第50卷第9期 中国科学技术大学学报 Vol. 50, No. 9
2020年9月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Sep. 2020

文章编号:0253-2778(2020)09-1266-11

A Cholesky factor model in correlation modeling for discrete longitudinal data

LI Yezhen, ZHANG Weiping

(*Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei* 230026, *China*)

Abstract: A joint mean-correlation regression model framework was proposed for a family of generic discrete responses either balanced or unbalanced, and a Cholesky decomposition method was used for statistically meaningful reparameterization of correlation structures. To overcome computational intractability in maximizing the full likelihood function of the model, a computationally efficient Monte Carlo expectation maximization (MCEM) approach was proposed. Theoretical properties were also established for the resulting estimators. Simulation studies and a real data analysis show that the proposed approach yields highly efficient estimators for the parameters.

Key words: discrete longitudinal data; Cholesky decomposition; mean-correlation regression model; Monte Carlo expectation maximization

CLC number: O212.1 **Document code**: A doi:10.3969/j.issn.0253-2778.2020.09.006

2010 Mathematics Subject Classification: 62F12

Citation: LI Yezhen, ZHANG Weiping. A Cholesky factor model in correlation modeling for discrete longitudinal data[J]. Journal of University of Science and Technology of China, 2020, 50(9): 1266-1276.

李叶蓁,张伟平. 一种离散纵向数据相依结构建模的 Cholesky 因子模型[J]. 中国科学技术大学学报, 2020,50(9):1266-1276.

一种离散纵向数据相依结构建模的 Cholesky 因子模型

李叶蓁,张伟平

(中国科学技术大学管理学院统计与金融系,安徽合肥 230026)

摘要:对一类响应变量为离散型的平衡或非平衡纵向数据,提出了均值-相关系数联合回归模型框架,并且使用 Cholesky 分解方法对模型的相关结构进行参数化,使其具有良好的统计解释性. 为了解决似然推断中高维积分计算的难题,提出了一种高效的蒙特卡罗期望最大化(MCEM)算法,并证明了参数估计的渐近性质. 模拟实验和实际数据分析表明提出的方法是高度有效的.

关键词:离散纵向数据;Cholesky 分解;均值-相关系数回归模型;蒙特卡罗期望最大化算法

Received: 2020-03-30; **Revised**: 2020-06-18

Foundation item: Supported by the NSFC of China (11671374, 71771203, 71631006).

Biography: LI Yezhen, female, born in 1995, master. Research field: Longitudinal data analysis, statistical inference. E-mail: 605916686@qq.com

Corresponding author: ZHANG Weiping, PhD/Prof. E-mail: zwp@ustc.edu.cn

0 Introduction

Longitudinal data, with repeated measurements collected from the same subject, are frequently encountered. Various models and methods for dealing with longitudinal data analysis are proposed in literature; see, among others, Refs. [1-2] for overviews of the methods in this area. It is well recognized that properly accounting for the correlation between these repeated measurements is important, not only for correct statistical inference, but also for efficiency in estimation. Recently, regression analysis of the covariance structure has attracted increasing attention. See Refs. [3-7] for related discussion. For discrete longitudinal data, however, there is no unified framework for the joint distribution of discrete longitudinal variables. Hence, modeling the within subject correlations/covariances are essentially case by case; see, for example, the Markov chain approach[8] for binary data, and other marginal approaches[9-10] for Poisson, multinomial, and other types of variables. The GEE approaches[11] and their variations have also been employed for studying discrete longitudinal data. For modeling a generic class of discrete and mixed types of variables in longitudinal data, Song et al.[12] proposed a Gaussian copula based approach to integrate separate one-dimensional generalized linear models into a joint regression analysis of continuous, discrete, and mixed correlated outcomes. Most recently, Tang et al.[13] proposed to model the discrete longitudinal responses with the Gaussian copula whose correlation matrix is modeled with the regression approach by using the hyperspherical parametrization[7].

We propose a copula-based joint mean-correlation modeling approach for discrete longitudinal data. The correlation structures for a family of generic discrete responses are decomposed by a moving average Cholesky decomposition. Unlike the hyperspherical parametrization in Ref. [13], the moving average Cholesky decomposition has a more direct interpretation of the statistical meaning and is particularly appealing because of the natural ordering of the variable in longitudinal data. Since the likelihood inference is computationally intractable in general, we develop a type of Monte Carlo expectation maximization (MCEM) based method for estimation.

The article is organized as follows. Section 1 introduces the joint mean-correlation model and the moving average Cholesky decomposition for correlation matrix. Section 2 provides the estimation procedure for our model. Section 3 presents extensive numerical simulations and a real data analysis. Conclusions are found in Section 4.

1 Models

1.1 Gaussian copula for discrete data

We denote by $\boldsymbol{y}_i=(y_{i1},\cdots,y_{im_i})^{\mathrm{T}}$ the m_i longitudinal measurements from the ith subject ($i=1,\cdots,n$), where the discrete response y_{ij} of interests is observed at time t_{ij}. Let $\boldsymbol{t}_i=(t_{i1},\cdots,t_{im_i})^{\mathrm{T}}$ and $\boldsymbol{x}_{ij}\in\mathbb{R}^p$ be the explanatory variable associated with the jth measurement of subject i. Here $\boldsymbol{t}_i$ and m_i can be subject specific so that the model is capable of handling unbalanced longitudinal data sets observed at irregularly spaced time points.

With multiple subjects, we denote the observations as $\{y_{ij},\boldsymbol{x}_{ij},t_{ij}\}$ ($i=1,\cdots,n$; $j=1,\cdots,m_i$). For categorical responses, we assume that y_{ij} follows an exponential family distribution so that the generalized linear models (GLMs) can be used for the discrete responses marginally[14]. That is, the probability mass function of y_{ij} takes the form $f(y)=c(y;\varphi)\exp\{[y\theta-\psi(\theta)]/\varphi\}$ with the canonical parameter θ and scale parameter φ.

For parsimoniously modeling the mean of y_{ij}, $\mu_{ij}=E(y_{ij})$, the traditional strategy of GLM is applied to incorporate the available explanatory variable $\boldsymbol{x}_{ij}$:

$$g(\mu_{ij})=\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} \tag{1}$$

with the known transformation $g(\cdot)$ and unknown model parameter $\boldsymbol{\beta}$.

