



Copula-based regression 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 15 September 2019

Received in revised form 1 July 2020

Accepted 1 July 2020

Available online 22 July 2020

AMS 2010 subject classifications:

primary 62G08

secondary 62H12

Keywords:

Calibration estimation

Generalized regression model

Missing at random (MAR)

Semiparametric copula

ABSTRACT

The existing literature of copula-based regression assumes that complete data are available, but this assumption is violated in many real applications. The present paper allows the regressand and regressors to be missing at random (MAR). We formulate a generalized regression model which unifies many prominent cases such as the conditional mean and quantile regressions. A semiparametric copula and the target regression curve are estimated via the calibration approach. The consistency and asymptotic normality of the estimated regression curve are proved. We show via Monte Carlo simulations that the proposed approach operates well in finite samples, while a benchmark equal-weight approach fails with substantial bias under MAR. An empirical application on revenues and R&D expenses of German manufacturing firms highlights a practical use of our approach.

© 2020 Elsevier Inc. All rights reserved.

1. Introduction

Regression is the most prevailing method for investigating the relationship between a regressand Y and regressors \mathbf{W} . Widely used regressions include conditional mean and quantile regressions. Noh et al. [27] proposed a novel approach to estimate a conditional mean regression function by exploiting copulas, where observations are assumed to be independently and identically distributed (*i.i.d.*) and completely observed. Their key insight is that the loss function expressed as a conditional expectation given \mathbf{W} can be rewritten as an unconditional expectation involving a parametric copula and nonparametric marginal distributions. The marginal distributions and the copula parameter are estimated via plug-in methods. The flexibility of the semiparametric copula alleviates model specification issues such as how to transform regressors and which cross-products of regressors to include.

Noh et al. [27] spurred extensive research on the copula-based regression. Noh et al. [28] applied the method of Noh et al. [27] to the quantile regression with *i.i.d.* or time series data that are completely observed. De Backer et al. [4] extended the method of Noh et al. [28] to the quantile regression with censored data. Kraus and Czado [19] studied the quantile regression with complete data, using D-vine copulas. Rémillard et al. [30] discussed the asymptotic connection between the estimators of Noh et al. [28] and Kraus and Czado [19]. Chang and Joe [3] proposed an algorithm for computing the conditional distribution function via the vine copula.

Nagler and Vatter [23] unified various copula-based regressions by formulating a general loss function which may not be continuously differentiable. Their generalized regression model includes the conditional mean regression of Noh et al. [27], the conditional quantile regression of Noh et al. [28], and the asymmetric least squares of Newey and Powell [26] as special cases. The unified framework enhances the systematic interpretation of the various existing regressions.

* 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).

A potential issue left in the existing literature of copula-based regression is that complete data are assumed to be available. There also exists the vast literature where copula itself is a primary target of estimation [11], but the availability of complete data is still assumed in most papers in that literature. The assumption of complete data is violated in many real applications such as applied microeconomics, corporate finance, and survey sampling. In survey sampling, for example, respondents may refuse to report their personal information such as age and salary.

To relax the assumption of complete data, this paper allows both the regressand Y and the regressors \mathbf{W} to be missing at random (MAR), a key concept originally explored by Rubin [31]. The MAR condition has been popularly used in econometrics and statistics to identify the parameter of interest [22]. Hamori et al. [13] is one of few works to deal with data missing at random in copula modeling. (Ding and Song [6] proposed an EM algorithm for estimating the Gaussian copula under the MAR condition. Emura et al. [8], Emura and Wang [9,10] considered the copula inference with truncated survival data. Guo et al. [12] studied the semiparametric estimation of copula models with nonignorable missing data.) Hamori et al. [13] use calibration weights proposed by Chan et al. [2] for both nonparametric marginal distributions and target copula parameters. The calibration estimation is a nonparametric method that balances the empirical moments of covariates between the observed and whole groups. It does not require an explicit specification of the missing mechanism, and delivers consistent inference under MAR.

Inspired by Hamori et al. [13], this paper adopts the calibration estimation to perform the copula-based regression with $\{Y, \mathbf{W}\}$ missing at random. As in Nagler and Vatter [23], we formulate the generalized regression model which unifies many prominent regressions. A semiparametric copula and the target regression curve are estimated via the calibration approach. The consistency and asymptotic normality of the estimated regression curve are proved. Our simulation study shows that the proposed approach performs well in finite samples, while a benchmark equal-weight approach fails with substantial bias under MAR.

As an empirical application, we regress the R&D expenses of German manufacturing firms onto their revenues. The revenue is observed for all 500 firms considered, while the R&D expense is observed for only 125 firms. The vast majority of the 375 firms with missing R&D have small revenues. The calibration approach delivers a plausible empirical result by assigning sufficiently large weights on firms with small revenues. The benchmark equal-weight approach, by contrast, delivers a misleading result by discarding the 375 firms with missing R&D and assigning the uniform weight on the 125 firms left. This contrast highlights how the calibration approach achieves valid inference under the MAR mechanism.

