Journal of Multivariate Analysis 173 (2019) 85–109

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Calibration estimation of semiparametric copula models with data missing at random

Shigeyuki Hamori [a], Kaiji Motegi [a,*], Zheng Zhang [b]

[a] Graduate School of Economics, Kobe University. Kobe, Hyogo 657-8501, Japan
[b] Institute of Statistics and Big Data, Renmin University of China. Haidian District, Beijing 100080, China

ARTICLE INFO

Article history:
Received 28 March 2018
Received in revised form 7 February 2019
Accepted 7 February 2019
Available online 15 February 2019

AMS 2010 subject classifications:
primary 62H12
secondary 62G05

Keywords:
Calibration estimation
Covariate balancing
Missing at random (MAR)
Semiparametric copula model

ABSTRACT

This paper investigates the estimation of semiparametric copula models with data missing at random. The maximum pseudo-likelihood estimation of Genest et al. (1995) is infeasible if there are missing data. We propose a class of calibration estimators for the nonparametric marginal distributions and the copula parameters of interest by balancing the empirical moments of covariates between observed and whole groups. Our proposed estimators do not require the estimation of the missing mechanism, and they enjoy stable performance even when the sample size is small. We prove that our estimators satisfy consistency and asymptotic normality. We also provide a consistent estimator for the asymptotic variance. We show via extensive simulations that our proposed method dominates existing alternatives.

1. Introduction

Copula models are a compelling tool for analyzing complex interdependence of multiple variables. A key characteristic of copula models is that, as Sklar [51] proved, any multivariate joint distribution can be recovered by inputting univariate marginal distributions to a correctly specified copula. The copula approach is capable of capturing a wide range of interdependence among variables with relatively small computational burden. There is a vast and growing literature applying copula models to economic and financial data, among others; see, e.g., [13,43], for extensive surveys and [37,39–41,50] for more recent contributions.

Especially popular are semiparametric copula inference approaches, which involve nonparametric marginal distributions and parametric copulas. Genest et al. [14] proposed the widely used maximum pseudo-likelihood estimator for the copula parameter. Chen and Fan [5] proposed pseudo-likelihood ratio tests for model selection. Chen and Fan [6] studied the estimation of a class of copula-based semiparametric stationary Markov models.

Most papers in the copula literature, including Genest et al. [14], assume complete data. In practice, missing data frequently appear in a broad range of research. In survey analysis, for example, respondents may refuse to report their personal information such as age, education, gender, race, salary, and weight. A primitive way of handling missing data is list-wise deletion, which picks individuals with complete data and treats them all equally. The list-wise deletion delivers consistent inference if data are Missing Completely At Random (MCAR), where target variables $\boldsymbol{Y}_i$ and their missing status $\boldsymbol{T}_i$ are independent of each other. Wang et al. [55] studied the estimation of Gaussian copulas under the MCAR condition. Hasler et al. [24] studied estimation in vine copula models under the MCAR assumption and monotone non-response. In practice, the MCAR condition is often violated, and in such a case list-wise deletion can deliver heavily biased estimators.

* Corresponding author.
E-mail addresses: hamori@econ.kobe-u.ac.jp (S. Hamori), motegi@econ.kobe-u.ac.jp (K. Motegi), zhengzhang@ruc.edu.cn (Z. Zhang).

https://doi.org/10.1016/j.jmva.2019.02.003
0047-259X/

It is therefore desirable to work under a more general assumption called Missing At Random (MAR), originally explored by Rubin [49], where $\boldsymbol{Y}_i$ and $\boldsymbol{T}_i$ are independent of each other given some observed covariates $\boldsymbol{X}_i$. Ding and Song [11] proposed an EM algorithm for estimating the Gaussian copula under the MAR condition. We are not aware of any systematic study on the estimation of general copula models with data MAR, and we fill that gap.

Missing data problems have a close connection with survey sampling, where the parameter of interest is the population total of a survey variable, while only a portion of response outcomes can be obtained. In survey sampling, population totals of certain auxiliary variables can be accurately ascertained from census data. Survey statisticians use auxiliary information in many ways to improve survey estimates, and calibration is one of the most popularly used techniques. Deville and Särndal [10] originally proposed a class of calibration estimators to improve the estimation of finite population totals by utilizing information from auxiliary data. A core insight of Deville and Särndal [10] is to make calibrated weights as close as possible to the original sampling design weights under a given distance measure, subject to a set of constraints. The calibration has been extensively studied in the survey sampling literature [3,9,29–31,33,34,36].

Applying calibration to missing data problems has attracted considerable research interest recently, and has brought many interesting results [8,18–22,45,46,52]. Despite those close connections, there still exists a major difference between survey sampling and missing data analysis: design weights are known in the former but not in the latter. Therefore, existing work that applies calibration to missing data problems is forced to parameterize design weights (namely propensity score functions). If the propensity score models are misspecified, the resulting estimators can be substantially biased.

In causal inference with binary treatments, Chan et al. [2] recently proposed a novel estimation technique to estimate the average treatment effects. They constructed a class of nonparametric calibration weights by balancing the moments of covariates among treated, controlled, and combined groups. Their method bypasses an explicit specification of a propensity score function. Moreover, calibration weights satisfy certain moment constraints in both finite and large samples, so that extreme weights are unlikely to arise. As a result, the calibration estimation attains significantly better finite-sample performance than other nonparametric approximation methods.

As is well known, causal inference with binary treatments is a variant of missing data problems since we can observe one and only one of potential outcomes. Being motivated by such an intimate connection, we extend the maximum pseudo-likelihood approach of Genest et al. [14] by adapting the calibration procedure of Chan et al. [2] in order to perform semiparametric copula inference with data MAR. Our estimator satisfies consistency and asymptotic normality. We also present a consistent estimator for the asymptotic variance of our estimator.

We show via extensive Monte Carlo simulations that our proposed estimator dominates existing alternatives. First, the list-wise deletion leads to severe bias under the MAR condition. Second, the parametric approach based on a specific functional form of propensity score suffers from substantial bias when the propensity score model is misspecified. Third, nonparametric estimators of Hirano et al. [26] exhibit serious sensitivity to the dimension of the approximation sieve, K. Our estimator achieves a remarkably sharp and stable performance compared with the other methods.

The remainder of this paper is organized as follows. In Section 2, we explain our notation and basic set-up. In Section 3, we propose our estimator and study its large-sample properties. In Section 4, we present a nonparametric consistent estimator for the asymptotic variance of our estimator. In Section 5, we propose a data-driven approach to determine the tuning parameter K. In Section 6, we perform Monte Carlo simulations. In Section 7, we provide some concluding remarks. Proofs of selected theorems are presented in technical appendices. Omitted proofs and complete simulation results are collected in the Online Supplement [17].

2. Notation and basic framework

Consider a d-dimensional variable of interest $\boldsymbol{Y}_i = (Y_{1i}, \ldots, Y_{di})^\top$, where $d \geq 2$. Suppose that $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$ form a random sample from distribution F^0. The marginal distributions of F^0, denoted by $F_1^0, \ldots, F_d^0$, are assumed to be continuous and differentiable. Sklar's characterization theorem [51] ensures the existence of a unique copula C^0 such that $F^0(y_1, \ldots, y_d) = C^0\{F_1^0(y_1), \ldots, F_d^0(y_d)\}$ for all $y_1, \ldots, y_d \in \mathbb{R}$. Assume that this copula C^0 has continuous partial derivatives. Then $f^0(y_1, \ldots, y_d) = c^0\{F_1^0(y_1), \ldots, F_d^0(y_d)\} f_1^0(y_1) \cdots f_d^0(y_d)$ for all $y_1, \ldots, y_d \in \mathbb{R}$, where f^0, f_j^0, and c^0 are the density functions of F^0, F_j^0, and C^0, respectively.

Genest et al. [14] pioneered the estimation of semiparametric copula models, here the copula belongs to a parametric family, i.e., $C^0 \in \{C(\cdot; \theta) : \theta \in \mathbb{R}^p\}$, while the marginal distributions $F_1^0, \ldots, F_d^0$ are left unspecified. See also [5,6,16] for more results on semiparametric copula models. Genest et al. [14] proposed the maximum pseudo-likelihood estimator for the target parameter θ, viz.

$$\tilde{\theta} = \arg\max_{\theta \in \Theta} \left[\frac{1}{N} \sum_{i=1}^{N} \ln c\{\tilde{F}_1(Y_{1i}), \ldots, \tilde{F}_d(Y_{di}); \theta\} \right], \tag{1}$$

where $c(\cdot; \theta)$ is the density of $C(\cdot; \theta)$, Θ is a compact subset of $\mathbb{R}^p$ containing the true value θ_0, and for each $j \in \{1, \ldots, d\}$ and $y \in \mathbb{R}$,

$$\tilde{F}_j(y) = \frac{1}{N+1} \sum_{i=1}^{N} \mathbf{1}(Y_{ji} \leq y)$$

is a rescaled empirical marginal distribution.

Genest et al. [14] and most subsequent papers assume complete data. A main goal of this paper is to generalize the maximum pseudo-likelihood estimator in (1), allowing for data missing at random. For each $i \in \{1, \ldots, N\}$, let $\boldsymbol{Y}_i = (\boldsymbol{Y}_{i,\text{obs}}^\top, \boldsymbol{Y}_{i,\text{mis}}^\top)^\top$, where the component $\boldsymbol{Y}_{i,\text{obs}} \in \mathbb{R}^{d_{\text{obs}}}$ is assumed to be always observed while the component $\boldsymbol{Y}_{i,\text{mis}} \in \mathbb{R}^{d_{\text{mis}}}$ may be missing, and $d = d_{\text{obs}} + d_{\text{mis}}$. Let $\boldsymbol{T}_i = (T_{1i}, \ldots, T_{d_{\text{mis}}i})^\top \in \{0,1\}^{d_{\text{mis}}}$ be a binary random vector indicating the missing status of $\boldsymbol{Y}_{i,\text{mis}} = (Y_{1i,\text{mis}}, \ldots, Y_{d_{\text{mis}}i,\text{mis}})^\top$, namely, $T_{ji} = 0$ (resp. $T_{ji} = 1$) if $Y_{ji,\text{mis}}$ is missing (resp. observed).

If $\boldsymbol{T}_i$ and $\boldsymbol{Y}_{i,\text{mis}}$ are independent of each other, the latter is then called missing completely at random (MCAR). Under the MCAR condition, an elementary approach of list-wise deletion, which merely picks individuals with complete observations and puts equal weights on them, is well known to deliver consistent inference. The MCAR condition, however, is an unrealistically strong assumption that is violated in many applications.

In this paper we impose a more realistic assumption called missing at random (MAR), which was put forward by Rubin [49]. Let $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{ri})^\top$ be an r-dimensional vector of data that are observable for all individuals $i \in \{1, \ldots, N\}$, where $\boldsymbol{X}_i \supset \boldsymbol{Y}_{i,\text{obs}}$ and $r \geq d_{\text{obs}}$. The MAR condition assumes that $\boldsymbol{T}_i$ and $\boldsymbol{Y}_{i,\text{mis}}$ are independent of each other given observed data $\boldsymbol{X}_i$.

Assumption 1 (*Missing at Random*). $\boldsymbol{T}_i \perp \boldsymbol{Y}_{i,\text{mis}} | \boldsymbol{X}_i$ for all $i \in \{1, \ldots, N\}$.

The MAR condition has been used in econometrics and statistics to identify the parameter of interest [7,47]. The MAR condition does not require the unconditional independence between $\boldsymbol{T}_i$ and $\boldsymbol{Y}_{i,\text{mis}}$. In many applications $\boldsymbol{T}_i$ and $\boldsymbol{Y}_{i,\text{mis}}$ are unconditionally correlated with each other through $\boldsymbol{X}_i$, and that violates MCAR but not MAR. To simplify notation without losing generality, in the rest of this article we assume that $\boldsymbol{Y}_{i,\text{obs}} = \emptyset$ and $\boldsymbol{Y}_{i,\text{mis}} = \boldsymbol{Y}_i$, which means that all components of $\boldsymbol{Y}_i$ are possibly missing. Then $d_{\text{obs}} = 0$, $d_{\text{mis}} = d$, and $0 < \Pr(T_{ji} = 1) < 1$ for all $j \in \{1, \ldots, d\}$.

3. Weighted two-step estimation

We assume throughout the paper that the true copula parameter θ_0 is a unique solution to

$$\theta_0 = \arg\max_{\theta \in \Theta} \mathrm{E}\left[\ln c\{F_1^0(Y_{1i}), \ldots, F_d^0(Y_{di}); \theta\}\right].$$

Using Assumption 1 and the law of iterated expectations, we can express θ_0 as follows:

$$\begin{aligned}
\theta_0 &= \arg\max_{\theta \in \Theta} \mathrm{E}\left[\mathrm{E}[\ln c\{F_1^0(Y_{1i}), \ldots, F_d^0(Y_{di}); \theta\} | \boldsymbol{X}_i]\right] \\
&= \arg\max_{\theta \in \Theta} \mathrm{E}\left[\mathrm{E}[\ln c\{F_1^0(Y_{1i}), \ldots, F_d^0(Y_{di}); \theta\} | \boldsymbol{X}_i] \times \mathrm{E}[\mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1)/\eta(\boldsymbol{X}_i) | \boldsymbol{X}_i]\right] \\
&= \arg\max_{\theta \in \Theta} \mathrm{E}\left[\mathrm{E}\left[\mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1) \ln[c\{F_1^0(Y_{1i}), \ldots, F_d^0(Y_{di}); \theta\}]/\eta(\boldsymbol{X}_i) \middle| \boldsymbol{X}_i\right]\right] \\
&= \arg\max_{\theta \in \Theta} \mathrm{E}\left[\mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1) \ln[c\{F_1^0(Y_{1i}), \ldots, F_d^0(Y_{di}); \theta\}]/\eta(\boldsymbol{X}_i)\right],
\end{aligned} \tag{2}$$

where $\eta(\boldsymbol{X}_i) = \Pr(T_{1i} = 1, \ldots, T_{di} = 1 | \boldsymbol{X}_i)$ is called a propensity score function.

In view of (2), we propose the weighted maximum pseudo-likelihood estimator for θ as follows. First, we estimate the marginal distributions $F_1^0, \ldots, F_d^0$ by $\hat{F}_1, \ldots, \hat{F}_d$, respectively. Second, we estimate the inverse probability weights $1/\{N\eta(\boldsymbol{X})\}$, denoted by $\hat{q}(\boldsymbol{X})$, and compute $\hat{\theta}$ via a sample version of (2), viz.

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_{i=1}^N \hat{q}(\boldsymbol{X}_i) \mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1) \ln c\{\hat{F}_1(Y_{1i}), \ldots, \hat{F}_d(Y_{di}); \theta\}.$$

The first step is elaborated in Section 3.1, and the second step is elaborated in Section 3.2.

3.1. Estimation of marginal distributions

Under Assumption 1, the marginal distribution F_j^0 can be represented, for each $j \in \{1, \ldots, d\}$, by

$$\begin{aligned}
F_j^0(y) &= \mathrm{E}\{\mathbf{1}(Y_{ji} \leq y)\} = \mathrm{E}\big[\mathrm{E}\{\mathbf{1}(Y_{ji} \leq y) | \boldsymbol{X}_i\}\big] = \mathrm{E}\big[\mathrm{E}\{\mathbf{1}(Y_{ji} \leq y) | \boldsymbol{X}_i\} \times \mathrm{E}\{T_{ji}/\pi_j(\boldsymbol{X}_i) | \boldsymbol{X}_i\}\big] \\
&= \mathrm{E}\big[\mathrm{E}\{\mathbf{1}(Y_{ji} \leq y) \times T_{ji}/\pi_j(\boldsymbol{X}_i) | \boldsymbol{X}_i\}\big] = \mathrm{E}\{\mathbf{1}(Y_{ji} \leq y) T_{ji}/\pi_j(\boldsymbol{X}_i)\}
\end{aligned} \tag{3}$$

for each $j \in \{1, \ldots, d\}$, where $\pi_j(x) = \Pr(T_{ji} = 1 | \boldsymbol{X}_i = x)$ is the propensity score function. If $\pi_j(x)$ were known, then it would be straightforward to estimate F_j^0 via a sample analogue of (3), defined for all $y \in \mathbb{R}$, by

$$\tilde{F}_j(y) = \frac{1}{N} \sum_{i=1}^N \frac{T_{ji}}{\pi_j(\boldsymbol{X}_i)} \mathbf{1}(Y_{ji} \leq y).$$

This estimator is known as the inverse probability weighting (IPW) estimator [27]. Since $\pi_j(x)$ is unknown in practice, it is typically estimated either parametrically [1,48] or nonparametrically [7,26]. Parametric methods are easy to implement, but will lead to erroneous results if the propensity score model is misspecified. Nonparametric methods such as kernel or sieve regression offer asymptotically robust estimators since they do not require the model assumption on the propensity score, but their small-sample performance is notoriously poor.

3.1.1. Calibration weighting estimator

A key property of the propensity score $\pi_j(\boldsymbol{X})$ is that, for any integrable function $u(\boldsymbol{X})$,

$$\mathrm{E}\{T_{ji} \times u(\boldsymbol{X}_i)/\pi_j(\boldsymbol{X}_i)\} = \mathrm{E}\{u(\boldsymbol{X}_i)\}. \tag{4}$$

The propensity score π_j balances all moments of the covariates between the observed group and the whole group, and it is characterized by the infinite moments condition (4). The calibration weights $\hat{p}_{1K}(\boldsymbol{X}), \ldots, \hat{p}_{dK}(\boldsymbol{X})$ are supposed to satisfy a sample analogue of (4), i.e., for all $j \in \{1, \ldots, d\}$,

$$\sum_{i=1}^{N} T_{ji}\hat{p}_{ji}(\boldsymbol{X}_i)u_{K_j}(\boldsymbol{X}_i) = \frac{1}{N}\sum_{i=1}^{N} u_{K_j}(\boldsymbol{X}_i), \tag{5}$$

where $u_{K_j}(\boldsymbol{X}) = (u_{K_j1}(\boldsymbol{X}), \ldots, u_{K_jK_j}(\boldsymbol{X}))^\top$ is the known basis function with dimension $K_j \in \mathbb{N}$. The functions $u_{K_j}(\boldsymbol{X})$ are called the approximation sieve and can be used to approximate any suitable functions $u(\boldsymbol{X})$ arbitrarily well as $K_j \to \infty$. Popularly used sieve functions include power series, splines, and wavelets. See [4] for a thorough discussion.

We now define $\hat{p}_{ji}(\boldsymbol{X}_i)$. Let $D(v, v_0)$ be a known distance measure that is continuously differentiable in $v \in \mathbb{R}$, non-negative, strictly convex in v, and $D(v_0, v_0) = 0$. The general idea of calibration put forward by Deville and Särndal [10] is to minimize the distance between the final weights and a given vector of design weights subject to a fixed number of moment constraints. Design weights (or inverse probability weights) are known in survey sampling, but not in missing data analysis since they are $(T_{j1}/\pi_j(\boldsymbol{X}_1), \ldots, T_{jN}/\pi_j(\boldsymbol{X}_N))$ containing the unknown function $1/\pi_j(\boldsymbol{X}_i)$.

Note that, although the function $1/\pi_j(\boldsymbol{X}_i)$ is unknown in practice, the population mean of the total design weights is equal to 1 since $\mathrm{E}\{T_{ji}/\pi_j(\boldsymbol{X}_i)\} = 1$. This motivates us to construct our calibration weights $T_{j1} \times Np_1, \ldots, T_{jN} \times Np_N$ by minimizing the distance from the population mean of the design weights (i.e., 1) subject to the constraints (5):

$$\min \sum_{i=1}^{N} D(T_{ji} \times Np_{ji}, 1), \quad \text{subject to } \sum_{i=1}^{N} T_{ji}p_{ji}u_{K_j}(\boldsymbol{X}_i) = \sum_{i=1}^{N} u_{K_j}(\boldsymbol{X}_i)/N, \tag{6}$$

where $K_j \to \infty$ as $N \to \infty$ yet with $K_j/N \to 0$. Some remarks on (6) are in order.

Remark 1. The formulation (6) is conceptually different from the existing calibration methods in survey sampling. The first important difference is that our proposed weights minimize a distance from the mean of the design weights, whereas the original survey calibration estimators minimize a distance from the design weights themselves, which are the unknown inverse propensity score weights for the evaluation problem. Hence, our formulation does not need the parametric estimation of the unknown propensity score function.