To fill the gap between the marginal distributions and the joint distributions, we consider the copula model. A function $C(u_1,\cdots,u_d)$ is called a copula function if it is a distribution function with each of its margins uniformly distributed on [0, 1]. That is,

$$C(\boldsymbol{u})=P(U_1\leqslant u_1,\cdots,U_d\leqslant u_d),$$

where each U_j ($j=1,\cdots,d$) is uniformly distributed on [0,1] and $\boldsymbol{u}=(u_1,\cdots,u_d)$. The Sklar's theorem ensures the existence of such multivariate function. In the context of longitudinal data, the joint distribution function of $\boldsymbol{y}_i$ and a copula $C(u_{i1},\cdots,u_{im_i})$ are connected via

$$\begin{aligned}F_{m_i}(\boldsymbol{y}_i)=P(Y_{i1}\leqslant y_{i1},\cdots,Y_{im_i}\leqslant y_{im_i})=\\P(F_{i1}(Y_{i1})\leqslant F_{i1}(y_{i1}),\cdots,\\F_{im_i}(Y_{im_i})\leqslant F_{im_i}(y_{im_i}))=\\F(F_{i1}^{-1}(u_{i1}),\cdots,F_{im_i}^{-1}(u_{im_i}))=\\C(u_{i1},\cdots,u_{im_i})\end{aligned}\tag{2}$$

where $u_{ij}=F_{ij}(y_{ij})$, $j=1,\cdots,m_i$ and $F_{ij}(\cdot)$ is the marginal cumulative distribution function of y_{ij}. In this paper, we use the Gaussian copula as it has the additional advantage of allowing a flexible parametric dependence structure. Then the joint distribution of $\boldsymbol{y}_i$ in (2) is

$$F_{m_i}(\boldsymbol{y}_i)=\Phi_{m_i}(z_{i1},\cdots,z_{im_i};\boldsymbol{R}_i)=\Phi_{m_i}(\boldsymbol{z}_i;\boldsymbol{R}_i)\tag{3}$$

where $\Phi_{m_i}(\cdots;\boldsymbol{R}_i)$ is the probability distribution function (PDF) of the m_i dimensional normal distribution with zero mean and correlation matrix $\boldsymbol{R}_i$, and $z_{ij}=\Phi^{-1}\{F_{ij}(y_{ij})\}$ for $j=1,\cdots,m_i$, with $\Phi(\cdot)$ being the PDF of the univariate standard normal distribution.

It should also be noticed that the entries of $\boldsymbol{R}_i$ are not directly the coefficients of correlation between the discrete observations, but they determine the dependence of the longitudinal observations via Eq. (3). Song[15] discussed the connection between the correlation coefficients in $\boldsymbol{R}_i$ and those of the observed variables explicitly.

1.2 Moving average Cholesky decomposition

Modeling correlation (and covariance) matrices can be challenging due to the positive-definiteness constraint. This problem can be removed by infusing regression-based ideas into Cholesky decomposition[16].

The standard Cholesky decomposition of an $m_i\times m_i$ positive definite covariance matrix is of the following form:

$$\boldsymbol{\Sigma}_i=\boldsymbol{C}_i\boldsymbol{C}_i^{\mathrm{T}}\tag{4}$$

where $\boldsymbol{C}_i=(c_{ijk})$ is a lower triangular matrix with positive diagonal elements and its entries are difficult to interpret. Pre-multiplying $\boldsymbol{C}_i$ by the inverse of $\boldsymbol{D}_i=\mathrm{diag}(c_{i11},c_{i22},\ldots,c_{im_im_i})$ leads to an alternative Cholesky decomposition (ACD)[17], and keeps $\boldsymbol{D}_i$ outside, and we have

$$\boldsymbol{\Sigma}_i=\boldsymbol{D}_i(\boldsymbol{D}_i^{-1}\boldsymbol{C}_i)(\boldsymbol{C}_i^{\mathrm{T}}\boldsymbol{D}_i^{-1})\boldsymbol{D}_i=\boldsymbol{D}_i\widetilde{\boldsymbol{L}}_i\widetilde{\boldsymbol{L}}_i^{\mathrm{T}}\boldsymbol{D}_i\tag{5}$$

where $\widetilde{\boldsymbol{L}}_i=\boldsymbol{D}_i^{-1}\boldsymbol{C}_i$ is obtained from a slightly different standardised $\boldsymbol{C}_i$, dividing each row by its corresponding diagonal entry.

For statistical interpretation of the below-diagonal entries of $\widetilde{\boldsymbol{L}}_i$, it is clear that $\boldsymbol{D}_i^{-1}(\boldsymbol{z}_i-\bar{\boldsymbol{z}}_i)$ has $\widetilde{\boldsymbol{L}}_i\widetilde{\boldsymbol{L}}_i^{\mathrm{T}}$ as the standard Cholesky decomposition of its covariance matrix and

$$\boldsymbol{\varepsilon}_i=(\boldsymbol{D}_i\widetilde{\boldsymbol{L}}_i)^{-1}(\boldsymbol{z}_i-\bar{\boldsymbol{z}}_i),$$

its vector of innovations, has $\mathrm{Cov}(\varepsilon_i)=\boldsymbol{I}_{m_i}$. Denote $\widetilde{\boldsymbol{L}}_i=(L_{ijk})$ and $\boldsymbol{D}_i=(\sigma_{ij})$, we obtain variable-order, moving average representation for the standardized $\dfrac{(z_{ij}-\bar{z}_{ij})}{\sigma_{ij}}$ as

$$\frac{(z_{ij}-\bar{z}_{ij})}{\sigma_{ij}}=\varepsilon_{ij}+\sum_{k=1}^{j-1}L_{ijk}\varepsilon_{ik}\tag{6}$$

Then we can prove, for any $1\leqslant j,k\leqslant m_i$, the correlation coefficient between z_{is} and z_{it} is given by

$$R_{ijk}=\mathrm{corr}(z_{ij},z_{ik})=\frac{\sum_{s=1}^{j\wedge k}L_{ijs}L_{iks}}{\sqrt{\sum_{s=1}^{j}L_{ijs}^2\sum_{s=1}^{k}L_{iks}^2}}\tag{7}$$

This property is a great motivation for modeling a correlation matrix.

In our approach, we parameterize the moving average parameters $\{L_{ijk}\}_{j>k}$ to overcome the over-parametrization problem. We propose to model these unconstrained parameters collectively via a regression model

$$L_{ijk}=\boldsymbol{w}'_{ijk}\gamma \tag{8}$$

where $\boldsymbol{w}_{ijk}\in\mathbb{R}^q$ is the explanatory variable and γ is the $q\times 1$ unknown parameters, we need to examine the covariates of the ith subject at the corresponding observations. We follow the convention of longitudinal data analysis by taking $\boldsymbol{w}_{ijk}$ as some function of the time lag $|t_{ij}-t_{ik}|$ between observations, which effectively ensures the correlation to be stationary. Other time-dependent covariates may also be meaningfully exploited. Thus our regression approach for the correlations can incorporate a broad class of covariates available for explaining the covariations between longitudinal measurements.

