Statistica Sinica **29** (2019), 245-264
doi:https://doi.org/10.5705/ss.202016.0135

COPULA-BASED QUANTILE REGRESSION FOR LONGITUDINAL DATA

Huixia Judy Wang[1], Xingdong Feng[2] and Chen Dong[2]

[1] *The George Washington University and* [2] *Shanghai University of Finance and Economics*

Abstract: Inference and prediction in quantile regression for longitudinal data are challenging without parametric distributional assumptions. We propose a new semiparametric approach that uses copula to account for intra-subject dependence and approximates the marginal distributions of longitudinal measurements, given covariates, through regression of quantiles. The proposed method is flexible, and it can provide not only efficient estimation of quantile regression coefficients but also prediction intervals for a new subject given the prior measurements and covariates. The properties of the proposed estimator and prediction are established theoretically, and assessed numerically through a simulation study and the analysis of a nursing home data.

Key words and phrases: Copula, estimating equation, longitudinal data, prediction, quantile regression.

1. Introduction

In many studies, it is common to observe longitudinal data where the outcomes are measured at multiple times for each subject. One interest in longitudinal studies is to predict the response based on a set of covariates and its past trajectory. Traditional projection methods focus on predicting the mean of the conditional response distribution. However, in some applications, researchers are interested in predicting tail quantiles, for instance, the low weight in children growth studies (Abrevaya (2001)), high expenses in insurance studies (Shi and Frees (2010)), or in modeling the entire conditional distribution, for instance, the children growth and blood pressure study discussed in Wu and Tian (2013).

Quantile regression provides a convenient tool for studying tail behaviors of the response conditional on covariates. Since its introduction by Koenker and Bassett (1978), quantile regression has been extensively studied for cross sectional data while less developed for longitudinal data. Some researchers considered marginal quantile regression models for analyzing longitudinal data; see for instance, Jung (1996), He, Fu and Fung (2003), Wei and He (2006), Wang (2009), Mu and Wei (2009), Tang and Leng (2011), and Leng and Zhang (2014). Marginal models focus on the covariate effects on the marginal distributions of the repeatedly measured responses and thus cannot be used for modeling their joint dependence. Considering a quantile regression model with a random intercept, Koenker (2004) proposed a L_1 regularization method to obtain a shrinkage estimator of the random subject effects. Some other researchers proposed Bayesian approaches for conditional quantile regression models, for instance, Geraci and Bottai (2007), Yuan and Yin (2010), Wang (2012), Geraci and Bottai (2014), Reich, Bondell and Wang (2010), Kim and Yang (2011). These methods all require some parametric or semiparametric modeling of the likelihood, and a parametric distributional assumption on the random effects.

In this paper, we propose a semiparametric copula-based quantile regression method, where copula functions are employed to accommodate the temporal dependence of longitudinal data. Copulas have been applied to longitudinal data analysis for generalized linear models (Meester and MacKay (1994); Lambert and Vandenhende (2002); Sun, Frees and Rosenberg (2008); Song (2000); Bai, Kang and Song (2014)). For time series data, Bouyé and Salmon (2008) and Chen, Koenker and Xiao (2009) studied nonlinear quantile autoregressive models implied by their copula specifications. Noh, Ghouch and Van Keilegom (2015) proposed a method for semiparametric quantile regression by modeling the joint distribution of the response and covariates through copulas. In an empirical study of longitudinal data, Shi and Frees (2010) considered a copula method for quantile regression by modeling the conditional marginals of the responses with an asymmetric Laplace (AL) distribution, but the validity of the method was not discussed. The AL distribution has a close connection with quantile regression because the maximum likelihood estimator under such a model coincides with the usual quantile regression estimator for cross sectional data. However, we shall show that the method based on asymmetric Laplace marginals is restrictive, and it could have detrimental effects on both quantile estimation and prediction under model misspecification; see numerical evidences in Sections 3 and 4. We propose a more flexible and theoretically justifiable approach that approximates the conditional marginals through regression of quantiles and models the dependence of the repeated measurements with copula functions. Instead of making parametric assumptions on the marginals, the proposed method only requires the marginal quantiles of the longitudinal responses to be linear in covariates and thus can be regarded as a semiparametric method. The proposed method cannot only give efficient estimation of coefficients in the marginal quantile regression model, but also provide prediction intervals of the response of a new subject given the prior measurements and other covariates.

2. Proposed Method

2.1. Notations and models

Let y_{ij} and $\mathbf{x}_{ij}$ be the response and p-dimensional covariate for the ith subject measured at the jth time point, $i = 1, \ldots, n$, $j = 1, \ldots, J_i$, where the subjects are assumed to be independent but repeated responses from the same subject may be dependent. Without loss of generality, we assume a balanced design with $J_i = J$ being finite. Throughout we assume that the first element of $\mathbf{x}_{ij}$ is one corresponding to the intercept. Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iJ})^T$ and $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iJ})^T$. Suppose $\{\mathbf{y}_i, \mathbf{x}_i, i = 1, \ldots, n\}$ is a random sample of $\{\mathbf{Y} = (Y_1, \ldots, Y_J)^T, \mathbf{X}\}$.

Let $G(y_1, \ldots, y_J|\mathbf{x})$ denote the joint distribution of $(Y_1, \ldots, Y_J)^T$ given $\mathbf{X} = \mathbf{x}$ with continuous conditional marginal distributions $F_1(\cdot|\mathbf{x}), \ldots, F_J(\cdot|\mathbf{x})$. By Sklar's theorem (Sklar (1959)), there exists a copula function C such that G can be uniquely represented as $G(y_1, \ldots, y_J|\mathbf{x}) = C\{F_1(y_1|\mathbf{x}), \ldots, F_J(y_J|\mathbf{x}); \mathbf{x}\}$. Throughout the paper all analyses are conditional on $\mathbf{X} = \mathbf{x}$ and we do not model the $\mathbf{X}$ distribution. For model parsimony, we consider a parametric copula function C and simplify the copula function by dropping the dependence on the covariates $\mathbf{x}$. That is, we consider the simplified copula model

$$G(y_1, \ldots, y_J|\mathbf{x}) = C\{F_1(y_1|\mathbf{x}), \ldots, F_J(y_J|\mathbf{x}); \boldsymbol{\theta}_0\}, \tag{2.1}$$

which assumes that the copula function is independent of covariates except through the conditional marginals $F_j(\cdot|\mathbf{x})$. There are many ways to construct a copula function; see for instance Joe (1996). One way is to extract from any J-dimensional joint distribution $\mathcal{F}(\cdot)$. For example, if $\mathcal{F}(\cdot)$ is a multivariate normal distribution $N_J(0, \Sigma)$, where Σ is the correlation matrix with ones on the diagonal, then $C(u_1, \ldots, u_J; \Sigma) = \mathcal{F}\{\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_J); \Sigma\}$ is the Gaussian copula with correlation matrix Σ. More discussions can be found in Remark 3. Here the matrix Σ is not the standard Pearson correlation matrix but rather some rank-based correlation measuring the nonlinear dependence of variables; see Song (2000) for more detailed interpretation of elements in Σ.

The simplified copula assumption in model (2.1) has been commonly used in the copula literature for modeling multivariate distributions; see for instance Haff, Aas and Frigessi (2010), Smith et al. (2010). Haff, Aas and Frigessi (2010) showed that the simplified copula serves as a good approximation even when the simplifying assumption is far from being satisfied. Our numerical investigation in Section 3 also confirmed the satisfying performance of the simplified copula even under some model misspecification. The framework and idea proposed in this paper can be extended to more general copula functions; see Section for some discussion.