Remark 2. The number of moment constraints is fixed in the conventional survey calibration, while $K_j \to \infty$ as $N \to \infty$ in our framework. Hellerstein and Imbens [25] showed that if the number of matching conditions is fixed, then an empirical likelihood calibration estimator with misspecified design weights is generally inconsistent. The growing number of moment conditions is necessary for removing asymptotic bias that is associated with misspecified design weights.

Remark 3. Since $u_{K_j}(\boldsymbol{X}_i)$ contains a constant term, one of the constraints in (6) is $T_{j1}p_{j1} + \cdots + T_{jN}p_{jN} = 1$. If we take $D(v, 1) = (v - 1)^2$, then the primal problem (6) reduces to finding the minimum-variance weights [56].

Remark 4. By minimizing the distance from their population mean, the dispersion of the resulting weights is well controlled and we can avoid extreme weights. It is well known that extreme weights cause instability in the IPW estimator.

Note that $D(T_{ji} \times Np_{ji}, 1) = T_{ji} \times D(Np_{ji}, 1) + (1 - T_{ji}) \times D(0, 1)$ and the second term does not depend on the variable p_{ji}. The problem (6) is therefore equivalent to

$$\min \sum_{i=1}^{N} T_{ji}D(Np_{ji}, 1), \quad \text{subject to } \sum_{i=1}^{N} T_{ji}p_{ji}u_{K_j}(\boldsymbol{X}_i) = \frac{1}{N}\sum_{i=1}^{N} u_{K_j}(\boldsymbol{X}_i), \tag{7}$$

where $K_j \to \infty$ as $N \to \infty$ yet with $K_j/N \to 0$. The problem (7) is a convex separable programming with linear constraints. The dual problem, by contrast, is an unconstrained convex maximization problem. The latter enhances the speed and stability of numerical optimization algorithms [53]. Hence we solve for the dual problem to compute calibration weights.

Let $f(v) = D(1-v, 1)$, and $f'(v) = \partial f(v)/\partial v$. When $T_{ji} = 1$, the dual solution of (7) is given by

$$\hat{p}_{jK}(\boldsymbol{X}_i) = \frac{1}{N}\,\rho'\{\hat{\lambda}_{jK}^{\top} u_{K_j}(\boldsymbol{X}_i)\}, \tag{8}$$

where ρ' is the first derivative of a strictly concave function $\rho(v) = f\{(f')^{-1}(v)\} + v - v(f')^{-1}(v)$ and $\hat{\lambda}_{jK} \in \mathbb{R}^{K_j}$ maximizes the following concave objective function

$$\hat{G}_{jK}(\lambda) = \frac{1}{N}\sum_{i=1}^{N}[T_{ji}\rho\{\lambda^{\top} u_{K_j}(\boldsymbol{X}_i)\} - \lambda^{\top} u_{K_j}(\boldsymbol{X}_i)].$$

In view of the first-order condition of the dual problem, it is straightforward to verify that the solution to the dual problem satisfies the linear constraints in primal problem (7).

The link between $f(v)$ and $\rho(v)$ is provided in Section 3 of the Online Supplement [17], where we show that the strict convexity of $D(\cdot, 1)$ is equivalent to the strict concavity of ρ. Since the primal and dual problems lead to the same solution and the latter is simpler to solve, we shall express the calibration estimator in terms of ρ.

The ρ function can be any increasing and strictly concave function. Some examples include $\rho(v) = -\exp(-v)$ for the exponential tilting [32], $\rho(v) = \ln(1+v)$ for the empirical likelihood [42], $\rho(v) = -(1-v)^2/2$ for the continuous updating of the generalized method of moments [23], and $\rho(v) = v - \exp(-v)$ for the inverse logistic.

3.1.2. Large-sample properties

Let $\|\cdot\|$ be the Frobenius norm defined by $\|A\| = \sqrt{\mathrm{tr}(AA^{\top})}$, where A is a real matrix. For any integer $K_j \in \mathbb{N}$, let $\zeta(K_j) = \sup_{x\in\mathcal{X}} \|u_{K_j}(x)\|$ be the supremum norm of approximation sieves $u_{K_j}(x)$. In general, this bound depends on the array of basis that is used. Newey [38] shows that $\zeta(K_j) \le CK_j$ for orthonormal polynomials, and $\zeta(K_j) \le C\sqrt{K_j}$ for B-splines, where $C > 0$ is a universal positive constant. The following conditions are sufficient to establish both L^{∞} and L^2-convergence rates of $N\hat{p}_{jK} \to 1/\pi_j$.

Assumption 2. The support of the covariate $\boldsymbol{X}$, denoted by $\mathcal{X}$, is a Cartesian product of r compact intervals.

Assumption 3. The smallest eigenvalue of $\mathrm{E}\{u_{K_j}(\boldsymbol{X})u_{K_j}(\boldsymbol{X})^{\top}\}$ is bounded away from zero uniformly in K_j.

Assumption 4. For any $j \in \{1, \ldots, d\}$, the inverse propensity score $1/\pi_j(x)$ is bounded above, i.e., there exists some constant η_1 such that $1 \le 1/\pi_j(x) \le \eta_1 < \infty$ for all $x \in \mathcal{X}$.

Assumption 5. There exists λ_{jK} in $\mathbb{R}^{K_j}$ and $\alpha > 0$ such that $\sup_{x\in\mathcal{X}} |(\rho')^{-1}\{1/\pi_{jK}(x)\} - \lambda_{jK}^{\top} u_{K_j}(x)| = O(K_j^{-\alpha})$ as $K_j \to \infty$.

Assumption 6. $\zeta(K_j)^2 K_j^4/N \to 0$ and $\sqrt{N}K_j^{-\alpha} \to 0$.

Assumption 7. The function ρ defined on $\mathbb{R}$ is strictly concave and three times continuously differentiable. Moreover, the range of ρ' contains $[1, \eta_1]$.

Assumption 2 restricts the covariates to be bounded. This condition is restrictive but convenient for computing the convergence rate under L^{∞} norm. It is commonly imposed in the nonparametric regression literature. This condition can be relaxed, however, if we restrict the tail distribution of $\boldsymbol{X}$. Assumption 3, which is also imposed in [38], essentially requires the sieve basis functions to be orthogonal. Assumption 4, a common condition in the missing data literature, ensures that a sufficient portion of marginal data are observed. Assumption 5 requires the sieve approximation error of $\rho'^{-1}\{1/\pi_j(x)\}$ to shrink at a polynomial rate. This condition is satisfied for a variety of sieve basis functions [38]. For example, if $\boldsymbol{X}$ is discrete, then the approximation error is zero for sufficient large K_j and in this case Assumption 5 is satisfied with $\alpha = \infty$. If some components of $\boldsymbol{X}$ are continuous, the polynomial rate depends positively on the smoothness of $\rho'^{-1}\{1/\pi_j(x)\}$ in continuous components and negatively on the number of the continuous components. We will show that the convergence rate of the estimated weight function is bounded by this polynomial rate. Assumption 6, another common assumption in nonparametric regression, restricts the smoothing parameter to balance the bias and variance. Assumption 7 is a mild restriction on ρ and is satisfied by all important special cases considered in the literature.

Theorem 1. *Under Assumptions 2–7, we have, for all $j \in \{1, \ldots, d\}$,*

$$\sup_{x\in\mathcal{X}} |N\hat{p}_{jK}(x) - 1/\pi_j(x)| = O_p\{\zeta(K_j)(K_j^{-\alpha} + \sqrt{K_j/N})\},$$

$$\int_{\mathcal{X}} |N\hat{p}_{jK}(x) - 1/\pi_j(x)|^2 dF_X(x) = O_p(K_j^{-2\alpha} + K_j/N),$$

$$\frac{1}{N}\sum_{i=1}^{N} |N\hat{p}_{jK}(\boldsymbol{X}_i) - 1/\pi_j(\boldsymbol{X}_i)|^2 = O_p(K_j^{-2\alpha} + K_j/N).$$

A proof of Theorem 1 is presented in Appendix A. A key step in the proof is to define an intermediate quantity $p^*_{jK}(x)$, which is the theoretical counterpart of $\hat{p}_{jK}(x)$. We shall establish the convergence rates of $Np^*_{jK} \to 1/\pi_j$ and $N\hat{p}_{jK} \to Np^*_{jK}$, respectively, as $N \to \infty$.

The calibration estimator of the marginal distribution $F^0_j(y)$ is defined as

$$\hat{F}_j(y) = \sum_{i=1}^{N} T_{ji}\hat{p}_{jK}(\boldsymbol{X}_i)\mathbf{1}(Y_{ji} \le y).$$

The following smoothness condition is required to establish the large-sample behavior of $\hat{F}_1, \ldots, \hat{F}_d$.

Assumption 8. For any $j \in \{1, \ldots, d\}$, the conditional distribution function $F_j(y|x) = \Pr(Y_{ji} \le y|\boldsymbol{X}_i = x)$ is continuously differentiable in x and is Lipschitz continuous in y.

Define the d-dimensional functions $\hat{\boldsymbol{F}} = (\hat{F}_1, \ldots, \hat{F}_d)^\top$ and $\boldsymbol{F}^0 = (F^0_1, \ldots, F^0_d)^\top$. The following theorem gives the asymptotically equivalent linear expression for $\sqrt{N}\{\hat{F}_j(y) - F^0_j(y)\}$ and the weak convergence result of $\sqrt{N}(\hat{\boldsymbol{F}} - \boldsymbol{F}^0)$.

Theorem 2. *Impose Assumptions 1–8. Then,*

(i) $\sup_{y\in\mathbb{R}} |\sqrt{N}\{\hat{F}_j(y) - F^0_j(y)\} - \sum_{i=1}^N \psi_j(Y_{ji}, \boldsymbol{X}_i, T_{ji}; y)| = o_p(1)/\sqrt{N}$ *for all* $j \in \{1, \ldots, d\}$, *where*

$$\psi_j(Y_{ji}, \boldsymbol{X}_i, T_{ji}; y) = T_{ji}\mathbf{1}(Y_{ji} \le y)/\pi_j(\boldsymbol{X}_i) - \{T_{ji}/\pi_j(\boldsymbol{X}_i) - 1\}F_j(y|\boldsymbol{X}_i) - F^0_j(y);$$

(ii) $\sqrt{N}(\hat{\boldsymbol{F}} - \boldsymbol{F}^0) \rightsquigarrow \Psi$, *where* $\rightsquigarrow$ *denotes weak convergence,* Ψ *is a* d*-dimensional Gaussian process with mean zero and covariance function*

$$\Omega(y_1, y_2) = \mathrm{E}\{\boldsymbol{\psi}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{T}_i; y_1)\boldsymbol{\psi}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{T}_i; y_2)^\top\},$$

and $\boldsymbol{\psi}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{T}_i; y) = (\psi_1(Y_{1i}, \boldsymbol{X}_i, T_{1i}; y), \ldots, \psi_d(Y_{di}, \boldsymbol{X}_i, T_{di}; y))^\top$.

See Appendix B for a sketched proof of Theorem 2. A complete proof is presented in Section 4 of the Online Supplement [17] in order to save space. Several remarks on Theorem 2 are in order.

Remark 5. A key step toward proving Theorem 2 is decomposing $\sum_{i=1}^N\{\hat{F}_j(y) - F^0_j(y) - \psi_j(Y_{ji}, \boldsymbol{X}_i, T_{ji}; y)\}/\sqrt{N}$ so that it can be rewritten as the sum of asymptotically negligible terms. See Appendix B for this important decomposition.

Remark 6. In the proof of Theorem 2, we use a weighted least squares projection of the conditional distribution $F_j(y|x)$ onto the approximation basis $u_K(x)$, where ρ only appears in the weights of the projection and not in the approximation basis. Our projection argument yields an asymptotically negligible residual term when the weights of the projection are bounded from above and below, which was established under our regularity conditions.

Remark 7. If there are no missing data, namely $T_{ji} = 1$ for all $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, d\}$, then

$$\sqrt{N}\{\hat{F}_j(y) - F^0_j(y)\} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{\mathbf{1}(Y_{ji} \le y) - F^0_j(y)\}$$

becomes the usual empirical process, and it weakly converges to a Gaussian process by Donsker's Theorem.

3.2. Estimation of the copula parameter

3.2.1. Consistency

In this section, we construct calibration weights which lead to consistent estimators for the inverse probability $1/\{N\eta(\boldsymbol{X}_i)\}$. We then obtain a consistent estimator for the target parameter θ_0 in accordance with (2).

By definition, $\eta(x) = \Pr(T_{1i} = \cdots = T_{di} = 1|\boldsymbol{X}_i = x)$ satisfies the following equation

$$\mathrm{E}\{\mathbf{1}(T_{1i} = \cdots = T_{di} = 1)u(\boldsymbol{X}_i)/\eta(\boldsymbol{X}_i)\} = \mathrm{E}\{u(\boldsymbol{X}_i)\}$$

for all integrable function $u(\boldsymbol{X})$. Similar to $\hat{p}_{jK}(\boldsymbol{X})$ in (7), we construct calibration weights $\hat{q}_K(\boldsymbol{X})$ by solving the following constrained optimization problem.

$$\begin{cases} \min \sum_{i=1}^{N} \mathbf{1}(T_{1i} = \cdots = T_{di} = 1)D(Nq_i, 1), \\ \text{subject to } \sum_{i=1}^{N} \mathbf{1}(T_{1i} = \cdots = T_{di} = 1)q_i u_{K_\eta}(\boldsymbol{X}_i) = \frac{1}{N}\sum_{i=1}^{N} u_{K_\eta}(\boldsymbol{X}_i). \end{cases} \tag{9}$$

Similar to (8), the dual solution of (9) is given by

$$\hat{q}_K(\boldsymbol{X}_i) = \frac{1}{N}\,\rho'\{\hat{\beta}_K^\top u_{K_\eta}(\boldsymbol{X}_i)\} \tag{10}$$

for values of i such that $T_{1i} = \cdots = T_{di} = 1$, where $\hat{\beta}_K$ maximizes the following concave objective function

$$\hat{H}_K(\beta) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(T_{1i} = \cdots = T_{di} = 1)\rho\{\beta^\top u_{K_\eta}(\boldsymbol{X}_i)\} - \frac{1}{N}\sum_{i=1}^{N} \beta^\top u_{K_\eta}(\boldsymbol{X}_i).$$

Recall that a complete derivation of the dual solution to (7) is provided in Section 3 of the Online Supplement [17]. Since the dual solution to (9) can be derived similarly, the derivation is omitted.

Assumption 9. The smallest eigenvalue of $\mathrm{E}\{u_{K_\eta}(\boldsymbol{X})u_{K_\eta}(\boldsymbol{X})^\top\}$ is bounded away from zero uniformly in K_η.

Assumption 10. There exists some constant $\eta_1 > 0$ such that $1 \le 1/\eta(x) \le \eta_1 < \infty$ for all $x \in \mathcal{X}$.

Assumption 11. There exist $\beta_K \in \mathbb{R}^{K_\eta}$ and $\alpha > 0$ such that $\sup_{x\in\mathcal{X}} |(\rho')^{-1}\{1/\eta(x)\} - \beta_K^\top u_{K_\eta}(x)| = O(K_\eta^{-\alpha})$ as $K_\eta \to \infty$.

Assumption 12. $\zeta(K_\eta)^2 K_\eta^4/N \to 0$ and $\sqrt{N}\,K_\eta^{-\alpha} \to 0$.

Assumptions 9–12 are natural counterparts of Assumptions 3–6, respectively, and they are used to establish the convergence rate of $N\hat{q}_K \to 1/\eta$. Similar to Theorem 1, the following result holds.

Theorem 3. *Under Assumptions 2, 7, 9–12, we have*

$$\sup_{x\in\mathcal{X}} |N\hat{q}_K(x) - 1/\eta(x)| = O_p\{\zeta(K_\eta)(K_\eta{}^{-\alpha} + \sqrt{K_\eta/N})\},$$

$$\int_{\mathcal{X}} \left|N\hat{q}_K(x) - 1/\eta(x)\right|^2 dF_X(x) = O_p(K_\eta{}^{-2\alpha} + K_\eta/N),$$

$$\frac{1}{N}\sum_{i=1}^{N} \left|N\hat{q}_K(\boldsymbol{X}_i) - 1/\eta(\boldsymbol{X}_i)\right|^2 = O_p(K_\eta{}^{-2\alpha} + K_\eta/N).$$

A proof of Theorem 3 is similar to the proof of Theorem 1 due to the similarity between $\hat{q}_K(x)$ and $\hat{p}_{jK}(x)$. Hence we refrain from presenting the proof of Theorem 3.

Finally, the proposed weighted maximum pseudo-likelihood estimator of θ_0 is defined by

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \left[\sum_{i=1}^{N} \mathbf{1}(T_{1i} = \cdots = T_{di} = 1)\hat{q}_K(\boldsymbol{X}_i)\ln c\{\hat{F}_1(Y_{1i}), \ldots, \hat{F}_d(Y_{di}); \theta\}\right].$$

Assumption 13. Let $U_{ji} = F_j^0(Y_{ji})$ and $\ell(v_1, \ldots, v_d; \theta) = \ln c(v_1, \ldots, v_d; \theta)$.

(i) $\ell(v_1, \ldots, v_d; \theta)$ is a continuous function of θ.
(ii) $\mathrm{E}\{\sup_{\theta\in\Theta} |\ell(U_{1i}, \ldots, U_{di}; \theta)|\} < \infty$.

Assumption 13 is an envelope condition that is sufficient for the applicability of the uniform law of large numbers. It is used for establishing the consistent estimation of $\hat{\theta}$.

Theorem 4. *Under Assumptions 1–13, we have* $\hat{\theta} \xrightarrow{p} \theta_0$.

See Appendix C for a proof of Theorem 4.

3.2.2. Asymptotic normality

Recall that $U_{ji} = F_j^0(Y_{ji})$ and $\ell(v_1, \ldots, v_d; \theta) = \ln c(v_1, \ldots, v_d; \theta)$. Define $\boldsymbol{U}_i = (U_{1i}, \ldots, U_{di})^\top$, $\ell_\theta(v_1, \ldots, v_d; \theta) = \partial\ell(v_1, \ldots, v_d; \theta)/\partial\theta$, $\ell_{\theta\theta}(v_1, \ldots, v_d; \theta) = \partial^2\ell(v_1, \ldots, v_d; \theta)/(\partial\theta\partial\theta^\top)$, $\ell_j(v_1, \ldots, v_d; \theta) = \partial\ell(v_1, \ldots, v_d; \theta)/\partial v_j$, as well as $\ell_{\theta j}(v_1, \ldots, v_d; \theta) = \partial^2\ell(v_1, \ldots, v_d; \theta)/(\partial\theta\partial v_j)$. As in [5], we impose the following conditions in order to establish the asymptotic normality of the proposed estimator $\hat{\theta}$.

Assumption 14. $\mathrm{E}\{\ell_\theta(U_{1i}, \ldots, U_{di}; \theta)|\boldsymbol{X}_i = x\}$ is continuously differentiable in x.

Assumption 15. $B = -\mathrm{E}\{\ell_{\theta\theta}(U_{1i}, \ldots, U_{di}; \theta_0)\}$ and $\Sigma = \mathrm{var}\{\varphi(\boldsymbol{T}_i, \boldsymbol{X}_i, \boldsymbol{U}_i; \theta_0) + \sum_{j=1}^{d} W_j(T_{ji}, \boldsymbol{X}_i, U_{ji}; \theta_0)\}$ are finite and positive definite, where

$$\varphi(\boldsymbol{T}_i, \boldsymbol{X}_i, \boldsymbol{U}_i; \theta_0) = \mathbf{1}(T_{1i} = \cdots = T_{di} = 1)\,\ell_\theta(U_{1i}, \ldots, U_{di}; \theta_0)/\eta(\boldsymbol{X}_i)$$

$$-\mathbf{1}(T_{1i} = \cdots = T_{di} = 1)\,\mathrm{E}[\ell_\theta(U_{1i}, \ldots, U_{di}; \theta_0)|\boldsymbol{X}_i]/\eta(\boldsymbol{X}_i)$$
$$+\mathrm{E}\{\ell_\theta(U_{1i}, \ldots, U_{di}; \theta_0)|\boldsymbol{X}_i\} - \mathrm{E}\{\ell_\theta(U_{1i}, \ldots, U_{di}; \theta_0)\},$$

$$W_j(T_{ji}, \boldsymbol{X}_i, U_{ji}; \theta_0) = \mathrm{E}[\ell_{\theta j}(U_{1s}, \ldots, U_{ds}; \theta_0)\{\phi_j(T_{ji}, \boldsymbol{X}_i, U_{ji}; U_{js}) - U_{js}\}|U_{ji}, \boldsymbol{X}_i, T_{ji}]$$

for $s \neq i$ and, for all $v \in [0, 1]$,

$$\phi_j(T_{ji}, \boldsymbol{X}_i, U_{ji}; v) = T_{ji}\mathbf{1}(U_{ji} \leq v)/\pi_j(\boldsymbol{X}_i) - T_{ji}\mathrm{E}\{\mathbf{1}(U_{ji} \leq v)|\boldsymbol{X}_i\}/\pi_j(\boldsymbol{X}_i) + \mathrm{E}\{\mathbf{1}(U_{ji} \leq v)|\boldsymbol{X}_i\}.$$

Assumption 16. (i) For each $(u_1, \ldots, u_d) \in (0, 1)^d$, $\ell_{\theta\theta}(u_1, \ldots, u_d; \theta)$ is continuous with respect to θ in a neighborhood of θ_0. (ii) $\mathrm{E}\{\sup_{\theta\in\Theta:\|\theta-\theta_0\|=o(1)} \|\ell_{\theta\theta}(U_{1i}, \ldots, U_{di}; \theta)\|\} < \infty$.