2 Estimation

We use the GLM for the responses marginally characterized by marginal parameters $\boldsymbol{\eta}=(\boldsymbol{\beta},\varphi)$, the copula model 3 for the joint distribution, and the Cholesky model for the correlation $\boldsymbol{R}_i$ with the parameters γ in (8). By combining all unknown parameters in this modeling framework, we write collectively the parameter vector of interest as $\boldsymbol{\theta}=(\boldsymbol{\beta},\varphi,\boldsymbol{\gamma})$. Then we are ready to develop the maximum likelihood estimators with the discrete longitudinal responses:

$$L(\theta)=\prod_{i=1}^{n}P(Y_{i1}=y_{i1},\cdots,Y_{im_i}=y_{im_i})=\prod_{i=1}^{n}P(y_{i1}-1<Y_{i1}\leqslant y_{i1},\cdots,y_{im_i}-1<Y_{im_i}\leqslant y_{im_i})=\prod_{i=1}^{n}\int\cdots\int_{z_i^-<\boldsymbol{u}\leqslant z_i}\phi_{m_i}(\boldsymbol{u};\boldsymbol{R}_i)\mathrm{d}\boldsymbol{u} \tag{9}$$

where $z_i=(z_{i1},\cdots,z_{im_i})^{\mathrm{T}}$ and $\boldsymbol{z}_i^-=(z_{i1}^-,\cdots,z_{im_i}^-)^{\mathrm{T}}$ with

$$z_{ij}=\Phi^{-1}\{F_{ij}(y_{ij})\},\ z_{ij}^-=\Phi^{-1}\{F_{ij}(y_{ij}-1)\},$$

and $z_{ij}^-=-\infty$ when y_{ij} takes the smallest possible value on its support.

However, evaluating the likelihood (9) involves the probability distribution functions of the m_i-dimensional normal distribution, whose analytical form is not available and numerical approximation is required. The computational cost of numerical approximation is high and may not scale easily to even a moderate number of repeat measurements. In fact, directly calculating the distribution function of each subject i specified by (2) requires 2^{m_i} summations of lower dimensional distribution functions as in the approach of Ref. [12], and the computational cost grows exponentially with m_i.

Inspired by Ref. [18], we implement a type of Monte Carlo expectation maximization (MCEM) algorithm[19] to estimate this integral. The expectation maximization (EM) algorithm[20] is a method to maximize the likelihood function in the presence of missing data $\boldsymbol{z}$. This is done iteratively. In the E-step one calculates the Q function, viz.

$$Q(\boldsymbol{\gamma},\hat{\boldsymbol{\gamma}}^{(m)})=\int_{\boldsymbol{z}}f(\boldsymbol{z}\mid\boldsymbol{y};\hat{\boldsymbol{\gamma}}^{(m)})\ln f(\boldsymbol{z};\boldsymbol{\gamma})\mathrm{d}\boldsymbol{z} \tag{10}$$

which is the expectation of the log likelihood with respect to the conditional predictive distribution $f(\boldsymbol{z}\mid\boldsymbol{y};\boldsymbol{R}(\hat{\boldsymbol{\gamma}}^{(m)}))$, under the current value of the model parameters $\hat{\boldsymbol{\gamma}}^{(m)}$ at the mth iteration. The Q function is then maximized in the M-step to find the new value of the model parameters, viz.

$$\hat{\boldsymbol{\gamma}}^{(m+1)}=\underset{\boldsymbol{\gamma}}{\arg\max}\,Q(\boldsymbol{\gamma},\hat{\boldsymbol{\gamma}}^{(m)}) \tag{11}$$

These steps are repeated iteratively until convergence. When the Q function is not available in closed form, a Monte Carlo estimate of the required expectation can be used instead. This is the MCEM algorithm. The Q function is replaced by

$$Q(\boldsymbol{\gamma},\hat{\boldsymbol{\gamma}}^{(m)})=\frac{1}{K}\sum_{k=1}^{K}\ln f(\boldsymbol{z}_k;\boldsymbol{R}(\hat{\boldsymbol{\gamma}}^{(m)})) \tag{12}$$

in the E-step, where $\boldsymbol{z}_1,\cdots,\boldsymbol{z}_K$ are drawn from $f(\boldsymbol{z}\mid\boldsymbol{y};\hat{\boldsymbol{\gamma}}^{(m)})$.

The Dunn-Smyth residuals[21] are a useful diagnostic tool for generalized linear modeling, which are used here as a device for numerical approximation of the integrand in Eq. (9). Let u_{ij} be independent draws from a standard uniform random variable U(0,1). We first define $u_{ij} = F_{ij}(y_{ij}) + u_{ij} f_{ij}(y_{ij})$, which are uniformly distributed on the (0,1) interval, if y_{ij} has the marginal distribution function F_{ij}. A Dunn-Smyth residual is then defined by $\zeta_{ij} = \Phi^{-1}(u_{ij})$. The distribution of these residuals, conditional on the data and marginal distributions, is a truncated multivariate normal with an identity covariance matrix. We can write the distribution of the vector of Dunn-Smyth residuals as

$$g(\boldsymbol{\zeta}_i) = \frac{\prod_{j=1}^{d} \phi(\zeta_{ij})}{\prod_{j=1}^{d} f_{ij}(y_{ij})} \tag{13}$$

This distribution has positive probability only in the region of integration of the likelihood defined in Eq. (9), making it a candidate for importance sampling to estimate this integral.

Theorem 2.1 The likelihood of the discrete Gaussian copula can be approximated by importance sampling with K sets of Dunn-Smyth residuals

$$L(\boldsymbol{\beta}, \varphi, \boldsymbol{\theta} \mid \boldsymbol{y}_1, \cdots, \boldsymbol{y}_n) = \prod_{i=1}^{n} \int_{z_i^- < z \leqslant z_i} \phi_{d_i}(\boldsymbol{z}; \boldsymbol{R}_i(\boldsymbol{\gamma})) \mathrm{d}z \simeq \prod_{i=1}^{n} \prod_{j=1}^{d_i} f_{ij}(y_{ij}) \prod_{i=1}^{n} \sum_{k=1}^{K} \frac{\phi_{d_i}(\zeta_i^k; \boldsymbol{R}_i(\boldsymbol{\gamma}))}{\prod_{j=1}^{d_i} \phi(\zeta_{ij}^k)},$$

where f_{ij} is the marginal density of variable j and observation i, and ζ_i are Dunn-Smyth residuals distributed by g.

