Statistica Sinica Preprint No: SS-2021-0422	
Title	Dynamic Copula-Based Nonparametric Estimation of Rank-Tracking Probabilities With Longitudinal Data
Manuscript ID	SS-2021-0422
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0422
Complete List of Authors	Xiaoyu Zhang, Mixia Wu and Colin O. Wu
Corresponding Authors	Mixia Wu
E-mails	wumixia@bjut.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

Dynamic copula-based nonparametric estimation of rank-tracking probabilities with longitudinal data

Xiaoyu Zhang[1], Mixia Wu[1,2], and Colin O. Wu[3]

[1] *Beijing University of Technology,* [2] *Beijing Institute for Scientific and Engineer Computing and* [3] *National Heart, Lung and Blood Institute*

Abstract:

The rank-tracking probability (RTP) is a useful statistical index for measuring the "tracking ability" of longitudinal disease risk factors in biomedical studies. A flexible nonparametric method for estimating the RTP is the two-step unstructured kernel smoothing estimator (Wu and Tian, 2013), which may be applied when there are time-invariant and categorical covariates. We propose in this article a dynamic copula-based smoothing method for estimating the RTP and show that it is both theoretically and practically superior to the unstructured smoothing method. We derive the asymptotic mean squared errors of the copula-based kernel smoothing estimators and demonstrate through a simulation study that the copula-based smoothing method has smaller empirical mean squared errors than the unstructured smoothing method. We apply the proposed estimation method to a longitudinal epidemiological study and show that it leads to clinically meaningful findings in biomedical applications.

Key words and phrases: dynamic copula model, longitudinal study, rank-tracking probability, risk factor, two-step smoothing, unstructured smoothing.

1. Introduction

An important objective of longitudinal analysis in biomedical studies is to investigate the effect of covariates on the response variables over time. Existing methods for longitudinal analysis are mostly focused on regression models based on the conditional mean and correlations. Examples of such methods can be found in Hoover et al. (1998), Rice and Wu (2001), Diggle et al., (2002), Fan et al. (2007), and Wu and Tian (2018), among others. These methods, although popular in practice, lack the ability to quantitatively track the persistence of disease risk factors over a time range of interest.

The need for longitudinal methods beyond conditional means and correlations can be demonstrated by the National Heart, Lung and Blood Institute Growth and Health Study (NGHS), e.g. NGHSRG (1992), NHBPEP (2004) and Obarzanek et al. (2010), in which many scientific questions could be answered by evaluating the conditional distributions rather than the conditional means or correlations. Designed as a prospective epidemiological study, the NGHS contains up to 10 annual follow-up observations from 1213 African American girls and 1166 Caucasian girls who were en-

rolled into the study at age 9 or 10 years. An important objective of the study is to determine whether a girl's cardiovascular disease risk factor has any tracking ability over an age range.

A class of longitudinal methods in the literature for evaluating tracking is based on modeling the serial correlations. However, serial correlation could be an inadequate measure of tracking ability when the conditional distribution function of the outcome variable is unknown. A practical approach for evaluating tracking ability is through the conditional distribution functions of the outcome variables, such as Hall, Wolff and Yao (1999) and Hall, Racine and Li (2004). Wu and Tian (2013) studied a class of conditional distributions known as the "rank-tracking probability" (RTP) to quantify the tracking ability of a longitudinal outcome variable and developed an unstructured kernel smoothing method to estimate the RTPs.

The unstructured smoothing estimation proposed by Wu and Tian (2013) could be impractical in many situations. The reason is that the RTPs involve joint probabilities at two time points which may not be appropriately estimated due to the lack of sufficient observations at these time points. This motivates the copula-based method for the estimation of the RTPs because, based on the Sklar's theorem, any multivariate distribution can be expressed by its marginal distributions and a copula function. Based

on a given copula model, we can estimate a multivariate distribution by separately estimating the marginal distributions and the copula function. If both the marginal distributions and the copula function are estimated by nonparametric estimators, such as the kernel estimators (Fermanian and Scaillet, 2003; Chen and Huang, 2007), we may encounter the problem of "curse of dimensionality." On the other hand, imposing a parametric structure on both the copula function and marginal distributions may lead to excessive bias due to the potential model misspecification (Hărdle, 2012). As a useful compromise, Joe (2014) suggested a flexible semiparametric copula approach, in which the copula function is modeled parametrically but the marginal distributions are estimated nonparametrically by their corresponding empirical distributions. To reduce the potential of model misspecification, Joe (1997) suggested a data-driven method which selects the copula function from a set of copula models based on a given model selection criterion.

We develop in this article a dynamic copula-based smoothing method to estimate the RTPs of a longitudinal outcome variable. This method is carried out through two estimation steps. First, we estimate the raw joint probabilities at a set of distinct design time points based on a copula model that is selected from a set of candidate copula models by maximizing the

likelihood functions. Second, we estimate the dynamic RTPs at any time points by smoothing the raw estimates using a kernel smoothing method. We compare our copula-based smoothing method with the unstructured smoothing method of Wu and Tian (2013) through a simulation study. Our simulation results suggest that the proposed copula-based estimation method is superior to the unstructured estimation method in the sense of having smaller empirical mean squared errors. We apply the proposed estimation method to the NGHS blood pressure data and show that it leads to clinical meaningful results for the tracking patterns of blood pressure levels for adolescent girls.

In Section 2, we introduce the longitudinal data structure and the definition of the RTPs. In Section 3, we present the two-step copula-based smoothing estimation procedure and derive the asymptotic properties of the raw and smoothing estimators. We conduct a simulation study in Section 4 and demonstrate the application of our estimation method to the NGHS blood pressure data in Section 5. We summarize the conclusions and some final remarks in Section 6.

2. Data Structure and Rank-Tracking Probabilities

2.1 Longitudinal Observations at Design Time Points

We consider the stochastic processes indexed by the time point $t \in \mathcal{T}$, where $\mathcal{T}$ is a bounded subset of $[0, \infty)$. At any given $t \in \mathcal{T}$, $Y(t) \in R$ is the real-valued outcome variable. For simplicity, our longitudinal sample of $\{Y(t); t \in \mathcal{T}\}$ is assumed to contain n independent subjects, and each subject is observed at a randomly selected subset of $K > 1$ distinct "design time points" $\mathscr{K} = \{t_{(1)}, \ldots, t_{(K)}\}$, where $t_{(k)} \in \mathcal{T}$. For the ith subject, $1 \leq i \leq n$, the outcome $Y_i(t_{ij}) = Y_{ij}$ is collected at time points $t = t_{ij} \in \mathscr{K} \subset \mathcal{T}, j = 1, \cdots, n_i$, where n_i is the number of observations for the ith subject and $N = \sum_{i=1}^{n} n_i$. At each $t_{(k)}$, $\mathscr{F}_k$ is the set of subjects with observations when $\{(Y_i(t_{(k)})); i \in \mathscr{F}_k,\ k = 1, \cdots K\}$ is the sample of $\{Y(t); t \in \mathcal{T}\}$, where $Y_i(t_{(k)})$ is the outcome for the ith subject. Let $n_k = \sharp\{i \in \mathscr{F}_k\}$ be the number of subjects in $\mathscr{F}_k$ and $n_{g,h} = \sharp\{i \in \mathscr{F}_g \bigcap \mathscr{F}_h\}$ the number of subjects in the intersection of $\mathscr{F}_g$ and $\mathscr{F}_h$.

This formulation of longitudinal samples is common in biomedical studies. In an epidemiological study, a subject's follow-up time is often chosen from a set of "design time points," which may lead to a large K and $n_{g,h} \ll \min\{n_g, n_h\}$. When the observed time points are not exactly con-

tained in $\mathscr{K}$, it is common to pool together the adjacent observed time points into $\mathscr{K}$ by a clinically meaningful criterion.