The rest of this paper is organized as follows: Section 2 explains our basic framework and notation. The calibration estimation is proposed in Section 3, and its large sample properties are derived in Section 4. Variance estimators and confidence intervals are constructed in Section 5. In Section 6, data-driven selection of a tuning parameter K for sieve basis functions is discussed. The simulation study is performed in Section 7. The empirical application is presented in Section 8. Conclusions are provided in Section 9. Proofs of main theorems are collected in Technical Appendices. In the Online Supplement [14], omitted technical and numerical details are provided.

2. Notation and set-up

2.1. Generalized regression model with missing data

Let Y be a regressand, and let $\mathbf{W} = (W_1, \dots, W_d)^\top$ be d -dimensional regressors. Consider the generalized regression model:

$$a_0(\mathbf{w}) = \arg \min_{a \in \mathbb{R}} E[L\{g(Y) - a\} | \mathbf{W} = \mathbf{w}], \quad (1)$$

where $\mathbf{w} = (w_1, \dots, w_d)^\top$; $L(\cdot)$ is a pre-specified loss function whose derivative, denoted as $L'(\cdot)$, exists almost everywhere; $g(Y)$ is a known function of Y . $L(\cdot)$ is not required to be continuously differentiable.

Let $\mathbf{1}(A)$ be the indicator function which equals 1 if event A occurs and 0 otherwise. The formulation (1) includes many prominent cases:

- $L(v) = v^2$ and $g(Y) = Y$, in which case $a_0(\mathbf{w}) = E[Y | \mathbf{W} = \mathbf{w}]$ is the conditional mean regression [27].
- $L(v) = v\{\tau - \mathbf{1}(v \leq 0)\}$ and $g(Y) = Y$, in which case $a_0(\mathbf{w})$ is the τ^{th} conditional quantile [28].
- $L(v) = v^2$ and $g(Y) = \mathbf{1}(Y \leq y)$, in which case $a_0(\mathbf{w}) = E\{\mathbf{1}(Y \leq y) | \mathbf{W} = \mathbf{w}\} = \Pr(Y \leq y | \mathbf{W} = \mathbf{w})$ is the conditional distribution function.
- $L(v) = v^2|\tau - \mathbf{1}(v \leq 0)|$ and $g(Y) = Y$, in which case $a_0(\mathbf{w})$ corresponds to the asymmetric least squares [26].

Nagler and Vatter [23] investigated the generalized regression model (1), but assumed that complete data were available. A goal of the present paper is to solve (1) with both Y and \mathbf{W} being possibly missing. Let $\mathbf{O}_{\text{obs}} \subset \mathbf{O} := \{Y, W_1, \dots, W_d\}$ be a set of components which are observed with probability 1. Let $\mathbf{O}_{\text{mis}} := \mathbf{O} \setminus \mathbf{O}_{\text{obs}}$ be a set of components which are missing with positive probability. Let d_{obs} (resp. d_{mis}) be the number of elements in \mathbf{O}_{obs} (resp. \mathbf{O}_{mis}), then $d + 1 = d_{\text{obs}} + d_{\text{mis}}$ by construction. Let $\mathbf{T}_i := (T_{1i}, \dots, T_{d_{\text{mis}},i})^\top \in \{0, 1\}^{d_{\text{mis}}}$ be binary indicators which represent the missing status of $\mathbf{O}_{i,\text{mis}} = (O_{1i,\text{mis}}, \dots, O_{d_{\text{mis}},i,\text{mis}})^\top$. For $j \in \{1, \dots, d_{\text{mis}}\}$ and $i \in \{1, \dots, N\}$, $T_{ji} = 0$ (resp. $T_{ji} = 1$) if $O_{ji,\text{mis}}$ is missing (resp. observed).

If \mathbf{T}_i and $\mathbf{O}_{i,\text{mis}}$ are independent of each other, then the latter is called missing completely at random (MCAR). List-wise deletion, an elementary approach which picks individuals with complete observations and assigns equal weights on them,

is well known to deliver consistent inference under MCAR. The MCAR condition, however, is too stringent to justify in many real applications.

In this paper, we impose a more realistic assumption called missing at random (MAR) [31]. Let $\mathbf{X}_i = (X_{i1}, \dots, X_{ri})^\top$ be r -dimensional covariates that are observable for all individuals $i \in \{1, \dots, N\}$, where $\mathbf{X}_i \supset \mathbf{O}_{i,\text{obs}}$ and hence $r \geq d_{\text{obs}}$. Under MAR, \mathbf{T}_i and $\mathbf{O}_{i,\text{mis}}$ are conditionally independent of each other given covariates \mathbf{X}_i .

