \beta_3 \)     | 0.0638(0.0346) 0.0533(0.0346) |
| \( \beta_{41} \)  | -0.7697(0.2073) -0.7556(0.2069) |
| \( \beta_{42} \)  | -1.5378(0.2127) -1.5339(0.2121) |
| \( \beta_{43} \)  | -1.6427(0.2452) -1.6345(0.2440) |
| \( \beta_{51} \)  | 0.7180(0.1226) 0.7899(0.1203) |
| \( \beta_{52} \)  | 1.2511(0.1300) 1.2665(0.1299) |
| \( \beta_{53} \)  | 1.9695(0.1346) 1.9977(0.1347) |
| \( \beta_{54} \)  | 2.1972(0.1776) 2.2626(0.1772) |
| \( \alpha_1 \)    | 0.7795(0.1039) 0.7184(0.1270) |
| QIC               | 4142.1 4140.5 |

4. Conclusions

In this paper we analyse the NorStOP data and find some significant scientific results. The statistical technique used is the copula-based GEE2 approach proposed by [1] that provides a valid way to model correlation coefficients for clustered binary data. For the NorStOP data, the binary response, i.e. depression status, is influenced linearly by some covariates, such as income status and PI level, and non-linearly by the effect of BMI. Our study shows that as the BMI increases, the probability that a human being suffers from depression would first decrease gradually, reach the lowest point, and then rise up. Other important finding includes that the PI level is positively associated with the depression status, while the income status is negatively correlated with it. On the other hand, our method suggests that there exist strong within-subject correlations between the repeated measurements of patients. Meanwhile, the QIC shows AR(1) is the likely structure of the latent correlation matrix. The correlation coefficient is positive, meaning that depression can be a kind of chronic coherent disease.

It is worth mentioning that, we also applied the proposed method to analysis of other real datasets arising in social science. The simulation studies in [1] and the real data analysis presented here prove the generalizability and effectiveness of our method. Due to space limit, we cannot present all the results here but will present the details if accepted to give a talk to the conference.

**Acknowledgement**

This research work is supported by the Natural Science Foundation of China under Grant 11871357.

**References**

[1] Peng C, Yang Y, Zhou J and Pan J 2019 Joint modeling of mean-correlation for longitudinal binary data [Manuscript]

[2] Thomas E, Mottram S and Peat G 2007 The effect of age on the onset of pain interference in a general population of older adults Prospective findings from the North Staffordshire Osteoarthritis Project (NorStOP) Pain 129 21-27

[3] Thomas E, Peat G, Harris L et al. 2004 The prevalence of pain and pain interference in a general population of older adults cross-sectional findings from the North Staffordshire Osteoarthritis Project (NorStOP) Pain 110 361-368
[4] Kuchibhatla M and Fillenbaum G G 2003 Comparison of methods for analyzing longitudinal binary outcomes: Cognitive status as an example Aging and Mental Health 7(6) 462-468
[5] Dale J R 1986 Global cross-ratio models for bivariate, discrete, ordered responses Biometrics 42(4) 909-917
[6] Prentice R L 1988 Correlated binary regression with covariates specific to each binary observation Biometrics 44(4) 1033-1048
[7] Liang K Y, Zeger S L and Qaqish B 1992 Multivariate regression analysis for categorical data (with discussion) Journal of the Royal Statistical Society: Series B 54 3-40
[8] Lipsitz S R, Laird N M and Harrington D P 1991 Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association Biometrika 78(1) 153-160
[9] Carey V and Diggle Z P 1993 Modelling multivariate binary data with alternating logistic regressions Biometrika 80(3) 517-526
[10] Cessie S L and Houwelingen J C V 1994 Logistic regression for correlated binary data Journal of the Royal Statistical Society: Series C 43(1) 95-108
[11] Chaganty N R and Joe H 2004 Efficiency of generalized estimating equations for binary responses Journal of the Royal Statistical Society: Series B 66(4) 851-860
[12] Fan J, Liu H, Ning Y and Zou H 2016 High dimensional semiparametric latent graphical model for mixed data Journal of the Royal Statistical Society: Series B 79(2) 405-421
[13] Liang K Y and Zeger S L 1986 Longitudinal data analysis using generalized linear models Biometrika 73(1) 13-22
[14] Prentice R L and Zhao L P 1991 Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses Biometrics 47(3) 825-839
[15] Wang Y G and Carey V 2003 Working correlation structure misspecification, estimation and covariate design: Implications for generalized estimating equations performance Biometrika 90(1) 29-41
[16] Fan J, Huang T and Li R 2007 Analysis of longitudinal data with semiparametric estimation of covariance function Journal of the American Statistical Association 102(478) 632-641
[17] Leng C, Zhang W and Pan J 2010 Semiparametric mean-covariance regression analysis for longitudinal data Journal of the American Statistical Association 105(489) 181-193
[18] Zeger S L and Diggle P J 1994 Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters Biometrics 50(3) 689-699
[19] Pan W 2001 Akaike’s information criterion in generalized estimating equations Biometrics 57(1) 120-125