2.2 The Rank-Tracking Probabilities

Suppose that, for $t \in \mathcal{T}$, the health status of a subject at time t is determined by whether $Y(t) \in A(t)$ for a prespecified subset $A(\cdot) \subseteq R$, which may change with t. The tracking ability of $\{Y(t); t \in \mathcal{K}\}$ at any two time points $t_1 < t_2$ can be measured by the conditional probability of $\{Y(t_2) \in A(t_2)\}$ given $\{Y(t_1) \in A(t_1)\}\}$, which is referred by Wu and Tian (2013) as the rank-tracking probability (RTP) based on $A(\cdot)$ at $t_1 < t_2$,

$$RTP_A(t_1, t_2) = P\{Y(t_2) \in A(t_2) \mid Y(t_1) \in A(t_1)\}, \tag{2.1}$$

where the choice of $A(\cdot)$ depends on the study questions and scientific objectives. As noted in Wu and Tian (2013), the "rank" in (2.1) does not necessarily refer to statistical ranking but is more generally used to characterize the ordinal "health status" of a subject in a biomedical study. A direct extension of (2.1) is to evaluate the probability that the subject's health status develops from status $A_1(\cdot)$ at time t_1 to status $A_2(\cdot)$ at time t_2. The RTP based on $A_1(\cdot)$ and $A_2(\cdot)$ at $t_1 < t_2$ is then

$$RTP_{A_1, A_2}(t_1, t_2) = P\{Y(t_2) \in A_2(t_2) \mid Y(t_1) \in A_1(t_1)\} \tag{2.2}$$

In biomedical studies, it is common to define $A_k(\cdot)$, $k = 1, 2$, using certain threshold values $y_k(t)$. A natural choice of $A_k(t)$ is $A_k(t) = (y_k(t), \infty)$, $k = 1, 2$, and a threshold-based RTP for (2.2) is

$$RTP_{A_1, A_2}(t_1, t_2) = P_{A_1, A_2}(t_1, t_2)/P_{A_1}(t_1), \tag{2.3}$$

where, $P_{A_1, A_2}(t_1, t_2) = P\{Y(t_1) > y_1(t_1), Y(t_2) > y_2(t_2)\}$ and $P_{A_1}(t_1) = P\{Y(t_1) > y_1(t_1)\}$ are the joint and marginal probabilities of $Y(t_2)$ and $Y(t_1)$, respectively. Furthermore, when $y_k(t)$ are certain quantile values, the corresponding RTP is referred by Wu and Tian (2013) as the " quantile-based RTP." Let $y_{\alpha_i}(t)$ be the $(100 \times \alpha_i)$ quantile of $Y(t)$, then the quantile-based RTP is

$$RTP_{\alpha_1, \alpha_2}(t_1, t_2) = P\{Y(t_2) > y_{\alpha_2}(t_2) \mid Y(t_1) > y_{\alpha_1}(t_1)\}. \tag{2.4}$$

Since thresholds are widely used for defining the status of diseases, we focus on estimating the threshold-based RTP (2.3), especially the quantile-based RTP (2.4), throughout this article.

3. Copula-based Smoothing Estimation

We develop the following two-step estimation procedure for the dynamic RTPs in (2.3) and (2.4) as functions of any two time points $t_1 < t_2 \in \mathcal{T}$: (a) obtaining the raw estimates for the joint probabilities at the design time

points under a copula family; (b) computing the smoothing estimates of the joint probabilities and RTPs at any two time points in $\mathcal{T}$.

3.1 Dynamic Copulas for the Joint Probabilities

Let $S_t(y) = P\{Y(t) > y\}$ be the "survival function" of $Y(t)$ at time point t. For any time points $t_1 < t_2 \in \mathcal{T}$, the joint "survival function" of $Y(t_1)$ and $Y(t_2)$ has the copula expression

$$P(y_1, y_2|t_1, t_2)=P\{Y(t_1) > y_1, Y(t_2) > y_2\}=C_{\theta(t_1,t_2)}\big(S_{t_1}(y_1), S_{t_2}(y_2)\big), \quad (3.5)$$

where $C_{\theta(t_1,\,t_2)}(s_1, s_2)$ is the " copula function" and $\theta(t_1, t_2)$ is the unknown time-varying copula parameter at time points t_1 and t_2. We assume that the copula model of $C_{\theta(t_1,\,t_2)}(s_1, s_2)$ is either known or unknown. When the copula model of $C_{\theta(t_1,\,t_2)}(s_1, s_2)$ is unknown, we assume that it is close to one of candidate copula models in term of the Kullback-Leibler distance. This assumption is practical in biomedical studies, because a suitable copula could be selected by evaluating the fitness of the copula model to the available data. Our objective is to estimate the RTPs as functions of (t_1, t_2) defined in (2.3) and (2.4) based on the time-varying copula (3.5) and a smoothing method.

At any given design time points $t_{(g)}\neq t_{(h)} \in \mathscr{K}$, we have, by (3.5), that $P_{A_1,\,A_2}(t_{(g)}, t_{(h)}) = C_{\theta(t_{(g)},\,t_{(h)})}\Big(S_{t_{(g)}}\big(y_1(t_{(g)})\big), S_{t_{(h)}}\big(y_2(t_{(h)})\big)\Big)$, which can be

estimated by substituting $S_{t_{(\cdot)}}(y)$ and $C_{\theta(t_{(g)},t_{(h)})}(\cdot,\cdot)$ with their corresponding consistent estimates. For the functional copula parameter $\theta(t_{(g)},t_{(h)})$, we consider two scenarios for its estimation: (a) $C_\theta(s_1,s_2)$ belongs to a known copula model that is specified by the unknown $\theta=\theta(\cdot)$; (b) both the copula model and the functional copula parameter are unknown, but an appropriate model for $C_\theta(s_1,s_2)$ can be selected from a set of candidate copula models.

3.2 Estimation with a Known Copula Model

When $C_\theta(s_1,s_2)$ belongs to a known copula model, we first estimate the univariate survival functions $S_{t_{(\cdot)}}(y)$ using the following empirical survival distribution of $Y(t_{(j)})>y$ for any design time $t_{(j)}$ and y,

$$\widetilde{S}_{t_{(j)}}(y)=\frac{1}{n_j}\sum_{i\in\mathscr{F}_j}1_{[Y_i(t_{(j)})>y]},\qquad j=g,h, \tag{3.6}$$

where $1_{[\ .\ ]}$ is an indicator function. For the copula parameter $\theta(t_{(g)},t_{(h)})$, we assume that $C_\theta(s_1,s_2)$ is differentiable with respect to s_1 and s_2, and define the derivative $c_\theta(s_1,s_2)=\partial^2 C_\theta(s_1,s_2)/\partial s_1\partial s_2$, where $\theta=\theta(t_{(g)},t_{(h)})$, $s_1=S_{t_{(g)}}\big(y(t_{(g)})\big)$ and $s_2=S_{t_{(h)}}\big(y(t_{(h)})\big)$. It then follows from (3.5) that we can define the following pseudo log-likelihood function for θ,

$$l_{g,h}(\theta\mid C)=\frac{1}{n_{g,h}}\sum_{k\in\left(\mathscr{F}_g\cap\mathscr{F}_h\right)}\log c_\theta\big(\widehat{S}_{kt_{(g)}},\widehat{S}_{kt_{(h)}}\big), \tag{3.7}$$

where $\widehat{S}_{kt_{(j)}}$ is the $n_j/(n_j+1)$ rescaled version of the empirical marginal survival function $\widetilde{S}_{t_{(j)}}(y)$ at $y = Y_k(t_{(j)})$ within the set $\mathscr{F}_j$, $j = g, h$. We use the above rescaling because it is necessary to avoid the potential unboundedness of $\log c_\theta(s_1, s_2)$ as s_1 or s_2 tend to one. Maximizing the pseudo log-likelihood $l_{g,h}(\theta \mid C)$ of (3.7) with respect to θ, the maximum likelihood estimator of $\theta(t_{(g)}, t_{(h)})$ is

$$\widehat{\theta}_C(t_{(g)}, t_{(h)}) = \arg\max_\theta l_{g,h}(\theta \mid C). \tag{3.8}$$

Numerical computation of (3.8) can be carried out using the procedure described in Genest et al. (1995).

Substituting $S_{t_{(g)}}\big(y_1(t_{(g)})\big)$, $S_{t_{(h)}}\big(y_2(t_{(h)})\big)$ and $\theta(t_{(g)}, t_{(h)})$ in (3.5) with their corresponding estimators (3.6) and (3.8), we obtain the following raw estimator of $P_{A_1,A_2}(t_{(g)}, t_{(h)})$ at any design time points $(t_{(g)}, t_{(h)})$,

$$\widetilde{P}_{A_1,\,A_2}\left(t_{(g)}, t_{(h)}\right) = C_{\widehat{\theta}_C(t_{(g)},\,t_{(h)})}\big(\widetilde{S}_{t_{(g)},1}, \widetilde{S}_{t_{(h)},2}\big), \tag{3.9}$$

where $\widetilde{S}_{t_{(g)},1} = \widetilde{S}_{t_{(g)}}\big(y_1\big(t_{(g)}\big)\big)$ and $\widetilde{S}_{t_{(h)},2} = \widetilde{S}_{t_{(h)}}\big(y_2\big(t_{(h)}\big)\big)$.

Specifically, when $y_1\big(t_{(g)}\big) = y_{\alpha_1}(t_{(g)})$ and $y_2\big(t_{(h)}\big) = y_{\alpha_2}(t_{(h)})$, we have $S_{t_{(g)}}\big(y_1\big(t_{(g)}\big)\big) = 1-\alpha_1$ and $S_{t_{(h)}}\big(y_2\big(t_{(h)}\big)\big) = 1-\alpha_2$. Thus, the joint probability $P_{A_1,\,A_2}\left(t_{(g)}, t_{(h)}\right) = P\{Y(t_{(g)}) > y_{\alpha_1}(t_{(g)}), Y(t_{(h)}) > y_{\alpha_2}(t_{(h)})\}$, denoted by $P_{\alpha_1,\,\alpha_2}\left(t_{(g)}, t_{(h)}\right)$, can be estimated by

$$\widetilde{P}_{\alpha_1,\,\alpha_2}\left(t_{(g)}, t_{(h)}\right) = C_{\widehat{\theta}_C(t_{(g)},\,t_{(h)})}(1-\alpha_1, 1-\alpha_2), \tag{3.10}$$

where α_1 and α_2 are given in (2.4).