Instead of making parametric assumptions on $F_j(\cdot|\mathbf{x})$, we propose to fit a quantile process by assuming the linear quantile regression model,

$$Q_\tau(Y_j|\mathbf{x}_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta}_0(\tau), j = 1, \ldots, J, \text{ for any } 0 < \tau < 1, \tag{2.2}$$

where $Q_\tau(Y_j|\mathbf{x}_{ij}) = \inf\{y : F_j(y|\mathbf{x}_{ij}) \geq \tau\}$ is the τth (marginal) quantile of Y_j given the covariate $\mathbf{x}_{ij}$. Model (2.2) was also considered in Jung (1996), He, Fu and Fung (2003), Mu and Wei (2009), Tang and Leng (2011), and so on, for analyzing clustered or longitudinal data. The conventional estimator of $\boldsymbol{\beta}_0(\tau)$ that completely ignores the intra-subject correlation can be obtained as

$$\tilde{\boldsymbol{\beta}}(\tau) = \underset{\mathbf{b}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n}\sum_{j=1}^{J} \rho_\tau(y_{ij} - \mathbf{x}_{ij}^T\mathbf{b}), \tag{2.3}$$

where $\rho_\tau(s) = s\{\tau - I(s < 0)\}$, and $I(\cdot)$ is the indicator function. Even though $\tilde{\boldsymbol{\beta}}(\tau)$ is a consistent estimator of $\boldsymbol{\beta}_0(\tau)$, its efficiency may be lost by ignoring the intra-subject correlation. In addition, sometimes we are interested in predicting the conditional quantiles of the response for a new subject, say at time J. It would be beneficial to take into account not only the covariate information but also the responses observed in the past time points, and such prediction is not feasible without modeling the joint distribution $G(y_1, \ldots, y_J|\mathbf{x})$. We propose to employ copula functions to accommodate temporal dependence, which can help not only the efficiency for estimating $\boldsymbol{\beta}_0(\tau)$ but also the prediction.

2.2. Three-step estimation

We can link the density function of $F_j(\cdot|\mathbf{x}_{ij})$, denoted by $f(\cdot|\mathbf{x}_{ij})$, with model (2.2) by the equation

$$f(y|\mathbf{x}_{ij}) = \lim_{\delta\to 0} \frac{\delta}{\mathbf{x}_{ij}^T\{\boldsymbol{\beta}_0(u_y+\delta) - \boldsymbol{\beta}_0(u_y)\}}, \tag{2.4}$$

where $u_y = \{\upsilon \in (0,1) : \mathbf{x}_{ij}^T\boldsymbol{\beta}_0(\upsilon) = y\}$. Here we write $f(y|\mathbf{x}_{ij})$ as $f\{y|\mathbf{x}_{ij};\boldsymbol{\beta}(\tau)\}$ to reflect its dependence on the quantile process $\boldsymbol{\beta}(\tau)$. The log likelihood function (conditional on $\{\mathbf{x}_{ij}\}$) can thus be written as

$$l\{\boldsymbol{\beta}(\tau), \boldsymbol{\theta}\} = \sum_{i=1}^{n}\sum_{j=1}^{J} \log(f\{y_{ij}|\mathbf{x}_{ij};\boldsymbol{\beta}(\tau)\}) + \sum_{i=1}^{n} \log(c(u_{i1}, \ldots, u_{iJ};\boldsymbol{\theta})), \tag{2.5}$$

where $c(\cdot;\boldsymbol{\theta})$ is the density associated with the copula $C(\cdot;\boldsymbol{\theta})$, and $u_{ij} = \{\upsilon \in (0,1) : \mathbf{x}_{ij}^T\boldsymbol{\beta}(\upsilon) = y_{ij}\} = P(Y_j \leq y_{ij}|\mathbf{x}_{ij}) = F_j(y_{ij}|\mathbf{x}_{ij})$.

We can estimate $(\boldsymbol{\beta}_0(\tau), \boldsymbol{\theta}_0)$ by maximizing the log likelihood function $l\{\boldsymbol{\beta}(\tau), \boldsymbol{\theta}\}$. However, direct maximization is challenging since $l\{\boldsymbol{\beta}(\tau), \boldsymbol{\theta}\}$ involves the entire quantile process $\boldsymbol{\beta}(\tau)$ and does not have an explicit form. Alternatively, we propose a three-step estimation procedure. In the first step, we obtain the consistent but not necessarily efficient estimator $\tilde{\boldsymbol{\beta}}(\tau)$ of the coefficient process $\boldsymbol{\beta}_0(\tau)$ by ignoring the intra-subject correlation. In the second step, we estimate u_{ij} by $\tilde{u}_{ij} = \{\upsilon : \mathbf{x}_{ij}^T\tilde{\boldsymbol{\beta}}(\upsilon) = y_{ij}\}$ and then estimate $\boldsymbol{\theta}_0$ by maximizing the copula likelihood. Finally, we obtain an efficient estimator of $\boldsymbol{\beta}_0(\tau)$ by taking into account the intra-subject correlation based on the estimated copula function. In this procedure $\boldsymbol{\beta}_0(\tau)$ is estimated at each quantile level τ separately, and the resulting estimator of $(\boldsymbol{\beta}_0(\tau), \boldsymbol{\theta}_0)$ is not the maximum likelihood estimator. The details of the procedure are as follows.

Step 1. Let $0 < \tau_1 < \cdots < \tau_{\kappa_n} < 1$ be a grid of quantile levels, where $\tau_k = k/(\kappa_n + 1)$. For $k = 1, \ldots, \kappa_n$, obtain $\tilde{\boldsymbol{\beta}}(\tau_k) = \text{argmin}_{\mathbf{b}\in\mathbb{R}^p} \sum_{i=1}^n \sum_{j=1}^J \rho_{\tau_k}(y_{ij} - \mathbf{x}_{ij}^T\mathbf{b})$. Define $\tilde{u}_{ij} = (1-\tilde{\alpha}_{ij})\tau_k + \tilde{\alpha}_{ij}\tau_{k+1},\ \tilde{\alpha}_{ij} = \{y_{ij} - \mathbf{x}_{ij}^T\tilde{\boldsymbol{\beta}}(\tau_k)\}/\mathbf{x}_{ij}^T\{\tilde{\boldsymbol{\beta}}(\tau_{k+1}) - \tilde{\boldsymbol{\beta}}(\tau_k)\}$, where $\mathbf{x}_{ij}^T\tilde{\boldsymbol{\beta}}(\tau_k) \leq y_{ij} < \mathbf{x}_{ij}^T\tilde{\boldsymbol{\beta}}(\tau_{k+1})$. In our implementation, we choose $\kappa_n = [4 + 3n^{0.4}]$, where $[\upsilon]$ denotes the integer part of υ.

Step 2. Estimate $\boldsymbol{\theta}_0$ by $\hat{\boldsymbol{\theta}}$, the maximizer of the pseudo-copula log-likelihood, $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \sum_{i=1}^n \log\{c(\tilde{u}_{i1}, \ldots, \tilde{u}_{iJ}; \boldsymbol{\theta})\}$.

Step 3. For any $0 < \tau < 1$, define $\hat{\boldsymbol{\beta}}(\tau)$ as the solution to the estimating equation

$$U_n(\mathbf{b}) = n^{-1}\sum_{i=1}^n \mathbf{x}_i^T\boldsymbol{\Gamma}_i(\mathbf{V}_i(\hat{\boldsymbol{\theta}}))^{-1}\boldsymbol{\psi}_\tau(\mathbf{y}_i - \mathbf{x}_i^T\mathbf{b}) = 0, \qquad (2.6)$$

where $\boldsymbol{\psi}_\tau(\mathbf{y}_i - \mathbf{x}_i^T\mathbf{b}) = \left(\psi_\tau(y_{i1} - \mathbf{x}_{i1}^T\mathbf{b}), \ldots, \psi_\tau(y_{iJ} - \mathbf{x}_{iJ}^T\mathbf{b})\right)^T$ with $\psi_\tau(u) = \tau - I(u < 0)$, $\mathbf{V}_i(\boldsymbol{\theta}) = \text{Cov}\left(\boldsymbol{\psi}_\tau\{\mathbf{y}_i - \mathbf{x}_i^T\boldsymbol{\beta}_0(\tau)\}|\mathbf{x}_i\right)$, and $\boldsymbol{\Gamma}_i = \text{diag}\{s_{i1}, \ldots, s_{iJ}\}$ with $s_{ij} = f\{\mathbf{x}_{ij}^T\boldsymbol{\beta}_0(\tau)|\mathbf{x}_{ij}\}$.

Alternatively, one could iteratively update $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}(\tau)$ in Steps 2 and 3 until convergence, but this does not affect the asymptotic efficiency of $\hat{\boldsymbol{\beta}}(\tau)$.

Remark 1. We propose to estimate $F_j(\cdot|\mathbf{x})$ through modeling the conditional quantile process in model (2.2). Instead of assuming parametric distributions, this approach requires the conditional quantiles of Y_{ij} to be linear in $\mathbf{x}$ and thus can be regarded as a semiparametric likelihood approach, which provides a balance between model parsimony and flexibility. Such global linearity assumptions were also employed to facilitate analyses in different contexts; see for instance, Portnoy (2003), Wei and Carroll (2009), Wang and Zhou (2010), Feng, Chen and He (2015).