Proof We can approximate the likelihood by importance sampling with K sets of Dunn-Smyth residuals

$$L_i(\boldsymbol{\beta}, \varphi, \boldsymbol{\gamma} \mid \boldsymbol{y}_i) = \int_{z_i^- < z \leqslant z_i} \phi_{d_i}(\boldsymbol{z}; \boldsymbol{R}_i(\boldsymbol{\gamma})) \mathrm{d}z = \int_{z_i^- < z \leqslant z_i} \phi_{d_i}(\boldsymbol{z}; \boldsymbol{R}_i(\boldsymbol{\gamma})) \frac{\prod_{j=1}^{d_i} f_{ij}(y_{ij})}{\prod_{j=1}^{d_i} \phi(z_{ij})} g(\boldsymbol{z}) \mathrm{d}z = \prod_{j=1}^{d_i} f_{ij}(y_{ij}) \int_{z_i^- < z \leqslant z_i} \frac{\phi_{d_i}(\boldsymbol{z}; \boldsymbol{R}_i(\boldsymbol{\gamma}))}{\prod_{j=1}^{d_i} \phi(z_{ij})} g(\boldsymbol{z}) \mathrm{d}z \tag{14}$$

which can be approximated using K samples from g, viz.

$$L_i(\boldsymbol{\beta}, \varphi, \boldsymbol{\gamma} \mid \boldsymbol{y}_i) \simeq \prod_{j=1}^{di} f_{ij}(y_{ij}) \sum_{k=1}^{K} \frac{\phi_{d_i}(\zeta_i^k; \boldsymbol{R}_i(\boldsymbol{\gamma}))}{\prod_{j=1}^{d_i} \phi(\zeta_{ij}^k)} \tag{15}$$

The parameters $(\boldsymbol{\beta}, \varphi, \boldsymbol{\gamma})$ are estimated via the algorithm in two steps. First, the marginal parameters $(\boldsymbol{\beta}, \varphi)$ are estimated by assuming independence, as with independence estimating equations. Second, these estimates $(\hat{\boldsymbol{\beta}}, \hat{\varphi})$ are plugged into the likelihood $L(\boldsymbol{\beta}, \varphi, \boldsymbol{\gamma} \mid \boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)$, defined in Eq. (14). The resulting plug-in likelihood $L(\boldsymbol{\gamma} \mid \hat{\boldsymbol{\beta}}, \hat{\varphi}, \boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)$ is maximized for correlation parameters $\boldsymbol{\gamma}$. Such algorithms have good asymptotic properties, including asymptotic efficiency relative to maximum likelihood[22].

We note that the observations $\boldsymbol{y}_i \sim f(\boldsymbol{y}, \eta, \boldsymbol{\gamma})$ for $i \in \{1, \cdots, n\}$. To establish the theoretical properties, we assume the following regularity conditions hold.

Condition C1 The parameter space Θ contains an open set of which the true parameter $\boldsymbol{\theta}_0$ is an interior point.

Condition C2 There exists an open subset of $\boldsymbol{\omega} \in \Omega$ containing $\boldsymbol{\eta}_0$ and an integrable function $M_r(\boldsymbol{y})$, such that for every $\boldsymbol{\eta} \in \omega$ and $\boldsymbol{y} \in \boldsymbol{y}$,

$$|\partial^3 \ln f(\boldsymbol{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}) / \partial^3 \eta_r| \leqslant M_r(\boldsymbol{y})$$

for $r \in \{1, \cdots, \dim(\boldsymbol{\eta})\}$, where $E_{\boldsymbol{\eta}_0}\{M_r(\boldsymbol{y})\} < \infty$.

Condition C3 For $r \in \{1, \cdots, \dim(\boldsymbol{\eta})\}$ there are bounded functions $V_r(\boldsymbol{y})$ such that in the neighborhood of $\boldsymbol{\eta}_0$ for any fixed γ,

$$\{\partial \ln f(\boldsymbol{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}) / \partial \eta_r\}^2 \leqslant V_r(\boldsymbol{y})$$

with $E_{\boldsymbol{\eta}_0}\{V_r(\boldsymbol{y})\} < \infty$.

Theorem 2.2 Under the conditions C1～C3, the estimators $(\hat{\boldsymbol{\beta}}, \hat{\varphi}, \hat{\boldsymbol{\gamma}})$ are consistent and asymptotically normally distributed.

The technical proof is given in Appendix.

3 Numerical studies

3.1 Simulations

In this section, we conduct simulation studies to investigate the finite performances of the proposed method. In each of the studies, we generated 500 data sets and took sample sizes $n=50$, 100 and 200, and the Monte Carlo sample size $K=200$. We also compared our method to the GEE method[23] and the generalized linear mixed-effects models(GLMM)[24] for estimating the parameters in the mean model and the dispersion. In GEE method, we assume unstructured correlations, and in GLMM, we consider time as random effects. All simulations were conducted in R.

Study 3.1 In this study we consider that the marginal distributions F_{ij} ($j=1,\ldots,m_i$) for n subjects as the negative binomial distribution $y_{ij}\sim \text{NegBin}(\delta,\mu_{ij})$ with mean μ_{ij} and variance $\mu_{ij}+\mu_{ij}^2/\varphi$, where $\varphi>0$ is the over-dispersion parameter and the numbers of repeated measurements m_i for each subject satisfies $m_i-1\sim \text{Binomial}(5, 0.8)$. The mean was then parameterized as $\mu_{ij}=\exp(x_{ij}^{\mathrm{T}}\beta)$ to allow dependence on explanatory variables, and the variance exceeds its mean (i. e. over-dispersion). The explanatory variables x_{ij1} and x_{ij2} were bivariate normal with correlation 0.5. The parameters in the correlation matrix was set as $L_{ijk}=\gamma_0+w_{ijk1}\gamma_1+w_{ijk2}\gamma_2$ for the moving-average (MA) structure as in Eq. (6) with $\boldsymbol{w}_{ijk}=\{1,t_{ij}-t_{ik},(t_{ij}-t_{ik})^2\}^{\mathrm{T}}$ and the measurement times t_{ij}s were uniformly distributed. The true parameters were taken as $\boldsymbol{\beta}=(\beta_0,\beta_1,\beta_2)=(1,-0.3,0.5)$, $\delta=4$ and $\boldsymbol{\gamma}=(\gamma_0,\gamma_1,\gamma_2)=(0.5,-0.3,0.3)$.

Tab. 1 shows the accuracy of the estimated parameters in terms of their mean biases (MB) and standard deviations (SD) in three methods. For the GLMM method, the dispersion parameter is fixed, because the estimation of it is always unstable. We can see the biases and the standard deviations decrease as the sample size increases, and all the biases are small especially when n is large. Compared to the GEE and GLMM estimates for the parameters in the mean model, our method have very competitive performances. In terms of dispersion parameter, the bias of our method is about one-half of the GEE, and of course GLMM. For the biases of other mean parameters, our method performs much better than GLMM, probably because we need to correctly model the random effect additionally in GLMM. And the SD of our method are slightly smaller than GLMM. Additionally, we can see that the biases are close but always smaller than GEE, and the SD of our method are much smaller than the SD of GEE. It shows that our method is more stable than GEE. All above show the advantage of proposed method.