Assumption 17. For $j \in \{1, \ldots, d\}$, $\ell_{\theta j}(u_1, \ldots, u_d; \theta_0)$ is well defined and continuous in $(u_1, \ldots, u_d) \in (0, 1)^d$. Furthermore,

(i) $\|\ell_\theta(u_1, \ldots, u_d; \theta_0)\| \leq \text{constant} \times \prod_{j=1}^d \{v_j(1 - v_j)\}^{-a_j}$ for some $a_j \geq 0$ such that $\mathrm{E}\big[\prod_{j=1}^d \{U_{ji}(1 - U_{ji})\}^{-2a_j}\big] < \infty$;

(ii) $\|\ell_{\theta k}(u_1, \ldots, u_d; \theta_0)\| \leq \text{constant} \times \{v_k(1 - v_k)\}^{-b_k} \prod_{j=1, j\neq k}^d \{v_j(1 - v_j)\}^{-a_j}$ for some $b_k > a_k$ such that, for some $\xi_k \in (0, 1/2)$,

$$\mathrm{E}\left[\{U_{ki}(1 - U_{ki})\}^{\xi_k - b_k} \prod_{j=1, j\neq k}^d \{U_{ji}(1 - U_{ji})\}^{-a_j}\right] < \infty.$$

.

Assumption 14 controls the approximation error. Assumption 15 guarantees the finiteness of the asymptotic variance. Assumption 16 guarantees the uniform convergence. Assumption 17 allows the score function and its partial derivatives with respect to the first d arguments to blow up at the boundaries, which occurs for many popular copula families such as Gaussian, Clayton, and t-copulas.

Theorem 5. *Under Assumptions 1–17, we have $\sqrt{N}(\hat{\theta} - \theta_0) \rightsquigarrow \mathcal{N}(0, V_0)$, where $V_0 = B^{-1}\Sigma B^{-1}$ and B and Σ are defined in Assumption 15.*

See Appendix D for a proof of Theorem 5. Admittedly, the calibration estimation does not exploit all available information as sample units with some missing values are left aside. Hence our estimator is not an efficient estimator of θ_0 in general. Efficient estimation of θ_0 is beyond the scope of this paper, and it will be pursued in future work.

We provide some remarks on Theorem 5.

Remark 8. The proof of Theorem 5 is basically in parallel with the proof of Proposition 2 in [5], but the latter rules out missing data. An extra complexity brought by missing data is that we need a different asymptotic representation for $\sqrt{N}\{\hat{F}_j(y) - F_j^0(y)\}$, which has been established in Theorem 2. Indeed, the large-sample behavior of $\sqrt{N}(\hat{\theta} - \theta_0)$ depends on that of $\sqrt{N}\{\hat{F}_j(y) - F_j^0(y)\}$.

Remark 9. If there are no missing data, then V_0 reduces to the asymptotic variance of the maximum pseudo-likelihood estimator derived by Genest et al. [14]. See Section 5 of the Online Supplement [17] for a detailed verification.

Remark 10. The proposed estimator $\hat{\theta}$ is a semiparametric estimator which involves the nonparametric estimation of marginal distributions. Our estimator would therefore suffer from the conventional issue of "curse of dimensionality" in nonparametric estimation as the dimension of covariates $\boldsymbol{X}_i$ is larger and larger. In Section 6.2.2, we perform simulation experiments with $r = 2$ covariates, and find that our estimator has a much sharper performance than existing estimators. An extension to high-dimensional $\boldsymbol{X}_i$ is nontrivial and will be pursued in future work. An application of the sufficient dimension reduction [28] seems to be an appealing solution.

4. Variance estimation

As shown in Theorem 4, the asymptotic variance of $\sqrt{N}(\hat{\theta} - \theta_0)$ is given by $V_0 = B^{-1}\Sigma B^{-1}$. In this section, we construct consistent estimators for B and Σ, which leads to a consistent estimator for V_0.

We first consider B. Using Assumption 1, B can be rewritten as

$$B = -\mathrm{E}\{\mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1)\,\ell_{\theta\theta}(U_{1i}, \ldots, U_{di}; \theta_0)/\eta(\boldsymbol{X}_i)\}.$$

Recall from Theorem 2 that $\hat{F}_j$ is consistent for F_j^0. Also recall from Theorem 3 that $N\hat{q}_K(x)$ is consistent for $1/\eta(x)$. Hence we define a plug-in estimator of B as

$$\hat{B} = -\sum_{i=1}^N \mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1)\hat{q}_K(\boldsymbol{X}_i)\ell_{\theta\theta}(\hat{U}_{1i}, \ldots, \hat{U}_{di}; \hat{\theta}), \tag{11}$$

where

$$\hat{U}_{ji} = \hat{F}_j(Y_{ji}) = \sum_{s=1}^{N} T_{js}\hat{p}_{jK}(\boldsymbol{X}_s)\mathbf{1}(Y_{js} \le Y_{ji}).$$

We next consider Σ. Under Assumption 1, Σ can be rewritten as

$$\Sigma = \mathrm{E}\left[\mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1)\left\{\varphi(\boldsymbol{T}_i, \boldsymbol{X}_i, \boldsymbol{U}_i; \theta_0) + \sum_{j=1}^{d} W_j(T_{ji}, \boldsymbol{X}_i, U_{ji}; \theta_0)\right\}^2 \Big/ \eta(\boldsymbol{X}_i)\right],$$

where $\varphi(\boldsymbol{T}_i, \boldsymbol{X}_i, \boldsymbol{U}_i; \theta)$ and $W_j(T_{ji}, \boldsymbol{X}_i, U_{ji}; \theta)$ are defined in Assumption 15. As in (11), we can define plug-in estimators of $\varphi(\boldsymbol{T}_i, \boldsymbol{X}_i, \boldsymbol{U}_i; \theta_0)$ and $W_j(T_{ji}, \boldsymbol{X}_i, U_{ji}; \theta_0)$ as

$$\begin{aligned}\hat{\varphi}(\boldsymbol{T}_i, \boldsymbol{X}_i, \boldsymbol{U}_i; \theta_0) = {} & \mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1)N\hat{q}_K(\boldsymbol{X}_i)\ell_\theta(\hat{U}_{1i}, \ldots, \hat{U}_{di}; \hat{\theta}) \\ & -\mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1)N\hat{q}_K(\boldsymbol{X}_i) \times \hat{\mathrm{E}}\{\ell_\theta(U_{1i}, \ldots, U_{di}; \theta_0)|\boldsymbol{X}_i\} \\ & +\hat{\mathrm{E}}\{\ell_\theta(U_{1i}, \ldots, U_{di}; \theta_0)|\boldsymbol{X}_i\} - \hat{\mathrm{E}}\{\ell_\theta(U_{1i}, \ldots, U_{di}; \theta_0)\}\end{aligned}$$

and

$$\hat{W}_j(T_{ji}, \boldsymbol{X}_i, U_{ji}; \theta_0) = \sum_{s=1}^{N} \mathbf{1}(T_{1s} = 1, \ldots, T_{ds} = 1)\hat{q}_K(\boldsymbol{X}_s)\ell_{\theta j}(\hat{U}_{1s}, \ldots, \hat{U}_{ds}; \hat{\theta})\{\hat{\phi}_j(T_{ji}, \boldsymbol{X}_i, U_{ji}; \hat{U}_{js}) - \hat{U}_{js}\},$$

where

$$\begin{aligned}\hat{\mathrm{E}}\{\ell_\theta(U_{1i}, \ldots, U_{di}; \theta_0)|\boldsymbol{X}_i\} = {} & \left\{\sum_{s=1}^{N} \mathbf{1}(T_{1s} = \cdots = T_{ds} = 1)\ell_\theta(\hat{U}_{1s}, \ldots, \hat{U}_{ds}; \hat{\theta})u_\ell(\boldsymbol{X}_s)\right\}^\top \\ & \times \left\{\sum_{\ell=1}^{N} \mathbf{1}(T_{1\ell} = \cdots = T_{d\ell} = 1)u_\ell(\boldsymbol{X}_l)u_\ell^\top(\boldsymbol{X}_l)\right\}^{-1} u_\ell(\boldsymbol{X}_i)\end{aligned}$$

is the least squares estimator of $\ell_\theta(\hat{U}_{1i}, \ldots, \hat{U}_{di}; \hat{\theta})$ based on the basis $u_\ell(\boldsymbol{X}_i)$. Further,

$$\hat{\mathrm{E}}\{\ell_\theta(U_{1i}, \ldots, U_{di}; \theta_0)\} = \sum_{s=1}^{N} \mathbf{1}(T_{1s} = \cdots = T_{ds} = 1)\hat{q}_K(\boldsymbol{X}_s)\ell_\theta(\hat{U}_{1s}, \ldots, \hat{U}_{ds}; \hat{\theta})$$

$$\begin{aligned}\hat{\phi}_j(T_{ji}, \boldsymbol{X}_i, U_{ji}; v) = {} & T_{ji}\{N\hat{p}_{jK}(\boldsymbol{X}_i)\}\mathbf{1}(\hat{U}_{ji} \le v) \\ & -T_{ji}\{N\hat{p}_{jK}(\boldsymbol{X}_i)\} \times \hat{\mathrm{E}}\{\mathbf{1}(U_{ji} \le v)|\boldsymbol{X}_i, T_{ji} = 1\} + \hat{\mathrm{E}}\{\mathbf{1}(U_{ji} \le v)|\boldsymbol{X}_i, T_{ji} = 1\}\end{aligned}$$

and

$$\hat{\mathrm{E}}\{\mathbf{1}(U_{ji} \le v)|\boldsymbol{X}_i, T_{ji} = 1\} = \left\{\sum_{s=1}^{N} \mathbf{1}(T_{js} = 1)\mathbf{1}(\hat{U}_{js} \le v)u_\ell(\boldsymbol{X}_s)\right\}^\top \left\{\sum_{\ell=1}^{N} \mathbf{1}(T_{j\ell} = 1)u_\ell(\boldsymbol{X}_\ell)u_\ell^\top(\boldsymbol{X}_l)\right\}^{-1} u_\ell(\boldsymbol{X}_i).$$

Under standard conditions in the nonparametric estimation literature such as Assumption 15.2 in [35] or assumptions in [38], it is well-known that the least squares projection estimators are consistent. Then together with the facts $\hat{U}_{ji} \xrightarrow{p} U_{ji}$ and $\hat{\theta} \xrightarrow{p} \theta_0$, the following results hold:

$$\sup_{x \in \mathcal{X}} |\hat{\mathrm{E}}\{\ell_\theta(U_{1i}, \ldots, U_{di}; \theta_0)|\boldsymbol{X}_i = x\} - \mathrm{E}\{\ell_\theta(U_{1i}, \ldots, U_{di}; \theta_0)|\boldsymbol{X}_i = x\}| \xrightarrow{p} 0, \tag{12}$$

$$\sup_{x \in \mathcal{X}} |\hat{\mathrm{E}}\{\mathbf{1}(U_{ji} \le v)|\boldsymbol{X}_i = x, T_{ji} = 1\} - \mathrm{E}\{\mathbf{1}(U_{ji} \le v)|\boldsymbol{X}_i = x, T_{ji} = 1\}| \xrightarrow{p} 0. \tag{13}$$

Finally, construct an estimator for V_0 as $\hat{V} = \hat{B}^{-1}\hat{\Sigma}\hat{B}^{-1}$. The following theorem establishes consistency as desired.

Theorem 6. *Impose Assumptions 1–17 and assume that (12) and (13) hold, then we have* $\|\hat{V} - V_0\| \xrightarrow{p} 0$.

See Section 6 of the Online Supplement [17] for a proof of Theorem 6.

5. Selection of tuning parameters

While our large-sample theory allows for a wide range of values for $K_1, \ldots, K_d$ and K_η, a practical question is how to choose those values. In this section, we present a data-driven approach that selects the optimal K_j^* and K_η^* in terms of

covariate balancing. A natural estimator of the joint distribution of $\boldsymbol{X}_i$, using the whole group of individuals, is given by

$$\hat{F}_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(X_{1i} \le x_1, \ldots, X_{ri} \le x_r).$$

An alternative estimator using the observed group of individuals is given by

$$\hat{F}_{\boldsymbol{X},K_j}(\boldsymbol{x}) = \sum_{i=1}^{N} T_{ji}\hat{p}_{jK_j}(\boldsymbol{X}_i)\mathbf{1}(X_{1i} \le x, \ldots, X_{ri} \le x_r),$$

which depends on K_j. In view of covariate balancing $\mathrm{E}\{T_{ji}\mathbf{1}(\boldsymbol{X}_i \le \boldsymbol{x})/\pi_j(\boldsymbol{X}_i)\} = \mathrm{E}\{\mathbf{1}(\boldsymbol{X}_i \le \boldsymbol{x})\}$, it is desired to select K_j such that $\hat{F}_{\boldsymbol{X}}$ and $\hat{F}_{\boldsymbol{X},K_j}$ are as close as possible. It is also desired to impose a penalty against having large K_j in order to avoid over-fitting. Hence we propose to select $K_j \in \{1, \ldots, \bar{K}_j\}$ that minimizes a penalized L^2 distance, viz.

$$d_K(\hat{F}_{\boldsymbol{X}}, \hat{F}_{\boldsymbol{X},K_j}) = \frac{1}{N(1-K_j^2/N)^2}\sum_{i=1}^{N}\{\hat{F}_{\boldsymbol{X}}(\boldsymbol{X}_i) - \hat{F}_{\boldsymbol{X},K_j}(\boldsymbol{X}_i)\}^2, \tag{14}$$

where $\bar{K}_j$ is a prespecified integer. The proposed approach is analogous to the generalized cross-validation in Section 15.2 of Li and Racine [35].

Similarly, we can compute a weighted distribution function of $\boldsymbol{X}_i$ by using the weights $\hat{q}_K(\boldsymbol{X}_i)$. Then, based on the covariate balancing equation $\mathrm{E}\{\mathbf{1}(T_{1i} = \cdots = T_{di} = 1)\mathbf{1}(\boldsymbol{X}_i \le \boldsymbol{x}_i)/\eta(\boldsymbol{X}_i)\} = \mathrm{E}\{\mathbf{1}(\boldsymbol{X}_i \le \boldsymbol{x}_i)\}$, we can construct an empirical criterion for selecting K_η^* like (14). Our approach is admittedly ad-hoc since one could use other distance measures or penalty terms. We, however, show in Section 6 that our approach achieves sharp performance in finite sample for both the nonparametric estimator of Hirano et al. [26] and the calibration estimator.

6. Monte Carlo simulations

In this section, we run Monte Carlo simulations in order to evaluate the finite sample performance of the proposed calibration estimator and other existing estimators. In Section 6.1, we investigate a benchmark scenario which has a relatively simple structure. In Section 6.2, we investigate more involved scenarios for completeness.

6.1. Benchmark scenario

6.1.1. Simulation design

Suppose that $\boldsymbol{Y}_i = (Y_{1i}, Y_{2i})^\top$ are bivariate target variables (i.e., $d = 2$) and X_i is a scalar covariate (i.e., $r = 1$). We specify the joint distribution of $\boldsymbol{Z}_i = (\boldsymbol{Y}_i^\top, X_i)^\top$ via two Archimedean copulas that are widely used in empirical applications. The first copula is the trivariate Clayton copula with a scalar parameter α_0, written as $C_3(\alpha_0)$. The second copula is the trivariate Gumbel copula with a scalar parameter γ_0, written as $G_3(\gamma_0)$.

As implied by Examples 1 and 2 in [15], Kendall's τ is given by $\tau = \alpha_0/(\alpha_0+2)$ for $C_3(\alpha_0)$ and $\tau = 1-1/\gamma_0$ for $G_3(\gamma_0)$. We consider two cases of $\tau \in \{0.45, 0.75\}$. Hence we set the true copula parameters to be $(\alpha_0, \gamma_0) = (1.636, 1.818)$ for $\tau = 0.45$ and $(\alpha_0, \gamma_0) = (6.000, 4.000)$ for $\tau = 0.75$. There exists relatively weak association among $\{Y_{1i}, Y_{2i}, X_i\}$ when $\tau = 0.45$ and relatively strong association when $\tau = 0.75$.

Inputs to the copulas are $u_1 = F_1^0(y_1)$, $u_2 = F_2^0(y_2)$, and $u_3 = F_X(x)$, where F_j^0 is the marginal distribution function of Y_{ji} and F_X is the marginal distribution function of X_i. We use the standard Gaussian distribution for F_1^0, F_2^0, and F_X. Since the standard Gaussian distribution has a tractable inverse distribution function, it is straightforward to draw (Y_{1i}, Y_{2i}, X_i) by first generating (U_{1i}, U_{2i}, U_{3i}) from the copulas and then transforming them to $Y_{1i} = (F_1^0)^{-1}(U_{1i})$, $Y_{2i} = (F_2^0)^{-1}(U_{2i})$, and $X_i = F_X^{-1}(U_{3i})$.

We next specify missing mechanisms. Assume for simplicity that Y_{1i} is always observed and only Y_{2i} can be missing with conditional probability:

$$\pi_2(x_i) = \Pr(T_{2i} = 1 \,|\, X_i = x_i) = 1/\{1 + \exp(a + bx_i)\}. \tag{15}$$

It is common in the missing data literature to use the logistic function (15) to specify missing probability [44]. We consider four cases for (a, b):

Case A: $(a, b) = (-1.385, 0.000)$, under which MCAR holds and $\mathrm{E}(T_{2i}) = 0.8$.
Case B: $(a, b) = (-1.430, 0.400)$, under which MAR holds and $\mathrm{E}(T_{2i}) = 0.8$.
Case C: $(a, b) = (-0.405, 0.000)$, under which MCAR holds and $\mathrm{E}(T_{2i}) = 0.6$.
Case D: $(a, b) = (-0.420, 0.400)$, under which MAR holds and $\mathrm{E}(T_{2i}) = 0.6$.

20% of the Y_{2i}s are missing on average in Cases A and B, since $\mathrm{E}(T_{2i}) = 0.8$. The crucial difference between the two cases is that $b = 0$ (i.e., MCAR) in Case A while $b \neq 0$ (i.e., MAR) in Case B. Under MCAR, the missing probability of Y_{2i} does

not depend on X_i although $\boldsymbol{Y}_i$ is (nonlinearly) related with X_i via the copula. Under MAR, the missing probability of Y_{2i} depends on X_i and $\boldsymbol{Y}_i$ is related with X_i via the copula. Similar structures apply for Cases C and D with a larger missing probability; 40% of the Y_{2i}s are missing on average since $\mathrm{E}(T_{2i}) = 0.6$.

We draw $J = 1000$ Monte Carlo samples with sample size $N \in \{250, 500, 1000\}$. We consider several estimation methods for comparison.

6.1.2. List-wise deletion

The first approach is semiparametric estimation with list-wise deletion. For each component $j \in \{1, 2\}$, we estimate the marginal distribution by

$$\hat{F}_j(y) = \frac{1}{N^* + 1} \sum_{i=1}^{N} \mathbf{1}(T_{1i} = 1, T_{2i} = 1)\mathbf{1}(Y_{ji} < y), \quad N^* = \sum_{i=1}^{N} \mathbf{1}(T_{1i} = 1, T_{2i} = 1).$$

We then compute the maximum likelihood estimator for the copula parameter based on the complete data. Taking the Clayton copula as an example, the maximum likelihood estimator is defined as follows. (The Gumbel case can be treated analogously.)

$$\hat{\alpha} = \arg \max_{\alpha \in (0,\infty)} \left[\sum_{i=1}^{N} \mathbf{1}(T_{1i} = 1, T_{2i} = 1) \ln c_2\{\hat{F}_1(Y_{1i}), \hat{F}_2(Y_{2i});\ \alpha\} \right],$$

where $c_2(u_1, u_2;\ \alpha)$ is the probability density function of the bivariate Clayton copula $C_2(\alpha)$. There is not a misspecification problem here due to a well-known property that any bivariate marginal distribution of $C_3(\alpha_0)$ is indeed $C_2(\alpha_0)$. This property holds for any Archimedean copula. It is a useful property when we perform simulations on copula models with data MAR. $\boldsymbol{Y}_i$ and X_i are associated with each other through a copula, and hence we can create a MAR situation using, say, (15). Moreover, the unconditional distribution of $\boldsymbol{Y}_i$ is tractable and hence we can compute the bias and related quantities of a maximum likelihood estimator.