Remark 3.1. Since $\widehat{S}_{kt_{(g)}}$, $\widehat{S}_{kt_{(h)}}$ and $\widehat{\theta}_C$ are calculated from datasets at different design time points $\mathscr{F}_g$, $\mathscr{F}_h$ and $\mathscr{F}_g \cap \mathscr{F}_h$, respectively, the proposed copula estimator of $P_{A_1,A_2}(t_{(g)}, t_{(h)})$ is less affected by the "unbalanced design" of the data, e.g., Wu and Tian (2018, Section 1.2.1), and perform better than the unstructured nonparametric estimator

$$\widetilde{P}^N_{A_1,A_2}(t_{(g)}, t_{(h)}) = \frac{1}{n_{g,h}} \sum_{i \in \left(\mathscr{F}_g \cap \mathscr{F}_h\right)} 1_{[Y_i(t_{(h)}) > y_1(t_{(g)}), Y_i(t_{(g)}) > y_2(t_{(h)})]}, \tag{3.11}$$

which is calculated without using the copula structure (3.5) and relies only on the observations in $\mathscr{F}_g \cap \mathscr{F}_h$. The advantage of (3.9) and (3.10) over (3.11) becomes obvious when $n_{g,h}$ is much smaller than $\min\{n_g, n_h\}$. In addition, for the quantile-based estimation of $P_{\alpha_1,\alpha_2}\left(t_{(g)}, t_{(h)}\right)$, the true percentile curves $y_{\alpha_1}(t)$ and $y_{\alpha_2}(t)$ in (3.11) are often unknown and need to be estimated, while the copula-based estimator (3.10) does not rely on the estimated quantile curves. In the simulation study, we use the true percentile curves for the nonparametric estimators, while the percentile curves in the application to the NGHS blood pressure data are estimated using the corresponding samples. □

3.3 Selection of Copula Models

In most biomedical studies, the exact copula model $C_{\theta(t_{(g)},t_{(h)})}(\cdot,\cdot)$ is unknown but may be selected from a set of copula models. A reasonable procedure for selecting an appropriate copula model that has been used in the literature is to maximize a pseudo likelihood function among all the candidate copula models, for detail see Joe (1997).

If $\mathcal{M}$ is the set of copula models, then the selected copula model is

$$C^*(\cdot) = \arg\max_{C\in\mathcal{M}} l_{g,h}(\widehat{\theta}_C|C) = \arg\max_{C\in\mathcal{M}} \left\{\max_{\theta} l_{g,h}(\theta|C)\right\}, \tag{3.12}$$

where $l_{g,h}(\theta|C)$ is defined by (3.7). Let $\widehat{\theta}_{C^*}(t_{(g)},t_{(h)})$ be the corresponding estimator derived from (3.8) under the selected copula model $C^*(\cdot)$. The approximated estimator of $P_{A_1,A_2}\left(t_{(g)},t_{(h)}\right)$ based on $C^*(\cdot)$ is

$$\widetilde{P}^*_{A_1,A_2}\left(t_{(g)},t_{(h)}\right) = C^*_{\widehat{\theta}_{C^*}(t_{(g)},t_{(h)})}\left(\widetilde{S}_{t_{(g)},1},\widetilde{S}_{t_{(h)},2}\right). \tag{3.13}$$

And the corresponding estimator of $P_{\alpha_1,\alpha_2}(t_{(g)},t_{(h)})$ is

$$\widetilde{P}^*_{\alpha_1,\alpha_2}(t_{(g)},t_{(h)}) = C^*_{\widehat{\theta}_{C^*}(t_{(g)},t_{(h)})}(1-\alpha_1,1-\alpha_2)\,. \tag{3.14}$$

Since the true joint probabilities may not necessarily belong to the selected copula model, we refer to (3.14) as an "approximated estimator" because the selected copula model may only be a reasonable approximation of the true copula model. The following remarks clarify some implications of relying on an approximated copula model in (3.14).

Remark 3.2. It follows from Joe (2014) that, if the Kullback-Leibler distance between the true copula model and the selected copula model is small, the difference between the joint probabilities derived from the true copula and the selected copula is also small. Thus, a good choice of candidate copula models $\mathcal{M}$ can lead to an appropriate estimator $\widetilde{P}^*_{A_1,A_2}(t_{(g)},t_{(h)})$ or $\widetilde{P}^*_{\alpha_1,\alpha_2}(t_{(g)},t_{(h)})$ that is close to $\widetilde{P}_{A_1,A_2}(t_{(g)},t_{(h)})$ or $\widetilde{P}_{\alpha_1,\alpha_2}(t_{(g)},t_{(h)})$ under the true copula. □

Remark 3.3. As discussed in Joe (2014), the similarity of copulas depends on the closeness of dependence in the tails. This suggests that the tail properties are useful for distinguishing distribution functions in copula model selection. In particular, the Frank copula is symmetric and has no tail dependence, while the Clayton and Gumbel copulas have strong lower and upper tail dependence, respectively. Since these three copula models can capture most of the dependence structures seen in real applications, they are widely used as candidate copula models in the literature. We demonstrate in the simulation study of Section 4 that the estimators based on the selected copula from these three candidate copula models give satisfactory performance in practice. □

3.4 Smoothing Estimators

For the smoothing step of the procedure, we use the raw estimates $\widetilde{P}_{A_1,\,A_2}(t_{(g)}, t_{(h)})$ at $t_{(g)} \neq t_{(h)} \in \mathcal{K}$ in (3.13) to estimate $P_{A_1,\,A_2}(t_1, t_2)$ at any time points $t_1 < t_2 \in \mathcal{T}$ through a kernel smoothing method. The smoothing estimator of $P_{A_1,\,A_2}(t_1, t_2)$ is then given by

$$\widehat{P}_{A_1,\,A_2}(t_1, t_2) = \sum_{g \neq h}^{J} W_{g,\,h}(t_1, t_2)\widetilde{P}_{A_1,\,A_2}(t_{(g)}, t_{(h)}), \tag{3.15}$$

where $W_{g,\,h}(t_1, t_2)$ is kernel-based weight function,

$$W_{g,\,h}(t_1, t_2) = \frac{n_{g,\,h}K\big((t_1 - t_{(g)})/h_1, (t_2 - t_{(h)})/h_2\big)}{\sum_{g \neq h} n_{g,\,h}K\big((t_1 - t_{(g)})/h_1, (t_2 - t_{(h)})/h_2\big)}, \tag{3.16}$$

$K(\cdot, \cdot)$ is a bivariate nonnegative kernel function, and h_1 and h_2 are the corresponding bandwidths.

Similarly, the kernel smoothing estimator of $P_{A_1,\,A_2}(t_1, t_2)$ based on the selected copula $C^*(\cdot)$ is given by

$$\widehat{P}^*_{A_1,\,A_2}(t_1, t_2) = \sum_{g \neq h}^{J} W_{g,\,h}(t_1, t_2)\widetilde{P}^*_{A_1,\,A_2}(t_{(g)}, t_{(h)}). \tag{3.17}$$

The smoothing estimator of marginal probability $P_{A_1}(t_1) = S_{t_1}(y_1(t_1))$, based on $\widetilde{S}_{t_{(k)}}(y_1(t_{(k)}))$, $k = 1, \cdots, K$, is given by

$$\widehat{S}_{t_1}(y_1(t_1)) = \sum_{k=1}^{K} W_k(t_1)\widetilde{S}_{t_{(k)}}(y_1(t_{(k)})), \tag{3.18}$$

where $W_k(t_1) = n_k K\big((t_l - t_{(k)})/h_0\big) \,/ \sum_{j=1}^{K} n_j K\big((t_1 - t_{(j)})/h_0\big)$, $K(\cdot)$ is a nonnegative kernel function, and h_0 is the corresponding bandwidth.