Tab. 1 Simulation results for Study 3.1. Mean bias (MB) and standard deviation (SD) of mean parameters

	our method			GEE			GLMM		
	50	100	200	50	100	200	50	100	200
MB_{β_0}	0.010	0.000	0.001	0.007	0.002	0.001	0.116	0.112	0.115
SD_{β_0}	0.085	0.060	0.041	0.128	0.129	0.137	0.096	0.067	0.047
MB_{β_1}	0.002	0.003	0.002	0.002	0.001	0.005	0.005	0.005	0.000
SD_{β_1}	0.052	0.038	0.023	0.096	0.089	0.109	0.053	0.040	0.025
MB_{β_2}	0.000	0.002	0.002	0.003	0.009	0.003	0.005	0.006	0.002
SD_{β_2}	0.055	0.038	0.026	0.119	0.158	0.093	0.057	0.040	0.028
MB_{ψ}	0.767	0.359	0.166	1.464	0.913	0.381	—	—	—
SD_{ψ}	1.817	1.075	0.660	2.395	1.916	1.401	—	—	—
MB_{γ_0}	0.001	0.006	0.004	—	—	—	—	—	—
SD_{γ_0}	0.151	0.103	0.073	—	—	—	—	—	—
MB_{γ_1}	0.072	0.003	0.001	—	—	—	—	—	—
SD_{γ_1}	0.917	0.613	0.438	—	—	—	—	—	—
MB_{γ_2}	0.081	0.010	0.008	—	—	—	—	—	—
SD_{γ_2}	1.151	0.764	0.532	—	—	—	—	—	—

Study 3.2 The data sets were generated from the model

$$y_{ij} \sim \text{Bernoulli}(p_{ij}),$$
$$\text{logit}(p_{ij}) = \beta_0 + x_{ij1}\beta_1 + x_{ij2}\beta_2,$$
$$L_{ijk} = \gamma_0 + w_{ijk1}\gamma_1 + w_{ijk2}\gamma_2,$$
$$i = 1, \cdots, n;\ j = 1, \cdots, m_i,$$

where again the measurement times t_{ij}'s were uniform and $m_i - 1 \sim \text{Binomial}(5, 0.8)$. The covariate $\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2})^{\mathrm{T}}$ was generated from a standard bivariate normal distribution with zero correlation. We took the covariates for the correlations as $\boldsymbol{w}_{ijk} = \{1, t_{ij} - t_{ik}, (t_{ij} - t_{ik})^2\}^{\mathrm{T}}$. The parameters were set as $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2) = (1.0, -0.3, 0.5)$ and

$$\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2) = (0.5, -0.3, 0.3).$$

Tab. 2 shows that the results are qualitatively similar to those in Study 3.1. When $n = 50$ and 100, the mean biases of GEE are slightly smaller than our method, but the standard deviations are lager than our method. And when $n = 200$, our method has the best performance. Overall, our method performs promisingly and indicates the potential benefit for estimating the mean model incorporating the correlations in the longitudinal data by using a parsimonious correlation model.

Tab. 2 Simulation results for Study 3.2. Mean bias (MB) and standard deviation (SD) of mean parameters

	our method			GEE			GLMM		
	50	100	200	50	100	200	50	100	200
MB_{β_0}	0.027	0.026	0.017	0.051	0.031	0.018	0.141	0.044	0.015
SD_{β_0}	0.225	0.149	0.105	0.235	0.153	0.106	1.092	0.501	0.106
MB_{β_1}	0.018	0.012	0.004	0.011	0.008	0.008	0.020	0.012	0.006
SD_{β_1}	0.169	0.116	0.089	0.174	0.128	0.090	0.197	0.122	0.094
MB_{β_2}	0.018	0.016	0.001	0.011	0.010	0.004	0.025	0.017	0.003
SD_{β_2}	0.186	0.119	0.090	0.193	0.123	0.090	0.198	0.132	0.096
MB_{γ_0}	0.063	0.057	0.016	—	—	—	—	—	—
SD_{γ_0}	0.299	0.207	0.128	—	—	—	—	—	—
MB_{γ_1}	0.139	0.113	0.076	—	—	—	—	—	—
SD_{γ_1}	1.904	1.219	0.771	—	—	—	—	—	—
MB_{γ_2}	0.238	0.076	0.029	—	—	—	—	—	—
SD_{γ_2}	2.385	1.460	1.460	—	—	—	—	—	—

3.2 Online shopping data

E-commerce application has become one of the most commonly used Internet applications all over the world. Research on behavior of shopping users, especially its impact on purchasing, is of great significance to deeply understand user's online purchase behavior, discover high potential users, and promote consumption.

In this section, we consider an application of our methods using online shopping logs data accumulated by Tmall.com. We analyze a randomly selected sample of $n = 1\ 000$ visits to Tmall.com from May to October in 2015. The response variable is the number of times a user make purchases. Users on average made 4.23 visits with a standard deviation 1.92, resulting in a highly unbalanced repeated measurement data set. We also consider a set of covariates that could explain the variation in users' consumption behavior, including users' personal information and shopping logs. Tab. 3 describes these explanatory variables in detail. In the application, the nonlinear effect of age is assumed including age^2 among the covariates.

The model described in Section 1 is fitted on this dataset. To account for the over-dispersion, we used a parametric negative binomial regression model for the mean,

$$\text{action}_{ij} \sim \text{NegBin}(\varphi, \mu_{ij}),$$
$$\ln(\mu_{ij}) = \ln(\text{time}_{ij}) + \beta_0 + \beta_1 \text{click}_{ij} + \beta_2 \text{favor}_{ij} + \beta_3 \text{age}_i + \beta_4 \text{age}_i^2 + \beta_5 \text{gender}_i,$$

where φ is the over-dispersion parameter, time_{ij} is the month of time from the start day. The $\ln(\text{time}_{ij})$ is needed to account for different observation periods.

Tab. 3　Variable descriptions

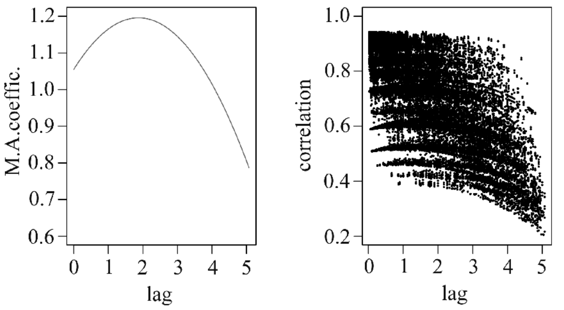
variable	description
action	number of times a user make purchases
age	1 for <18; 2 for [18,24]; 3 for [25,29] 4 for [30,34]; 5 for [35,39]; 6 for [40,49] 7 for ≥50
age²	square of age
gender	0 for female, 1 for male
merchant	number of merchants clicked
click	number of items clicked
favor	number of items added to favourite

Tab. 4 reports the estimates of the regression parameters. And also as a comparison, a GLMM approach with visit-time as random effect was implemented. From the results, we notice that all the explanatory variables are statistically significant. The variable click is a positively significant variable, indicating that users who click more will have more chance to make purchases, given other explanatory variables. The times users make purchases is negatively related to favor and gender. Since gender takes 1 for a man and 0 for a woman, it is obvious that a man tends to make less purchases. And the fact that favor is negatively correlated to purchase suggests users who add an item to favorites may be less likely to buy it at the same time. As for nonlinear relations, the quadratic effects of age is negatively significant, which means that middle-aged people will make more purchases. The over-dispersion parameter $\hat{\delta}=0.682$ is significant, suggesting that the counts are over-dispersed.