Assumption 1 (Missing at Random). $\mathbf{T}_i \perp \mathbf{O}_{i,\text{mis}} \mid \mathbf{X}_i$.

The MAR condition is popularly used in econometrics and statistics to identify the parameter of interest. It does not require the unconditional independence between \mathbf{T}_i and $\mathbf{O}_{i,\text{mis}}$. In many real applications, \mathbf{T}_i and $\mathbf{O}_{i,\text{mis}}$ are correlated with each other through \mathbf{X}_i , and that violates MCAR but not MAR.

To simplify notation without losing generality, we assume hereafter that $\mathbf{O}_{\text{obs}} = \emptyset$ and $\mathbf{O}_{\text{mis}} = \mathbf{O} = \{Y, \mathbf{W}\}$ (i.e., all of the regressand and regressors are missing with positive probability). Then $d_{\text{obs}} = 0$, $d_{\text{mis}} = d + 1$, and $0 < \Pr(T_{ji} = 1) < 1$ for all $j \in \{0, 1, \dots, d\}$ and $i \in \{1, \dots, N\}$, where T_{0i} indicates the missing status of Y_i and T_{ji} with $j \in \{1, \dots, d\}$ indicates the missing status of W_{ji} . Assume further that the observations $\{\mathbf{T}_i, \mathbf{X}_i, \mathbf{W}_i, Y_i\}_{i=1}^N$ are i.i.d.

Remark 1. Data missing at random are a specific class of missing data, and copula-based regression with other types of missing data has not been explored in the literature yet. Missing data of particular relevance include censored data [4,8] and truncated data [9,10]. A brief discussion on these cases is provided in Section 2 of the Online Supplement [14]; further investigations are left as future work.

2.2. Copula-based regression

Let $f_{Y, \mathbf{W}}(\cdot)$ and $f_{\mathbf{W}}(\cdot)$ be the joint density functions of $\{Y, \mathbf{W}\}$ and \mathbf{W} , respectively. Let $F_0(\cdot)$ and $F_j(\cdot)$ be the cumulative distribution functions of Y and W_j for $j \in \{1, \dots, d\}$, respectively. Let $f_0(\cdot)$ and $f_j(\cdot)$ be the probability density functions of Y and W_j , respectively. Let $c(\cdot)$ and $c_{\mathbf{W}}(\cdot)$ be the copula densities of $\{Y, \mathbf{W}\}$ and \mathbf{W} , respectively. Sklar's Theorem [32] ensures that

$$f_{Y, \mathbf{W}}(y, \mathbf{w}) = c\{F_0(y), F_1(w_1), \dots, F_d(w_d)\}f_0(y) \prod_{j=1}^d f_j(w_j) \quad \text{and} \quad f_{\mathbf{W}}(\mathbf{w}) = c_{\mathbf{W}}\{F_1(w_1), \dots, F_d(w_d)\} \prod_{j=1}^d f_j(w_j).$$

See [24] for a comprehensive treatment of the copula literature. Define the propensity score functions as

$$\pi_j(\mathbf{x}) = \Pr(T_{ji} = 1 \mid \mathbf{X}_i = \mathbf{x}), \quad j \in \{0, 1, \dots, d\}, \quad (2)$$

$$\eta(\mathbf{x}) = \Pr(T_{0i} = T_{1i} = \dots = T_{di} = 1 \mid \mathbf{X}_i = \mathbf{x}). \quad (3)$$

Use Assumption 1, Sklar's Theorem, (2), and (3) to identify $a_0(\mathbf{w})$ in (1) as follows:

$$\begin{aligned} a_0(\mathbf{w}) &= \arg \min_{a \in \mathbb{R}} \mathbb{E}[L\{g(Y) - a\} \mid \mathbf{W} = \mathbf{w}] = \arg \min_{a \in \mathbb{R}} \int L\{g(y) - a\} \frac{f_{Y, \mathbf{W}}(y, \mathbf{w})}{f_{\mathbf{W}}(\mathbf{w})} dy \\ &= \arg \min_{a \in \mathbb{R}} \int L\{g(y) - a\} \frac{c\{F_0(y), F_1(w_1), \dots, F_d(w_d)\}f_0(y) \prod_{j=1}^d f_j(w_j)}{c_{\mathbf{W}}\{F_1(w_1), \dots, F_d(w_d)\} \prod_{j=1}^d f_j(w_j)} dy \\ &= \arg \min_{a \in \mathbb{R}} \mathbb{E}[L\{g(Y) - a\}c\{F_0(Y), F_1(w_1), \dots, F_d(w_d)\}]] \end{aligned} \quad (4)$$

$$= \arg \min_{a \in \mathbb{R}} \mathbb{E}\left[\frac{T_0}{\pi_0(\mathbf{X})} L\{g(Y) - a\}c\{F_0(Y), F_1(w_1), \dots, F_d(w_d)\}\right]. \quad (5)$$