It follows from (3.15) and (3.18) that the kernel smoothing estimator of $RTP_{A_1,A_2}(t_1,t_2)$ in (2.3) is

$$\widehat{RTP}_{A_1,A_2}(t_1,t_2) = \widehat{P}_{A_1,A_2}(t_1,t_2)/\widehat{S}_{t_1}(y_1(t_1)) \tag{3.19}$$

when the true copula model $C(\cdot)$ is known, and

$$\widehat{RTP}^{*}_{A_1,A_2}(t_1,t_2) = \widehat{P}^{*}_{A_1,A_2}(t_1,t_2)/\widehat{S}_{t_1}(y_1(t_1)) \tag{3.20}$$

when the copula model $C^*(\cdot)$ is selected from the set of copula models $\mathcal{M}$. For the quantile-based RTP (2.4), since $P_{\alpha_i}(t_i) = P(Y(t_i) > y_{\alpha_i}(t_i)) = \alpha_i$ for each $t_i \in \mathcal{T}$, $i = 1, 2$, the kernel smoothing estimators (3.19) and (3.20) can be simplified as

$$\widehat{RTP}_{\alpha_1,\alpha_2}(t_1,t_2) = \widehat{P}_{\alpha_1,\alpha_2}(t_1,t_2)/\alpha_1 \tag{3.21}$$

and

$$\widehat{RTP}^{*}_{\alpha_1,\alpha_2}(t_1,t_2) = \widehat{P}^{*}_{\alpha_1,\alpha_2}(t_1,t_2)/\alpha_1, \tag{3.22}$$

respectively, where

$$\widehat{P}_{\alpha_1,\alpha_2}(t_1,t_2) = \sum_{g\neq h}^{J} W_{g,h}(t_1,t_2)\widetilde{P}_{\alpha_1,\alpha_2}(t_{(g)},t_{(h)}), \tag{3.23}$$

and

$$\widehat{P}^{*}_{\alpha_1,\alpha_2}(t_1,t_2) = \sum_{g\neq h}^{J} W_{g,h}(t_1,t_2)\widetilde{P}^{*}_{\alpha_1,\alpha_2}(t_{(g)},t_{(h)}). \tag{3.24}$$

Remark 3.4. It is well-known in the literature that, in practice, the appropriateness of kernel smoothing estimators is mostly affected by the bandwidth choices, while the kernel functions are relatively less important. Thus, a number of kernel functions may be used to compute (3.21) and (3.22). For simplicity, we use the product kernels $K(u_1, u_2) = K_1(u_1)K_2(u_2)$ with the same bandwidth $h_1 = h_2 = h$ for all the numerical computations in this article. Specifically, we use the Gaussian kernel $K_1(u_1) = (2\pi)^{-1/2}\exp\{-(u_1^2/2)\}$ for $\widehat{P}_{\alpha_1}(t_1)$ and the product Gaussian kernel $K(u_1, u_2) = (2\pi)^{-1}\exp\{-(u_1^2+u_2^2)/2\}$ for $\widetilde{P}^*_{\alpha_1,\alpha_2}(t_1, t_2)$. For data-driven bandwidth choices of h, we use the "leave-one-subject-out cross-validation" (LSCV) described in Wu and Tian (2018, Section 12.3.5). □

3.5 Asymptotic Properties

We present in this section the consistency and asymptotic normality of the raw estimators, consistency of the dynamic copula function estimators and the asymptotic mean squared risks of the kernel smoothing RTP estimators.

3.5.1 Asymptotic properties of Raw Probability Estimators

We first consider asymptotic properties of the raw estimator (3.9) and (3.10) when the copula model is known. Assume that the conditional joint survival

distribution function of $Y(t_1)$ and $Y(t_2)$ for any $t_1 < t_2 \in \mathcal{T}$ belongs to a known copula model, which satisfies the following assumptions:

C1. For all $t \in \mathcal{T}$, $S_t(y)$ is twice continuously differentiable with respect to t.

C2. For $l = 1, 2$, the partial derivatives $\partial \log c_\theta(s_1, s_2)/\partial\theta$, $\partial \log c_\theta(s_1, s_2)/\partial s_l$, $\partial^2 \log c_\theta(s_1, s_2)/\partial\theta^2$, $\partial^2 \log c_\theta(s_1, s_2)/\partial\theta\partial s_l$ and $\partial^3 \log c_\theta(s_1, s_2)/(\partial\theta^2\partial s_l)$ are all continuous and bounded for any $(s_1, s_2) \in (0, 1)^2$.

C3. For any fixed $t_1 < t_2 \in \mathcal{T}$, the functional Fisher information $I(\theta|t_1, t_2) = -\mathrm{E}_\theta \left\{ \frac{\partial \log c_\theta\left[S_{t_1}(Y(t_1)), S_{t_2}(Y(t_2))\right]}{\partial\theta} \right\}^2$ is finite and bounded away from zero, i.e. $0 < I(\theta|t_1, t_2) < \infty$. □

The following theorem shows that the raw estimator $\widetilde{P}_{A_1,A_2}\left(t_{(g)}, t_{(h)}\right)$ of (3.9) is consistent and $\widetilde{P}_{\alpha_1,\alpha_2}\left(t_{(g)}, t_{(h)}\right)$ of (3.10) is asymptotically normal when the number of subjects with observations at the design time points $(t_{(g)}, t_{(h)})$ is large.

Theorem 1. *If* C1 - C3 *hold, then, for any two design time points* $t_{(g)} < t_{(h)} \in \mathscr{K}$*, we have that, as* $n_{g,h} \to \infty$, $\widetilde{P}_{A_1,A_2}\left(t_{(g)}, t_{(h)}\right) \xrightarrow{p} P_{A_1,A_2}\left(t_{(g)}, t_{(h)}\right)$ *and* $\sqrt{n_{g,h}}\left\{\widetilde{P}_{\alpha_1,\alpha_2}\left(t_{(g)}, t_{(h)}\right) - P_{\alpha_1,\alpha_2}\left(t_{(g)}, t_{(h)}\right)\right\} \xrightarrow{L} N\left(0, \sigma_{g,h}\right)$, *where* $\sigma_{g,h}$ *is the asymptotical variance.* □

Proof of the theorem and explicit expression of $\sigma_{g,h}$ are given in the online supplemental material.

3.5.2 Consistency of Dynamic Copula Estimators

When the copula model $C(\cdot)$ is unknown, we show that, under the best copula model $C^*(\cdot)$ from the collection of copula models $\mathcal{M}$, the approximated raw estimator (3.14) is consistent when the number of subjects with observations at the design time points $(t_{(g)}, t_{(h)})$ is large.

First, by the derivations in Joe (2014), we have the following lemma, which shows that, when the selected copula $C^*(\cdot)$ is "close" to the true copula $C(\cdot)$, the estimated functional copula parameter converges to the functional copula parameter under $C^*(\cdot)$.

Lemma 1. *If $c_\theta(s_1, s_2)$ and $c^*_{\theta^*}(s_1, s_2)$ are the densities of the true copula $C(\cdot)$ and the selected copula $C^*(\cdot)$ at time points $(t_{(g)}, t_{(h)})$, respectively, then $\widehat{\theta}_{C^*}(t_{(g)}, t_{(h)}) \xrightarrow{P} \theta_{C^*}(t_{(g)}, t_{(h)})$ in probability as $n \to \infty$, where $\theta_{C^*}(t_{(g)}, t_{(h)}) = \underset{\theta^*}{\operatorname{argmax}} \iint c_\theta(s_1, s_2) \log c^*_{\theta^*}(s_1, s_2)\, \mathrm{d}s_1\, \mathrm{d}s_2$, that is, $\theta_{C^*}(t_{(g)}, t_{(h)})$ is the copula parameter such that $c^*_{\theta_{C^*}(t_{(g)}, t_{(h)})}(s_1, s_2)$ is the closest copula density to $c_\theta(s_1, s_2)$ among all copula densities of the form $c^*_{\theta^*}(s_1, s_2)$ in the Kullback-Leibler divergence.* □

Let $P^*_{A_1,A_2}(t_{(g)}, t_{(h)}) = C^*_{\theta_{C^*}}(S_{t_{(g)}}(y_1(t_{(g)})), S_{t_{(h)}}(y_2(t_{(h)})))$ and $P^*_{\alpha_1,\alpha_2}(t_{(g)},$

$t_{(h)}) = C^*_{\theta_{C^*(t_{(g)},t_{(h)})}}(1-\alpha_1, 1-\alpha_2)$. Using the continuous mapping theorem, the following theorem shows that, under the selected copula model, the raw approximated estimator (3.14) is a consistent estimator of $P^*_{\alpha_1,\alpha_2}(t_{(g)}, t_{(h)})$ at the design time points $(t_{(g)}, t_{(h)})$.

Theorem 2. *If the selected copula model $C^*(\cdot)$ satisfies the assumptions* C1-C3*, then, for any two design time points $t_{(g)} < t_{(h)} \in \mathscr{K}$, as $n_{g,h} \to \infty$, $\widetilde{P}^*_{A_1,A_2}\left(t_{(g)}, t_{(h)}\right) \xrightarrow{P} P^*_{A_1,A_2}\left(t_{(g)}, t_{(h)}\right)$ and $\widetilde{P}^*_{\alpha_1,\alpha_2}(t_{(g)}, t_{(h)}) \xrightarrow{p} P^*_{\alpha_1,\alpha_2}(t_{(g)}, t_{(h)})$.* □

This consistency result suggests that, even if the true copula model is unknown, the raw approximated estimator (3.13) (or (3.14)) could still be close to the target conditional probability $P^*_{A_1,A_2}\left(t_{(g)}, t_{(h)}\right)$ (or $P^*_{\alpha_1,\alpha_2}(t_{(g)}, t_{(h)})$) when the number of observations in $\mathscr{K}$ is large.

3.5.3 Asymptotic Risk of the Smoothing RTP Estimators

We now derive the asymptotic mean squared errors of the kernel smoothing RTP estimators (3.21) and (3.22) under the following assumptions:

C4. For all $t_1, t_2 \in \mathcal{T}$, the positive density function $f(t_1, t_2)$ is continuously differentiable with respect to t_1 and t_2.