Tab. 4　Estimates of the parameters

	our method			GLMM		
	Est.	SE	p-value	Est.	SE	p-value
(intercept)	−3.693	0.514	<0.01	−3.688	0.482	<0.01
click	0.045	0.004	<0.01	0.034	0.004	<0.01
favor	−0.244	0.051	<0.01	−0.189	0.043	<0.01
age	1.133	0.219	<0.01	1.195	0.204	<0.01
age²	−0.098	0.023	<0.01	−0.109	0.021	<0.01
gender	−0.113	0.049	0.021	−0.120	0.047	0.01
φ	0.682	0.028	<0.01	0.832	−	−

For the parameters in the correlation, we obtained $\hat{\gamma}_0=1.055$, $\hat{\gamma}_1=0.142$, $\hat{\gamma}_2=-0.037$. We summarize the fitting with some plots. Fig. 1(a) shows the plots of the fitted M. A. coefficient versus the time lag, suggesting that a polynomial model for correlations is reasonable. The curved pattern between the correlation and time in Fig. 1(b) is interesting, which may be due to the fact that purchases made within a short period of time are more correlated to each other but such an effect becomes weaker in a relative longer term.



(a) plot of fitted M.A. coefficient versus time lag　(b) plot of fitted correlations versus time lag

Fig. 1　The online shopping data

4 Conclusion

With discrete longitudinal data, we propose a mean-correlation model based on moving average Cholesky decomposition. For this class of models, computing the full likelihood is often infeasible. Therefore, we propose a computationally efficient MCEM approach for model estimation. Our method can deal with any set of marginal distributions in Gaussian copula model, and is simple and flexible to implement. We assess the performance by a series of simulation studies which show that our approach has very competitive performances. Overall, the proposed framework is revealing some informative features from the statistical modeling for generic data with temporal dependence. It is benefits to discover interesting dependence properties of the covariance structures. Topics for future research include the feature selection for the mean-correlation model and the model diagnostic tools for assessing model adequacy.

References

[1] DIGGLE P J, HEAGERTY P, LIANG K Y, et al. Analysis of Longitudinal Data[M]. 2nd ed. Oxford, UK: Oxford University Press, 2002.

[2] FITZMAURICE G, DAVIDIAN M, VERBEKE G, et al. Longitudinal Data Analysis [M]. London: Chapman and Hall, 2009.

[3] POURAHMADI M. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation[J]. Biometrika, 1999, 86: 677-690.

[4] YE H, PAN J. Modelling covariance structures in generalized estimating equations for longitudinal data [J]. Biometrika, 2006, 93: 927-941.

[5] DANIELS M J, POURAHMADI M. Modeling covariance matrices via partial autocorrelations [J]. Journal of Multivariate Analysis, 2009, 100: 2352-2363.

[6] ZHANG W P, LENG C L. A moving average Cholesky factor model in covariance modeling for longitudinal data[J]. Biometrika, 2012, 99: 141-150.

[7] ZHANG W P, LENG C L, TANG C Y. A joint modeling approach for longitudinal studies[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology),2015, 77: 219-238.

[8] ZEGER S L, LIANG K Y, SELF S G. The analysis of binary longitudinal data with time independent covariates[J]. Biometrika, 1985, 72: 31-38.

[9] MOLENBERGHS G, VERBEKE G. Models for Discrete Longitudinal Data [M]. New York: Springer, 2005.

[10] SUTRADHAR B C. Longitudinal Categorical Data Analysis[M]. New York: Springer, 2014.

[11] LIANG K Y, ZEGER S L. Longitudinal data analysis using generalized linear models[J]. Biometrika, 1986, 73: 13-22.

[12] SONG P X K, LI M, YUAN Y. Joint regression analysis of correlated data using Gaussian copulas[J]. Biometrics, 2009, 65: 60-68.

[13] TANG C Y, ZHANG W P, LENG C L. Discrete longitudinal data modeling with a mean-correlation regression approach [J]. Statistica Sinica, 2019, 29 (2): 283-302.

[14] MCCULLAGH P, NELDER J A. Generalized Linear Models[M]. London: Chapman and Hall, 1989.

[15] SONG P X K. Multivariate dispersion models generated from Gaussian copula [J]. Scandinavian Journal of Statistics, 2000, 27: 305-320.

[16] POURAHMADI M. High-Dimensional Covariance Estimation: With High-Dimensional Data [M]. Hoboken, NJ: Wiley, 2013.

[17] CHEN Z, DUNSON D B. Random effects selection in linear mixed models[J]. Biometrics, 2003, 59(4): 762-769.

[18] POPOVIC G C, HUI F K C, WARTON D I. A general algorithm for covariance modeling of discrete data[J]. Journal of Multivariate Analysis, 2018, 165: 86-100.

[19] WEI G C, TANNER M A. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms[J]. Journal of the American Statistical Association, 1990, 85: 699-704.

[20] DEMPSTER A P, LAIRD N M, RUBIN D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1977, 39: 1-38.

[21] DUNN P K, SMYTH G K. Randomized quantile residuals[J]. Journal of Computational and Graphical Statistics, 1996, 5: 236-244.

[22] JOE H. Asymptotic efficiency of the two-stage estimation method for copula-based models[J]. Journal of Multivariate Analysis, 2005, 94: 401-419.

[23] ZEGER S L, LIANG K Y. Longitudinal data analysis for discrete and continuous outcomes[J]. Biometrics, 1986, 42: 121-130.

[24] JIANG J. Linear and Generalized Linear Mixed Models and Their Applications [M]. New York: Springer, 2007.

[25] FERGUSON T S. A Course in Large Sample Theory [M]. London: Chapman and Hall/CRC, 1996.

Appendix

Our method is a two-step estimation procedure, where we estimate $\boldsymbol{\beta}$ and φ from a marginal likelihood and then maximize the conditional likelihood given these estimates. By treating $\boldsymbol{\beta}$ and φ as nuisance parameters, we can get consistent estimators of parameters $\boldsymbol{\gamma}$.

Proof of Theorem 2.2

Let $\boldsymbol{\theta}_0=(\boldsymbol{\eta}_0, \boldsymbol{\gamma}_0)$ be the true parameters, and $\hat{\boldsymbol{\eta}}$ be the estimated coefficients where $\hat{\boldsymbol{\eta}}_j$ is found by maximizing the jth marginal likelihood,

$$\hat{\boldsymbol{\eta}}_j=\underset{\boldsymbol{\eta}_j}{\operatorname{argmax}} \ln L_j(\boldsymbol{y}_j, \boldsymbol{\eta}_j) \tag{A1}$$

Because maximizing Eq. (A1) is equivalent to using independence estimating equations in the GEE framework[23], which is consistent under Condition C1 and Condition C2, we now state a result, without proof, concerning the consistency of the marginal parameters.

Lemma A.1 Under some regular conditions C1~C3, $\hat{\boldsymbol{\eta}}\to\boldsymbol{\eta}_0$ in probability as $n\to\infty$.