Note that $a_0(\mathbf{w})$ is expressed as the conditional expectation given $\mathbf{W} = \mathbf{w}$ in (1), while it is expressed as the unconditional expectation involving the copula density in (4). This is the key transformation utilized in [27]. Further, compare (4) and (5) to see how missing data are handled. The objective function is evaluated on the whole group in (4), while it is evaluated on the observed group with respect to Y in (5). The transformation from (4) to (5) requires individuals in the observed group to be weighted by $\pi_0(\mathbf{X})^{-1}$, a core insight of the inverse probability weighting [16].

Assume that the copula density of $\{Y, \mathbf{W}\}$ admits a parametric model $c(u_0, u_1, \dots, u_d) = c(u_0, u_1, \dots, u_d; \theta_0)$. Then, the target regression curve $a_0(\mathbf{w})$ is identified as follows:

$$a_0(\mathbf{w}) = \arg \min_{a \in \mathbb{R}} \mathbb{E}\left[\frac{T_0}{\pi_0(\mathbf{X})} L\{g(Y) - a\}c\{F_0(Y), F_1(w_1), \dots, F_d(w_d); \theta_0\}\right]. \quad (6)$$

To estimate $a_0(\mathbf{w})$, consider a sample counterpart to (6):

$$\tilde{a}(\mathbf{w}) = \arg \min_{a \in \mathbb{R}} \sum_{i=1}^N \frac{T_{0i}}{N\pi_0(\mathbf{X}_i)} L\{g(Y_i) - a\}c\{F_0(Y_i), F_1(W_{1i}), \dots, F_d(W_{di}); \theta_0\}. \quad (7)$$

$\tilde{a}(\mathbf{w})$ is an infeasible estimator for $a_0(\mathbf{w})$, since $\pi_0(\mathbf{X}_i)$, $F_0(Y_i)$, $F_1(W_{1i})$, ..., $F_d(W_{di})$, and θ_0 are all unknown. These quantities should be replaced with feasible estimators in order to derive a feasible estimator for $a_0(\mathbf{w})$.

Using [Assumption 1](#) and the law of iterated expectations, rewrite $F_0(y)$ and $F_j(w)$ with $j \in \{1, \dots, d\}$ as follows:

$$F_0(y) = E\{\mathbf{1}(Y_i \leq y)\} = E\left\{\frac{T_{0i}}{\pi_0(\mathbf{X}_i)} \mathbf{1}(Y_i \leq y)\right\} \quad \text{and} \quad F_j(w) = E\{\mathbf{1}(W_{ji} \leq w)\} = E\left\{\frac{T_{ji}}{\pi_j(\mathbf{X}_i)} \mathbf{1}(W_{ji} \leq w)\right\}. \quad (8)$$

If $\mathbf{W} = \mathbf{X}$, then F_j can be estimated straightforwardly. Otherwise, the last transformation in (8) is required in order to handle \mathbf{W} missing at random.

The sample counterparts to (8) are respectively given by

$$\tilde{F}_0(y) = \sum_{i=1}^N T_{0i} \left\{ \frac{1}{N\pi_0(\mathbf{X}_i)} \right\} \mathbf{1}(Y_i \leq y) \quad \text{and} \quad \tilde{F}_j(w) = \sum_{i=1}^N T_{ji} \left\{ \frac{1}{N\pi_j(\mathbf{X}_i)} \right\} \mathbf{1}(W_{ji} \leq w), \quad j \in \{1, \dots, d\}. \quad (9)$$

$\{\tilde{F}_j\}_{j=0}^d$ are infeasible estimators for $\{F_j\}_{j=0}^d$, since $\{\pi_j(\mathbf{X}_i)\}_{j=0}^d$ are unknown.

Further, the estimation of θ_0 involves yet another unknown quantity. To see this, assume that θ_0 is identified as the maximizer of the log-likelihood function:

$$\theta_0 = \arg \max_{\theta \in \Theta} E[\ln c\{F_0(Y_i), F_1(W_{1i}), \dots, F_d(W_{di}); \theta\}],$$

where Θ is a compact subset of \mathbb{R}^p which contains the true value θ_0 . Using [Assumption 1](#) and the law of iterated expectations, θ_0 can be rewritten as

$$\theta_0 = \arg \max_{\theta \in \Theta} E\left[\frac{\mathbf{1}(T_{0i} = \dots = T_{di} = 1)}{\eta(\mathbf{X}_i)} \ln c\{F_0(Y_i), F_1(W_{1i}), \dots, F_d(W_{di}); \theta\}\right]. \quad (10)$$

To estimate θ_0 , consider a sample counterpart to (10):

$$\tilde{\theta} = \arg \max_{\theta \in \Theta} \sum_{i=1}^N \frac{\mathbf{1}(T_{0i} = \dots = T_{di} = 1)}{N\eta(\mathbf{X}_i)} \ln c\{F_0(Y_i), F_1(W_{1i}), \dots, F_d(W_{di}); \theta\}. \quad (11)$$

$\tilde{\theta}$ is an infeasible estimator for θ_0 since $\eta(\mathbf{X}_i)$ and $\{F_j(\cdot)\}_{j=0}^d$ are unknown.