C5. Let $\boldsymbol{u} = (u_1, u_2)$. The bivariate kernel $K(u_1, u_2) = K_1(u_1)K_2(u_2)$ is a symmetric probability density on a bounded set assumed to be

$[-1,1]^2$. It satisfies $\int \boldsymbol{u}K(\boldsymbol{u})d\boldsymbol{u} = \mathbf{0}$, $R(K) = \int K^2(\boldsymbol{u})d\boldsymbol{u}$, $\int u_1^2 K_1(u_1)du_1 = \mu_{(21)}$ and $\int u_2^2 K_2(u_2)du_2 = \mu_{(22)}$ for some positive constants $\mu_{(21)}$ and $\mu_{(22)}$.

C6. h satisfies that $h \to 0$ and $nh^2 \to \infty$ as $n \to \infty$.

C7. For all $i = 1, \ldots, n$ and $j_1 \neq j_2$, $|t_{ij_1} - t_{ij_2}| > h$. □

The following theorem shows the consistency of the proposed estimator (3.19) and the asymptotic expressions of the bias and variance of the estimator (3.21).

Theorem 3. *If the number of subjects n is large, $t_1 < t_2$ are interior points within the support of $f(\cdot,\cdot)$, and the assumptions* C1-C7 *are satisfied, then*

$$\widehat{RTP}_{A_1,A_2}(t_1,t_2) \xrightarrow{P} RTP_{A_1,A_2}(t_1,t_2) \quad \text{as } n \to \infty. \tag{3.25}$$

Specifically, $\widehat{RTP}_{\alpha_1,\alpha_2}(t_1,t_2) \xrightarrow{P} RTP_{\alpha_1,\alpha_2}(t_1,t_2)$ as $n \to \infty$, and the asymptotic bias and variance of $\widehat{RTP}_{\alpha_1,\alpha_2}(t_1,t_2)$ in (3.21) *are, respectively,*

$$\begin{aligned}
&Bias\left[\widehat{RTP}_{\alpha_1,\alpha_2}(t_1,t_2)\right] \\
&= \left[\mu_{(21)}\left(\frac{\partial P_{\alpha_1,\alpha_2}(t_1,t_2)}{\partial t_1}\frac{\partial f(t_1,t_2)}{\partial t_1} + \frac{1}{2}\frac{\partial^2 P_{\alpha_1,\alpha_2}(t_1,t_2)}{\partial t_1^2}f(t_1,t_2)\right)\right. \\
&\quad \left.+\mu_{(22)}\left(\frac{\partial P_{\alpha_1,\alpha_2}(t_1,t_2)}{\partial t_2}\frac{\partial f(t_1,t_2)}{\partial t_2} + \frac{1}{2}\frac{\partial^2 P_{\alpha_1,\alpha_2}(t_1,t_2)}{\partial t_2^2}f(t_1,t_2)\right)\right] \\
&\quad \times \alpha^{-1} f(t_1,t_2)^{-1} h^2\big(1+o(1)\big)
\end{aligned}$$

and $Var\Big[\widehat{RTP}_{\alpha_1,\alpha_2}(t_1,t_2)\Big] = \alpha^{-2}\, f(t_1,t_2)^{-3}\, N^{-1}h^{-2}\, R(K)\,\big(1+o(1)\big)$, *where* $R(K) = \int K^2(\boldsymbol{u})d\boldsymbol{u}$. □

The next theorem shows the convergency of the proposed estimator (3.20) and the asymptotic expressions of the bias and variance of the proposed estimator (3.22) under the selected copula model.

Theorem 4. *If the selected copula model* $C^*(\cdot)$ *satisfies the assumptions* $C1-C4$ *and the assumptions* $C5-C7$ *hold, then*

$$\widehat{RTP}^{*}_{A_1,A_2}(t_1,t_2) \xrightarrow{P} RTP^{*}_{A_1,A_2}(t_1,t_2) \;\; as\; n\to\infty.$$

Specifically, $\widehat{RTP}^{*}_{\alpha_1,\alpha_2}(t_1,t_2) \xrightarrow{P} RTP^{*}_{\alpha_1,\alpha_2}(t_1,t_2)$ *as* $n\to\infty$, *and the asymptotic bias and variance of* $\widehat{RTP}^{*}_{\alpha_1,\alpha_2}(t_1,t_2)$ *in* (3.22) *are, respectively,*

$$\begin{aligned} &Bias\Big[\widehat{RTP}^{*}_{\alpha_1,\alpha_2}(t_1,t_2)\Big] \\ &= \Big[\mu_{(21)}\Big(\frac{\partial P^{*}_{\alpha_1,\alpha_2}(t_1,t_2)}{\partial t_1}\frac{\partial f(t_1,t_2)}{\partial t_1} + \frac{1}{2}\frac{\partial^2 P^{*}_{\alpha_1,\alpha_2}(t_1,t_2)}{\partial t_1^2} f(t_1,t_2)\Big) \\ &\quad +\mu_{(22)}\Big(\frac{\partial P^{*}_{\alpha_1,\alpha_2}(t_1,t_2)}{\partial t_2}\,\frac{\partial f(t_1,t_2)}{\partial t_2} + \frac{1}{2}\frac{\partial^2 P^{*}_{\alpha_1,\alpha_2}(t_1,t_2)}{\partial t_2^2} f(t_1,t_2)\Big)\Big] \\ &\quad \times\alpha_1^{-1}\, f(t_1,t_2)^{-1}\, h^2\big(1+o(1)\big) + \alpha_1^{-1}\, b(t_1,t_2) \end{aligned}$$

and $Var\Big[\widehat{RTP}^{*}_{\alpha_1,\alpha_2}(t_1,t_2)\Big] = \alpha_1^{-2}\, f(t_1,t_2)^{-3}\, N^{-1}h^{-2}\, R\big(K\big)\,\big(1+o(1)\big)$, *where* $b(t_1,t_2) = P^{*}_{\alpha_1,\alpha_2}(t_1,t_2) - P_{\alpha_1,\alpha_2}(t_1,t_2)$. □

Remark 3.5. It follows from Theorems 3 and 4 that, if the true copula model is known or $P^{*}_{A_1,A_2}(t_1,t_2) = P_{A_1,A_2}(t_1,t_2)$, the copula-based kernel

smoothing estimator (3.19) or (3.20) of the RTP is asymptotically unbiased. In fact, the proposed estimator (3.20) could be close to $RTP_{A_1,A_2}(t_1,t_2)$ only if the Kullback-Leibler distance between the true copula model and one of the candidate copula models is small. In addition, $\widehat{RTP}_{\alpha_1,\alpha_2}(t_1,t_2)$ and $\widehat{RTP}^*_{\alpha_1,\alpha_2}(t_1,t_2)$ have the same asymptotic variance, while the asymptotic biases of (3.21) and (3.22) differ by a fraction of $b(t_1,t_2)$, which depends on the difference between $P^*_{\alpha_1,\alpha_2}(t_1,t_2)$ and $P_{\alpha_1,\alpha_2}(t_1,t_2)$. □

4. Simulation Study

We conduct a simulation study to investigate the finite sample properties of the proposed copula-based smoothing estimators and compare them with the unstructured nonparametric smoothing estimators proposed by Wu and Tian (2013). Let $\{t_{(1)},\cdots,t_{(40)}\}=\{0.25,0.5,\cdots,10\}$ be the design time points. We generate 1000 subjects with 10 visits per subject in each sample. The jth visit time of the ith subject t_{ij} is corresponding to $t_{(k_{ij})}$ in the set of design time points. For each subject $i=1,\ldots,1000$ at the jth time point t_{ij}, we generate the observation $Y_{ij}=Y_i(t_{ij})$ from

$$Y_{ij}=21.5+0.7\big(t_{ij}-5\big)-0.05\big(t_{ij}-5\big)^2+\epsilon_{ij},\quad j=1,\cdots,n_i=40, \tag{4.26}$$

where $\epsilon_{ij}=z_{ik_{ij}}$ is the k_{ij}th element of $(z_{i1},\cdots,z_{i40})$, which are generated from the 40-dimensional t-copula $C^t_{\mathbf{R},v}(u_1,\cdots,u_{40})$ with the dispersion

structure

$$\mathbf{R} = \begin{pmatrix} 1 & \rho & \cdots & \rho^{39} \\ \rho & 1 & \cdots & \rho^{38} \\ \vdots & \vdots & & \vdots \\ \rho^{39} & \rho^{38} & \cdots & 1 \end{pmatrix}$$

and $v = 4$ degrees of freedom, and k_{ij} is generated by the ceiling of a random number from the uniform distribution $U[4(j-1), 4j]$ for $j = 1, 2, \cdots, 10$.

We consider three most commonly used Archimedean copulas (Joe, 2014), namely the Frank, Clayton and Gumbel copulas, as the choice of candidate copula models. In our simulation, the copula model used in the smoothing estimators is not necessarily from the true t-copula model but an Archimedean copula that is "closest"

to the true t-copula from which the data are generated. The simulation has 1000 replications.