Remark 2. The estimated quantiles in Step 1 may have a crossing issue in finite samples: the conditional quantile $Q_\tau(Y_j|\mathbf{x}_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta}_0(\tau)$ at some upper quantile may be estimated to be smaller than that at a lower quantile. To avoid this, we employ the quantile rearrangement procedure proposed in Chernozhukov, Fernández-Val and Galichon (2010) that constructs monotone quantile estimates by sorting or monotone rearranging $\mathbf{x}_{ij}^T\tilde{\boldsymbol{\beta}}(\tau_k)$.

The quantity $s_{ij} = f\{\mathbf{x}_{ij}^T\boldsymbol{\beta}_0(\tau)|\mathbf{x}_{ij}\}$ measures the dispersion of $\epsilon_{ij}(\tau) = y_{ij} - \mathbf{x}_{ij}^T\boldsymbol{\beta}_0(\tau)$. In our implementation, we estimate s_{ij} by using the quotient estimation method of Hendricks and Koenker (1992), $\hat{s}_{ij} = 2h_n/\mathbf{x}_{ij}^T\{\tilde{\boldsymbol{\beta}}(\tau+h) - \tilde{\boldsymbol{\beta}}(\tau-h)\}$, where h_n is a positive bandwidth such that $h_n \to 0$ as $n \to \infty$. We choose $h_n = 1.57n^{-1/3}(1.5\phi^2\{\Phi^{-1}(\tau)\}/[2\{\Phi^{-1}(\tau)\}^2+1])^{1/3}$ following the rule suggested in Hall and Sheather (1988), where $\Phi(\cdot)$ and $\phi(\cdot)$ are the distribution and density functions of the standard normal distribution.

The matrix $\mathbf{V}_i(\boldsymbol{\theta}) = \mathrm{Cov}\big(\tau - I(u_{i1} < \tau), \ldots, \tau - I(u_{iJ} < \tau)\big)$ is the covariance for the score vector $\boldsymbol{\psi}_\tau\{\mathbf{y}_i - \mathbf{x}_i^T\boldsymbol{\beta}_0(\tau)\}$, capturing the intra-subject correlation. Define $\lambda_\tau(u_{ij}, u_{ik}) = \mathrm{Cov}\{\tau - I(u_{ij} < \tau), \tau - I(u_{ik} < \tau)\} = P(u_{ij} < \tau, u_{ik} < \tau) - \tau^2$, which equals $\tau - \tau^2$ for $j = k$. For a given multivariate copula $C(\cdot;\boldsymbol{\theta})$, $\lambda_\tau(u_{ij}, u_{ik}) = C_{jk}(\tau,\tau;\boldsymbol{\theta}) - \tau^2$ since u_{ij} and u_{ik} are uniformly distributed on $(0,1)$, where $C_{jk}(\tau,\tau;\boldsymbol{\theta})$ is induced by setting the jth and kth elements of $C(\cdot;\boldsymbol{\theta})$ to τ and the rest to 1.

Remark 3. Commonly used multivariate copulas include the elliptical and Archimedean copulas. Elliptical copulas may incorporate a specific correlation structure, for instance, the exchangeable, autoregressive, Toeplitz, and unstructured correlation structures, and thus can capture time-dependent intra-subject correlation, a typical feature of longitudinal data. In contrast, Archimedean copulas can only capture exchangeable correlation across time and may not be useful for longitudinal studies with large time dimension. For high dimensional problems, we can also consider vine copulas that are constructed from a series of bivariate copulas. More discussions of multivariate copulas and vine copulas can be found in Joe (1996), Aas et al. (2009), Smith et al. (2010).

2.3. Induced smoothed estimator of regression parameters

The estimating function $U_n(\cdot)$ in Step 3 of the three-step estimation in Section 2.2 involves an indicator function. This nonsmoothness not only makes it difficult to solve the estimating equation but it also challenges the estimation of the asymptotic covariance of $\hat{\boldsymbol{\beta}}(\tau)$, which is often sensitive to the choice of smoothing parameters involved in estimating the unknown density function $f(\cdot|\mathbf{x}_{ij})$. To bypass these challenges, we consider an induced smoothed estimator. The induced smoothing method was first proposed by Brown and Wang (2005) for robust regression, and later extended to quantile regression by Wang, Shao and Zhu (2009) for cross sectional data, by Fu and Wang (2012) and Leng and Zhang (2014) for longitudinal data, and by Pang, Lu and Wang (2012) for censored data. The idea of the induced smoothing is as follows. By the asymptotic normality of $\hat{\boldsymbol{\beta}}(\tau)$ in Theorem 1, we can regard $\hat{\boldsymbol{\beta}}(\tau)$ as a random perturbation of $\boldsymbol{\beta}_0(\tau)$ by writting $\hat{\boldsymbol{\beta}}(\tau) = \boldsymbol{\beta}_0(\tau) + n^{-1/2}\mathbf{H}^{1/2}\mathbf{Z}$, where $\mathbf{Z} \sim N(0, I_{p\times p})$ and $\mathbf{H}$ is defined in the Assumption (A7). We consider the smoothed estimating function

$$\begin{aligned}\tilde{U}_n(\mathbf{b}, \mathbf{H}) &= E_{\mathbf{Z}}\{U_n(\mathbf{b} + n^{-1/2}\mathbf{H}^{1/2}\mathbf{Z})\}\\ &= n^{-1}\sum_{i=1}^{n}\mathbf{x}_i^T\boldsymbol{\Gamma}_i(\mathbf{V}_i(\hat{\boldsymbol{\theta}}))^{-1}\tilde{\boldsymbol{\psi}}_\tau(\mathbf{y}_i - \mathbf{x}_i^T\mathbf{b}, \mathbf{H}), \qquad (2.7)\end{aligned}$$

where $\tilde{\boldsymbol{\psi}}_\tau(\mathbf{y}_i - \mathbf{x}_i^T\mathbf{b}, \mathbf{H}) = (\tau - \Phi((\mathbf{x}_{i1}^T\mathbf{b} - y_{i1})/h_{i1}), \ldots, \tau - \Phi((\mathbf{x}_{iJ}^T\mathbf{b} - y_{iJ})/h_{iJ}))^T$, and $h_{ij} = \sqrt{\mathbf{x}_{ij}^T\mathbf{H}\mathbf{x}_{ij}/n}$. When $\mathbf{H}$ is known, the induced smoothed estimator $\hat{\boldsymbol{\beta}}_s(\tau)$ can be obtained by solving $\tilde{U}_n(\mathbf{b}, \mathbf{H}) = 0$. In practice, since $\mathbf{H}$ is unknown, we estimate $\hat{\boldsymbol{\beta}}_s(\tau)$ and $\mathbf{H}$ through the following iterating procedure.

Step 3.1. Let $\hat{\boldsymbol{\beta}}_s^{(0)}(\tau) = \tilde{\boldsymbol{\beta}}(\tau)$, the estimator obtained by assuming working independence, and $\hat{\mathbf{H}}^{(0)} = \mathbf{I}_p$, the $p \times p$ identity matrix.

Step 3.2. Given $\hat{\boldsymbol{\beta}}_s^{(k)}(\tau)$ and $\mathbf{H}^{(k)}$ from the kth iteration, update $\hat{\boldsymbol{\beta}}_s^{(k+1)}(\tau)$ and $\mathbf{H}^{(k+1)}$ by

$$\hat{\boldsymbol{\beta}}_s^{(k+1)}(\tau) = \hat{\boldsymbol{\beta}}_s^{(k)}(\tau) - \left\{\left.\frac{\partial\tilde{U}_n(\mathbf{b}, \hat{\mathbf{H}}^{(k)})}{\partial\mathbf{b}}\right|_{\hat{\boldsymbol{\beta}}_s^{(k)}(\tau)}\right\}^{-1}\tilde{U}_n\{\hat{\boldsymbol{\beta}}_s^{(k)}(\tau), \hat{\mathbf{H}}^{(k)}\} \text{ and}$$

$$\begin{aligned}\hat{\mathbf{H}}^{(k+1)} = &\left\{\left.\frac{\partial\tilde{U}_n(\mathbf{b}, \hat{\mathbf{H}}^{(k)})}{\partial\mathbf{b}}\right|_{\hat{\boldsymbol{\beta}}_s^{(k+1)}(\tau)}\right\}^{-1} A\{\hat{\boldsymbol{\beta}}_s^{(k+1)}(\tau), \hat{\mathbf{H}}^{(k)}\}\\ &\left\{\left.\frac{\partial\tilde{U}_n(\mathbf{b}, \hat{\mathbf{H}}^{(k)})}{\partial\mathbf{b}}\right|_{\hat{\boldsymbol{\beta}}_s^{(k+1)}(\tau)}\right\}^{-1},\end{aligned}$$

$$A(\mathbf{b}, \mathbf{H}) = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i^T\boldsymbol{\Gamma}_i(\mathbf{V}_i(\hat{\boldsymbol{\theta}}))^{-1}\tilde{\boldsymbol{\psi}}_\tau(\mathbf{y}_i - \mathbf{x}_i^T\mathbf{b}, \mathbf{H})\tilde{\boldsymbol{\psi}}_\tau^T(\mathbf{y}_i - \mathbf{x}_i^T\mathbf{b}, \mathbf{H})\{\mathbf{V}_i(\hat{\boldsymbol{\theta}})\}^{-1}\boldsymbol{\Gamma}_i\mathbf{x}_i.$$

Step 3.3. Repeat Step 3.2 till convergence. Denote the coefficient estimate and the covariance estimate at convergence as $\hat{\boldsymbol{\beta}}_s(\tau)$ and $\hat{\mathbf{H}}$, respectively.