6.1.3. Parametric estimation

The second approach estimates the propensity score function $\pi_2(x)$ parametrically. Define

$$\pi_2(x; a, b) = 1/\{1 + \exp(a + bx)\}, \tag{16}$$

then the log-likelihood function of $(T_{21}, X_1), \ldots, (T_{2N}, X_N)$ is given by

$$\ell(a, b) = \sum_{i=1}^{N} [T_{2i} \ln \pi_2(X_i; a, b) + (1 - T_{2i}) \ln\{1 - \pi_2(X_i; a, b)\}].$$

Compute the maximum likelihood estimator $(\hat{a}, \hat{b})$, and then calculate $\hat{p}_2(X_i) = \hat{q}(X_i) = 1/\{N \times \pi_2(X_i; \hat{a}, \hat{b})\}$. Marginal distributions are estimated as

$$\hat{F}_j(y) = \sum_{i=1}^{N} \mathbf{1}(T_{ji} = 1)\hat{p}_j(X_i)\mathbf{1}(Y_{ji} < y) \tag{17}$$

and the copula parameter is estimated as

$$\hat{\alpha} = \arg \max_{\alpha \in (0,\infty)} \left[\sum_{i=1}^{N} \mathbf{1}(T_{1i} = 1, T_{2i} = 1)\hat{q}(X_i) \ln c_2\{\hat{F}_1(Y_{1i}), \hat{F}_2(Y_{2i});\ \alpha\} \right]. \tag{18}$$

Note that (16) is correctly specified relative to the true propensity score function (15). For comparison, we also use a misspecified model

$$\pi_2(x; a, b) = 1/\{1 + \exp(bx)\}. \tag{19}$$

Model (19) is misspecified since $a \neq 0$ in each of Cases A–D. We are supposed to get consistent estimators when (16) is used and inconsistent estimators when (19) is used.

6.1.4. Nonparametric estimation

The third approach estimates the propensity score function $\pi_2(x)$ nonparametrically based on [26]. To this end, define $\pi_{2K}(X_i; \lambda) = 1/[1 + \exp\{-\lambda^\top u_{K_2}(X_i)\}]$, where $u_{K_2}(X_i) = (1, X_i, \ldots, X_i^{K_2-1})^\top$ is the approximation sieve also used in the calibration estimation. The log-likelihood function of $(T_{21}, X_1), \ldots, (T_{2N}, X_N)$ is written as

$$\ell(\lambda) = \sum_{i=1}^{N} [T_{2i} \ln \pi_{2K}(X_i; \lambda) + (1 - T_{2i}) \ln\{1 - \pi_{2K}(X_i; \lambda)\}].$$

Table 1
Benchmark simulation results on Clayton copula with $\alpha_0 = 6.000$ (Kendall's $\tau = 0.75$).

MCAR with $E(T_{2i}) = 0.6$			
	$N = 250$ Bias, Stdev, RMSE	$N = 500$ Bias, Stdev, RMSE	$N = 1000$ Bias, Stdev, RMSE
List-wise deletion	−0.115, 0.754, 0.763	−0.046, 0.539, 0.541	−0.052, 0.379, 0.382
Param (correct)	−0.442, 0.663, 0.797	−0.283, 0.464, 0.544	−0.172, 0.337, 0.378
Param (misspec)	−1.687, 0.647, 1.807	−1.570, 0.456, 1.635	−1.470, 0.338, 1.509
Nonparam ($K_2 = 3$)	−0.369, 0.649, 0.747	−0.239, 0.482, 0.538	−0.151, 0.364, 0.394
Nonparam ($K_2 = 4$)	−0.630, 1.985, 2.083	−0.801, 1.945, 2.104	−0.921, 2.153, 2.342
Nonparam (CB)	−0.307, 0.659, 0.727	−0.185, 0.481, 0.515	−0.116, 0.347, 0.366
Calibration ($K_2 = 3$)	−0.365, 0.648, 0.744	−0.245, 0.456, 0.517	−0.145, 0.307, 0.339
Calibration ($K_2 = 4$)	−0.304, 0.646, 0.714	−0.175, 0.460, 0.492	−0.130, 0.324, 0.349
Calibration (CB)	−0.317, 0.643, 0.717	−0.213, 0.462, 0.509	−0.122, 0.319, 0.342
MAR with $E(T_{2i}) = 0.6$			
	$N = 250$ Bias, Stdev, RMSE	$N = 500$ Bias, Stdev, RMSE	$N = 1000$ Bias, Stdev, RMSE
List-wise deletion	0.420, 0.840, 0.940	0.453, 0.594, 0.747	0.472, 0.408, 0.624
Param (correct)	−0.248, 0.650, 0.696	−0.152, 0.454, 0.479	−0.089, 0.322, 0.334
Param (misspec)	−1.367, 0.610, 1.497	−1.273, 0.420, 1.340	−1.194, 0.330, 1.239
Nonparam ($K_2 = 3$)	−0.248, 0.612, 0.661	−0.152, 0.434, 0.460	−0.057, 0.322, 0.327
Nonparam ($K_2 = 4$)	−0.495, 1.916, 1.978	−0.723, 2.358, 2.467	−0.972, 2.470, 2.654
Nonparam (CB)	−0.167, 0.670, 0.691	−0.125, 0.452, 0.469	−0.062, 0.333, 0.338
Calibration ($K_2 = 3$)	−0.266, 0.605, 0.661	−0.161, 0.450, 0.477	−0.110, 0.322, 0.340
Calibration ($K_2 = 4$)	−0.274, 0.649, 0.705	−0.165, 0.456, 0.485	−0.092, 0.327, 0.340
Calibration (CB)	−0.259, 0.653, 0.703	−0.142, 0.472, 0.493	−0.076, 0.315, 0.324

In this table we report bias, standard deviation, and root mean squared error (RMSE) of each estimator with respect to $\alpha_0 = 6.000$ after $J = 1000$ Monte Carlo trials. "Param (correct)" signifies the parametric estimator based on a correctly specified propensity score model. "Param (misspec)" signifies the parametric estimator based on a misspecified propensity score model. "Nonparam" signifies the nonparametric estimator of Hirano et al. [26]. "Calibration" signifies our proposed calibration estimator. For the nonparametric and calibration estimators, approximation sieves are constructed from the power series of X_i. The dimension of the approximation sieve is either fixed at $K_2 \in \{3, 4\}$ or automatically selected from $K_2 \in \{1, \ldots, 5\}$ based on the covariate balancing (CB) principle.

Compute the maximum likelihood estimator $\hat{\lambda}$, and calculate $\hat{p}_{2K}(X_i) = \hat{q}_K(X_i) = 1/\{N \times \pi_{2K}(X_i; \hat{\lambda})\}$. Then use (17) and (18) to complete the procedure.

The difference between the parametric and nonparametric approaches is that the former requires an explicit specification of propensity score functions while the latter does not. The nonparametric approach, however, requires a selection of K_2. We use $u_{K_2}(X_i) = (1, X_i, X_i^2)^\top$ (i.e., $K_2 = 3$) and $u_{K_2}(X_i) = (1, X_i, X_i^2, X_i^3)^\top$ (i.e., $K_2 = 4$) in order to see how results change across different values of K_2. We also perform the data-driven selection of K_2^* with an upper bound $\bar{K}_2 = 5$ as described in Section 5.

6.1.5. Calibration estimation

The fourth approach is our proposed calibration estimation. For the second component, we estimate the marginal distribution by (17), where $\hat{p}_{2K}(X_i)$ is given in (8). We then compute the maximum likelihood estimator for the copula parameter from (18), where $\hat{q}_K(X_i)$ is given in (10). As in the nonparametric approach, we use fixed $K_2 \in \{3, 4\}$ and the data-driven selection of K_2^* with upper bound $\bar{K}_2 = 5$.

6.1.6. Simulation results

In Table 1, we report simulation results on the Clayton copula. To save space, we only present the most informative results with $\alpha_0 = 6.000$ and Cases C–D. Those cases correspond to the larger values of copula parameter and missing probability. Similarly, Table 2 reports results on the Gumbel copula with $\gamma_0 = 4.000$ and Cases C–D. See Tables 1–6 of the Online Supplement [17] for complete results including $\alpha_0 = 1.636$, $\gamma_0 = 1.818$, and Cases A–B.

First, the list-wise-deletion estimator is consistent under MCAR and inconsistent under MAR, as expected. See, for example, Table 1 with $N = 1000$. The bias of the list-wise-deletion estimator is −0.052 under MCAR and 0.472 under MAR. In Table 2, we observe that similar results hold for the Gumbel copula; the bias is −0.007 under MCAR and −0.274 under MAR. Another interesting finding is that the list-wise deletion results in positive bias under the Clayton copula and negative bias under the Gumbel copula. Those results are essentially a consequence of our simulation design and the tail-dependence properties of the two copulas. See Section 7.1 of the Online Supplement [17] for a precise reason for the positive bias under Clayton and the negative bias under Gumbel.

Second, the parametric approach produces extremely large bias for all cases when the propensity score model is misspecified. See, for example, the top half of Table 2, where the bias is $\{-2.593, -2.618, -2.621\}$ for $N \in \{250, 500, 1000\}$, respectively. Those results confirm that the parametric approach utterly fails if the model is misspecified. Note also that,

Table 2
Benchmark simulation results on Gumbel copula with $\gamma_0 = 4.000$ (Kendall's $\tau = 0.75$).

MCAR with $E(T_{2i}) = 0.6$			
	$N = 250$ Bias, Stdev, RMSE	$N = 500$ Bias, Stdev, RMSE	$N = 1000$ Bias, Stdev, RMSE
List-wise deletion	0.033, 0.374, 0.375	0.007, 0.266, 0.266	−0.007, 0.186, 0.186
Param (correct)	−0.231, 0.307, 0.384	−0.113, 0.232, 0.258	−0.047, 0.168, 0.174
Param (misspec)	−2.593, 0.249, 2.605	−2.618, 0.137, 2.622	−2.621, 0.088, 2.622
Nonparam ($K_2 = 3$)	−0.211, 0.293, 0.361	−0.113, 0.253, 0.277	−0.046, 0.178, 0.184
Nonparam ($K_2 = 4$)	−0.587, 0.960, 1.125	−0.631, 1.104, 1.271	−0.797, 1.239, 1.474
Nonparam (CB)	−0.198, 0.311, 0.369	−0.080, 0.227, 0.240	−0.041, 0.162, 0.167
Calibration ($K_2 = 3$)	−0.075, 0.323, 0.332	−0.042, 0.232, 0.236	−0.026, 0.158, 0.160
Calibration ($K_2 = 4$)	−0.059, 0.336, 0.341	−0.037, 0.234, 0.237	−0.029, 0.167, 0.170
Calibration (CB)	−0.062, 0.347, 0.352	−0.045, 0.240, 0.244	−0.018, 0.161, 0.162
MAR with $E(T_{2i}) = 0.6$			
	$N = 250$ Bias, Stdev, RMSE	$N = 500$ Bias, Stdev, RMSE	$N = 1000$ Bias, Stdev, RMSE
List-wise deletion	−0.243, 0.348, 0.424	−0.256, 0.244, 0.353	−0.274, 0.173, 0.324
Param (correct)	−0.313, 0.323, 0.450	−0.153, 0.223, 0.271	−0.081, 0.158, 0.177
Param (misspec)	−2.610, 0.204, 2.618	−2.620, 0.124, 2.623	−2.623, 0.082, 2.625
Nonparam ($K_2 = 3$)	−0.318, 0.315, 0.448	−0.165, 0.233, 0.285	−0.074, 0.172, 0.188
Nonparam ($K_2 = 4$)	−0.645, 0.923, 1.126	−0.682, 1.081, 1.278	−0.785, 1.233, 1.462
Nonparam (CB)	−0.323, 0.308, 0.446	−0.140, 0.224, 0.264	−0.073, 0.164, 0.179
Calibration ($K_2 = 3$)	−0.140, 0.324, 0.353	−0.093, 0.234, 0.252	−0.058, 0.168, 0.177
Calibration ($K_2 = 4$)	−0.129, 0.342, 0.366	−0.094, 0.243, 0.260	−0.046, 0.163, 0.169
Calibration (CB)	−0.142, 0.333, 0.362	−0.100, 0.240, 0.260	−0.064, 0.172, 0.183

In this table we report bias, standard deviation, and root mean squared error (RMSE) of each estimator with respect to $\gamma_0 = 4.000$ after $J = 1000$ Monte Carlo trials. "Param (correct)" signifies the parametric estimator based on a correctly specified propensity score model. "Param (misspec)" signifies the parametric estimator based on a misspecified propensity score model. "Nonparam" signifies the nonparametric estimator of Hirano et al. [26]. "Calibration" signifies our proposed calibration estimator. For the nonparametric and calibration estimators, approximation sieves are constructed from the power series of X_i. The dimension of the approximation sieve is either fixed at $K_2 \in \{3, 4\}$ or automatically selected from $K_2 \in \{1, \ldots, 5\}$ based on the covariate balancing (CB) principle.

even if the model is correctly specified, the parametric approach can suffer from large bias in small sample. See again the top half of Table 2, where the bias of the correctly-specified parametric estimator is $\{-0.231, -0.113, -0.047\}$ for $N \in \{250, 500, 1000\}$. Those values are in fact larger than the bias of the list-wise-deletion estimator. A potential reason for the poor performance of the correctly-specified parametric estimator in small sample is that the propensity score model is highly nonlinear in X_i.

Third, the performance of the nonparametric estimator is extremely sensitive to the choice of K_2. When $K_2 = 3$, the performance of the nonparametric estimator is roughly comparable with the correctly-specified parametric estimator. When $K_2 = 4$, the nonparametric estimator is substantially biased for all cases. In the bottom half of Table 2, the bias of the nonparametric estimator with $K_2 = 4$ is $\{-0.645, -0.682, -0.785\}$ for $N \in \{250, 500, 1000\}$, respectively. It is a practical disadvantage that the nonparametric estimator exhibits a substantial variation across different values of K_2. The data-driven selection of K_2^* alleviates the sensitivity of the nonparametric estimator successfully. Focusing on the sample example, the bias is now $\{-0.323, -0.140, -0.073\}$ for $N \in \{250, 500, 1000\}$. (Another compelling way to alleviate the sensitivity to K_2 is reconstructing $u_{K_2}(X)$ via B-splines instead of power series. See Section 7.2 of the Online Supplement [17] for extra simulations using B-splines.)

Fourth and most importantly, the calibration estimator shows a strikingly sharp and stable performance. The performance of the calibration estimator remains almost the same whether $K_2 = 3$, $K_2 = 4$, or K_2^* is used, and the bias and standard deviation are small enough for all cases. Robustness against different values of K_2 is a great advantage of the calibration estimator relative to the nonparametric estimator. When both nonparametric and calibration estimators are assisted by the data-driven K_2^*, they perform as well as each other in many cases. Note, however, that the calibration estimator with K_2^* clearly dominates the nonparametric estimator with K_2^* under the Gumbel copula with $N = 250$ (Table 2); the bias of the calibration and nonparametric estimators is $\{-0.062, -0.198\}$ under MCAR and $\{-0.142, -0.323\}$ under MAR, respectively. Overall, the benchmark simulation experiment highlights the superior performance of the calibration estimator.

6.2. Extended scenarios

In the benchmark simulation, we found that the list-wise deletion fails under MAR and the parametric approach fails when the model is misspecified. In this section, we make a further inspection of the nonparametric approach and the calibration approach via more involved scenarios. In Section 6.2.1, we change the propensity score function so that the

conditional probability of observing Y_{2i} depends on not only X_i and but also Y_{1i}. In Section 6.2.2, we assume that there are $r = 2$ covariates.

6.2.1. Misspecified missing mechanism

We keep the benchmark set-up except for changing (15) as follows:

$$\pi_2(x_i, y_{1i}) = \Pr(T_{2i} = 1 \,|\, X_i = x_i, Y_{1i} = y_{1i}) = 1/\{1 + \exp(-0.42 + 0.2x_i + 0.2y_{1i})\}. \tag{20}$$

As in Case D of the benchmark case, (20) implies MAR with $\mathrm{E}(T_{2i}) = 0.6$. A unique feature of (20) is that the propensity score depends on not only X_i but also Y_{1i}. This might challenge the nonparametric and calibration approaches when the approximation sieves consist of the power series of X_i only.

Simulation results are shown in the top half of Table 3, where the true copula is Clayton with $\alpha_0 = 6.000$. Results with $C_3(1.636)$, $G_3(1.818)$, and $G_3(4.000)$ are collected in Tables 11–14 of the Online Supplement [17] in order to save space, and they yield qualitatively similar implications. Interestingly, adding Y_{1i} to the propensity score does not have an adverse impact on either the nonparametric estimator or the calibration estimator. When we choose K_2^* among $\{1, \ldots, 5\}$, the two estimators are comparable for each sample size, and the larger sample size naturally leads to sharper inference. When $N = 1000$, the bias is as small as -0.059 for the nonparametric estimator and -0.069 for the calibration estimator. Those results suggest that, in our set-up, the extra impact of Y_{1i} on the propensity score is well captured by the power series of X_i. That is not a surprising result since Y_{1i} and X_i are associated with each other through the copula.

The nonparametric estimator with $K_2 = 4$ has a substantial bias, and its magnitude even increases as sample size grows: -0.378, -0.686, and -0.827 for $N \in \{250, 500, 1000\}$, respectively. The calibration estimator, by contrast, always exhibits stable performance across $K_2 \in \{2, 3, 4\}$, and the bias diminishes as sample size grows. Taking $K_2 = 4$ as an example, the bias of the calibration estimator is $\{-0.218, -0.123, -0.068\}$ for $N \in \{250, 500, 1000\}$. Those results augment the benchmark simulation results that the calibration estimator is more stable than the nonparametric estimator.

In Section 7.3.1 of the Online Supplement [17], we provide further evidence of the high performance of the calibration estimator by comparing it with two more estimators. The first one is the parametric estimator whose propensity score model is correctly specified relative to (20). The second one is the calibration estimator whose approximation sieve consists of power series of X_i and Y_{1i}. Those estimators are expected to have sharp performances by construction. Interestingly, the calibration estimator based on the power series of X_i performs as well as those two competitors in terms of bias and standard deviation. See [17] for more details.

6.2.2. Two covariates

Revisit the benchmark scenario, and let $r = 2$ here in order to check the robustness of the nonparametric and calibration estimators against large dimensionality. Specifically, we add the second covariate X_{2i} whose marginal distribution is the standard Gaussian. We draw $\boldsymbol{U}_i = (Y_{1i}, Y_{2i}, X_{1i}, X_{2i})^\top$ jointly from a copula. The missing mechanism is set to be $\pi_2(\boldsymbol{X}_i) = 1/\{1 + \exp(-0.42 + 0.2X_{1i} + 0.2X_{2i})\}$. It is similar to Case D of the benchmark scenario in that the missing mechanism is MAR with $\mathrm{E}(T_{2i}) = 0.6$. Note, however, that both covariates affect the conditional probability of observing Y_{2i} in the present scenario.

Since there are two covariates, we reconstruct the approximation sieve $u_{K_2}(\boldsymbol{X}_i)$. Define

$$u_{10}(\boldsymbol{X}_i) = (1, X_{1i}, X_{2i}, X_{1i}^2, X_{2i}^2, X_{1i}X_{2i}, X_{1i}^3, X_{2i}^3, X_{1i}^2X_{2i}, X_{1i}X_{2i}^2)^\top.$$

For $K_2 \in \{1, \ldots, 10\}$, let $u_{K_2}(\boldsymbol{X}_i)$ be the first K_2 elements of $u_{10}(\boldsymbol{X}_i)$. We use $K_2 = 3$ (i.e., only the first moments of $\boldsymbol{X}_i$), $K_2 = 6$ (i.e., the second moments added), and $K_2 = 10$ (i.e., the third moments added). We also use the data-driven K_2^* with a choice set $K_2 \in \{1, \ldots, 10\}$.