We first consider the quantile-based RTPs. Let $A_1(t_1) = (y_{\alpha_1}(t_1), \infty)$ and $A_2(t_2) = (y_{\alpha_2}(t_2), \infty)$ with $\alpha_1 = \alpha_2 = 0.8$. We calculate the empirical biases and root mean squared errors (RMSE) for each estimator (3.22) of $RTP_{0.8,\,0.8}(t_2 - 3, t_2)$ and $RTP_{0.8,\,0.8}(3, t_2)$ at a sequence of t_2 values, $t_1 = t_2 - 3$ and $t_1 = 3$. Here $RTP_{0.8,\,0.8}(t_2 - 3, t_2)$ and $RTP_{0.8,\,0.8}(3, t_2)$ represent the "three-year tracking ability" and the "first $(t_2 - 3)$-year tracking ability" for the simulated samples, respectively. For comparison, we also present the

Table 1: Empirical biases and RMSEs of the copula-based smoothing estimates and the unstructured nonparametric smoothing estimates obtained from 1000 simulation replications based on $A_1(t_1) = (y_{\alpha_1}(t_1), \infty)$ and $A_2(t_2) = (y_{\alpha_2}(t_2), \infty)$ with $\alpha_1 = \alpha_2 = 0.8$.

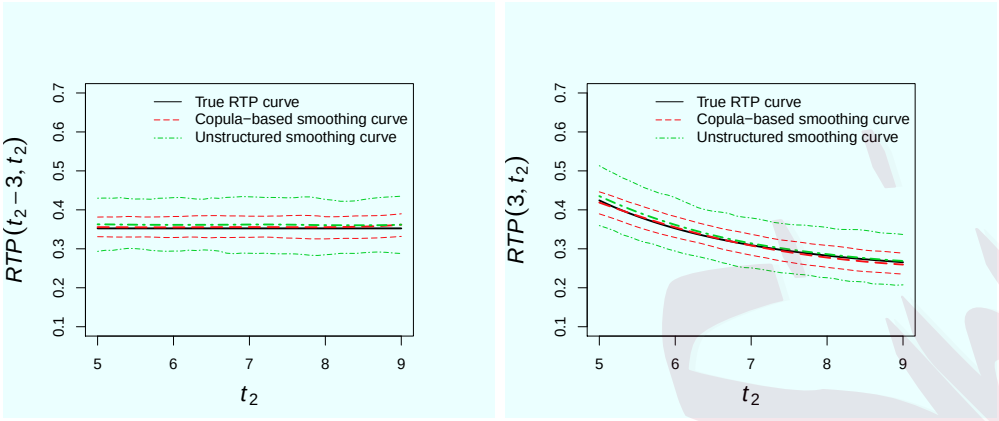
RTP	t_2	True	Copula Smoothing		Unstructured Smoothing	
			Bias	RMSE	Bias	RMSE
$RTP_{0.8,0.8}(t_2-3, t_2)$	5	0.352	0.003	0.014	0.009	0.038
	6	0.352	0.003	0.014	0.008	0.038
	7	0.352	0.003	0.015	0.009	0.039
	8	0.352	0.003	0.015	0.007	0.040
	9	0.352	0.010	0.018	0.007	0.040
$RTP_{0.8,0.8}(3, t_2)$	5	0.424	-0.005	0.015	0.011	0.043
	6	0.352	0.003	0.014	0.008	0.038
	7	0.309	-0.001	0.014	0.004	0.035
	8	0.283	-0.005	0.015	0.003	0.036
	9	0.266	-0.006	0.015	0.003	0.034

true RTP values and compute the empirical biases and RMSEs from the unstructured smoothing estimates.

Table 1 shows the the true RTP values and the empirical biases and RMSEs obtained from the copula-based and unstructured smoothing esti-

mates at several selected time points. The entries of Table 1 demonstrate that both the copula-based and unstructured smoothing estimates have small biases. But the copula-based smoothing estimates are superior to the unstructured smoothing estimates in the sense that they have smaller RMSEs for all the time points in Table 1. This superiority of the copula-based smoothing estimates even holds under the scenario that the candidate Archimedean copula models do not include the true t-copula model and the unstructured smoothing estimates were calculated using the true percentile curve $y_{0.8}(t)$.

Figure 1 shows the averages of the estimated $RTP_{0.8,\,0.8}(t_2-3, t_2)$ in Figure 1(a) and $RTP_{0.8,\,0.8}(3, t_2)$ in Figure 1(b) and their corresponding lower and upper 2.5th percentiles computed from the 1000 simulated samples using the copula-based smoothing estimator and the unstructured smoothing estimator. The plots in the figure demonstrate that the averages of the estimated curves from both estimators are close to the true RTP curves. But the widths between the lower and upper 2.5th percentile curves of the copula-based smoothing estimates are much smaller than that of the unstructured smoothing estimates. The narrower widths of the percentile curves suggest that the copula-based smoothing estimator has lower variability than the unstructured smoothing estimator. This result is consistent



(a) $RTP_{0.8,\,0.8}(t_2-3,t_2)$ (b) $RTP_{0.8,\,0.8}(3,t_2)$

Figure 1: The true RTP curves, the averages of the estimated $RTP_{\alpha_1,\alpha_2}(t_1,t_2)$ with $\alpha_1 = \alpha_2 = 0.8$ using the copula-based smoothing estimator and the unstructured smoothing estimator, and the lower and upper 2.5th percentiles computed from the 1000 simulated samples generated from (4.26).

with the empirical RMSE results in Table 1.

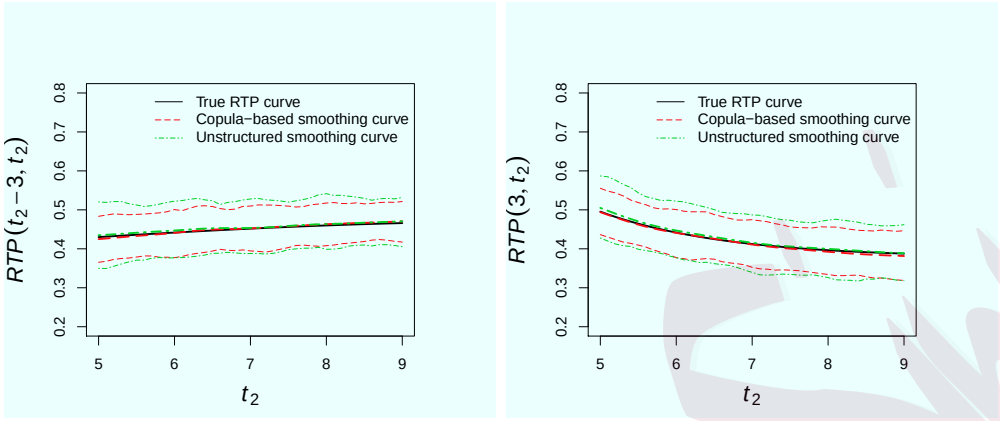
We next consider the threshold-based RTPs. Let $A_1(t_1) = (28,\infty)$ and $A_2(t_2) = (28,\infty)$. The corresponding empirical biases, RMSEs, averages and percentiles of the estimator (3.20) and the unstructured smoothing estimator are summarized in Table 2 and Figure 2. These results again suggest that the copula-based estimator (3.20) leads to similar empirical biases but smaller RMSEs and variabilities compared with the correspond-

Table 2: Empirical biases and RMSEs of the copula-based smoothing estimates and the unstructured nonparametric smoothing estimates obtained from 1000 simulation replications based on $A_1(t_1) = (28, \infty)$ and $A_2(t_2) = (28, \infty)$.

RTP	t_2	True	Copula Smoothing		Unstructured Smoothing	
			Bias	RMSE	Bias	RMSE
$RTP_{A_1,A_2}(t_2-3,t_2)$	5	0.430	-0.005	0.032	0.005	0.043
	6	0.442	-0.001	0.031	0.005	0.039
	7	0.452	0.001	0.029	0.002	0.038
	8	0.460	0.003	0.028	0.003	0.035
	9	0.466	0.004	0.027	0.005	0.033
$RTP_{A_1,A_2}(3,t_2)$	5	0.495	-0.001	0.032	0.010	0.041
	6	0.442	-0.001	0.031	0.005	0.039
	7	0.412	-0.001	0.031	0.003	0.038
	8	0.396	-0.004	0.032	0.003	0.039
	9	0.388	-0.006	0.032	0.001	0.037

ing unstructured smoothing estimator.

The results of Tables 1 and 2 and Figures 1 and 2 demonstrate that the copula-based smoothing estimator is superior to the unstructured smoothing estimator for both scenarios of estimating the quantile-based RTPs and



(a) $RTP_{A_1,A_2}(t_2-3,t_2)$ (b) $RTP_{A_1,A_2}(3,t_2)$

Figure 2: The true RTP curves, the averages of the estimated $RTP_{A_1,A_2}(t_1,t_2)$ with $A_1(t)=(28,\infty), A_2(t)=(28,\infty)$ using the copula-based smoothing estimator and the unstructured smoothing estimator, and the lower and upper 2.5th percentiles computed from the 1000 simulated samples generated from (4.26).

the threshold-based RTPs. The only difference between these two scenarios is that the marginal probabilities $S(t_1)$ and $S(t_2)$ are equal to $1-\alpha_1$ and $1-\alpha_2$ in (3.22), while they need to be estimated in (3.20).