For estimating $\boldsymbol{\beta}_0(\tau)$, we can solve the equation $\tilde{U}_n(\mathbf{b}, \mathbf{M}) = 0$ for $\mathbf{b}$ with any known positive definite matrix $\mathbf{M}$ such that $\|\mathbf{M}\| = O(1)$, so $\tilde{U}_n(\mathbf{b}, \mathbf{M})$ can be viewed as a smoothed estimating function with Gaussian kernel and subject-specific bandwidth of order $n^{-1/2}$. By the proof of Theorem 1 in the supplementary file, this will also lead to an estimator that is asymptotically equivalent to $\hat{\boldsymbol{\beta}}(\tau)$. However, in our procedure, we also solve for the unknown $\mathbf{H}$ iteratively to obtain a consistent estimator for the asymptotic covariance of $n^{1/2}\hat{\boldsymbol{\beta}}(\tau)$.

2.4. Prediction

The proposed copula regression can provide a predictive distribution of the response variable Y incorporating the covariates and the prior Y measurements. The predictive distribution can be used to predict the mean or any quantiles of the future response, or to construct prediction intervals.

For instance, suppose that for a new subject we have observed $\mathbf{x}_1^*, \ldots, \mathbf{x}_J^*$ and $y_1^*, \ldots, y_{J-1}^*$ at the first $J-1$ time points. We would like to predict the outcome at time J, y_J^*. Under model (2.1), $F(y_1^*, \ldots, y_{J-1}^*, y_J^* | \mathbf{x}_1^*, \ldots, \mathbf{x}_J^*) = C\big(u_1^*, \ldots, u_{J-1}^*, u_J^*; \boldsymbol{\theta}_0\big)$, where $u_j^* = F_j(y_j^* | \mathbf{x}_j^*), j = 1, \ldots, J$. With some standard calculus derivations, we can show that

$$F(y_J^* | \mathbf{x}_1^*, \ldots, \mathbf{x}_J^*, y_1^*, \ldots, y_{J-1}^*) = h(u_J^* | \mathbf{x}_1^*, \ldots, \mathbf{x}_J^*, u_1^*, \ldots, u_{J-1}^*; \boldsymbol{\theta}_0), \text{ where}$$

$$h(u_J^* | \mathbf{x}_1^*, \ldots, \mathbf{x}_J^*, u_1^*, \ldots, u_{J-1}^*; \boldsymbol{\theta}_0) = \frac{\partial^{J-1} C(u_1^*, \ldots, u_{J-1}^*, u_J^*; \boldsymbol{\theta}_0) / \partial u_1^* \cdots \partial u_{J-1}^*}{\partial^{J-1} C(u_1^*, \ldots, u_{J-1}^*, 1; \boldsymbol{\theta}_0) / \partial u_1^* \cdots \partial u_{J-1}^*}.$$

Under model (2.2), we can estimate u_j^* by $\hat{u}_j^* = \{u : \mathbf{x}_j^{*T} \hat{\boldsymbol{\beta}}(u) = y_j^*\}$, $j = 1, \ldots, J$, and estimate the predictive distribution by

$$\hat{F}(y_J^* | \mathbf{x}_1^*, \ldots, \mathbf{x}_J^*, y_1^*, \ldots, y_{J-1}^*) = h(\hat{u}_J^* | \mathbf{x}_1^*, \ldots, \mathbf{x}_J^*, \hat{u}_1^*, \ldots, \hat{u}_{J-1}^*; \hat{\boldsymbol{\theta}}).$$

For any $\tau^* \in (0, 1)$, the τ^*th conditional quantile of y_J^* can thus be estimated by numerically solving the equation

$$\hat{F}(y | \mathbf{x}_1^*, \ldots, \mathbf{x}_J^*, y_1^*, \ldots, y_{J-1}^*) = \tau^*. \tag{2.8}$$

For any $0 < \alpha < 1$, the $(1-\alpha)$ prediction interval can be constructed by the $(\alpha/2)$th and the $(1-\alpha/2)$th conditional quantiles of y_J^*.

Remark 4. If the research interest is on the marginal quantiles as in model (2.2), we can also estimate the intra-subject correlation directly based on the estimated residuals $y_{ij} - \mathbf{x}_{ij}^T \tilde{\boldsymbol{\beta}}(\tau)$, or approximate $\mathbf{V}_i^{-1}$ by a linear combination of basis functions and then obtain a more efficient estimator of $\boldsymbol{\beta}(\tau)$ by solving weighted estimating equations similar to (2.6); see for instance Wang (2009), Leng and Zhang (2014). The resulting estimators have similar properties as does our proposed estimator $\hat{\boldsymbol{\beta}}(\tau)$. The bigger advantage of the copula-based approach is that it provides a convenient way to model the joint distribution of $(Y_1, \ldots, Y_J)^T$ conditional on covariates, which can be used to obtain more accurate prediction based on both covariates and responses in the past time points.

2.5. Asymptotic properties

Throughout, let $\boldsymbol{\beta}_0(\tau)$ and $\boldsymbol{\theta}_0$ denote the true values of $\boldsymbol{\beta}(\tau)$ and $\boldsymbol{\theta}$. Theorem 1 presents the asymptotic normality of $\hat{\boldsymbol{\beta}}(\tau)$ and $\hat{\boldsymbol{\beta}}_s(\tau)$, and the consistency of $\hat{\mathbf{H}}$.

Theorem 1. *Under models* (2.1)-(2.2) *and Assumptions* (A1)–(A8) *in the supplementary file, if* $\kappa_n^{2a+1}/n^{1-2q} \to 0$ *and* $\kappa_n \to \infty$ *as* $n \to \infty$, *we have (i)* $n^{1/2}\{\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}_0(\tau)\} \xrightarrow{d} N(0, \mathbf{H})$; *(ii)* $n^{1/2}\{\hat{\boldsymbol{\beta}}_s(\tau) - \boldsymbol{\beta}_0(\tau)\} \xrightarrow{d} N(0, \mathbf{H})$, *and (iii)* $\hat{\mathbf{H}} \xrightarrow{p} \mathbf{H}$, *where* $\mathbf{H}^{-1} = \lim_{n\to\infty} n^{-1}\sum_{i=1}^n \mathbf{x}_i^T \boldsymbol{\Gamma}_i\{\mathbf{V}_i(\boldsymbol{\theta}_0)\}^{-1}\boldsymbol{\Gamma}_i\mathbf{x}_i$.