Simulation results are shown in the bottom half of Table 3, where the true copula is $C_4(6.000)$. Results with $C_4(1.636)$, $G_4(1.818)$, and $G_4(4.000)$ are collected in Tables 15–16 of the Online Supplement [17] in order to save space, and they yield qualitatively similar implications. The nonparametric estimator assisted by the data-driven K_2^* has small bias but large variance. When $N = 1000$, the bias, standard deviation, and RMSE of the nonparametric estimator are $\{-0.108, 1.265, 1.270\}$, respectively. The calibration estimator assisted by the data-driven K_2^*, by contrast, has small bias and remarkably small variance. When $N = 1000$, the bias, standard deviation, and RMSE of the calibration estimator are $\{-0.098, 0.346, 0.360\}$. Those results highlight that the calibration estimator is more robust against multivariate covariates than the nonparametric estimator.

The calibration estimator keeps its high performance when fixed $K_2 \in \{3, 6, 10\}$ are used. The nonparametric estimator, in contrast, is substantially biased when $K_2 = 10$; the bias is $\{-0.929, -1.127, -1.579\}$ for $N \in \{250, 500, 1000\}$. Those results are consistent with the results of the benchmark simulation with a single covariate.

7. Conclusion

Copula models are a useful tool for capturing complex interdependence of multiple variables. Genest et al. [14] proposed the maximum pseudo-likelihood estimator for semiparametric copula models. While there exists a vast literature on copula models, most papers including [5,14] assume complete data. In this article, we propose a new estimator for semiparametric copula models with data missing at random. We extend the maximum likelihood estimator

Table 3
Extra simulation results on Clayton copula with $\alpha_0 = 6.000$ (Kendall's $\tau = 0.75$).

Scenario (1): misspecified missing mechanism			
	$N = 250$ Bias, Stdev, RMSE	$N = 500$ Bias, Stdev, RMSE	$N = 1000$ Bias, Stdev, RMSE
Nonparam ($K_2 = 2$)	−0.231, 0.637, 0.677	−0.121, 0.464, 0.479	−0.049, 0.339, 0.342
Nonparam ($K_2 = 3$)	−0.230, 0.622, 0.663	−0.134, 0.471, 0.490	−0.058, 0.338, 0.343
Nonparam ($K_2 = 4$)	−0.378, 2.650, 2.677	−0.686, 2.031, 2.144	−0.827, 2.395, 2.534
Nonparam (CB)	−0.181, 0.641, 0.666	−0.140, 0.454, 0.475	−0.059, 0.327, 0.332
Calibration ($K_2 = 2$)	−0.290, 0.619, 0.683	−0.195, 0.458, 0.498	−0.125, 0.339, 0.362
Calibration ($K_2 = 3$)	−0.231, 0.641, 0.682	−0.148, 0.452, 0.475	−0.080, 0.313, 0.322
Calibration ($K_2 = 4$)	−0.218, 0.637, 0.674	−0.123, 0.463, 0.479	−0.068, 0.322, 0.330
Calibration (CB)	−0.212, 0.628, 0.663	−0.147, 0.453, 0.476	−0.069, 0.323, 0.330
Scenario (2): two covariates			
	$N = 250$ Bias, Stdev, RMSE	$N = 500$ Bias, Stdev, RMSE	$N = 1000$ Bias, Stdev, RMSE
Nonparam ($K_2 = 3$)	−0.266, 0.626, 0.681	−0.150, 0.461, 0.484	−0.086, 0.324, 0.335
Nonparam ($K_2 = 6$)	−0.252, 0.688, 0.733	−0.160, 0.531, 0.554	−0.153, 0.476, 0.500
Nonparam ($K_2 = 10$)	−0.929, 2.725, 2.879	−1.127, 4.028, 4.183	−1.579, 3.013, 3.402
Nonparam (CB)	−0.318, 0.671, 0.743	−0.106, 1.153, 1.158	−0.108, 1.265, 1.270
Calibration ($K_2 = 3$)	−0.308, 0.655, 0.724	−0.253, 0.458, 0.523	−0.178, 0.330, 0.375
Calibration ($K_2 = 6$)	−0.260, 0.619, 0.671	−0.147, 0.446, 0.470	−0.084, 0.331, 0.342
Calibration ($K_2 = 10$)	−0.265, 0.646, 0.698	−0.143, 0.455, 0.477	−0.076, 0.331, 0.340
Calibration (CB)	−0.301, 0.619, 0.688	−0.200, 0.458, 0.500	−0.098, 0.346, 0.360

In this table we report bias, standard deviation, and root mean squared error (RMSE) of each estimator with respect to $\alpha_0 = 6.000$ after $J = 1000$ Monte Carlo trials. In Scenario (1), the missing mechanism of Y_{2i} is specified as $\pi_2(X_i, Y_{1i}) = 1/\{1+\exp(-0.42+0.2X_i+0.2Y_{1i})\}$. In Scenario (2), there are two covariates $\boldsymbol{X}_i = (X_{1i}, X_{2i})^\top$, and the missing mechanism of Y_{2i} is specified as $\pi_2(\boldsymbol{X}_i) = 1/\{1 + \exp(-0.42 + 0.2X_{1i} + 0.2X_{2i})\}$. In both scenarios, the missing mechanism is MAR with $\mathrm{E}(T_{2i}) = 0.6$. "Nonparam" signifies the nonparametric estimator of Hirano et al. [26]. "Calibration" signifies our proposed calibration estimator. For both estimators, approximation sieves are constructed from the power series of covariate (s). In Scenario (1), the dimension of the approximation sieve is either fixed at $K_2 \in \{2, 3, 4\}$ or automatically selected from $K_2 \in \{1, \ldots, 5\}$ based on the covariate balancing (CB) principle. In Scenario (2), the dimension of the approximation sieve is either fixed at $K_2 \in \{3, 6, 10\}$ or automatically selected from $K_2 \in \{1, \ldots, 10\}$ based on the CB principle.

of Genest et al. [14] by adapting the calibration estimator of Chan et al. [2]. Under the MAR condition, our estimator satisfies consistency and asymptotic normality. We also present a consistent estimator for the asymptotic variance of our estimator. We show via extensive simulations that our proposed estimator dominates the list-wise deletion, parametric estimators, and nonparametric estimators of Hirano et al. [26].

Acknowledgments

We thank two referees, an Associate Editor, and the Editor-in-Chief, Christian Genest, for constructive comments which led to substantial improvement of the paper. In particular, they made an insightful suggestion of adding the simulation studies on the extended scenarios (Section 6.2). We also thank Qu Feng, Eric Ghysels, Jonathan B. Hill, Wei Huang, Toru Kitagawa, Guodong Li, Yasutomo Murasawa, Daisuke Nagakura, and Ke Zhu, seminar participants at Nanyang Technological University, Konan University, the University of Hong Kong, Keio University, Renmin University of China, and the University of North Carolina at Chapel Hill, and conference participants at EcoSta 2018 for helpful comments. The first author, Shigeyuki Hamori, is grateful for the financial support of JSPS, Japan KAKENHI Grant Number 17H00983. The second author, Kaiji Motegi, is grateful for the financial supports of Japan Center for Economic Research and Nihon Hoseigakkai Foundation. The third author, Zheng Zhang, acknowledges the financial support from Renmin University of China through the project 297517501221, and the fund for building world-class universities (disciplines) of Renmin University of China. All authors contributed to the paper equally.

Appendix A. Proof of Theorem 1

For a given sieve basis $u_{K_j}(x)$, we can approximate a function $f_j : \mathbb{R}^r \to \mathbb{R}$ by $\lambda^\top u_{K_j}(x)$. For a non-degenerate matrix A_{K_j}, $\lambda^\top u_{K_j}(\boldsymbol{X}) = \lambda^\top A_{K_j}^{-1} A_{K_j} u_{K_j}(\boldsymbol{X})$ and hence we can also use $\tilde{u}_{K_j}(\boldsymbol{X}) = A_{K_j} u_{K_j}(\boldsymbol{X})$ as a new basis for approximation. Specifically, by choosing $A_{K_j} = \mathrm{E}\{u_{K_j}(\boldsymbol{X}) u_{K_j}^\top(\boldsymbol{X})\}^{-1/2}$, which is non-degenerate by Assumption 3, we can obtain a system of orthonormal sieve basis $\tilde{u}_{K_j}(\boldsymbol{X})$. Without loss of generality, we assume the original sieve basis $u_{K_j}(\boldsymbol{X})$ are orthonormal, viz.

$$\mathrm{E}\{u_{K_j}(\boldsymbol{X}) u_{K_j}^\top(\boldsymbol{X})\} = I_{K_j}. \tag{A.1}$$

Also let $\zeta(K_j) = \sup_{x\in\mathcal{X}} \|u_{K_j}(x)\|$. Above condition and notation are also adopted in [38]. For $j \in \{1, \ldots, d\}$, we define the theoretical counterparts of $\hat{G}_{jK}(\lambda)$, $\hat{\lambda}_{jK}$, and $\hat{p}_{jK}(x)$, viz.

$$G^*_{jK}(\lambda) = \mathrm{E}\{\hat{G}_{jK}(\lambda)\} = \mathrm{E}[\pi_j(\boldsymbol{X}_i)\rho\{\lambda^\top u_{K_j}(\boldsymbol{X}_i)\} - \lambda^\top u_{K_j}(\boldsymbol{X}_i)], \tag{A.2}$$

$$\lambda^*_{jK} = \arg\max_{\lambda\in\mathbb{R}^{K_j}} G^*_{jK}(\lambda), \quad p^*_{jK}(x) = \frac{1}{N}\,\rho'\{(\lambda^*_{jK})^\top u_{K_j}(x)\}.$$

Theorem 1 is an immediate consequence of Lemmas A and B presented below.

A.1. Lemma A

Lemma A gives the approximation rate of the function $1/\pi_j(x)$ by $Np^*_{jK}(x)$.

Lemma A. *Under Assumptions 2–7, we have, for any $j \in \{1, \ldots, d\}$,*

$$\sup_{x\in\mathcal{X}} |Np^*_{jK}(x) - 1/\pi_j(x)| = O\{\zeta(K_j)K_j^{-\alpha}\}, \tag{A.3}$$

$$\int_{\mathcal{X}} |Np^*_{jK}(x) - 1/\pi_j(x)|^2 dF_X(x) = O(K_j^{-2\alpha}), \tag{A.4}$$

$$\frac{1}{N}\sum_{i=1}^{N} |Np^*_{jK}(\boldsymbol{X_i}) - 1/\pi_j(\boldsymbol{X_i})|^2 = O_p(K_j^{-2\alpha}). \tag{A.5}$$

Proof. We first prove (A.3). By Assumptions 4 and 7, $1/\pi_j(x) \in [1, \eta_1]$, and $(\rho')^{-1}$ is strictly decreasing. We can thus define two finite constants, viz.

$$\overline{\gamma} = \sup_{x\in\mathcal{X}}(\rho')^{-1}\{1/\pi_j(x)\} \le (\rho')^{-1}(1), \quad \underline{\gamma} = \inf_{x\in\mathcal{X}}(\rho')^{-1}\{1/\pi_j(x)\} \ge (\rho')^{-1}(\eta_1).$$

By Assumption 5, there exist some constants $C > 0$ and $\lambda_{jK} \in \mathbb{R}^{K_j}$ such that $\sup_{x\in\mathcal{X}} |(\rho')^{-1}\{1/\pi_j(x)\} - \lambda_{jK}^\top u_{K_j}(x)| \le CK_j^{-\alpha}$. Then we have, for all $x \in \mathcal{X}$,

$$\lambda_{jK}^\top u_{K_j}(x) \in \left((\rho')^{-1}\{1/\pi_j(x)\} - CK_j^{-\alpha}, (\rho')^{-1}\{1/\pi_j(x)\} + CK_j^{-\alpha}\right) \subseteq [\underline{\gamma} - CK_j^{-\alpha}, \overline{\gamma} + CK_j^{-\alpha}] \tag{A.6}$$

and for all $x \in \mathcal{X}$,

$$\rho'\{\lambda_{jK}^\top u_{K_j}(x) + CK_j^{-\alpha}\} - \rho'\{\lambda_{jK}^\top u_{K_j}(x)\} < 1/\pi_j(x) - \rho'\{\lambda_{jK}^\top u_{K_j}(x)\} < \rho'\{\lambda_{jK}^\top u_{K_j}(x) - CK_j^{-\alpha}\} - \rho'\{\lambda_{jK}^\top u_{K_j}(x)\}.$$

By the Mean Value Theorem, for large enough K_j, there exist

$$\xi_{j1}(x) \in (\lambda_{jK}^\top u_{K_j}(x), \lambda_{jK}^\top u_{K_j}(x) + CK_j^{-\alpha}) \subseteq [\underline{\gamma} - CK_j^{-\alpha}, \overline{\gamma} + 2CK_j^{-\alpha}] \subseteq \Gamma_1,$$
$$\xi_{j2}(x) \in (\lambda_{jK}^\top u_{K_j}(x) - CK_j^{-\alpha}, \lambda_{jK}^\top u_{K_j}(x)) \subseteq [\underline{\gamma} - 2CK_j^{-\alpha}, \overline{\gamma} + CK_j^{-\alpha}] \subseteq \Gamma_1\ ,$$

such that

$$\rho'\{\lambda_{jK}^\top u_{K_j}(x) + CK_j^{-\alpha}\} - \rho'\{\lambda_{jK}^\top u_{K_j}(x)\} = \rho''\{\xi_{j1}(x)\}CK_j^{-\alpha} \ge -aCK_j^{-\alpha},$$
$$\rho'\{\lambda_{jK}^\top u_{K_j}(x) - CK_j^{-\alpha}\} - \rho'\{\lambda_{jK}^\top u_{K_j}(x)\} = -\rho''\{\xi_{j2}(x)\}CK_j^{-\alpha} \le aCK_j^{-\alpha},$$

where $\Gamma_1 = [\underline{\gamma} - 1, \overline{\gamma} + 1]$ and $a = \sup_{\gamma\in\Gamma_1} \left|\rho''(\gamma)\right| = \sup_{\gamma\in\Gamma_1} \left\{-\rho''(\gamma)\right\}$, because $\rho'' < 0$. Since Γ_1 is compact and independent of x, a is a finite constant. Therefore,

$$\sup_{x\in\mathcal{X}} \left|1/\pi_j(x) - \rho'\{\lambda_{jK}^\top u_{K_j}(x)\}\right| < aCK_j^{-\alpha}. \tag{A.7}$$

By (A.1), (A.2) and (A.7), we can deduce that

$$\begin{aligned}
\left\|(G^*_{jK})'(\lambda_{jK})\right\| &= \left\|\mathrm{E}\left[\pi_j(\boldsymbol{X})[\rho'\{\lambda_{jK}^\top u_{K_j}(\boldsymbol{X})\} - 1/\pi_j(\boldsymbol{X})]u_{K_j}(\boldsymbol{X})\right]\right\| \\
&= \left\|\mathrm{E}\left[\pi_j(\boldsymbol{X})\{\rho'\{\lambda_{jK}^\top u_{K_j}(\boldsymbol{X})\} - 1/\pi_j(\boldsymbol{X})\}u_{K_j}(\boldsymbol{X})^\top\right] \times \mathrm{E}\{u_{K_j}(\boldsymbol{X})u_{K_j}(\boldsymbol{X})^\top\}^{-1}u_{K_j}(\boldsymbol{X})\right\|_{L^2} \\
&\le \left\|\pi_j(\boldsymbol{X})[\rho'\{\lambda_{jK}^\top u_{K_j}(\boldsymbol{X})\} - 1/\pi_j(\boldsymbol{X})]\right\|_{L^2} \le \left\|\rho'\{\lambda_{jK}^\top u_{K_j}(\boldsymbol{X})\} - 1/\pi_j(\boldsymbol{X})\right\|_{L^2} \le aCK_j^{-\alpha},
\end{aligned} \tag{A.8}$$

where the first inequality follows from the fact that

$$\mathrm{E}\left[\pi_j(\boldsymbol{X})[\rho'\{\lambda_{jK}^\top u_{K_j}(\boldsymbol{X})\} - 1/\pi_j(\boldsymbol{X})]u_{K_j}(\boldsymbol{X})^\top\right] \times \mathrm{E}\{u_{K_j}(\boldsymbol{X})u_{K_j}(\boldsymbol{X})^\top\}^{-1}u_{K_j}(\boldsymbol{X})$$

is the $L^2(dF_X)$-projection of $\pi_j(\boldsymbol{X})[\rho'\{\lambda_{jK}^\top u_{K_j}(\boldsymbol{X})\} - 1/\pi_j(\boldsymbol{X})]$ on the space spanned by $u_{K_j}(\boldsymbol{X})$.

For some fixed $C_2 > 0$ (to be chosen later), define

$$\Lambda_{jK} = \{\lambda \in \mathbb{R}^{K_j} : \|\lambda - \lambda_{jK}\| \le C_2 K_j^{-\alpha}\}.$$

For sufficiently large K_j, by (A.6) and Assumption 6, we have, for all $\lambda \in \Lambda_{jK}$ and all $x \in \mathcal{X}$,

$$|\lambda^\top u_{K_j}(x) - \lambda_{jK}^\top u_{K_j}(x)| = |(\lambda - \lambda_{jK})^\top u_{K_j}(x)| \le \|\lambda - \lambda_{jK}\| \times \|u_{K_j}(x)\| \le C_2 K_j^{-\alpha}\zeta(K_j),$$

which implies that for all $\lambda \in \Lambda_{jK}$,

$$\begin{aligned}\lambda^\top u_{K_j}(x) &\in (\lambda_{jK}^\top u_{K_j}(x) - C_2 K_j^{-\alpha}\zeta(K_j), \lambda_{jK}^\top u_{K_j}(x) + C_2 K_j^{-\alpha}\zeta(K_j)) \\ &\subseteq \left[\underline{\gamma} - CK_j^{-\alpha} - C_2 K_j^{-\alpha}\zeta(K_j), \overline{\gamma} + CK_j^{-\alpha} + C_2 K_j^{-\alpha}\zeta(K_j)\right] \subseteq \Gamma_1 . \end{aligned} \tag{A.9}$$

For any $\lambda \in \partial\Lambda_{jK}$, i.e., $\|\lambda - \lambda_{jK}\| = C_2 K_j^{-\alpha}$, using the Mean Value Theorem we can deduce that

$$\begin{aligned} G_{jK}^*(\lambda) - G_{jK}^*(\lambda_{jK}) &= (\lambda - \lambda_{jK})^\top (G_{jK}^*)'(\lambda_{jK}) + (\lambda - \lambda_{jK})^\top (G_{jK}^*)''(\bar{\lambda})(\lambda - \lambda_{jK})/2 \\ &\le \|\lambda - \lambda_{jK}\| \times \|(G_{jK}^*)'(\lambda_{jK})\| + (\lambda - \lambda_{jK})^\top \mathrm{E}[\pi_j(\boldsymbol{X})\rho''\{\bar{\lambda}^\top u_{K_j}(\boldsymbol{X})\}u_{K_j}(\boldsymbol{X})u_{K_j}^\top(\boldsymbol{X})](\lambda - \lambda_{jK})/2 \\ &\le \|\lambda - \lambda_{jK}\| \times \|(G_{jK}^*)'(\lambda_{jK})\| - a_1(\lambda - \lambda_{jK})^\top \mathrm{E}\{u_{K_j}(\boldsymbol{X})u_{K_j}^\top(\boldsymbol{X})\}(\lambda - \lambda_{jK})/(2\eta_1) \\ &\le \|\lambda - \lambda_{jK}\| \times \{aCK_j^{-\alpha} - a_1 C_2 K_j^{-\alpha}/(2\eta_1)\}, \end{aligned}$$

where $\bar{\lambda}$ lies on the line joining from λ to λ_{jK}, $a_1 = \inf_{y\in\Gamma_1}\{-\rho''(y)\} > 0$ is a finite positive constant, and the last inequality follows from (A.8). By choosing $C_2 > 2aC\eta_1/a_1$, we can obtain that $G_{jK}^*(\lambda) < G_{jK}^*(\lambda_{jK})$ for any $\lambda \in \partial\Lambda_{jK}$. Because $G_{jK}^*(\lambda)$ is continuous, there is a local maximum of G_{jK}^* in the interior of Λ_{jK}. Furthermore, G_{jK}^* is a strictly concave function with a unique global maximum point λ_{jK}^*, therefore we can claim $\lambda_{jK}^* \in \Lambda_{jK}^\circ$, i.e.,

$$\|\lambda_{jK}^* - \lambda_{jK}\| \le C_2 K_j^{-\alpha}. \tag{A.10}$$

By the Mean Value Theorem, for large enough K_j, there exists $\xi^*(x)$ lying between $(\lambda_{jK}^*)^\top u_{K_j}(x)$ and $\lambda_{jK}^\top u_{K_j}(x)$. Then we have $\xi^*(x) \in \Gamma_1$ by (A.9) and, for any $x \in \mathcal{X}$,

$$|\rho'\{\lambda_{jK}^\top u_{K_j}(x)\} - \rho'\{(\lambda_{jK}^*)^\top u_{K_j}(x)\}| \le |\rho''\{\xi^*(x)\}| \times \|\lambda_{jK} - \lambda_{jK}^*\| \times \|u_{K_j}(x)\| \le a\, C_2 K_j^{-\alpha}\zeta(K_j). \tag{A.11}$$