Eqs. (7), (9), and (11) motivate the estimation of $\{N\pi_0(\mathbf{X}_i)\}^{-1}$, ..., $\{N\pi_d(\mathbf{X}_i)\}^{-1}$, and $\{N\eta(\mathbf{X}_i)\}^{-1}$. Directly estimating $\pi_j(\mathbf{X}_i)$ and $\eta(\mathbf{X}_i)$ and substituting $\{N\hat{\pi}_j(\mathbf{X}_i)\}^{-1}$ and $\{N\hat{\eta}(\mathbf{X}_i)\}^{-1}$ often performs poorly in practice. Parametric modeling of $\pi_j(\mathbf{X}_i)$ and $\eta(\mathbf{X}_i)$, on one hand, may suffer from misspecification. Nonparametric estimation such as kernel smoothing, on the other hand, is unstable in finite samples; $\hat{\pi}_j(\mathbf{X}_i)$ and $\hat{\eta}(\mathbf{X}_i)$ can take extremely close values to 0, which destroys the entire computation. Indeed, the inverse probability weighting estimators are notoriously sensitive to the estimated propensity score [18]. In the next section, the estimation of the propensity score functions is discussed in detail.

3. Calibration estimation

3.1. Calibration weights

This paper adopts the covariate balancing principle of Chan et al. [2] and Hamori et al. [13] to estimate $\{N\pi_j(\mathbf{X}_i)\}^{-1}$ and $\{N\eta(\mathbf{X}_i)\}^{-1}$. Their key insight is that the following equation holds for any integrable function $u(\mathbf{X})$ and $j \in \{0, 1, \dots, d\}$:

$$E\left[T_{ji} \left\{ \frac{1}{\pi_j(\mathbf{X}_i)} \right\} u(\mathbf{X}_i)\right] = E\{u(\mathbf{X}_i)\}, \quad (12)$$

$$E\left[\mathbf{1}(T_{0i} = \dots = T_{di} = 1) \left\{ \frac{1}{\eta(\mathbf{X}_i)} \right\} u(\mathbf{X}_i)\right] = E\{u(\mathbf{X}_i)\}. \quad (13)$$

The estimator of $\{N\pi_j(\mathbf{X})\}^{-1}$, denoted as $\hat{p}_{jk}(\mathbf{X})$, should satisfy the sample counterpart of (12):

$$\sum_{i=1}^N T_{ji} \hat{p}_{jk}(\mathbf{X}_i) u_K(\mathbf{X}_i) = \frac{1}{N} \sum_{i=1}^N u_K(\mathbf{X}_i), \quad (14)$$

where $u_K(\mathbf{X}) = \{u_{K,1}(\mathbf{X}), \dots, u_{K,K}(\mathbf{X})\}^\top$ is a known sieve basis function that can approximate any suitable function $u(\mathbf{X})$ arbitrarily well, and $K \rightarrow \infty$ as $N \rightarrow \infty$. Common specifications of $u_K(\cdot)$ include orthonormal polynomials, B-splines, and wavelets. Define the supremum of the Frobenius norm of $u_K(\cdot)$ as

$$\zeta(K) = \sup_{\mathbf{x} \in \mathcal{X}} \text{tr}\{u_K(\mathbf{x})u_K(\mathbf{x})^\top\}^{\frac{1}{2}}, \quad (15)$$

where \mathcal{X} is the support of the covariates \mathbf{X} . In general, this bound depends on the array of basis used. Newey [25] shows that $\zeta(K) \leq CK$ for orthonormal polynomials, and $\zeta(K) \leq C\sqrt{K}$ for B-splines, where $C > 0$ is a universal positive constant. $\zeta(K)$ shall be used when we characterize large sample properties of $\hat{p}_{jk}(\mathbf{X})$.