Analogously to the proof of standard maximum likelihood estimation in Ref. [25], we define

$$\tau(\boldsymbol{\gamma})=\frac{1}{n}\ln\frac{L(\boldsymbol{y}; \hat{\boldsymbol{\eta}}, \boldsymbol{\gamma})}{L(\boldsymbol{y}; \hat{\boldsymbol{\eta}}, \boldsymbol{\gamma}_0)}=\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f(\boldsymbol{y}_i; \hat{\boldsymbol{\eta}},\boldsymbol{\gamma})}{f(\boldsymbol{y}_i; \hat{\boldsymbol{\eta}},\boldsymbol{\gamma}_0)} \tag{A2}$$

However, we cannot use the law of large numbers directly to show this converges to its expectation under $\boldsymbol{\theta}_0$ as each summand of $\tau(\boldsymbol{\gamma})$ is a function of all the data, through $\hat{\boldsymbol{\eta}}$. Instead, we develop the following result.

Lemma A.2 Let $l_n(\boldsymbol{\theta})=l_n(\boldsymbol{\eta},\boldsymbol{\gamma})=\ln L(\boldsymbol{y};\boldsymbol{\eta},\boldsymbol{\gamma})=\sum_{i=1}^{n}\ln f(\boldsymbol{y}_i; \boldsymbol{\eta},\boldsymbol{\gamma})$, we have $l_n(\hat{\boldsymbol{\eta}}, \boldsymbol{\gamma})/n\to E_{\boldsymbol{\theta}_0}\ln f(\boldsymbol{y},\boldsymbol{\eta}_0,\boldsymbol{\gamma})$ in probability as $n\to\infty$.

Proof Under Conditions C1 ~ C3 one has that, for any fixed $\boldsymbol{\gamma}$, the Taylor expansion of the standardized likelihood around $\boldsymbol{\eta}_0$ is

$$\frac{1}{n}l_n(\hat{\boldsymbol{\eta}}, \boldsymbol{\gamma})=\frac{1}{n}l_n(\boldsymbol{\eta}_0, \boldsymbol{\gamma})+\frac{1}{n}(\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0)^{\mathrm{T}}\frac{\partial}{\partial\boldsymbol{\eta}}l_n(\boldsymbol{\eta}, \boldsymbol{\gamma})\big|_{\tilde{\boldsymbol{\eta}}} \tag{A3}$$

where $\tilde{\boldsymbol{\eta}}$ is between $\hat{\boldsymbol{\eta}}$ and $\boldsymbol{\eta}_0$. By the Cauchy-Schwarz inequality, the last term is

$$\left\|\frac{1}{n}(\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0)^{\mathrm{T}}\frac{\partial}{\partial\boldsymbol{\eta}}l_n(\boldsymbol{\eta}, \boldsymbol{\gamma})\big|_{\tilde{\boldsymbol{\eta}}}\right\|\leqslant\frac{1}{n}\|\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0\|\times\left\|\frac{\partial}{\partial\boldsymbol{\eta}}l_n(\boldsymbol{\eta}, \boldsymbol{\gamma})\big|_{\tilde{\boldsymbol{\eta}}}\right\| \tag{A4}$$

By Lemma A.1, we know $\|\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0\|=o_p(1)$. We then look at the square of the last term, viz.

$$\left\|\frac{\partial}{\partial\boldsymbol{\eta}}l_n(\boldsymbol{\eta}, \boldsymbol{\gamma})\big|_{\tilde{\boldsymbol{\eta}}}\right\|^2=\sum_{r=1}^{\dim(\boldsymbol{\eta})}\left\{\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\eta}_r}\ln f(\boldsymbol{y}_i,\boldsymbol{\eta},\boldsymbol{\gamma})\big|_{\tilde{\boldsymbol{\eta}}}\right\}^2=O_p(n^2) \tag{A5}$$

which follows from the regularity conditions. Hence

$$\left\|\frac{\partial}{\partial\boldsymbol{\eta}}l_n(\boldsymbol{\eta}, \boldsymbol{\gamma})\big|_{\tilde{\boldsymbol{\eta}}}\right\|=O_p(n) \tag{A6}$$

So the remainder term in Eq. (A3) is given by

$$\left\|\frac{1}{n}(\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0)^{\mathrm{T}}\frac{\partial}{\partial\boldsymbol{\eta}}l_n(\boldsymbol{\eta},\boldsymbol{\gamma})\big|_{\tilde{\boldsymbol{\eta}}}\right\|\leqslant\frac{1}{n}o_p(1)O_p(n)=o_p(1) \tag{A7}$$

This in turn implies

$$\frac{1}{n}l_n(\hat{\boldsymbol{\eta}}, \boldsymbol{\gamma})=\frac{1}{n}l_n(\boldsymbol{\eta}_0,\boldsymbol{\gamma})+o_p(1) \tag{A8}$$

Therefore for any $\boldsymbol{\gamma}$, $l_n(\hat{\boldsymbol{\eta}}, \boldsymbol{\gamma})/n\to E_{\boldsymbol{\theta}_0}\ln f(\boldsymbol{y},\boldsymbol{\eta}_0,\boldsymbol{\gamma})$ in probability. The proof is completed.

Now we can return to the standard proof, it can be easily shown that

$$\tau(\boldsymbol{\gamma})=\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f(\boldsymbol{y}_i;\ \hat{\boldsymbol{\eta}},\boldsymbol{\gamma})}{f(\boldsymbol{y}_i;\ \hat{\boldsymbol{\eta}},\boldsymbol{\gamma}_0)}\to E_{\boldsymbol{\theta}_0}\{\ln\frac{f(\boldsymbol{y},\boldsymbol{\eta}_0,\boldsymbol{\gamma})}{f(\boldsymbol{y},\boldsymbol{\eta}_0,\boldsymbol{\gamma}_0)}\}=-K(\boldsymbol{\theta}_0,\ \boldsymbol{\theta})<0 \tag{A9}$$

unless $f(\boldsymbol{y},\boldsymbol{\theta})=f(\boldsymbol{y};\ \boldsymbol{\theta}_0)$, and $\hat{\boldsymbol{\theta}}\to\boldsymbol{\theta}_0$ in probability and hence $\hat{\boldsymbol{\gamma}}\to\boldsymbol{\gamma}_0$ in probability. Then through the standard process, we can easily obtain the asymptotic normality.