Therefore, by (A.7) and (A.11) we have

$$\begin{aligned}\sup_{x\in\mathcal{X}}\left|1/\pi_j(x) - \rho'\{(\lambda_{jK}^*)^\top u_{K_j}(x)\}\right| &\le \sup_{x\in\mathcal{X}}\left|1/\pi_j(x) - \rho'\{\lambda_{jK}^\top u_{K_j}(x)\}\right| + \sup_{x\in\mathcal{X}}\left|\rho'\{\lambda_{jK}^\top u_{K_j}(x)\} - \rho'\{(\lambda_{jK}^*)^\top u_{K_j}(x)\}\right| \\ &\le aCK_j^{-\alpha} + aC_2 K_j^{-\alpha}\zeta(K_j) \le (aC + aC_2)K_j^{-\alpha}\zeta(K_j).\end{aligned}$$

Next, we prove (A.4). Similarly, by (A.1), (A.7), (A.9), and (A.10), we can deduce that

$$\begin{aligned}&\int_{\mathcal{X}} \left|\pi_j(x)^{-1} - Np_{jK}^*(x)\right|^2 dF_X(x) \\ &\le 2\int_{\mathcal{X}} |\pi_j(x)^{-1} - \rho'\{\lambda_{jK}^\top u_{K_j}(x)\}|^2 dF_X(x) + 2\int_{\mathcal{X}} \left|\rho'\{\lambda_{jK}^\top u_{K_j}(x)\} - \rho'\{(\lambda_{jK}^*)^\top u_{K_j}(x)\}\right|^2 dF_X(x) \\ &\le 2\sup_{x\in\mathcal{X}}|\pi_j(x)^{-1} - \rho'\{\lambda_{jK}^\top u_{K_j}(x)\}|^2 + 2\int_{\mathcal{X}} |\rho''\{\xi^*(x)\}|^2 \times |(\lambda_{jK} - \lambda_{jK}^*)^\top u_{K_j}(x)|^2 dF_X(x) \\ &\le 2\sup_{x\in\mathcal{X}}|\pi_j(x)^{-1} - \rho'\{\lambda_{jK}^\top u_{K_j}(x)\}|^2 + 2\sup_{\gamma\in\Gamma_1}|\rho''(\gamma)|^2 \times (\lambda_{jK} - \lambda_{jK}^*)^\top \int_{\mathcal{X}} u_{K_j}(x)u_{K_j}(x)^\top dF_X(x) \times (\lambda_{jK} - \lambda_{jK}^*) \\ &= 2\sup_{x\in\mathcal{X}}|\pi_j(x)^{-1} - \rho'\{\lambda_{jK}^\top u_{K_j}(x)\}|^2 + 2\sup_{\gamma\in\Gamma_1}|\rho''(\gamma)|^2 \times \|\lambda_{jK} - \lambda_{jK}^*\|^2 \\ &= O(K_j^{-2\alpha}) + O(1) \times O(K_j^{-2\alpha}) = O(K_j^{-2\alpha}).\end{aligned}$$

Finally, we prove (A.5). We can also obtain

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^N \left|\pi_j(\boldsymbol{X}_i)^{-1} - Np_{jK}^*(\boldsymbol{X}_i)\right|^2 \\ &\le \frac{2}{N}\sum_{i=1}^N \left|\pi_j(\boldsymbol{X}_i)^{-1} - \rho'\{\lambda_{jK}^\top u_{K_j}(\boldsymbol{X}_i)\}\right|^2 + \frac{2}{N}\sum_{i=1}^N \left|\rho'\{\lambda_{jK}^\top u_{K_j}(\boldsymbol{X}_i)\} - \rho'\{(\lambda_{jK}^*)^\top u_{K_j}(\boldsymbol{X}_i)\}\right|^2 \\ &= \frac{2}{N}\sum_{i=1}^N \left|\pi_j(\boldsymbol{X}_i)^{-1} - \rho'\{\lambda_{jK}^\top u_{K_j}(\boldsymbol{X}_i)\}\right|^2 + \frac{2}{N}\sum_{i=1}^N \left|\rho''\{\xi^*(\boldsymbol{X}_i)\}^\top(\lambda_{jK} - \lambda_{jK}^*)u_{K_j}(\boldsymbol{X}_i)\right|^2\end{aligned}$$

$$\leq 2\sup_{x\in\mathcal{X}}\left|\pi_j(x)^{-1}-\rho'\{\lambda_{jK}^\top u_{K_j}(x)\}\right|^2+2\sup_{\gamma\in\Gamma_1}|\rho''(\gamma)|^2\times(\lambda_{jK}-\lambda_{jK}^*)^\top\left\{\frac{1}{N}\sum_{i=1}^N u_{K_j}(\boldsymbol{X}_i)u_{K_j}(\boldsymbol{X}_i)^\top\right\}(\lambda_{jK}-\lambda_{jK}^*)$$

$$\leq 2\sup_{x\in\mathcal{X}}\left|\pi_j(x)^{-1}-\rho'\left(\lambda_{jK}^\top u_{K_j}(x)\right)\right|^2+2\sup_{\gamma\in\Gamma_1}|\rho''(\gamma)|^2\times\lambda_{\max}\left\{\frac{1}{N}\sum_{i=1}^N u_{K_j}(\boldsymbol{X}_i)u_{K_j}(\boldsymbol{X}_i)^\top\right\}\|\lambda_{jK}-\lambda_{jK}^*\|^2$$

$$=O(K_j^{-2\alpha})+O(1)\times O_p(1)\times O(K_j^{-2\alpha})=O_p(K_j^{-2\alpha}),$$

where $\lambda_{\max}(A)$ denotes the largest eigenvalue of a matrix A; the second equality follows from (A.7), (A.10) and the fact that $\lambda_{\max}\{N^{-1}\sum_{i=1}^N u_K(\boldsymbol{X}_i)u_{K_j}(\boldsymbol{X}_i)^\top\}\xrightarrow{p}\lambda_{\max}\{\mathrm{E}[u_{K_j}(\boldsymbol{X})u_{K_j}(\boldsymbol{X})^\top]\}<\infty$ because

$$\mathrm{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N u_{K_j}(\boldsymbol{X}_i)u_{K_j}(\boldsymbol{X}_i)^\top-\mathrm{E}\{u_{K_j}(\boldsymbol{X})u_{K_j}(\boldsymbol{X})^\top\}\right\|^2\right]=\mathrm{E}\left[\left\|u_{K_j}(\boldsymbol{X})u_{K_j}(\boldsymbol{X})^\top-\mathrm{E}\{u_{K_j}(\boldsymbol{X})u_{K_j}(\boldsymbol{X})^\top\}\right\|^2\right]/N$$

$$=\mathrm{E}\left[\mathrm{tr}\{u_{K_j}(\boldsymbol{X})u_{K_j}(\boldsymbol{X})^\top u_{K_j}(\boldsymbol{X})u_{K_j}(\boldsymbol{X})^\top\}\right]/N-\mathrm{tr}\left[\mathrm{E}\{u_{K_j}(\boldsymbol{X})u_{K_j}(\boldsymbol{X})^\top\}\times\mathrm{E}\{u_{K_j}(\boldsymbol{X})u_{K_j}(\boldsymbol{X})^\top\}\right]/N$$

$$\leq\zeta(K_j)^2\mathrm{E}\{\|u_{K_j}(\boldsymbol{X})\|^2\}/N=\zeta(K_j)^2K_j/N\to 0, \tag{A.12}$$

where the last line is justified by Assumption 6. Thus the proof of Lemma A is complete. □

A.2. *Lemma* B

Lemma B gives the approximation rate of $Np_{jK}^*(x)$ by $N\hat{p}_{jK}(x)$.

Lemma B. *Under Assumptions 2–7, we have, for all $j\in\{1,\ldots,d\}$,*

$$\|\hat{\lambda}_{jK}-\lambda_{jK}^*\|=O_p(\sqrt{K_j/N}), \tag{A.13}$$

$$\sup_{x\in\mathcal{X}}|N\hat{p}_{jK}(x)-Np_{jK}^*(x)|=O_p\{\zeta(K_j)\sqrt{K_j/N}\}, \tag{A.14}$$

$$\int_{\mathcal{X}}|N\hat{p}_{jK}(x)-Np_{jK}^*(x)|^2dF_X(x)=O_p(K_j/N), \tag{A.15}$$

$$\frac{1}{N}\sum_{i=1}^N|N\hat{p}_{jK}(\boldsymbol{X}_i)-Np_{jK}^*(\boldsymbol{X}_i)|^2=O_p(K_j/N). \tag{A.16}$$

Proof. First we prove (A.13). Define $\hat{S}_{jN}=\sum_{i=1}^N T_{ji}u_{K_j}(\boldsymbol{X}_i)u_{K_j}(\boldsymbol{X}_i)^\top/N$. Obviously, $\hat{S}_{jN}$ is a symmetric matrix and $\mathrm{E}(\hat{S}_{jN})=\mathrm{E}\{\pi_j(\boldsymbol{X})u_{K_j}(\boldsymbol{X})u_{K_j}^\top(\boldsymbol{X})\}$. We have

$$\mathrm{E}\left[\|\hat{S}_{jN}-\mathrm{E}\{\pi_j(\boldsymbol{X})u_{K_j}(\boldsymbol{X})u_{K_j}^\top(\boldsymbol{X})\}\|^2\right]$$

$$=\mathrm{tr}\left[\mathrm{E}(\hat{S}_{jN}\hat{S}_{jN})-2\mathrm{E}(\hat{S}_{jN})\times\mathrm{E}\{\pi_j(\boldsymbol{X})u_{K_j}(\boldsymbol{X})u_{K_j}^\top(\boldsymbol{X})\}+\mathrm{E}\{\pi_j(\boldsymbol{X})u_{K_j}(\boldsymbol{X})u_{K_j}^\top(\boldsymbol{X})\}\mathrm{E}\{\pi_j(\boldsymbol{X})u_{K_j}(\boldsymbol{X})u_{K_j}^\top(\boldsymbol{X})\}\right]$$

and the right-hand side can be expanded as follows:

$$\mathrm{tr}\left[\mathrm{E}\left\{\frac{1}{N^2}\sum_{i=1}^N T_{ji}^2u_K(\boldsymbol{X}_i)u_{K_j}^\top(\boldsymbol{X}_i)u_{K_j}(\boldsymbol{X}_i)u_{K_j}^\top(\boldsymbol{X}_i)\right\}+\mathrm{E}\left\{\frac{1}{N^2}\sum_{i,k=1,i\neq k}^N T_{ji}T_{jk}u_{K_j}(\boldsymbol{X}_i)u_{K_j}^\top(\boldsymbol{X}_i)u_{K_j}(\boldsymbol{X}_j)u_{K_j}^\top(\boldsymbol{X}_j)\right\}\right.$$

$$\left.-\mathrm{E}\{\pi_j(\boldsymbol{X})u_{K_j}(\boldsymbol{X})u_{K_j}^\top(\boldsymbol{X})\}\mathrm{E}\{\pi_j(\boldsymbol{X})u_{K_j}(\boldsymbol{X})u_{K_j}^\top(\boldsymbol{X})\}\right]$$

$$=\frac{1}{N}\mathrm{E}\{\pi_j(\boldsymbol{X})u_{K_j}(\boldsymbol{X})^\top u_{K_j}(\boldsymbol{X})u_{K_j}(\boldsymbol{X})^\top u_{K_j}(\boldsymbol{X})\}-\frac{1}{N}\mathrm{tr}\left[\mathrm{E}\{\pi_j(\boldsymbol{X})u_{K_j}(\boldsymbol{X})u_{K_j}^\top(\boldsymbol{X})\}\times\mathrm{E}\{\pi_j(\boldsymbol{X})u_{K_j}(\boldsymbol{X})u_{K_j}^\top(\boldsymbol{X})\}\right].$$

Therefore,

$$\mathrm{E}\left[\|\hat{S}_{jN}-\mathrm{E}\{\pi_j(\boldsymbol{X})u_{K_j}(\boldsymbol{X})u_{K_j}^\top(\boldsymbol{X})\}\|^2\right]\leq\zeta(K_j)^2K_j/N, \tag{A.17}$$

given that $u_{K_j}^\top(\boldsymbol{X})u_{K_j}(\boldsymbol{X})=\|u_{K_j}(\boldsymbol{X})\|^2\leq\zeta(K_j)^2$, $1/\eta_1<\pi_j(x)<1$, and $\mathrm{E}\{u_{K_j}^\top(\boldsymbol{X})u_{K_j}(\boldsymbol{X})\}=K_j$. Consider the event

$$E_{jN}=\left\{(\lambda-\lambda_{jK}^*)^\top\hat{S}_{jN}(\lambda-\lambda_{jK}^*)>(\lambda-\lambda_{jK}^*)^\top\left[\mathrm{E}\{\pi_j(\boldsymbol{X})u_{K_j}(\boldsymbol{X})u_{K_j}^\top(\boldsymbol{X})\}-\frac{1}{2\eta_1}I_{K_j}\right](\lambda-\lambda_{jK}^*),\lambda\neq\lambda_{jK}^*\right\}.$$

By Chebyshev's inequality and (A.17), we have

$$\Pr\left[\left|(\lambda-\lambda_{jK}^*)^\top\hat{S}_{jN}(\lambda-\lambda_{jK}^*)-(\lambda-\lambda_{jK}^*)^\top\mathrm{E}\{\pi_j(\boldsymbol{X})u_{K_j}(\boldsymbol{X})u_{K_j}^\top(\boldsymbol{X})\}(\lambda-\lambda_{jK}^*)\right|\geq\frac{1}{2\eta_1}\|\lambda-\lambda_{jK}^*\|^2,\lambda\neq\lambda_{jK}^*\right]$$
$$\leq 4\eta_1^2\|\lambda-\lambda_{jK}^*\|^4\mathrm{E}\left[\|\hat{S}_{jN}-\mathrm{E}\{\pi_j(\boldsymbol{X})u_{K_j}(\boldsymbol{X})u_{K_j}^\top(\boldsymbol{X})\}\|^2\right]/\|\lambda-\lambda_{jK}^*\|^4\leq 4\eta_1^2\zeta(K_j)^2K_j/N.$$

Hence, for any $\epsilon>0$, there exists $N_0(\epsilon)\in\mathbb{N}$ such that for $N>N_0(\epsilon)$ large enough

$$\Pr\{(E_{jN})^\complement\}<\epsilon/2. \tag{A.18}$$

By definition, λ_{jK}^* is the unique maximizer of $G_{jK}^*(\lambda)$. Thus,

$$(G_{jK}^*)'(\lambda_{jK}^*)=\mathrm{E}\big[[\pi_j(\boldsymbol{X})\rho'\{(\lambda_{jK}^*)^\top u_{K_j}(\boldsymbol{X})\}-1]u_{K_j}(\boldsymbol{X})\big]=\mathrm{E}\big[[T_{ji}\rho'\{(\lambda_{jK}^*)^\top u_{K_j}(\boldsymbol{X}_i)\}-1]u_{K_j}(\boldsymbol{X})\big]=0.$$

Note that $\hat{G}_{jK}'(\lambda_{jK}^*)=\sum_{i=1}^N[T_{ji}\rho'\{(\lambda_{jK}^*)^\top u_{K_j}(\boldsymbol{X}_i)\}-1]u_{K_j}(\boldsymbol{X}_i)/N$. Then for large K_j we have

$$\begin{aligned}\mathrm{E}\{\|\hat{G}_{jK}'(\lambda_{jK}^*)\|^2\}&=\mathrm{E}\{\hat{G}_{jK}'(\lambda_{jK}^*)^\top\hat{G}_{jK}(\lambda_{jK}^*)\}=\frac{1}{N}\mathrm{E}\big[[T_{ji}\rho'\{(\lambda_{jK}^*)^\top u_{K_j}(\boldsymbol{X}_i)\}-1]^2u_K(\boldsymbol{X}_i)^\top u_{K_j}(\boldsymbol{X}_i)\big]+\frac{N-1}{N}\times 0\\&\leq\frac{1}{N}\mathrm{E}\left[[2+2|\rho'\{(\lambda_{jK}^*)^\top u_{K_j}(\boldsymbol{X}_i)\}|^2]u_{K_j}(\boldsymbol{X}_i)^\top u_{K_j}(\boldsymbol{X}_i)\right]\\&\leq C_4^2\,\mathrm{E}\{u_{K_j}(\boldsymbol{X}_i)^\top u_{K_j}(\boldsymbol{X}_i)\}/N=C_4^2\times K_j/N,\end{aligned} \tag{A.19}$$

where the second equality holds since $\mathrm{E}\{\hat{G}_{jK}'(\lambda_{jK}^*)\}=(G_{jK}^*)'(\lambda_{jK}^*)=0$ and $C_4=\{2\sup_{\gamma\in\Gamma_1}|\rho'(\gamma)|^2+2\}^{1/2}$ is a finite constant.

Let $\epsilon>0$, fix $C_5(\epsilon)>0$ (to be chosen later), and define

$$\hat{\Lambda}_{jK}(\epsilon)=\{\lambda\in\mathbb{R}^{K_j}:\|\lambda-\lambda_{jK}^*\|\leq C_5(\epsilon)C_4\sqrt{K_j/N}\}.$$

For all $\lambda\in\hat{\Lambda}_{jK}(\epsilon)$ and all $x\in\mathcal{X}$, and large enough N, by (A.9) and Assumption 6 we have

$$\begin{aligned}&|\lambda^\top u_{K_j}(x)-(\lambda_{jK}^*)^\top u_{K_j}(x)|\leq\|\lambda-\lambda_{jK}^*\|\times\|u_{K_j}(x)\|\leq C_5(\epsilon)C_4\sqrt{K_j/N}\,\zeta(K_j)\quad\Rightarrow\\&\lambda^\top u_{K_j}(x)\in\left[(\lambda_{jK}^*)^\top u_{K_j}(x)-C_5(\epsilon)C_4\zeta(K_j)\sqrt{K_j/N},\ (\lambda_{jK}^*)^\top u_{K_j}(x)+C_5(\epsilon)C_4\zeta(K_j)\sqrt{K_j/N}\right]\\&\subseteq\left[\underline{\gamma}-CK_j^{-\alpha}-C_2K_j^{-\alpha}\zeta(K_j)-C_5(\epsilon)C_4\zeta(K_j)\sqrt{K_j/N},\overline{\gamma}+CK_j^{-\alpha}+C_2K_j^{-\alpha}\zeta(K_j)+C_5(\epsilon)C_4\zeta(K_j)\sqrt{K_j/N}\right]\\&\subseteq\Gamma_2(\epsilon),\end{aligned} \tag{A.20}$$

where $\Gamma_2(\epsilon)=[\underline{\gamma}-1-C_5(\epsilon),\overline{\gamma}+1+C_5(\epsilon)]$ is a compact set and independent of x.