5. Application to NGHS Blood Pressure Data

We apply our estimation method to the NGHS blood pressure (BP) data. This dataset has been previously described and analyzed by Wu and Tian

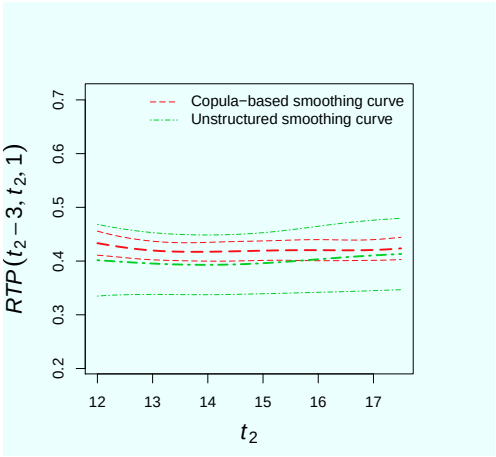
(2013b) and Wu and Tian (2018, Chapter 12) using the unstructured non-parametric smoothing estimator. Given that an important objective of the NGHS is to evaluate the racial differences of BP distributions, we estimate the RTPs for Caucasian girls and African American girls separtely.

Since age (in years) in the dataset is rounded up to two decimal places, there are a large number of distinct time points and the numbers of subjects observed at each of these distinct time point are small. A practical approach that is clinically meaningful and has been used in the literature is "data binning" (e.g. Wu and Tian, 2018, Section 12.2), which pools the observations at the adjacent time points to create a set $\mathscr{K}$ of design time points that are clinically interpretable. We consider the age range $T = [9.00, 19.00)$ and specify 4 design time points at each age, which leads to a total of $K = 40$ equally spaced age bins $[9.00, 9.25), ..., [18.75, 19.00)$ corresponding to the design time points $\mathscr{K} = \{9.00, 9.25, ..., 18.75\}$. If the ith girl is observed within the age bin $[t_{(j)}, t_{(j+1)})$, her corresponding design time point is $t_{(j)}$. This age binning has adequate clinical interpretation for pre-teens and adolescents since two girls who were born within three months have approximately the same age.

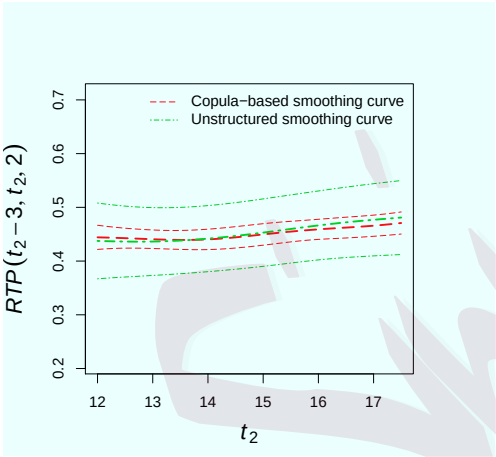
Let $Y(t)$ be a girl's systolic blood pressure (SBP) at time point t years of age, and let $A_1(t_1) = (y_{0.8}(t_1), \infty)$ and $A_2(t_2) = (y_{0.8}(t_2), \infty)$ be the 80th

percentile SBP ranges at ages t_1 and t_2 years. We estimate the rank tracking probabilities $RTP_{A_1,A_2}(t_2-3,t_2)$ and $RTP_{A_1,A_2}(t_1,t_2)$ at a sequence of t_2 values with $t_1 = t_2 - 3$ and $t_1 = 10$ using the copula-based smoothing method and the unstructured smoothing method for Caucasian and African American girls, respectively. These RTPs give quantitative measures for the chance of a girl's SBP is above the 80th percentile at age t_2 years given that her SBP is known to be above the 80th percentile at age t_1 years. Further details about clinical interpretations of $RTP_{\cdot,\cdot}(t_2-3,t_2)$ and $RTP_{\cdot,\cdot}(t_1,t_2)$ for epidemiology studies can be found in Wu and Tian (2018, Chapter 12). We compute both the copula-based and unstructured smoothing estimates and their corresponding bootstrap 95% empirical quantile point-wise confidence intervals using resampling-subject bootstrap with 500 replications, e.g. Wu and Tian (2018, Section 12.3.6).

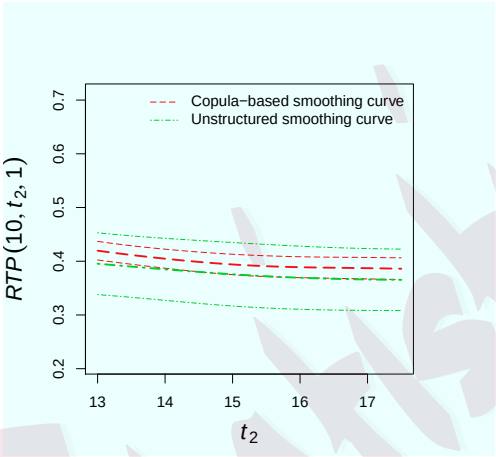
Figure 3 shows the estimates of $RTP_{0.8,0.8}(t_1,t_2)$ and their bootstrap 95% point-wise confidence intervals for Caucasian and African American girls over different age ranges. The top panels of Figure 3 show the estimated probabilities that, given a Caucasian (Figure 3a) or an African American (Figure 3b) girl's SBP was higher than the age-specific 80th percentiles at ages 9 to 14.5 years, her SBP is also higher than the age-specific 80th percentiles at three years later. The bottom panels of Figure 3 show the



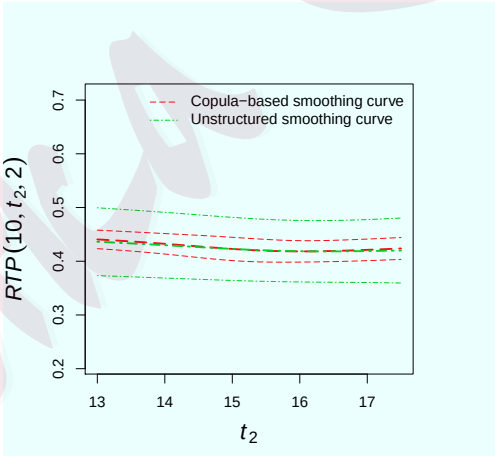
(a) Caucasian Girls



(b) African American Girls



(c) Caucasian Girls



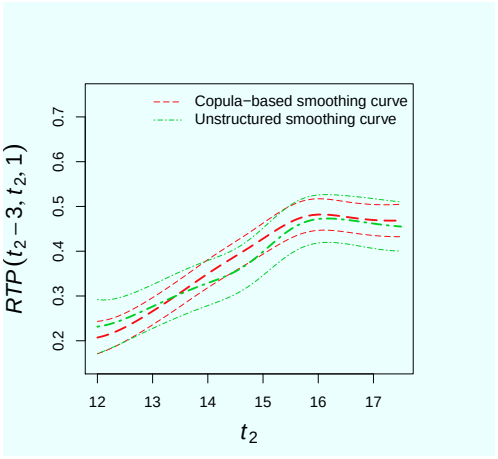
(d) African American Girls

Figure 3: The estimated $RTP_{\alpha_1,\alpha_2}(t_1, t_2)$ curves with $\alpha_1 = \alpha_2 = 0.8$ using the copula-based smoothing estimator and the unstructured smoothing estimator, and the corresponding bootstrap 95% empirical quantile point-wise confidence intervals.

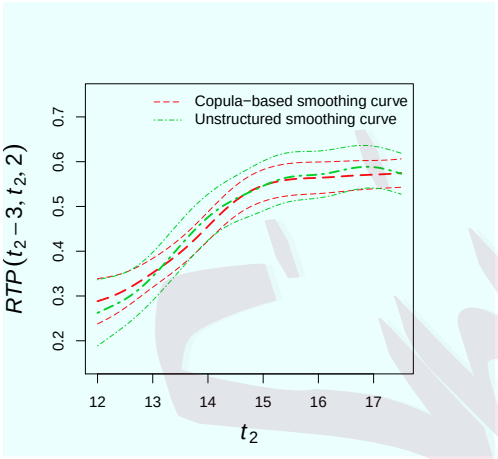
estimated probabilities that a Caucasian (Figure 3c) or an African American (Figure 3c) girl's SBP is higher than the age-specific 80th percentiles at ages 13 to 17.5 years given that her SBP was already higher than the 80th percentile at age 10 years. All four panels of Figure 3 suggests that the rank-tracking probabilities for both Caucasian and African American girls vary approximately around the range of 40% to 45% for different age ranges, which are much higher than the anticipated value of 20% if there were no tracking ability for the SBP of this population. These results, which are consistent with the finding in Wu and Tian (2013), indicates that the percentile SBP values for a girl at different ages have positive tracking ability, hence, are positively correlated.