Remark 5. Consider the weighted estimation equation $U_{\mathbf{W}}(\mathbf{b}) = n^{-1}\sum_{i=1}^n \mathbf{x}_i^T \mathbf{W}_i \boldsymbol{\psi}_\tau(\mathbf{y}_i - \mathbf{x}_i^T\mathbf{b}) = 0$, where $\mathbf{W}_i$ are any given $J \times J$ matrices. Since $E\{\boldsymbol{\psi}_\tau(\mathbf{y}_i - \mathbf{x}_i^T\boldsymbol{\beta}_0)|\mathbf{x}_i\} = 0$, $U_{\mathbf{W}}(\mathbf{b})$ is an unbiased estimating function. Consequently the resulting estimator, denoted by $\hat{\boldsymbol{\beta}}_{\mathbf{W}}(\tau)$, is consistent to $\boldsymbol{\beta}_0(\tau)$ with asymptotic covariance matrix $\mathbf{H}_{\mathbf{W}} = \lim_{n\to\infty}\Delta_{n\mathbf{W}}^{-1}\Lambda_{n\mathbf{W}}\Delta_{n\mathbf{W}}^{-1}$, where $\Delta_{n\mathbf{W}} = n^{-1}\sum_{i=1}^n \mathbf{x}_i^T\mathbf{W}_i \boldsymbol{\Gamma}_i\mathbf{x}_i$, and $\Lambda_{n\mathbf{W}} = n^{-1}\sum_{i=1}^n \mathbf{x}_i^T\mathbf{W}_i\mathbf{V}_i(\boldsymbol{\theta}_0)\mathbf{W}_i^T\mathbf{x}_i$. We can show that $\mathbf{H}_{\mathbf{W}} - \mathbf{H}$ is nonnegative definite. Therefore, the estimator $\hat{\boldsymbol{\beta}}(\tau)$, obtained with $\mathbf{W}_i = \boldsymbol{\Gamma}_i\{\mathbf{V}_i(\boldsymbol{\theta}_0)\}^{-1}$ is optimal within such a class, including the estimator $\tilde{\boldsymbol{\beta}}(\tau)$ obtained by assuming working independence. The proof is similar to that of Theorem 3 in Jung (1996).

Theorem 1 suggests that the induced smoothed estimator $\hat{\boldsymbol{\beta}}_s(\tau)$ is asymptotically equivalent to the unsmoothed estimator $\hat{\boldsymbol{\beta}}(\tau)$. Since the computation based on induced smoothing estimation equation is more efficient, we use the induced smoothing estimator throughout our numerical studies.

With copulas we can specify the joint distribution of responses from the same subject, and thus obtain the conditional distribution of any measurement given the others for each subject, as discussed in Section 2.4. Theorem 2 shows that the conditional distribution can be estimated consistently.

Theorem 2. *Under the conditions of Theorem* 1, *for any* $y_J^* \in \mathbb{R}$, *we have*

$$\hat{F}(y_J^*|\mathbf{x}_1^*,\ldots,\mathbf{x}_J^*,y_1^*,\ldots,y_{J-1}^*)\xrightarrow{p}F(y_J^*|\mathbf{x}_1^*,\ldots,\mathbf{x}_J^*,y_1^*,\ldots,y_{J-1}^*).$$

Theorem 2 implies that the solution of (2.8), denoted by $\hat{q}_J(\tau^*)$, converges in probability to the τ^*th conditional quantile of y_J^* given $\mathbf{x}_1^*,\ldots,\mathbf{x}_J^*,y_1^*,\ldots,y_{J-1}^*$, denoted by $q_J(\tau^*)$, the solution to $F(y|\mathbf{x}_1^*,\ldots,\mathbf{x}_J^*,y_1^*,\ldots,y_{J-1}^*)=\tau^*$. This ensures the asymptotic validity of the prediction intervals constructed by $\hat{q}_J(\tau^*)$.

Remark 6. The consistency of $\hat{\boldsymbol{\beta}}(\tau)$ requires the correct specification of model (2.2) only at the quantile level τ of interest. The misspecification of the global marginal linear quantile regression model (2.2) and the copula model (2.1) does not affect the consistency of $\hat{\boldsymbol{\beta}}(\tau)$, but only its efficiency and the consistency of the estimated conditional distribution in Theorem 2. However, our numerical studies in Section 3 suggest that even under copula model misspecification, the constructed prediction intervals still maintain the coverage probabilities well. In practice, we can check the adequacy of model (2.2) by applying some goodness-of-fit test, for instance, the method in He and Zhu (2003). Our experience suggests that the proposed method performs reasonably well unless the global linearity assumption is severely violated, in which case more flexible models such as polynomial quantile regression can be considered in Step 1.

3. Simulation Study

The data was generated from the model

$$y_{ij}=-0.5+0.5x_{ij1}+x_{ij2}+(1+\gamma x_{ij1})\epsilon_{ij}, i=1,\ldots,n=500, j=1,\ldots,4,$$

where x_{ij1} are i.i.d. $Bernoulli(1,0.5)$, and x_{ij2} are i.i.d. $N(0,1)$. We considered Case 1 (heteroscedastic normal error) with $\gamma=1$, $\boldsymbol{\epsilon}_i=(\epsilon_{i1},\ldots,\epsilon_{i4})^T\sim N(0,\Sigma)$; Case 2 (homoscedastic t error) with $\gamma=0$, $\boldsymbol{\epsilon}_i$ from a multivariate t_3 distribution with covariance Σ; and Case 3 (heteroscedastic lognormal error) with $\gamma=1$, $\boldsymbol{\epsilon}_i$ from a multivariate lognormal distribution with mean zero and covariance Σ_i. In Cases 1–2, the covariance Σ is common and covariate-independent, and it has an AR(1) structure in Case 1 and an exchangeable correlation structure in Case 2 with variance 1 and correlation ϱ. In Case 3, we let $\Sigma_i=(\sigma_{i,j,j'})_{j,j'=1}^4$ with $\sigma_{i,j,j'}=\varrho^{|j-j'|+(x_{ij1}+x_{ij'1})/2}$ for $j\neq j'$ and $\sigma_{i,j,j'}=1$ for $j=j'$. We took $\varrho=0.3$, 0.5 and 0.8. The assumed models (2.1) and (2.2) hold in both Cases 1 and 2, but in Case 3 the copula model (2.1) is misspecified as the true copula parameter $\boldsymbol{\theta}$ is in fact covariate-dependent.

3.1. Quantile coefficient estimation

We compare three different types of estimators of $\boldsymbol{\beta}(\tau)=(\beta_0(\tau),\beta_1(\tau),$

$\beta_2(\tau))^T$: the estimator $\tilde{\boldsymbol{\beta}}(\tau)$ obtained by assuming working independence (WI); the parametric copula regression estimator assuming asymmetric Laplace marginal distributions (AL); the estimator proposed by Leng and Zhang (2014) (LZ); the proposed semiparametric copula-based quantile regression (CQR) estimator $\hat{\boldsymbol{\beta}}_s(\tau)$. For all copula-based methods, we considered multivariate Gaussian copula and t-copula with exchangeable, first-order autoregressive AR(1) and unstructured correlation structures. To save space, we only report the results from Gaussian copula as the t-copula gave similar results. For each scenario, the simulation was repeated 500 times.

Table 1 summarizes the relative efficiency of the copula-based and LZ estimators with respect to the WI estimator, and the coverage probabilities of 95% confidence intervals in Cases 1–3 at $\tau = 0.25$. Results for $\tau = 0.5$ are in the supplementary file. The confidence intervals were constructed based on normality and the asymptotic variance estimated by the induced smoothing method for the proposed estimator, and the Hessian matrix for the AL approach. Results show that the proposed method is quite insensitive to the choice of correlation structures in the copula function, so here we only report the results based on copula with an exchangeable correlation structure and leave the rest in the supplement.

Across all scenarios considered, the proposed CQR estimator shows higher efficiency than the WI estimator, and the efficiency gain is more obvious when there is a stronger intra-subject dependence. The induced smoothing method gives reasonable variance estimation; the coverage probabilities of the CQR intervals are close to 95% in all three cases. The CQR method performs well even under the misspecification of the copula function (Case 2) and of the copula model (2.1) in Case 3. The AL estimator for $\boldsymbol{\beta}_2(\tau)$ appears to have competitive efficiency, but its estimation for the other two coefficients can have very low efficiency in some scenarios, and the coverage probabilities are in general poor. A closer examination shows that the AL estimator has large bias that is caused by the misspecification of the marginal distributions. These results suggest that inference of copula regression based on parametric marginals can be misleading under model misspecifications. The LZ estimator performs similarly as CQR, but the latter tends to be more efficient for estimating $\boldsymbol{\beta}_2(\tau)$, especially in Cases 1 and 3 with heteroscedastic errors.

3.2. Prediction

We assessed the performance of the proposed method for predicting the τth conditional quantile of y_{n4} conditioning on the covariates and $y_{nj}, j = 1, 2, 3$.

Table 1. Relative efficiency (RE) with respect to the working independence estimator $\tilde{\beta}(\tau)$ and coverage probability (CovP) of 95% confidence intervals from different methods at $\tau = 0.25$ in Cases 1–3.