By the Mean Value Theorem, for any $\lambda\in\partial\hat{\Lambda}_{jK}(\epsilon)$, there exists $\bar{\lambda}$ that lies on the line joining from λ to λ_{jK}^* such that

$$\hat{G}_{jK}(\lambda)=\hat{G}_{jK}(\lambda_{jK}^*)+(\lambda-\lambda_{jK}^*)^\top\hat{G}_{jK}'(\lambda_{jK}^*)+(\lambda-\lambda_{jK}^*)^\top\hat{G}_{jK}''(\bar{\lambda})(\lambda-\lambda_{jK}^*)/2. \tag{A.21}$$

Considering the second order term of (A.21), by (A.20) we have, for large enough N,

$$\begin{aligned}(\lambda-\lambda_{jK}^*)^\top\hat{G}_{jK}''(\bar{\lambda})(\lambda-\lambda_{jK}^*)&=\frac{1}{N}\sum_{i=1}^N T_{ji}\rho''\{\bar{\lambda}^\top u_{K_j}(\boldsymbol{X}_i)\}(\lambda-\lambda_K^*)^\top u_{K_j}(\boldsymbol{X}_i)u_{K_j}(\boldsymbol{X}_i)^\top(\lambda-\lambda_{jK}^*)\\&\leq-\bar{b}(\epsilon)\frac{1}{N}\sum_{i=1}^N T_{ji}(\lambda-\lambda_{jK}^*)^\top u_{K_j}(\boldsymbol{X}_i)u_{K_j}(\boldsymbol{X}_i)^\top(\lambda-\lambda_{jK}^*)=-\bar{b}(\epsilon)\times(\lambda-\lambda_{jK}^*)^\top\hat{S}_{jN}(\lambda-\lambda_{jK}^*),\end{aligned} \tag{A.22}$$

where $-\bar{b}(\epsilon)=\sup_{\gamma\in\Gamma_2(\epsilon)}\rho''(\gamma)$ is negative and finite since $\Gamma_2(\epsilon)$ is a compact set and ρ is a concave function. Then on the event $E_{j,N}$ with large enough N, we have, for all $\lambda\in\partial\hat{\Lambda}_{jK}(\epsilon)$,

$$\begin{aligned}\hat{G}_{jK}(\lambda)-\hat{G}_{jK}(\lambda_{jK}^*)&=(\lambda-\lambda_{jK}^*)^\top\hat{G}_{jK}'(\lambda_{jK}^*)+(\lambda-\lambda_{jK}^*)^\top\hat{G}_{jK}''(\bar{\lambda})(\lambda-\lambda_{jK}^*)/2\\&\leq\|\lambda-\lambda_{jK}^*\|\times\|\hat{G}_{jK}'(\lambda_{jK}^*)\|-\bar{b}(\epsilon)(\lambda-\lambda_{jK}^*)^\top\hat{S}_{jN}(\lambda-\lambda_{jK}^*)/2\\&\leq\|\lambda-\lambda_{jK}^*\|\times\|\hat{G}_{jK}'(\lambda_{jK}^*)\|-\bar{b}(\epsilon)(\lambda-\lambda_{jK}^*)^\top\left[\mathrm{E}\{\pi_j(\boldsymbol{X})u_{K_j}(\boldsymbol{X})u_{K_j}^\top(\boldsymbol{X})\}-\frac{1}{2\eta_1}I_{K_j}\right](\lambda-\lambda_{jK}^*)/2\\&\leq\|\lambda-\lambda_{jK}^*\|\times\|\hat{G}_{jK}'(\lambda_{jK}^*)\|-\bar{b}(\epsilon)(\lambda-\lambda_{jK}^*)^\top\left(\frac{1}{\eta_1}I_{K_j}-\frac{1}{2\eta_1}I_K\right)(\lambda-\lambda_{jK}^*)/2\end{aligned}$$

$$= \|\lambda - \lambda^*_{jK}\| \left\{ \|\hat{G}'_{jK}(\lambda^*_{jK})\| - \frac{\bar{b}(\epsilon)}{4\eta_1}\|\lambda - \lambda^*_{jK}\| \right\}, \tag{A.23}$$

where the first inequality follows from (A.22). By Chebyshev's inequality and (A.19), for sufficiently large N,

$$\Pr\left\{ \|\hat{G}'_{jK}(\lambda^*_{jK})\| \geq \frac{\bar{b}(\epsilon)}{4\eta_1}\|\lambda - \lambda^*_{jK}\| \right\} \leq \frac{16\eta_1^2}{\bar{b}(\epsilon)^2 C_5^2(\epsilon)} \leq \epsilon/2, \tag{A.24}$$

where the last inequality holds by choosing $C_5(\epsilon) \geq 4\eta_1 b(\epsilon)^{-1}\sqrt{2/\epsilon}$. Therefore, for sufficiently large N, by (A.18) and (A.24) we can derive

$$\Pr\left\{ (E_{jN})^{\complement} \text{ or } \|\hat{G}'_{jK}(\lambda^*_{jK})\| \geq \frac{\bar{b}(\epsilon)}{4\eta_1}\|\lambda - \lambda^*_{jK}\| \right\} \leq \epsilon/2 + \epsilon/2 = \epsilon \Rightarrow$$
$$\Pr\left\{ E_{jN} \text{ and } \|\hat{G}'_{jK}(\lambda^*_{jK})\| < \frac{\bar{b}(\epsilon)}{4\eta_1}\|\lambda - \lambda^*_{jK}\| \right\} > 1 - \epsilon. \tag{A.25}$$

Then by (A.23) and (A.25), for sufficiently large N we can get

$$\Pr\{\forall_{\lambda \in \partial \hat{\Lambda}_{jK}} \hat{G}_{jK}(\lambda) - \hat{G}_{jK}(\lambda^*_{jK}) < 0\} \geq 1 - \epsilon\ .$$

Note that the event $\{\forall_{\lambda \in \partial \hat{\Lambda}_{jK}(\epsilon)} \hat{G}_{jK}(\lambda^*_{jK}) > \hat{G}_{jK}(\lambda)\}$ implies that there exists a local maximum point in the interior of $\hat{\Lambda}_{jK}(\epsilon)$. Furthermore, the function $\hat{G}_{jK}$ is strictly concave and $\hat{\lambda}_{jK}$ is the unique maximizer of $\hat{G}_{jK}$, then we get

$$\Pr\{\hat{\lambda}_{jK} \in \hat{\Lambda}_{jK}(\epsilon)\} > 1 - \epsilon\ , \tag{A.26}$$

which implies (A.13).

Next we prove (A.14). By the Mean Value Theorem, we have

$$N\hat{p}_{jK}(x) - Np^*_{jK}(x) = \rho'\{\hat{\lambda}_{jK}^\top u_{K_j}(x)\} - \rho'\{(\lambda^*_{jK})^\top u_{K_j}(x)\} = \rho''\{\tilde{\lambda}_{jK}^\top u_{K_j}(x)\}(\hat{\lambda}_{jK} - \lambda^*_{jK})^\top u_{K_j}(x),$$

where $\tilde{\lambda}_{jK}$ lies on the line joining $\hat{\lambda}_{jK}$ and λ^*_{jK}. From (A.20) and (A.26), we have

$$\sup_{x\in\mathcal{X}} |\rho''\{\tilde{\lambda}_{jK}^\top u_{K_j}(x)\}| = O_p(1). \tag{A.27}$$

Hence we can obtain that

$$\sup_{x\in\mathcal{X}} |N\hat{p}_{jK}(x) - Np^*_{jK}(x)| \leq \sup_{x\in\mathcal{X}} |\rho''\{\tilde{\lambda}_{jK}^\top u_{K_j}(x)\}| \times \|\hat{\lambda}_{jK} - \lambda^*_{jK}\| \times \sup_{x\in\mathcal{X}} \|u_{K_j}(x)\| = O_p\{\zeta(K_j)\sqrt{K_j/N}\}.$$

Next, we prove (A.15). By the Mean Value Theorem, (A.1) and (A.27), we have

$$\begin{aligned}
\int_{\mathcal{X}} |N\hat{p}_{jK}(x) - Np^*_{jK}(x)|^2 dF_X(x) &= \int_{\mathcal{X}} |\rho''\{\tilde{\lambda}_{jK}^\top u_{K_j}(x)\} \times (\hat{\lambda}_{jK} - \lambda^*_{jK})^\top u_{K_j}(x)|^2 dF_X(x) \\
&\leq \sup_{x\in\mathcal{X}} |\rho''\{\tilde{\lambda}_{jK}^\top u_{K_j}(x)\}|^2 \times (\hat{\lambda}_{jK} - \lambda^*_{jK})^\top \times \int_{\mathcal{X}} u_{K_j}(x)u_{K_j}(x)^\top dF_X(x) \times (\hat{\lambda}_{jK} - \lambda^*_{jK}) \\
&= \sup_{x\in\mathcal{X}} |\rho''\{\tilde{\lambda}_{jK}^\top u_{K_j}(x)\}|^2 \times \|\hat{\lambda}_{jK} - \lambda^*_{jK}\|^2 = O_p(1) \times O_p(K_j/N) = O_p(K_j/N).
\end{aligned}$$

Finally, we prove (A.16). By the Mean Value Theorem, (A.12) and (A.27), we have

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^N |N\hat{p}_{jK}(\boldsymbol{X}_i) - Np^*_{jK}(\boldsymbol{X}_i)|^2 &= \frac{1}{N}\sum_{i=1}^N |\rho''\{\tilde{\lambda}_{jK}^\top u_{K_j}(\boldsymbol{X}_i)\} \times (\hat{\lambda}_{jK} - \lambda^*_{jK})^\top u_{K_j}(\boldsymbol{X}_i)| \\
&\leq \sup_{x\in\mathcal{X}} \left|\rho''\{\tilde{\lambda}_{jK}^\top u_{K_j}(x)\}\right|^2 \times (\hat{\lambda}_{jK} - \lambda^*_{jK})^\top \times \left\{ \frac{1}{N}\sum_{i=1}^N u_{K_j}(\boldsymbol{X}_i)u_{K_j}(\boldsymbol{X}_i)^\top \right\} (\hat{\lambda}_{jK} - \lambda^*_{jK}) \\
&\leq \sup_{x\in\mathcal{X}} |\rho''\{\tilde{\lambda}_{jK}^\top u_{K_j}(x)\}|^2 \times \|\hat{\lambda}_{jK} - \lambda^*_{jK}\|^2 \times \lambda_{\max}\left\{ \frac{1}{N}\sum_{i=1}^N u_{K_j}(\boldsymbol{X}_i)u_{K_j}(\boldsymbol{X}_i)^\top \right\} \\
&\leq O_p(1) \times O_p(K_j/N) \times O_p(1) = O_p(K_j/N).
\end{aligned}$$

This completes the proof of Lemma B. □

Appendix B. Sketched proof of Theorem 2

See Section 4 of the Online Supplement [17] for a complete proof of Theorem 2. In this section, we present an asymptotic expansion of $\sqrt{N}\{\hat{F}_j(y) - F_j^0(y) - \sum_{i=1}^N \psi_j(Y_{ji}, \boldsymbol{X}_i, T_{ji}; y)/N\}$, which is a key step in our proof. We begin with introducing some notation:

$$\begin{aligned}
\Sigma_{jK} &= \mathrm{E}\left[\pi_j(\boldsymbol{X})\rho''\{(\lambda_{jK}^*)^\top u_{K_j}(\boldsymbol{X})\}u_{K_j}(\boldsymbol{X})u_K^\top(\boldsymbol{X})\right], \\
\Psi_{jK}(y) &= -\mathrm{E}\left[F_j(y|\boldsymbol{X}_i)\pi_j(\boldsymbol{X}_i)\rho''\{(\lambda_{jK}^*)^\top u_{K_j}(\boldsymbol{X}_i)\}u_{K_j}(\boldsymbol{X}_i)\right], \\
Q_{jK}(x,y) &= \Psi_{jK}(y)^\top \Sigma_{jK}^{-1} \times u_{K_j}(x).
\end{aligned}$$

Note that $Q_{jK}(x,y)$ is a weighted L^2-projection (with respect to the weighted measure $-\rho''\{(\lambda_{jK}^*)^\top u_{K_j}(x)\}dF_X(x)$ of $-F_j(y|x)$ onto the space linearly spanned by $u_{K_j}(x)$. We also define

$$\begin{aligned}
\tilde{\Sigma}_{jK} &= \frac{1}{N}\sum_{i=1}^N T_{ji}\rho''\{(\tilde{\lambda}_{jK})^\top u_{K_j}(\boldsymbol{X}_i)\}u_{K_j}(\boldsymbol{X}_i)u_{K_j}^\top(\boldsymbol{X}_i), \\
\tilde{\Psi}_{jK}(y) &= -\int_{\mathcal{X}} F_j(y|x) \times \pi_j(x) \times \rho''\{\tilde{\lambda}_{jK}^\top u_{K_j}(x)\}u_{K_j}(x)dF_X(x), \\
\tilde{Q}_{jK}(x,y) &= \tilde{\Psi}_{jK}(y)^\top \tilde{\Sigma}_{jK} u_{K_j}(x),
\end{aligned}$$

where $\tilde{\lambda}_{jK}$ lies on the line joining λ_{jK}^* and $\hat{\lambda}_{jK}$ such that the Mean Value Theorem holds:

$$\begin{aligned}
0 &= \frac{1}{N}\sum_{i=1}^N T_{ji}\rho'\{\hat{\lambda}_{jK}^\top u_{K_j}(\boldsymbol{X}_i)\}u_{K_j}(\boldsymbol{X}_i) - \frac{1}{N}\sum_{i=1}^N u_{K_j}(\boldsymbol{X}_i) \\
&= \frac{1}{N}\sum_{i=1}^N T_{ji}\rho'\{(\lambda_{jK}^*)^\top u_{K_j}(\boldsymbol{X}_i)\}u_{K_j}(\boldsymbol{X}_i) - \frac{1}{N}\sum_{i=1}^N u_{K_j}(\boldsymbol{X}_i) + \frac{1}{N}\sum_{i=1}^N T_{ji}\rho''\{\tilde{\lambda}_{jK}^\top u_{K_j}(\boldsymbol{X}_i)\}u_{K_j}(\boldsymbol{X}_i)u_{K_j}(\boldsymbol{X}_i)^\top \times (\hat{\lambda}_{jK} - \lambda_{jK}).
\end{aligned}$$

For each $j \in \{1, \ldots, d\}$, a key decomposition holds as follows:

$$\begin{aligned}
&\sqrt{N}\left[\hat{F}_j(y) - F_j^0(y) - \frac{1}{N}\sum_{i=1}^N \psi_j(Y_{ji}, \boldsymbol{X}_i, T_{ji}; y)\right\} \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^N \left\{NT_{ji}\hat{p}_{jK}(\boldsymbol{X}_i)\mathbf{1}(Y_{ji} \le y) - F_j^0(y) - \left[\frac{T_{ji}}{\pi_j(\boldsymbol{X}_i)}\mathbf{1}(Y_{ji} \le y) - F_j(y|\boldsymbol{X}_i)\left\{\frac{T_{ji}}{\pi_j(\boldsymbol{X}_i)} - 1\right\} - F_j^0(y)\right]\right] \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^N \left[\{N\hat{p}_{jK}(\boldsymbol{X}_i) - Np_{jK}^*(\boldsymbol{X}_i)\}T_{ji}\mathbf{1}(Y_{ji} \le y) - \int_{\mathcal{X}} F_j(y|x)\pi_j(x)\{N\hat{p}_{jK}(x) - Np_{jK}^*(x)\}dF_X(x)\right] \qquad \text{(B.1)} \\
&\quad + \frac{1}{\sqrt{N}}\sum_{i=1}^N \left[\{Np_{jK}^*(\boldsymbol{X}_i) - 1/\pi_j(\boldsymbol{X}_i)\}T_{ji}\mathbf{1}(Y_{ji} \le y) - \mathrm{E}\left[F_j(y|\boldsymbol{X}_i)\pi_j(\boldsymbol{X}_i)\{Np_{jK}^*(\boldsymbol{X}_i) - 1/\pi_j(\boldsymbol{X}_i)\}\right]\right] \qquad \text{(B.2)} \\
&\qquad\qquad + \sqrt{N}\,\mathrm{E}\left[F_j(y|\boldsymbol{X}_i)\pi_j(\boldsymbol{X}_i)\{Np_{jK}^*(\boldsymbol{X}_i) - 1/\pi_j(\boldsymbol{X}_i)\}\right] \qquad \text{(B.3)} \\
&\quad + \sqrt{N}\left[\int_{\mathcal{X}} F_j(y|x)\pi_j(x)\{N\hat{p}_{jK}(x) - Np_{jK}^*(x)\}dF_X(x) - \frac{1}{N}\sum_{i=1}^N [T_{ji}\rho'\{(\lambda_{jK}^*)^\top u_{K_j}(\boldsymbol{X}_i)\} - 1]\tilde{Q}_{jK}(\boldsymbol{X}_i, y)\right] \qquad \text{(B.4)} \\
&\quad + \frac{1}{\sqrt{N}}\sum_{i=1}^N [T_{ji}\rho'\{(\lambda_{jK}^*)^\top u_{K_j}(\boldsymbol{X}_i)\} - 1]\{\tilde{Q}_{jK}(\boldsymbol{X}_i, y) - Q_{jK}(\boldsymbol{X}_i, y)\} \qquad \text{(B.5)} \\
&\quad + \frac{1}{\sqrt{N}}\sum_{i=1}^N \left[[T_{ji}\rho'\{(\lambda_{jK}^*)^\top u_{K_j}(\boldsymbol{X}_i)\} - 1]Q_{jK}(\boldsymbol{X}_i, y) + F_j(y|\boldsymbol{X}_i)\{T_{ji}/\pi_j(\boldsymbol{X}_i) - 1\}\right]. \qquad \text{(B.6)}
\end{aligned}$$

We will show the terms (B.1)–(B.6) are of $o_p(1)$ uniformly in $y \in \mathbb{R}$. A core part of the proof is to show that the term (B.6) is $o_p(1)$, as it links all the unknown functions, i.e., $\pi_j(x)$ and $F_j(y|x)$, with the calibration weights and balancing moment conditions. This is done by a weighted projection argument. Then, by Lemma A.2 in [12], the collection $\{\psi_j(Y_{ji}, \boldsymbol{X}_i, T_{ji}; y) : y \in \mathbb{R}\}$ is a Donsker class for each $j \in \{1, \ldots, d\}$. The Cartesian product of d Donsker classes of functions is also a Donsker class; see p. 270 in [54]. Hence, by Donsker's Theorem or the Functional Central Limit Theorem, Theorem 2 holds. □

Appendix C. Proof of Theorem 4

Using Assumption 1, the definitions $U_{ji} = F^0(Y_{ji})$ and $\eta(\boldsymbol{X}_i) = \Pr(T_{1i} = \cdots = T_{di} = 1|\boldsymbol{X}_i)$, we can write the true copula C^0 as follows:

$$\begin{aligned}
C^0(u_1, \ldots, u_d) &= \mathrm{E}\left[\mathbf{1}\{F_1^0(Y_{1i}) \le u_1, \ldots, F_d^0(Y_{di}) \le u_d\}\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\mathbf{1}\{F_1^0(Y_{1i}) \le u_1, \ldots, F_d^0(Y_{di}) \le u_d\}|\boldsymbol{X}_i\right]\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1)/\eta(\boldsymbol{X}_i)|\boldsymbol{X}_i\right] \times \mathrm{E}\left[\mathbf{1}\{F_1^0(Y_{1i}) \le u_1, \ldots, F_d^0(Y_{di}) \le u_d\}|\boldsymbol{X}_i\right]\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1)\mathbf{1}\{F_1^0(Y_{1i}) \le u_1, \ldots, F_d^0(Y_{di}) \le u_d\}/\eta(\boldsymbol{X}_i)|\boldsymbol{X}_i\right]\right] \\
&= \mathrm{E}\left[\mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1)\mathbf{1}\{F_1^0(Y_{1i}) \le u_1, \ldots, F_d^0(Y_{di}) \le u_d\}/\eta(\boldsymbol{X}_i)\right] \\
&= \mathrm{E}\left[\mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1)\mathbf{1}\{U_{1i} \le u_1, \ldots, U_{di} \le u_d\}/\eta(\boldsymbol{X}_i)\right].
\end{aligned}$$

By Theorem 3, $N\hat{q}_K(x)$ is a consistent estimator for $\eta(x)$. Then we define the empirical copula

$$\hat{C}(u_1, \ldots, u_d) = \sum_{i=1}^{N} \mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1)\hat{q}_K(\boldsymbol{X}_i)\mathbf{1}\{\hat{F}_1(Y_{1i}) \le u_1, \ldots, \hat{F}_d(Y_{di}) \le u_d\}.$$

Recall the definitions of $\hat{\theta}$ and θ_0, viz.

$$\hat{\theta} = \arg\max_{\theta\in\Theta}\left\{\int_{[0,1]^d} \ell(u_1, \ldots, u_d; \theta)d\hat{C}(u_1, \ldots, u_d)\right\}, \quad \theta_0 = \arg\max_{\theta\in\Theta}\left\{\int_{[0,1]^d} \ell(u_1, \ldots, u_d; \theta)dC^0(u_1, \ldots, u_d)\right\}.$$

By Theorem 2, $\sqrt{N}(\hat{F}_j - F_j)$ weakly converges to a Gaussian process in $L^\infty([0, 1])$. Then, in light of Lemma 1(a) in [5], $\sqrt{N}(\hat{C} - C^0)$ weakly converges to a Gaussian process in $L^\infty([0, 1]^d)$. By Example 19.8 in [54], Assumption 13 implies the bracketing number $N_{[\times]}\{\delta, \mathcal{H}, L_1(C^0)\} < \infty$ for the class of functions $\mathcal{H} = \{\ell(u_1, \ldots, u_d; \theta) : \theta \in \Theta\}$. Then by Lemma 1(c) in [5], we obtain

$$\sup_{\theta\in\Theta}\left|\int_{[0,1]^d} \ell(u_1, \ldots, u_d; \theta)d\hat{C}(u_1, \ldots, u_d) - \int_{[0,1]^d} \ell(u_1, \ldots, u_d; \theta)dC^0(u_1, \ldots, u_d)\right| \xrightarrow{p} 0,$$

which implies that $\hat{\theta} \xrightarrow{p} \theta_0$. This completes the proof of Theorem 4. □

Appendix D. Proof of Theorem 5

By the definition of $\hat{\theta}$ and the Mean Value Theorem, we have

$$\begin{aligned}
0 &= \sqrt{N}\sum_{i=1}^{N} \mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1)\hat{q}_K(\boldsymbol{X}_i)\ell_\theta\{\hat{F}_1(Y_{1i}), \ldots, \hat{F}_d(Y_{di}); \hat{\theta}\} \\
&= \sqrt{N}\sum_{i=1}^{N} \mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1)\hat{q}_K(\boldsymbol{X}_i)\ell_\theta\{\hat{F}_1(Y_{1i}), \ldots, \hat{F}_d(Y_{di}); \theta_0\} \qquad \text{(D.1)} \\
&\quad + \sum_{i=1}^{N} \mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1)\hat{q}_K(\boldsymbol{X}_i)\ell_{\theta\theta}\{\hat{F}_1(Y_{1i}), \ldots, \hat{F}_d(Y_{di}); \bar{\theta}\} \times \sqrt{N}(\hat{\theta} - \theta_0), \qquad \text{(D.2)}
\end{aligned}$$

where $\bar{\theta}$ lies on the line joining $\hat{\theta}$ and θ_0. Concerning the term (D.2), two facts hold:

(a) by Example 19.8 in [54], Assumption 16 implies the bracketing number $N_{[\times]}\{\delta, \mathcal{H}, L_1(C^0)\} < \infty$ for the class of functions $\mathcal{H} = \{\ell_{\theta\theta}(u_1, \ldots, u_d; \theta) : \theta \in \Theta, \|\theta - \theta_0\| = o(1)\}$.
(b) $\sqrt{N}(\hat{C} - C^0)$ converges weakly to a Gaussian process in $\ell^\infty([0, 1]^d)$.