Figure 4 shows the estimated $RTP_{A_1,A_2}(t_1,t_2)$ with $A_1(t_1) = (115,\infty)$ and $A_2(t_2) = (115,\infty)$ at a sequence of t_2 values with $t_1 = t_2 - 3$ and $t_1 = 10$ using the copula-based smoothing method and the unstructured smoothing method. Since $A_1(t_1)$ and $A_2(t_2)$ depends on the fixed SBP level of 115 mmHg, the estimated RTP curves in Figure 4a to Figure 4d are all increasing with age, suggesting that the girls' SBP levels are increasing with age. These finding are consistent with the ones observed in Figure 3.

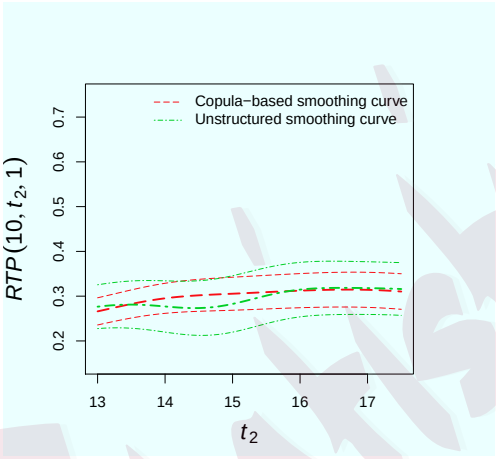
Comparing the bootstrap 95% point-wise confidence intervals in Figures 3 and 4, we observe that, for all the panels in Figure 3a to Figure



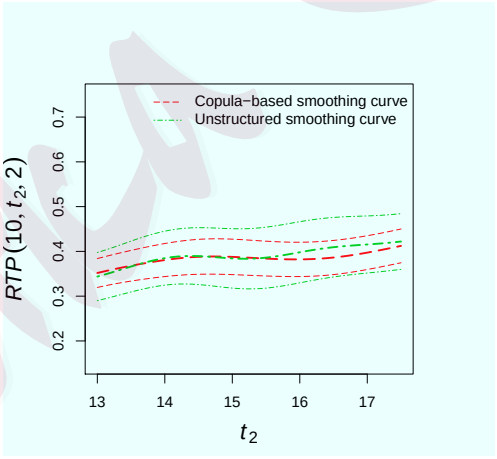
(a) Caucasian Girls



(b) African American Girls



(c) Caucasian Girls



(d) African American Girls

Figure 4: The estimated $RTP_{A_1, A_2}(10, t_2)$ curves with $A_1(t) = A_2(t) = (115, \infty)$ using the copula-based smoothing estimator and the unstructured smoothing estimator, and the corresponding bootstrap 95% empirical quantile point-wise confidence intervals.

3d and Figure 4a to Figure 4d, the widths of the confidence intervals for the copula-based smoothing estimates are narrower than the ones for the unstructured smoothing estimates. These results are similar to the simulation results summarized in Tables 1 and 2 and Figures 1 and 2, where it was demonstrated that the copula-based smoothing estimator generally has smaller standard errors than the unstructured smoothing estimator.

6. Discussion

The RTP has been shown to be a useful measure of tracking abilities of time-varying disease status and risk factors in longitudinal studies. We developed in this article a copula-based smoothing method for the estimation of the RTPs in longitudinal studies either without covariates or with time-invariant categorical covariates. Our theoretical and simulation results demonstrate that this copula-based smoothing method has major advantages over the unstructured smoothing method that is currently used for this situation. The proposed smoothing method consists of two steps: (a) computing the raw estimates at "design time points" based on a known copula model or a set of candidate copula models; (b) obtaining the functional RTP estimates at any time point by smoothing the raw estimates using a kernel smoothing method. We provide theoretical justifications for the

proposed copula-based smoothing method by deriving the asymptotic mean squared errors of the estimators, and demonstrate its finite sample superiority over the unstructured smoothing method through a simulation study under the robust scenario that the copula model is not known but is selected from a set of candidate copula models. Our application to the NGHS SBP data also demonstrates that this copula-based smoothing method leads to clinically meaningful results in epidemiological studies.

A main limitation of the proposed estimation method is that it applies only to the situation that either there is no covariate or the covariates are time-invariant and categorical. When it is necessary to include time-varying or continuous covariates, additional modeling assumptions are required for the copula models so that their dependence structures on the covariates can be specified. Further research is warranted to establish a flexible and clinically meaningful copula model to include time-varying and continuous covariates. In addition, since the RTPs are often sensitive to the tail dependence structures, the copula models considered in this article may not be enough to handle all the possible situations in practice. Therefore, it is preferable to extend our method to include more flexible models for the distribution functions. Examples of such an extension may include mixtures of copulas and vine copulas. Finally, the RTPs considered in this paper and

Wu and Tian (2013, 2018) are defined at two different time points. Given that there is practical interest to study pediatric blood pressure distributions at more than three time points (e.g. NHBPEP, 2004), other statistical models for RTPs at three or more time points also deserve substantial future research.

Supplementary Material

The online Supplementary Material provides detailed derivations of proofs for Theorems 1-4.

Acknowledgements

The NGHS data is available upon request from the BioLINCC website (https://biolincc.nhlbi.nih.gov/studies/nghs/) of the National Heart, Lung and Blood Institute. The authors thank the investigators and participants of the NGHS study. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the NHLBI, the National Institutes of Health, or the US Department of Health and Human Services. Research of X. Y. Zhang and M. X. Wu is supported by National Natural Science Foundation of China (No: 11771032). The authors are grateful to the editors and the two referees for their detailed suggestions

which considerably improved the quality of the paper.

References

Chen, S. X., and Huang, T.M. (2007). Nonparametric Estimation of Copula Functions for Dependence Modelling. *Canadian Journal of Statistics* **35**, 265-282.

Diggle, P. J., P. Heagerty, K. Y. Liang, and S. L. Zeger. (2002). *Analysis of Longitudinal Data*, 2nd edition. Oxford, UK: Oxford University Press.

Fan, J., Huang, T. and Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of the American Statistical Association* **102**, 632-641.

Fermanian, J. D. and Scaillet, O. (2003). Nonparametric estimation of copulas for time series. *Journal of Risk* **5**, 25-54.

Genest, C., Ghoudi, K., Rivest, L. P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* **82**, 543-552.

Hall, P., Racine, J. S., and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association* **99**, 1015-1026.

Hall, P., Wolff, R. C. L., and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association* **94**, 154-163.

Härdle, W. K., Muller, M., Sperlich, S. and Werwatz, A. (2012). *Nonparametric and Semi-parametric Models*, Springer Science & Business Media.

Hoover, D. R., Rice, J., Wu, C. O., and Yang, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809-822.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*, Volume 73. London: Chapman & Hall.

Joe, H. (2014). *Dependence modeling with copulas*, CRC press.

National High Blood Pressure Education Program Working Group on High Blood Pressure in Children and Adolescents (NHBPEP Working Group) (2004). The fourth report on the diagnosis, evaluation, and treatment of high blood pressure in children and adolescents. *Pediatrics* **114**, 555-576.

National Heart, Lung, and Blood Institute Growth and Health Research Group (NGHSRG) (1992). Obesity and Cardiovascular Disease Risk Factors in Black and White Girls: The NHLBI Growth and Health Study. *American Journal of Public Health* **82**, 1613-1620.

Obarzanek, E., Wu, C. O., Cutler, J. A., Kavey, R.-E. W., Pearson, G. D. and Daniels, S. R. (2010). Prevalence and incidence of hypertension in adolescent girls. *The Journal of Pediatrics* **157**, 461-467.

Rice, J., and Wu, C. O. (2001). Nonparametric Mixed Effects Models for Unequally Sampled Noisy Curves. *Biometrics* **57**, 253-259.

Sklar, A. (1959). Fonctions de Réprtition à n dimensions et leurs marges. *Publications de I'Institut Statistique de I'Université de Paris* **8**, 229-231.

Wu, C. O. and Tian, X. (2013). Nonparametric estimation of conditional distribution functions and rank-tracking probabilities with longitudinal data. *Journal of Statistical Theory and Practice* **7**, 259-284.

Wu, C. O. and Tian, X. (2018). *Nonparametric Models for Longitudinal Data: With Implementation in R*. Chapman and Hall/CRC.

College of Statistics and Data Science, Faculty of Science, Beijing University of Technology, Beijing 100124, China

E-mail: (zhangxiaoyu006@126.com)

College of Statistics and Data Science, Faculty of Science, Beijing University of Technology, and Beijing Institute for Scientific and Engineer Computing, Beijing 100124, China

E-mail: (wumixia@bjut.edu.cn)

Office of Biostatistics Research, Division of Intramural Research, National Heart, Lung and Blood Institute, Bethesda, Maryland 20892, USA

E-mail: (wuc@nhlbi.nih.gov)