			RE			CovP		
Case	ϱ	Method	$\beta_0(\tau)$	$\beta_1(\tau)$	$\beta_2(\tau)$	$\beta_0(\tau)$	$\beta_1(\tau)$	$\beta_2(\tau)$
1	0.30	LZ	1.09	1.06	1.06	0.94	0.94	0.94
		AL	0.44	0.86	1.10	0.55	0.56	0.69
		CQR	1.04	1.07	1.26	0.94	0.94	0.92
	0.50	LZ	1.09	1.12	1.16	0.94	0.94	0.94
		AL	0.21	0.54	1.32	0.27	0.41	0.71
		CQR	1.08	1.11	1.29	0.95	0.92	0.91
	0.80	LZ	1.14	1.40	1.68	0.93	0.93	0.94
		AL	0.07	0.17	2.71	0.05	0.03	0.82
		CQR	1.18	1.33	1.79	0.94	0.92	0.93
2	0.30	LZ	1.09	1.14	1.08	0.91	0.94	0.94
		AL	0.08	1.34	1.50	0.02	0.72	0.78
		CQR	1.07	1.11	1.18	0.94	0.93	0.94
	0.50	LZ	1.09	1.27	1.23	0.91	0.94	0.93
		AL	0.05	2.12	1.91	0.01	0.81	0.81
		CQR	1.13	1.31	1.32	0.95	0.93	0.95
	0.80	LZ	1.10	2.01	2.21	0.92	0.94	0.94
		AL	0.03	5.02	5.16	0.00	0.90	0.90
		CQR	1.25	2.01	2.10	0.95	0.94	0.94
3	0.30	LZ	1.05	1.04	1.03	0.92	0.95	0.94
		AL	0.77	0.93	1.07	0.87	0.84	0.89
		CQR	1.05	1.04	1.26	0.94	0.94	0.93
	0.50	LZ	1.07	1.08	1.09	0.94	0.93	0.94
		AL	0.52	0.94	1.12	0.78	0.82	0.88
		CQR	1.07	1.08	1.26	0.94	0.93	0.92
	0.80	LZ	1.07	1.28	1.47	0.92	0.94	0.95
		AL	0.28	1.19	1.66	0.47	0.82	0.93
		CQR	1.16	1.32	1.68	0.95	0.91	0.93

LZ: the method from Leng and Zhang (2014) based on quadratic inference assuming an exchangeable working correlation structure; AL: copula regression method assuming asymmetric Laplace marginal distributions; CQR: the proposed copula-based quantile regression method. Both AL and CQR are based on Gaussian copula with exchangeable correlation structure.

For comparison, we also included the prediction from copula regression assuming asymptotic Laplace marginals, and the methods from conventional linear quantile regression and from Leng and Zhang (2014) that use only covariate information for prediction. For copula-based methods, we used the Gaussian copula with an exchangeable correlation structure. For each method, we report the mean prediction error at $\tau = 0.25$ and 0.5, defined as $1/500 \sum_{k=1}^{500} \rho_\tau(y_{n4,k} - \hat{q}_{n4,k}(\tau))$,

Table 2. The mean prediction error (MPE) at $\tau = 0.25$ and 0.5, and coverage probability (CovP) and mean length (ML) of 90% prediction intervals in Cases 1–3 with $\varrho = 0.5$. Values in the parentheses are standard errors.

		MPE			
Case	Method	$\tau = 0.25$	$\tau = 0.5$	CovP	ML
1	QR	0.987 (0.041)	0.599 (0.023)	0.902	4.946 (0.074)
	LZ	0.984 (0.041)	0.598 (0.023)	0.904	4.916 (0.074)
	AL	0.907 (0.037)	0.559 (0.021)	0.892	4.727 (0.070)
	CQR	0.925 (0.038)	0.564 (0.021)	0.908	4.640 (0.071)
2	QR	0.848 (0.039)	0.550 (0.026)	0.888	4.710 (0.013)
	LZ	0.844 (0.039)	0.550 (0.026)	0.889	4.658 (0.012)
	AL	0.678 (0.034)	0.438 (0.022)	0.857	3.027 (0.030)
	CQR	0.708 (0.032)	0.444 (0.022)	0.910	4.594 (0.096)
3	QR	1.275 (0.085)	0.782 (0.050)	0.900	7.516 (0.116)
	LZ	1.267 (0.085)	0.775 (0.049)	0.901	7.478 (0.116)
	AL	1.181 (0.075)	0.742 (0.042)	0.631	5.210 (0.066)
	CQR	1.236 (0.082)	0.782 (0.049)	0.912	7.941 (0.280)

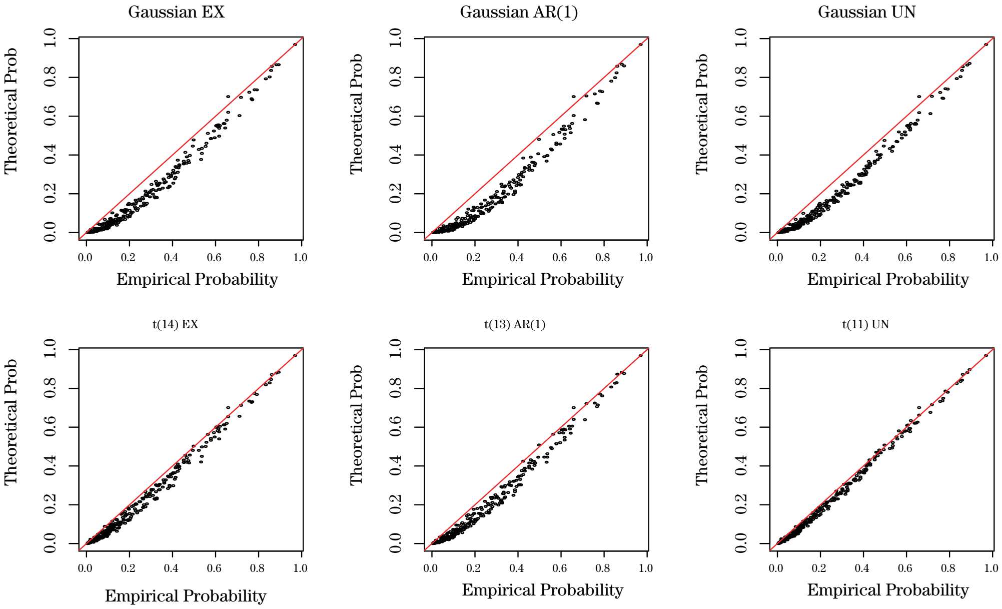
QR: conventional quantile regression method that uses only covariate information for prediction; LZ: prediction based on the estimator of Leng and Zhang (2014); AL: the copula regression based on asymmetric Laplace marginals; CQR: the proposed copula-based quantile regression method.

where $y_{n4,k}$ is the actual response of the nth subject at the fourth time point from the kth simulation and $\hat{q}_{n4,k}(\tau)$ is the predicted τth conditional quantile of $y_{n4,k}$. In addition, we consider the coverage probabilities and mean lengths of 90% prediction intervals, constructed by the predicted 5th and 95th percentiles of $y_{n4,k}$ from different methods.

Table 2 summarizes the prediction results of the four methods in Cases 1–3 with $\varrho = 0.5$ (results for $\varrho = 0.3$ and 0.8 are provided in the supplementary file). The proposed CQR method gives more accurate predictions than the QR and LZ methods in most scenarios, even when the correlation structure, or copula function, or copula model are misspecified. The prediction intervals from QR, LZ and CQR have coverage probabilities close to 90%, but the intervals from the CQR method are in general narrower. On the appearance, the mean prediction errors from the AL method are comparable to those from the CQR method, but predictions from the AL method may be misleading, manifested by the low coverage of the 90% prediction intervals in Cases 2–3.

4. Analysis of the New Jersey Nursing Home Data

To illustrate the proposed method, we analyzed the New Jersey nursing



EX: exchangeable correlation structure; AR(1): first–order autoregressive correlation structure; UN: unstructured correlation.

Figure 1. Copula probability-probability plots from Gaussian copula and t-copula with different correlation structures.

home data set from report years 2009 to 2013, available at Centers for Medicare and Medicaid Services (`www.cms.gov`). The data set contains the information from 286 nursing facilities. To quantify the utilization of nursing home care, we took the annual occupation rate as the ratio of total residents and the number of certified beds for each facility. We defined the response variable y_{ij} as the logistic transformed occupation rate of the ith facility at year j, where $i = 1, \ldots, n = 286$ and $j = 1, \ldots, 5$. We considered five covariates, including two indicator variables indicating whether the facility is a non-profit or government-owned (with private party being the baseline), the reporting year (after subtracting 2008) and the reported total nurse staffing hours per resident per day (TOTHRS). Our preliminary analysis suggests that the data has temporal correlations above 0.6, so it is important to accommodate such correlation in both estimation and prediction.