《中国科学技术大学学报》征稿简则

《中国科学技术大学学报》由郭沫若、华罗庚和严济慈等一大批老一辈科学家亲手于 1965 年在北京创刊，先后有 30 多位院士担任编委．由中国科学院主管，中国科学技术大学主办，为综合性自然科学国家重点级核心学术期刊（月刊，国内外公开发行），主要刊登具有创新性、高水平的学术论文以及由科学大家或知名教授撰写的反映学科前沿的评述，并且开辟专家论坛，就一些重大科学研究问题进行有益的讨论．

欢迎国内外学者投稿，中英文稿均可．

1　栏目

本刊设科研论文、综述文章、专家观点、科普文章等栏目．

1.1　科研论文　介绍某一课题高水平研究成果．来稿要求内容充实，推论严谨，数据可靠、完整，文字精练，结论正确．欢迎发表系列论文和团队论文．

1.2　综述文章　综述某一重要研究领域的代表性成果，评论研究现状，提出尚待解决的问题，并指明今后研究方向．一般约请科学大家或知名教授撰写，作者亦可向编辑部自荐．

1.3　专家意见　就科学研究重大基础问题、前沿问题或热点问题提出解决问题的新思路，发表不同的见解或进行必要的有益讨论．

1.4　科普文章　关注社会热点，关注最新的科研动态，向公众宣传和普及科技知识．

2　投稿要求和注意事项

2.1　正文书写顺序　标题（一般不超过 20 个汉字）、作者姓名、作者单位，所在城市及邮政编码、中文摘要、关键词（3～8 条）、中图分类号（数学稿还须提供 AMS Subject Classification，参照 http://just.ustc.edu.cn）、与中文相对应的英文标题、作者姓名（汉语拼音，姓前名后，姓全大写，名首字母大写）、作者单位译名、英文摘要、英文关键词、正文、参考文献．若为英文稿，题名不超过 100 个字符，书写顺序同上．

在文稿首页地脚处注明基金资助项目名称及项目号（将作为论文评审时参考的重要背景资料），并对第一作者（姓名，性别，出生年，学位/职称，目前主要从事的研究方向及 E-mail）与通讯作者（姓名，学位（博士以上才注）/职称（教授以上才注），E-mail 及必要的联系电话）简要介绍．通讯作者是课题负责人或导师，要负责及时对读者的问题给予解答．

2.2　对摘要的要求　摘要内容应包括有与论文同等量的主要信息，应说明研究目的、采用的方法、研究成果及结论四个部分．中英文摘要需对应．中文摘要约 250 个汉字，英文摘要约 1500 个字符．请参照 EI，SCI 要求，避免使用“This paper，in this paper（本文）”或“I（我）”等，用词要客观，尽量减少不必要的修饰．

2.3　对量、单位及符号的要求　文中物理量、计量单位及符号的使用必须符合国际标准和国家标准（GB3100-93～GB3102-93）．正确书写易混淆的外文字母的文种、大小写、正斜体、黑白体及上下角标．

2.4　对图、照片、表的要求　文中图要直观、简明、清晰．图中的文字、符号、纵横坐标必须写清，并与正文保持一致．

图版、照片必须图像清晰，层次分明，请提供矢量图或线条图，不接收扫描图；可根据作者需要印刷彩页．

表的格式采用三线表，必要时可加辅助线，所用文字、符号、单位要与正文一致．

图、图版、照片、表均要求提供中、英文对照的图题、表题．

2.5　对参考文献的要求　参考文献必须标全并注意引用国内外及本刊的最新文献．按在文中出现的先后次序列于文后，用数字加方括号表示，如[1]，[2]，…，与正文中的指示序号一致．

本刊执行国家标准《文后参考文献著录规则》（GB/T 7714-2015）和《中国学术期刊（光盘版）检索与评价数据规范》，并参照 EI，SCI 要求．中文参考文献请先按下列格式完整列出其中文，然后完整著录出英文．各类参考文献条目的编排格式如下：

期刊类　作者名．引文题目[J]．期刊名，出版年份，卷（期）：起止页码．

专著类 作者名.书名[M].版本(第一版不写).出版地:出版者,出版年:起止页码.

论文集类(或会议论文集类) 引文作者名.引文题目[C]//论文集主编名.论文集名.出版地(会址):出版者,出版年:起止页码.

学位论文类 作者名.题名[D].保存地点:保存单位,年份.

专利类 专利所有者.专利题名:专利国别,专利号[P/文献载体标志].出版日期.

标准类 标准编号,标准名称[S].

电子文献类 作者名.电子文献题名[文献类型标志/文献载体标志].[引用日期].获取和访问路径.

作者(或编者、译者)不超过3人时全部写出,超过者只写前3人,后加"等"或"et al",书写外文作者或编者时,姓前名后,名用缩写.

所引文献必须是作者直接阅读过的、最主要的、发表在正式出版物上的文献,请核实所著录的文献;尚未公开发表的论文、预印本等,一律不列入正式文献,如有必要可作注释处理,一般在当页下脚加注.

3 投稿约定

3.1 文稿要求一式两份(最好是打印稿,也接受电子投稿),并提交一份填好的来稿登记表(来稿登记表可从 http://just.ustc.edu.cn 下载,具体投稿程序和要求也请登录我们的网页).对拟刊用的稿件要求提供电子文档(方正排版文件、*.DOC、*.TXT、*.TEX 文件均可;图件请存为 eps, vsd, tif, jpg 格式,矢量图可以直接放到 word 文档中),同时收取版面费.来稿一经发表,即按篇酌致稿酬(包含了纸版、光盘版和网络电子版稿酬),并赠送当期期刊2册.

3.2 稿件实行严格的同行专家评审制度,"不计资历,不抑后起,不亲近疏远,以质论稿",须经同行专家评审、编委会审定决定取舍.为了确保与客观表明在本刊发表的论文的水平与学术档次,这一评审过程是按照严格、客观、公正的要求以及一定程序进行的,因而评审是需要一个过程的,在这期间作者不得将该稿投往他处,编辑部负责及时(一般在3个月内)通报审稿意见并要求作者及时吸收这些意见修改稿件、提高水平.录用的稿件将存档保存,未录用的稿件原稿不再返还,请作者自留底稿.除主动约请外,本刊只刊载首发稿,若发现作者一稿多投或有剽窃、抄袭或造假等行为,编辑部将追补对本刊造成的损失并通报作者单位.

3.3 稿件文责自负.编辑部对来稿有权做技术性和文字修改,但实质性内容的修改须征得作者同意.

3.4 本刊现为纸张印刷版—光盘版—网络版三位一体的出版模式,编辑部自文稿发表之日起取得文稿的各种专有出版权和独家使用与代理权,有权进行各种方式的出版或复制,进一步传播论文,扩大论文影响.本刊已被国内外众多有影响的数据库列入收录刊源,为了扩大论文影响与传播范围,本刊发表的论文将被编入《中国学术期刊(光盘版)》和"万方数据(China Info)系统科技期刊群"、《中文科技期刊数据库》、"龙源国际名刊网"及"台湾华艺电子期刊资料库"等数据库和有关的文摘或进入个别有影响的正规出版的全文文选,并向国外有关数据库、检索机构报送.因此,向本刊投稿者则视同认可并授权本刊上述做法.

3.5 作者务必严格遵守国家有关保密规定,不得泄露国家秘密.

对投稿有关事项需进一步了解者,可向合肥市金寨路96号中国科学技术大学学报编辑部询问,邮政编码:230026;电话:(0551)63601961,63607694,63606890;E-mail:just@ustc.edu.cn;网址:http://just.ustc.edu.cn.

中国科学技术大学学报编辑部