Similarly, the estimator of $\{N\eta(\mathbf{X})\}^{-1}$, denoted as $\hat{q}_K(\mathbf{X})$, should satisfy the sample counterpart of (13):

$$\sum_{i=1}^N \mathbf{1}(T_{0i} = \dots = T_{di} = 1) \hat{q}_K(\mathbf{X}_i) u_K(\mathbf{X}_i) = \frac{1}{N} \sum_{i=1}^N u_K(\mathbf{X}_i). \quad (16)$$

The dimension K can vary across $\{\hat{p}_{0,K_0}(\mathbf{X}), \dots, \hat{p}_{d,K_d}(\mathbf{X}), \hat{q}_{K_\eta}(\mathbf{X})\}$ without any conceptual difficulty. For the sake of notational brevity, we use a single value K for all components hereafter. Data-driven selections of $\mathbf{K} = (K_0, \dots, K_d, K_\eta)^\top$ are discussed in Section 6.

Multiple values of $\hat{p}_{jK}(\mathbf{X}_i)$ and $\hat{q}_K(\mathbf{X}_i)$ satisfy (14) and (16), respectively. Among them, the calibration approach chooses the one closest to a uniform weight given some distance measure in order to enhance stable performance in finite samples. As shown in [13], the resulting estimator of $\{N\pi_j(\mathbf{X})\}^{-1}$ is

$$\hat{p}_{jK}(\mathbf{X}) = \frac{1}{N} \rho' \left\{ \hat{\lambda}_{jK}^\top u_K(\mathbf{X}) \right\}, \quad j \in \{0, 1, \dots, d\}, \quad (17)$$

where $\rho(\cdot)$ is an increasing and strictly concave function on \mathbb{R} which is three times continuously differentiable; $\rho'(\cdot)$ is the derivative of $\rho(\cdot)$; $\hat{\lambda}_{jK} \in \mathbb{R}^K$ is the maximizer of the concave objective function:

$$\hat{G}_{jK}(\lambda) = \frac{1}{N} \sum_{i=1}^N T_{ji} \rho\{\lambda^\top u_K(\mathbf{X}_i)\} - \frac{1}{N} \sum_{i=1}^N \lambda^\top u_K(\mathbf{X}_i). \quad (18)$$

It is straightforward to derive (17) by taking the first order condition on (18). Similarly, the estimator of $\{N\eta(\mathbf{X})\}^{-1}$ is

$$\hat{q}_K(\mathbf{X}) = \frac{1}{N} \rho' \left\{ \hat{\beta}_K^\top u_K(\mathbf{X}) \right\}, \quad (19)$$

where $\hat{\beta}_K \in \mathbb{R}^K$ is the maximizer of the concave objective function:

$$\hat{H}_K(\beta) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(T_{0i} = \dots = T_{di} = 1) \rho\{\beta^\top u_K(\mathbf{X}_i)\} - \frac{1}{N} \sum_{i=1}^N \beta^\top u_K(\mathbf{X}_i). \quad (20)$$

$\hat{p}_{jK}(\mathbf{X})$ can be interpreted as a generalized empirical likelihood (GEL) estimator of $\{N\pi_j(\mathbf{X})\}^{-1}$. Indeed, Chan et al. [2] and Hamori et al. [13] show that the unconstrained maximization problem (18) is the dual problem of the constrained minimization of the distance between $\hat{p}_{jK}(\mathbf{X})$ and the uniform weight subject to the covariate balancing condition (14). Similarly, $\hat{q}_K(\mathbf{X})$ can be interpreted as a GEL estimator of $\{N\eta(\mathbf{X})\}^{-1}$; (20) is the dual problem of minimizing the distance between $\hat{q}_K(\mathbf{X})$ and the uniform weight subject to (16).

The increasing and strictly concave function $\rho(\cdot)$ corresponds to a well-defined distance measure used in the dual problems. In the Monte Carlo simulation (Section 7), we compare alternative specifications of $\rho(\cdot)$, and find that the inverse logistic $\rho(v) = v - \exp(-v)$ leads to the best finite sample performance in many cases.

3.2. Calibration estimation of the target $a_0(\mathbf{w})$

Use (17) in (9) to estimate the marginal distributions $\{F_j(\cdot)\}_{j=0}^d$:

$$\hat{F}_{0,K}(y) := \sum_{i=1}^N T_{0i} \hat{p}_{0K}(\mathbf{X}_i) \mathbf{1}(Y_i \leq y) \text{ and } \hat{F}_{j,K}(w) = \sum_{i=1}^N T_{ji} \hat{p}_{jK}(\mathbf{X}_i) \mathbf{1}(W_{ji} \leq w), \quad j \in \{1, \dots, d\}. \quad (21)$$

Use (19) and (21) in (11) to construct a feasible estimator for θ_0 :

$$\hat{\theta}_K := \arg \max_{\theta \in \Theta} \sum_{i=1}^N \mathbf{1}(T_{0i} = \dots = T_{di} = 1) \hat{q}_K(\mathbf{X}_i) \ln c \left\{ \hat{F}_{0,K}(Y_i), \hat{F}_{1,K}(W_{1i}), \dots, \hat{F}_{d,K}(W_{di}); \theta \right\}. \quad (22)$$