Before carrying out the copula quantile regression analysis, we checked the adequacy of model (2.2) by applying the lack-of-fit test of He and Zhu (2003) at 19 quantile levels $\tau = 0.05, 0.1, \ldots, 0.95$. The minimum p-value was 0.03, suggesting a reasonable fit of model (2.2) after multiple test adjustment. In addition, we examined the copula probability-probability plots from Gaussian copula and t-copula with exchangeable, first-order autoregressive or unstructured correlation structures in Figure 1. In each plot, the x-axis shows the empirical copula probabilities $\{C_n(u_{i1},\ldots,u_{i5}), i=1,\ldots,n\}$, where $C_n(u_1,\ldots,u_5)=n^{-1}\sum_{i=1}^{n} I(\tilde{u}_{i1}\leq u_1,\ldots,\tilde{u}_{i5}\leq u_5)$, $\tilde{u}_{ij}$ are obtained as in Step 1 of Section 2.2, and the y-axis shows the corresponding probabilities from the estimated parametric copulas. The departure of a probability-probability plot from the 45 degree line indicates that the parametric copula does not agree with the empirical copula and thus may not be a good choice (Mendes and De Melo (2010)). Figure 1 suggests that t-copula (degrees of freedom estimated as 11) with unstructured correlation has better agreement with the empirical copula. Therefore, in the sequel we focus on t-copula with unstructured correlation.

Table 3 summarizes the estimated covariate effects on the τth conditional quantiles of y_{ij} from different methods at $\tau=0.05, 0.5$, and 0.95. Results from the CQR method suggest that, compared to private facilities, non-profit facilities tend to have higher occupation rate at all three quantiles, while government-owned ones show no significant difference; the occupation rate tends to be decreasing over years and this effect is significant at the median; the total of nurse staffing hours has a significantly negative effect at the median and the lower quantile of the occupation rate, but the effect is not significant at $\tau=0.95$. In general, the AL method gives estimates with larger standard errors, making it miss the significance of the Non-profit and TOTHRS variables. Compared to AL, the CQR gives estimation and significance results that are more in line with those from QR and LZ, except that the LZ method misses the significance of the Non-profit effect at $\tau=0.95$ due to a larger standard error.

To assess the prediction accuracy of the different methods, we carried out a leave-one-out cross validation. For each $i=1,\ldots,n$, we left out the data from the ith subject, obtained the parameter estimation based on the rest of the data and used the estimation to predict the conditional quantiles of the response of the ith subject in year 2013 given the covariates and the responses in previous four years. Table 4 summarizes the mean prediction error at median, the coverage probability and mean length of 90% prediction intervals formed by the predicted 5th and 95th conditional quantiles from three methods: the methods from conventional linear quantile regression method and Leng and Zhang (2014) that use only covariates information for prediction, the copula regression method based on AL marginals and the proposed copula-based CQR method. The proposed method

Table 3. Estimated effects of covariates on the τth conditional quantile of the logit occupation rate from different methods. Values in the parentheses are standard errors.

		Variable			
τ	Method	Non-profit	Government	Year	TOTHRS
0.05	QR	0.54 (0.16)	0.13 (0.26)	−0.04 (0.03)	−0.40 (0.11)
	LZ	0.45 (0.13)	0.04 (0.23)	−0.03 (0.03)	−0.30 (0.09)
	AL	0.19 (0.29)	0.20 (0.42)	0.01 (0.03)	−0.15 (0.14)
	CQR	0.51 (0.19)	0.17 (0.31)	−0.06 (0.04)	−0.30 (0.11)
0.5	QR	0.30 (0.10)	0.39 (0.28)	−0.04 (0.02)	−0.17 (0.08)
	LZ	0.29 (0.10)	0.42 (0.28)	−0.05 (0.02)	−0.17 (0.07)
	AL	0.39 (0.12)	0.64 (0.37)	−0.05 (0.01)	−0.12 (0.07)
	CQR	0.28 (0.11)	0.44 (0.29)	−0.04 (0.02)	−0.17 (0.07)
0.95	QR	3.00 (0.84)	0.28 (0.35)	−0.11 (0.08)	−0.12 (0.22)
	LZ	1.42 (0.94)	0.15 (0.31)	−0.15 (0.07)	−0.49 (0.23)
	AL	0.48 (1.13)	0.90 (0.69)	−0.10 (0.05)	−0.08 (0.21)
	CQR	3.00 (0.79)	0.28 (0.40)	−0.11 (0.08)	−0.12 (0.24)

QR: the conventional quantile regression estimator; LZ: the estimator of Leng and Zhang (2014) based on quadratic inference assuming an exchangeable working correlation structure; AL: the copula regression assuming asymmetric Laplace marginal distributions; CQR: the proposed copula-based quantile regression estimator; TOTHRS: reported total nurse staffing hours per resident per day.

Table 4. Results of the cross validation study of the nursing home data. The values in the parentheses are standard errors.

	MPE ($\tau = 0.5$)	CovP	ML
QR	0.387 (0.025)	0.892	3.599 (0.048)
LZ	0.377 (0.025)	0.877	3.349 (0.053)
AL	0.115 (0.014)	0.727	1.288 (0.069)
CQR	0.163 (0.023)	0.883	0.869 (0.060)

MPE: the mean prediction error at median; CovP: the coverage probability of 90% prediction intervals; ML: the mean length of 90% prediction intervals; QR: the standard quantile regression method that uses only covariates information for prediction; LZ: prediction based on covariates and the quantile coefficient estimator in Leng and Zhang (2014); AL: the copula regression assuming asymmetric Laplace distribution; CQR: the proposed copula quantile regression.

shows clear advantage; it gives more accurate median prediction than the QR and LZ methods, and narrower prediction intervals than all the other three methods with coverage close to 90%.

5. Discussion

For notational simplicity, we assumed in model (2.2) that $\boldsymbol{\beta}_0(\tau)$ was common across measurement time j, but the proposed method can be easily modified to allow time-dependent coefficients, and the estimation efficiency can still be improved through the proposed three-step procedure as long as some elements of $\boldsymbol{\beta}_0(\tau)$ are common across time. The proposed method can be easily extended to accommodate unbalanced designs. For instance, suppose that one subject has measurements obtained at time points $1, \ldots, J-1$. Then the assumed copula model (2.1) for the $J-1$ responses becomes $P(Y_1 < y_1, \ldots, Y_{J-1} < y_{J-1}|\mathbf{x}) = C\{F_1(y_1|\mathbf{x}), \ldots, F_{J-1}(y_{J-1}|\mathbf{x}), 1; \boldsymbol{\theta}_0\}$ and the corresponding density function is $f(y_1, \ldots, y_{J-1}|\mathbf{x}) = \partial^{J-1}C\{u_1, \ldots, u_{J-1}, 1; \boldsymbol{\theta}_0\}/\partial u_1 \cdots \partial u_{J-1} \prod_{j=1}^{J-1} f_j(y_j|\mathbf{x})$.

In this paper, the dependence of longitudinal measurements is modeled through copulas to perform prediction and improve the estimation efficiency. For model parsimony, we considered a simplified copula in model (2.1), which assumes that the copula function is independent of covariates except through the conditional marginals $F_j(\cdot|\mathbf{x})$ and that the copula parameter $\boldsymbol{\theta}$ is common across $\mathbf{x}$. For more flexibility, we can extend the proposed idea of semiparametric marginals to accommodate more general copulas, for instance, copulas with $\boldsymbol{\theta}$ depending on $\mathbf{x}$ parametrically or nonparametrically, or nonparametric estimation of covariate-dependent copulas, such as those studied in Tsukahara (2005), Abegaz, Gijbels and Veraverbeke (2012), Omelka, Gijbels and Veraverbeke (2009) and Veraverbeke, Omelka and Gijbels (2011) for two-dimensional responses and univariate $\mathbf{x}$. Research in this direction for longitudinal data deserves further investigation.

Some applications may involve correlated outcomes of mixed types, including both continuous and discrete outcomes, for instance the burn injury study reported in Fan and Gijbels (1996). The proposed semiparametric method can be extended to analyze such data by adapting the joint modeling idea in Song, Li and Yuan (2009). Specifically, we can model the marginal distributions of continuous outcomes through fitting quantile regression processes, while modeling the marginals of discrete outcomes through fitting generalized linear models.