Combining the facts (a) and (b) and Lemma 1(c) in [5], we can conclude that

$$\begin{aligned}
&\sup_{\theta\in\Theta:\|\theta-\theta_0\|=o(1)}\left\|\sum_{i=1}^{N} \mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1)\hat{q}_K(\boldsymbol{X}_i)\ell_{\theta\theta}\{\hat{F}_1(Y_{1i}), \ldots, \hat{F}_d(Y_{di}); \theta\} - \mathrm{E}\{\ell_{\theta\theta}(U_{1i}, \ldots, U_{di}; \theta_0)\}\right\| \\
&= \sup_{\theta\in\Theta:\|\theta-\theta_0\|=o(1)}\left\|\int_{\boldsymbol{u}\in[0,1]^d} \ell_{\theta\theta}(\boldsymbol{u}; \theta)d\{\hat{C}(\boldsymbol{u}) - C^0(\boldsymbol{u})\}\right\| \xrightarrow{p} 0, \qquad \text{(D.3)}
\end{aligned}$$

where $\boldsymbol{u} = (u_1, \ldots, u_d)^\top \in [0, 1]^d$. Combining (D.3) with the consistency result $\|\bar{\theta} - \theta_0\| \le \|\hat{\theta} - \theta_0\| \xrightarrow{p} 0$, we find that

$$\sum_{i=1}^{N} \mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1)\hat{q}_K(\boldsymbol{X}_i)\ell_{\theta\theta}\{\hat{F}_1(Y_{1i}), \ldots, \hat{F}_d(Y_{di}); \bar{\theta}\} = -B + o_p(1), \qquad \text{(D.4)}$$

where $B = -\mathrm{E}\{\ell_{\theta\theta}(U_{1i}, \ldots, U_{di}; \theta_0)\}$ with $U_{ji} = F_j^0(Y_{ji})$.

We next consider the term (D.1). Since $\mathrm{E}\{\ell_\theta(U_{1i}, \ldots, U_{di}; \theta_0)\} = 0$, by the Mean Value Theorem we deduce that

$$\text{(D.1)} = \sqrt{N}\left[\sum_{i=1}^{N} \mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1)\hat{q}_K(\boldsymbol{X}_i) \times \ell_\theta\{F_1^0(Y_{1i}), \ldots, F_d^0(Y_{di}); \theta_0\} - \mathrm{E}\{\ell_\theta(U_{1i}, \ldots, U_{di}; \theta_0)\}\right] \tag{D.5}$$

$$+ \sum_{i=1}^{N}\sum_{j=1}^{d} \mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1)\hat{q}_K(\boldsymbol{X}_i) \times \ell_{\theta j}\{\bar{F}_1(Y_{1i}), \ldots, \bar{F}_d(Y_{di}); \theta_0\} \times \sqrt{N}\,\{\hat{F}_j(Y_{ji}) - F_j^0(Y_{ji})\}, \tag{D.6}$$

where $\bar{F}_j(Y_{ji})$ lies between $F_j^0(Y_{ji})$ and $\hat{F}_j(Y_{ji})$. Similar to Theorem 2 (1), replacing T_{ji} by $\mathbf{1}(T_{1i} = \cdots = T_{di} = 1)$, replacing $\hat{p}_{jK}(x)$ by $\hat{q}_K(x)$, and replacing $\mathbf{1}(Y_{ji} \le y)$ by $\ell_\theta\{F_1^0(Y_{1i}), \ldots, F_d^0(Y_{di}); \theta_0\}$, we have the following asymptotic linear representation for the term (D.5):

$$\text{(D.5)} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \varphi(\boldsymbol{T}_i, \boldsymbol{X}_i, \boldsymbol{U}_i; \theta_0) + o_p(1), \tag{D.7}$$

where

$$\begin{aligned}\varphi(\boldsymbol{T}_i, \boldsymbol{X}_i, \boldsymbol{U}_i; \theta_0) = {} & \mathbf{1}(T_{1i} = \cdots = T_{di} = 1)\ell_\theta(\boldsymbol{U}_i; \theta_0)/\eta(\boldsymbol{X}_i) \\ & - \mathrm{E}\{\ell_\theta(\boldsymbol{U}_i; \theta_0)|\boldsymbol{X}_i\}\{\mathbf{1}(T_{1i} = \cdots = T_{di} = 1)/\eta(\boldsymbol{X}_i) - 1\} - \mathrm{E}\{\ell_\theta(\boldsymbol{U}_i; \theta_0)\}.\end{aligned}$$

For the term (D.6), we can deduce from Theorem 2 that

$$\begin{aligned}\sqrt{N}\{\hat{F}_j(y) - F_j^0(y)\} &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[T_{ji}\mathbf{1}(Y_{ji} \le y)/\pi_j(\boldsymbol{X}_i) - \mathrm{E}\{\mathbf{1}(Y_{ji} \le y)|\boldsymbol{X}_i\}\{T_{ji}/\pi_j(\boldsymbol{X}_i) - 1\} - F_j^0(y)\right] + o_p(1) \\ &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{T_{ji}\mathbf{1}\{U_{ji} \le F_j^0(y)\}/\pi_j(\boldsymbol{X}_i) - \mathrm{E}[\mathbf{1}\{U_{ji} \le F_j^0(y)\}|\boldsymbol{X}_i]\{T_{ji}/\pi_j(\boldsymbol{X}_i) - 1\} - F_j^0(y)\right\} + o_p(1) \\ &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}[\phi_j\{T_{ji}, \boldsymbol{X}_i, U_{ji}; F_j^0(y)\} - F_j^0(y)] + o_p(1),\end{aligned} \tag{D.8}$$

where, for all $v \in [0, 1]$,

$$\phi_j(T_{ji}, \boldsymbol{X}_i, U_{ji}; v) = \frac{T_{ji}}{\pi_j(\boldsymbol{X}_i)}\mathbf{1}(U_{ji} \le v) - \frac{T_{ji}}{\pi_j(\boldsymbol{X}_i)}\mathrm{E}\{\mathbf{1}(U_{ji} \le v)|\boldsymbol{X}_i\} + \mathrm{E}\{\mathbf{1}(U_{ji} \le v)|\boldsymbol{X}_i\}.$$

By Theorems 2 and 3, and (D.8), we can deduce that

$$\begin{aligned}\text{(D.6)} &= \sum_{i=1}^{N}\sum_{j=1}^{d} \mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1)\hat{q}_K(\boldsymbol{X}_i) \times \ell_{\theta j}\{\bar{F}_1(Y_{1i}), \ldots, \bar{F}_d(Y_{di}); \theta_0\} \times \sqrt{N}\,\{\hat{F}_j(Y_{ji}) - F_j^0(Y_{ji})\} \\ &= \left[\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d} \frac{\mathbf{1}(T_{1i} = 1, \ldots, T_{di} = 1)}{\eta(\boldsymbol{X}_i)} \times \ell_{\theta j}\{F_1^0(Y_{1i}), \ldots, F_d^0(Y_{di}); \theta_0\}\right] \\ &\qquad \times \frac{1}{\sqrt{N}}\sum_{k=1}^{N}\{\phi_j(T_{jk}, \boldsymbol{X}_k, U_{jk}; U_{ji}) - U_{ji}\} + o_p(1) \\ &= \frac{1}{\sqrt{N}}\sum_{k=1}^{N}\sum_{j=1}^{d} \mathrm{E}\left[\ell_{\theta j}(U_{1s}, \ldots, U_{ds}; \theta_0)\{\phi_j(T_{jk}, \boldsymbol{X}_k, U_{jk}; U_{js}) - U_{js}\}|U_{jk}, \boldsymbol{X}_k, T_{jk}\right] + o_p(1)\ (s \ne k) \\ &= \frac{1}{\sqrt{N}}\sum_{k=1}^{N}\sum_{j=1}^{d} W_j(T_{jk}, \boldsymbol{X}_k, U_{jk}; \theta_0) + o_p(1)\,,\end{aligned} \tag{D.9}$$

where the third equality follows from the Law of Large Numbers.

Combining (D.5), (D.6), (D.7), and (D.9) yields

$$\text{(D.1)} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\varphi(\boldsymbol{T}_i, \boldsymbol{X}_i, \boldsymbol{U}_i; \theta_0) + \sum_{j=1}^{d} W_j(T_{ji}, \boldsymbol{X}_i, U_{ji}; \theta_0)\right\} + o_p(1),$$

then in light of (D.2) and (D.4), we have

$$\sqrt{N}\,(\hat{\theta}-\theta_0) = B^{-1} \times \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ \varphi(\boldsymbol{T}_i, \boldsymbol{X}_i, \boldsymbol{U}_i; \theta_0) + \sum_{j=1}^{d} W_j(T_{ji}, \boldsymbol{X}_i, U_{ji}; \theta_0) \right\} + o_p(1).$$

Finally, by the Central Limit Theorem, $\sqrt{N}\,(\hat{\theta}-\theta_0) \rightsquigarrow B^{-1}\Sigma B^{-1}$ as $N \to \infty$, where

$$\Sigma = \mathrm{var} \left\{ \varphi(\boldsymbol{T}_i, \boldsymbol{X}_i, \boldsymbol{U}_i; \theta_0) + \sum_{j=1}^{d} W_j(T_{ji}, \boldsymbol{X}_i, U_{ji}; \theta_0) \right\}.$$

This completes the proof of Theorem 5. □

Appendix E. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2019.02.003.

References

[1] H. Bang, J.M. Robins, Doubly robust estimation in missing data and causal inference models, Biometrics 61 (2005) 962–973.
[2] K.C.G. Chan, S.C.P. Yam, Z. Zhang, Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting, J. R. Stat. Soc. Ser. B Stat. Methodol. 78 (2016) 673–700.
[3] T. Chang, P.S. Kott, Using calibration weighting to adjust for nonresponse under a plausible model, Biometrika 95 (2008) 555–571.
[4] X. Chen, Large sample sieve estimation of semi-nonparametric models, Handb. Econometrics 6 (2007) 5549–5632.
[5] X. Chen, Y. Fan, Pseudo-likelihood ratio tests for semiparametric multivariate copula model selection, Canad. J. Statist. 33 (2005) 389–414.
[6] X. Chen, Y. Fan, Estimation of copula-based semiparametric time series models, J. Econometrics 130 (2006) 307–335.
[7] X. Chen, H. Hong, A. Tarozzi, Semiparametric efficiency in GMM models with auxiliary data, Ann. Statist. 36 (2008) 808–843.
[8] S.X. Chen, D.H. Leung, J. Qin, Improving semiparametric estimation by using surrogate data, J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (4) (2008) 803–823.
[9] J. Chen, R. Sitter, A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys, Statist. Sinica (1999) 385–406.
[10] J.-C. Deville, C.-E. Särndal, Calibration estimators in survey sampling, J. Amer. Statist. Assoc. 87 (1992) 376–382.
[11] W. Ding, P.X.-K. Song, EM algorithm in Gaussian copula with missing data, Comput. Statist. Data Anal. 101 (2016) 1–11.
[12] S.G. Donald, Y.-C. Hsu, Estimation and inference for distribution functions and quantile functions in treatment effect models, J. Econometrics 178 (2014) 383–397.
[13] Y. Fan, A.J. Patton, Copulas in econometrics, Annu. Rev. Econ. 6 (2014) 179–200.
[14] C. Genest, K. Ghoudi, L.-P. Rivest, A semiparametric estimation procedure of dependence parameters in multivariate families of distributions, Biometrika 82 (1995) 543–552.
[15] C. Genest, J. Nešlehová, N. Ben Ghorbal, Estimators based on Kendall's tau in multivariate copula models, Aust. N. Z. J. Stat. 53 (2011) 157–177.
[16] C. Genest, B.J.M. Werker, Conditions for the asymptotic semiparametric efficiency of an omnibus estimator of dependence parameters in copula models, in: C.M. Cuadras, J. Fortiana, J.A. Rodríguez Lallena (Eds.), Distributions with Given Marginals and Statistical Modelling, Kluwer, Dordrecht, The Netherlands, 2002, pp. 103–112.
[17] S. Hamori, K. Motegi, Z. Zhang, Online Supplement for "Calibration estimation of semiparametric copula models with data missing at random", Kobe University and Renmin University of China, 2019.
[18] P. Han, Multiply robust estimation in regression analysis with missing data, J. Amer. Statist. Assoc. 109 (507) (2014) 1159–1173.
[19] P. Han, Combining inverse probability weighting and multiple imputation to improve robustness of estimation, Scand. J. Stat. 43 (1) (2016) 246–260.
[20] P. Han, Intrinsic efficiency and multiple robustness in longitudinal studies with drop-out, Biometrika 103 (3) (2016) 683–700.
[21] P. Han, Calibration and multiple robustness when data are missing not at random, Statist. Sinica (2018).
[22] P. Han, L. Wang, Estimation with missing data: beyond double robustness, Biometrika 100 (2) (2013) 417–430.
[23] L.P. Hansen, J. Heaton, A. Yaron, Finite-sample properties of some alternative GMM estimators, J. Bus. Econom. Statist. 14 (1996) 262–280.
[24] C. Hasler, R.V. Craiu, L.-P. Rivest, Vine copulas for imputation of monotone non-response, Internat. Statist. Rev. 0 (2018) 1–24, http://dx.doi.org/10.1111/insr.12263.
[25] J.K. Hellerstein, G.W. Imbens, Imposing moment restrictions from auxiliary data by weighting, Rev. Econ. Statist. 81 (1999) 1–14.
[26] K. Hirano, G.W. Imbens, G. Ridder, Efficient estimation of average treatment effects using the estimated propensity score, Econometrica 71 (2003) 1161–1189.
[27] D.G. Horvitz, D.J. Thompson, A generalization of sampling without replacement from a finite universe, J. Amer. Statist. Assoc. 47 (1952) 663–685.
[28] M.-Y. Huang, K.C.G. Chan, Joint sufficient dimension reduction and estimation of conditional and average treatment effects, Biometrika 104 (3) (2017) 583–596.
[29] J.K. Kim, Calibration estimation using empirical likelihood in survey sampling, Statist. Sinica (2009) 145–157.
[30] J.K. Kim, Calibration estimation using exponential tilting in sample surveys, Surv. Methodol. 36 (2) (2010) 145.
[31] J.K. Kim, M. Park, Calibration estimation in survey sampling, Internat. Statist. Rev. 78 (1) (2010) 21–39.
[32] Y. Kitamura, M. Stutzer, An information-theoretic alternative to generalized method of moments estimation, Econometrica 65 (1997) 861–874.
[33] P.S. Kott, T. Chang, Using calibration weighting to adjust for nonignorable unit nonresponse, J. Amer. Statist. Assoc. 105 (491) (2010) 1265–1275.
[34] P.S. Kott, D. Liao, Calibration weighting for nonresponse that is not missing at random: Allowing more calibration than response-model variables, J. Surv. Statist. Methodol. 5 (2) (2017) 159–174.
[35] Q. Li, J.S. Racine, Nonparametric Econometrics: Theory and Practice, Princeton University Press, 2007.
[36] S. Lundström, C.-E. Särndal, Calibration as a standard method for treatment of nonresponse, J. Official Statist. 15 (2) (1999) 305–327.
[37] G. Marra, K. Wyszynski, Semi-parametric copula sample selection models for count responses, Comput. Statist. Data Anal. 104 (2016) 110–129.
[38] W.K. Newey, Convergence rates and asymptotic normality for series estimators, J. Econometrics 79 (1) (1997) 147–168.
[39] D.H. Oh, A.J. Patton, High-dimensional copula-based distributions with mixed frequency data, J. Econometrics 193 (2016) 349–366.

[40] D.H. Oh, A.J. Patton, Modeling dependence in high dimensions with factor copulas, J. Bus. Econom. Statist. 35 (2017) 139–154.
[41] D.H. Oh, A.J. Patton, Time-varying systemic risk: Evidence from a dynamic copula model of CDS spreads, J. Bus. Econom. Statist. 36 (2018) 181–195.
[42] A.B. Owen, Empirical likelihood ratio confidence intervals for a single functional, Biometrika 75 (1988) 237–249.
[43] A.J. Patton, A review of copula models for economic time series, J. Multivariate Anal. 110 (2012) 4–18.
[44] J. Qin, D. Leung, J. Shao, Estimation with survey data under nonignorable nonresponse or informative sampling, J. Amer. Statist. Assoc. 97 (2002) 193–200.
[45] J. Qin, J. Shao, B. Zhang, Efficient and doubly robust imputation for covariate-dependent missing responses, J. Amer. Statist. Assoc. 103 (482) (2008) 797–810.
[46] J. Qin, B. Zhang, Empirical-likelihood-based inference in missing response problems and its application in observational studies, J. R. Stat. Soc. Ser. B Stat. Methodol. 69 (1) (2007) 101–122.
[47] J.M. Robins, A. Rotnitzky, Semiparametric efficiency in multivariate regression models with missing data, J. Amer. Statist. Assoc. 90 (1995) 122–129.
[48] J.M. Robins, A. Rotnitzky, L.P. Zhao, Estimation of regression coefficients when some regressors are not always observed, J. Amer. Statist. Assoc. 89 (1994) 846–866.
[49] D.B. Rubin, Inference and missing data, Biometrika 63 (1976) 581–592.
[50] I.D.L. Salvatierra, A.J. Patton, Dynamic copula models and high frequency data, J. Empir. Finance 30 (2015) 120–135.
[51] A. Sklar, Fonctions de répartition à n dimensions et leurs marges, Publ. Inst. Statist. Univ. Paris 8 (1959) 229–231.
[52] Z. Tan, Bounded, efficient and doubly robust estimation with inverse weighting, Biometrika 97 (3) (2010) 661–682.
[53] P. Tseng, D.P. Bertsekas, Relaxation methods for problems with strictly convex separable costs and linear constraints, Math. Program. 38 (1987) 303–321.
[54] A.W. van der Vaart, Asymptotic Statistics, Cambridge University Press, 1998.
[55] H. Wang, F. Fazayeli, S. Chatterjee, A. Banerjee, Gaussian copula precision estimation with missing values, in: Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, 2014, pp. 978–986.
[56] S. Yiu, L. Su, Covariate association eliminating weights: a unified weighting framework for causal effect estimation, Biometrika 105 (2018) 709–722.