Remark 2. If $\Pr(T_{ji} = 1) = 1$ for all $j \in \{0, 1, \dots, d\}$ and $i \in \{1, \dots, N\}$ (i.e., complete data), then the calibration estimator $\hat{\theta}_K$ in (22) coincides with the pseudo-maximum likelihood estimator of Genest et al. [11]. It is well known that the pseudo-maximum likelihood estimator is a special case of the Z-estimator of Tsukahara [33]. It is not necessarily the case that the calibration estimator coincides with the Z-estimator under complete data. Missing data are not allowed in either [11] or [33]. See Section 3 of the Online Supplement [14] for more details.

Finally, the target regression curve $a_0(\mathbf{w})$ in (6) is estimated via

$$\hat{a}(\mathbf{w}) = \arg \min_{a \in \mathbb{R}} \sum_{i=1}^N T_{0i} \hat{p}_{0K}(\mathbf{X}_i) L\{g(Y_i) - a\} c \left\{ \hat{F}_{0,K}(Y_i), \hat{F}_{1,K}(w_1), \dots, \hat{F}_{d,K}(w_d); \hat{\theta}_K \right\}. \quad (23)$$

Remark 3. Under the conditional mean regression of Noh et al. [27], $L\{g(Y_i) - a\} = (Y_i - a)^2$ and hence $\hat{a}(\mathbf{w})$ has a closed-form solution:

$$\hat{a}(\mathbf{w}) = \frac{\sum_{i=1}^N T_{0i} \hat{p}_{0K}(\mathbf{X}_i) Y_i c \left\{ \hat{F}_{0,K}(Y_i), \hat{F}_{1,K}(w_1), \dots, \hat{F}_{d,K}(w_d); \hat{\theta}_K \right\}}{\sum_{i=1}^N T_{0i} \hat{p}_{0K}(\mathbf{X}_i) c \left\{ \hat{F}_{0,K}(Y_i), \hat{F}_{1,K}(w_1), \dots, \hat{F}_{d,K}(w_d); \hat{\theta}_K \right\}}. \quad (24)$$

Remark 4. In the present paper, a semiparametric copula with only one or few parameters governs the entire dependence among the $(d + 1)$ variables $\{Y, \mathbf{W}\}$, which may be too restrictive when d is large. A potential solution to this issue is adopting vine copulas to the proposed method [17]. A specific procedure for adopting vine copulas is described in Section 4 of the Online Supplement [14]. Investigating the theoretical properties of the vine-copula approach is beyond the scope of the present paper, and left as a future task.

Remark 5. The proposed methodology is based on the parametric copula assumption. In addition to potential misspecification, the parametric copula is not able to model non-monotonic regressions [5]. The model suggested by De Backer et al. [4] could be a solution to this problem.

4. Large sample properties

It is evident from (22)–(23) that the proposed estimator $\hat{a}(\mathbf{w})$ depends on $\hat{p}_{jK}(\mathbf{X})$, $\hat{q}_K(\mathbf{X})$, and $\hat{\theta}_K$. Large sample properties of $\hat{p}_{jK}(\mathbf{X})$, $\hat{q}_K(\mathbf{X})$, and $\hat{\theta}_K$ are established by Hamori et al. [13], and we provide a brief review of their results in Section 4.1. Then we establish large sample properties of the proposed estimator $\hat{a}(\mathbf{w})$ in Section 4.2.

4.1. Large sample properties of $\hat{p}_{jK}(\mathbf{X})$, $\hat{q}_K(\mathbf{X})$, and $\hat{\theta}_K$

The convergence rate of $\hat{p}_{jK}(\mathbf{x})$, the convergence rate of $\hat{q}_K(\mathbf{x})$, and the asymptotic normality of $\hat{\theta}_K$ are established in Theorems 1, 3, and 5 of Hamori et al. [13], respectively. Several key assumptions which underpin these theorems are replicated and explained here. First, the sieve approximation errors of $(\rho')^{-1}\{1/\pi_j(\mathbf{x})\}$ and $(\rho')^{-1}\{1/\eta(\mathbf{x})\}$ are assumed to shrink at a polynomial rate:

Assumption 2. There exist $\lambda_{jK}, \beta_K \in \mathbb{R}^K$ and $\alpha > 0$ such that $\sup_{\mathbf{x} \in \mathcal{X}} |(\rho')^{-1}\{1/\pi_j(\mathbf{x})\} - \lambda_{jK}^\top u_K(\mathbf{x})| = O(K^{-\alpha})$ and $\sup_{\mathbf{x} \in \mathcal{X}} |(\rho')^{-1}\{1/\eta(\mathbf{x})\} - \beta_K^\top u_K(\mathbf{x})| = O(K^{-\alpha})$ as $K \rightarrow \infty$.