Supplementary Materials

Proofs for Theorems 1–2, and some additional simulation results are provided in the online supplementary material.

Acknowledgment

The authors would like to thank the Editor, an associate editor, and two anonymous reviewers for their constructive comments that have significantly improved the paper, and Drs. Chenlei Leng and Weiping Zhang for providing their code. Wang's research was supported by National Science Foundation CAREER Award DMS-1149355 and OSR-2015-CRG4-2582 grant from KAUST. Feng's research is supported by National Natural Science Foundation of China (11571218, 11690012), the State Key Program in the Major Research Plan of National Science Foundation of China (91546202), and Program for Innovative Research Team of SUFE.

References

Aas, K., Czado, C., Frigessi, A. and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics* **44**, 182–198.

Abegaz, F., Gijbels, I. and Veraverbeke, N. (2012). Semiparametric estimation of conditional copulas. *Journal of Multivariate Analysis* **110**, 43–73.

Abrevaya, J. (2001). The effect of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics* **26**, 247–259.

Bai, Y., Kang, J. and Song, P. X. (2014). Efficient pairwise composite likelihood estimation for spatial-clustered data. *Biometrics* **70**, 661–670.

Bouyé, E. and Salmon, M. (2008). Dynamic copula quantile regressions and tail area dynamic dependence in forex markets. Web: `http://dx.doi.org/10.2139/ssrn.1129855`.

Brown, B. and Wang, Y. (2005). Standard errors and covariance matrices for smoothed rank estimators. *Biometrika* **92**, 149–158.

Chen, X., Koenker, R. and Xiao, Z. (2009). Copula-based nonlinear quantile autoregression. *Econometrics* **12**, S50–S67.

Chernozhukov, C., Fernández-Val, I. and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica* **78**, 1093–1125.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman and Hall, London.

Feng, Y., Chen, Y. and He, X. (2015). Bayesian quantile regression with approximate likelihood. *Bernoulli* **21**, 832–850.

Fu, L. and Wang, Y. G. (2012). Quantile regression for longitudinal data with a working correlation model. *Computational Statistics and Data Analysis* **56**, 2526–2538.

Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics* **8**, 140–154.

Geraci, M. and Bottai, M. (2014). Linear quantile mixed models. *Statistics and Computing* **24**, 461–479.

Haff, I. H., Aas, K. and Frigessi, A. (2010). On the simplified pair-copula construction – simply useful or too simplistic? *Journal of Multivariate Analysis* **101**, 1296–1310.

Hall, P. and Sheather, S. J. (1988). On the distribution of a studentized quantile. *Journal of the Royal Statistical Society. Series B Statistical Methodology* **50**, 381–391.

He, X., Fu, B. and Fung, W. (2003). Median regression for longitudinal data. *Statistics in Medicine* **22**, 3655–3669.

He, X. and Zhu, L. (2003). A Lack-of-fit test for quantile regression. *Journal of the American Statistical Association* **98**, 1013–1022.

Hendricks, W. and Koenker, R. (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American Statistical Association* **87**, 58–68.

Joe, H. (1996). Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters. Hayward, CA: Institute of Mathematical Statistics, *Lecture Notes–Monograph Series* **28**, 120–141.

Jung, S. H. (1996). Quasi-likelihood for median regression models. *Journal of the American Statistical Association* **91**, 251–257.

Kim, M. and Yang, Y. (2011). Semiparametric approach to a random effects quantile regression model. *Journal of the American Statistical Association* **106**, 1405–1417.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis* **91**, 74–89.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica* **46**, 33–50.

Lambert, P. and Vandenhende, F. (2002). A copula-based model for multivariate non-normal longitudinal data: Analysis of a dose titration safety study on a new antidepressant. *Statistics in Medicine* **21**, 3197–3217.

Leng, C. and Zhang, W. (2014). Smoothing combined estimating equations in quantile regression for longitudinal data. *Statistics and Computing* **24**, 123–136.

Meester, S. G. and MacKay, J. (1994). A parametric model for cluster correlated categorical data. *Biometrics* **50**, 954–963.

Mendes, B. V. M. and De Melo, E. F. L. (2010). Local estimation of dynamic copula models. *International Journal of Theoretical and Applied Finance* **13**, 241–258.

Mu, Y. and Wei, Y. (2009). A dynamic quantile regression transformation model for longitudinal data. *Statistica Sinica* **19**, 1137–1153.

Noh, H., Ghouch, A. E. and Van Keilegom, I. (2015). Semiparametric conditional quantile estimation through copula-based multivariate models. *Journal of Business and Economic Statistics* **33**, 167–178.

Omelka, M., Gijbels, I. and Veraverbeke, N. (2009). Improved kernel estimation of copulas: weak convergence and goodness-of-fittesting. *The Annals of Statistics* **37**, 3023–3058.

Pang, L., Lu, W. and Wang, H. (2012). Variance estimation in censored quantile regression via induced smoothing. *Computational Statistics and Data Analysis* **56**, 785–796.

Portnoy, S. (2003). Censored regression quantiles. *Journal of the American Statistical Association* **98**, 1001–1012.

Reich, B. J., Bondell, H. D. and Wang, H. (2010). Flexible Bayesian quantile regression for independent and clustered data. *Biostatistics* **11**, 337–352.

Shi, P. and Frees, E. W. (2010). Long-tail longitudinal modeling of insurance company expenses. *Insurance: Mathematics and Economics* **47**, 303–314.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Universite dé Paris* **8**, 229–231.

Smith, M., Min, A., Almeida, C. and Czado, C. (2010). Modeling longitudinal data using a pair-copula decomposition of serial dependence. *Journal of the American Statistical Association* **105**, 1467–1479.

Song, P. X. K. (2000). Multivariate dispersion models generated from Gaussian copula. *Scan-*

dinavian Journal of Statistics **27**, 305–320.

Song, P. X. K., Li, M. and Yuan, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics* **65**, 60–68.

Sun, J., Frees, E. W. and Rosenberg, M. A. (2008). Heavy-tailed longitudinal data modeling using copulas. *Insurance: Mathematics and Economics* **42**, 817–830.

Tang, C. Y. and Leng, C. (2011). Empirical likelihood and quantil regression in longitudinal data analysis. *Biometrika* **98**, 1001–1006.

Tsukahara, H. (2005). Semiparametric estimation in copula models. *Canadian Journal of Statistics* **33**, 357–375.

Veraverbeke, N., Omelka, M. and Gijbels, I. (2011). Estimation of a conditional copula and association measures. *Scandinavian Journal of Statistics* **38**, 766–780.

Wang, H. (2009). Inference on quantile regression for heteroscedastic mixed models. *Statistica Sinica* **19**, 1247–1261.

Wang, H. and Zhou, X. (2010). Estimation of the retransformed conditional mean in health care cost studies. *Biometrika* **97**, 147–158.

Wang, J. (2012). Bayesian quantile regression for parametric nonlinear mixed effects models. *Statistical Methods and Applications* **21**, 279–295.

Wang, Y. G., Shao, Q. and Zhu, M. (2009). Quantile regression without the curse of unsmoothness. *Computational Statistics and Data Analysis* **53**, 3696–3705.

Wei, Y. and Carroll, R. J. (2009). Quantile regression with measurement error. *Journal of the American Statistical Association* **104**, 1129–1143.

Wei, Y. and He, X. (2006). Conditional growth charts (with discussions). *The Annals of Statistics* **34**, 2069–2097 and 2126–2131.

Wu, C. O. and Tian, X. (2013). Nonparametric estimation of conditional distributions and rank-tracking probabilities with time-varying transformation models in longitudinal studies. *Journal of the American Statistical Association* **108**, 971–982.

Yuan, Y. and Yin, G. S. (2010). Bayesian quantile regression for longitudinal studies with nonignorable missing data. *Biometrics* **66**, 105–114.

Department of Statistics, George Washington University, Washington, District of Columbia 20052, USA.

E-mail: judywang@email.gwu.edu

Institute of Data Science and Statistics, and School of Statistics and Management, Shanghai University of Finance and Economics, and Key Laboratory of Mathematical Economics(SUFE), Ministry of Education, Shanghai 200433, China.

E-mail: feng.xingdong@mail.shufe.edu.cn

Institute of Data Science and Statistics, and School of Statistics and Management, Shanghai University of Finance and Economics, and Key Laboratory of Mathematical Economics(SUFE), Ministry of Education, Shanghai 200433, China.

E-mail: dongchen@live.sufe.edu.cn

(Received March 2016; accepted July 2017)