Assumption 2 is satisfied for a variety of sieve basis functions [25]. If \mathbf{X} is discrete, then the approximation error is zero for sufficiently large K , satisfying Assumption 2 with $\alpha = \infty$. If \mathbf{X} is continuous, then the polynomial rate depends positively on the smoothness of $(\rho')^{-1}\{1/\pi_j(\mathbf{x})\}$ and $(\rho')^{-1}\{1/\eta(\mathbf{x})\}$ in the continuous components and negatively on the number of the continuous components; indeed, for power series and B-splines, $\alpha = -s/r$, where s is the smoothness of approximand and r is the dimension of \mathbf{X} .

The approximation errors $\{N\hat{p}_{jK}(\mathbf{x}) - \pi_j(\mathbf{x})^{-1}\}$ and $\{N\hat{q}_K(\mathbf{x}) - \eta(\mathbf{x})^{-1}\}$ consist of the approximation bias and variance. The bias component depends on both the approximand and the basis function, which is of the rate $O(K^{-\alpha})$. The variance component is proportional to the complexity of the implied model (i.e., the dimension of basis function K). A large K leads to a small approximation bias but a large variance, resulting in over-fitting. A small K leads to a small variance but a large bias, resulting in under-fitting. To balance the bias–variance trade-off in a way that $\hat{\theta}_K$ satisfies the \sqrt{N} -normality, the following restriction on K is imposed:

Assumption 3. $\zeta(K)^2 K^4 / N \rightarrow 0$ and $\sqrt{N} K^{-\alpha} \rightarrow 0$, where $\zeta(K)$ is defined in (15).

Theorems 1, 3, and 5 of Hamori et al. [13] are summarized as Proposition 1, and complete regularity conditions including Assumptions 2–3 are collected in Section 5 of the Online Supplement [14]. See Appendices A and D of Hamori et al. [13] for proofs.

Proposition 1. Under the regularity conditions imposed in [13], the following are true. (i) For $j \in \{0, 1, \dots, d\}$, the convergence rate of $\hat{p}_{jK}(\mathbf{x})$ is characterized as follows:

$$\sup_{\mathbf{x} \in \mathcal{X}} |N\hat{p}_{jK}(\mathbf{x}) - \pi_j(\mathbf{x})^{-1}| = O_p \left\{ \zeta(K) \left(K^{-\alpha} + \sqrt{\frac{K}{N}} \right) \right\} \quad \text{and} \quad \int_{\mathcal{X}} |N\hat{p}_{jK}(\mathbf{x}) - \pi_j(\mathbf{x})^{-1}|^2 dF_{\mathbf{X}}(\mathbf{x}) = O_p \left(K^{-2\alpha} + \frac{K}{N} \right).$$

(ii) The convergence rate of $\hat{q}_K(\mathbf{x})$ is characterized as follows:

$$\sup_{\mathbf{x} \in \mathcal{X}} |N\hat{q}_K(\mathbf{x}) - \eta(\mathbf{x})^{-1}| = O_p \left\{ \zeta(K) \left(K^{-\alpha} + \sqrt{\frac{K}{N}} \right) \right\} \quad \text{and} \quad \int_{\mathcal{X}} |N\hat{q}_K(\mathbf{x}) - \eta(\mathbf{x})^{-1}|^2 dF_{\mathbf{X}}(\mathbf{x}) = O_p \left(K^{-2\alpha} + \frac{K}{N} \right).$$

(iii) It follows that

$$\sqrt{N}(\hat{\theta}_k - \theta_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \xi_i + o_p(1),$$

where $\{\xi_i\}_{i=1}^N$ are i.i.d. random variables with zero mean and finite covariance matrix which is defined in Theorem 5 of Hamori et al. [13] and also replicated in Section 5 of the Online Supplement [14].

4.2. Large sample properties of $\hat{a}(\mathbf{w})$

In this section, we derive the consistency and the asymptotic normality of the proposed estimator $\hat{a}(\mathbf{w})$. The following assumption is imposed to establish the consistency.

Assumption 4. (i) The parameter space $\mathcal{A} \subset \mathbb{R}$ is a compact set and the true parameter $a_0(\mathbf{w})$ lies in the interior of \mathcal{A} ; (ii) $E[\sup_{a \in \mathcal{A}} |L(g(Y) - a)|^2] < \infty$.

Condition (i) is often imposed in the regression literature. Condition (ii) is an envelope condition that is sufficient for the applicability of the uniform law of large numbers.

Theorem 1. Impose Assumptions 1–4 and the regularity conditions which underpin Proposition 1. Then, for each fixed \mathbf{w} , $\hat{a}(\mathbf{w}) \xrightarrow{p} a_0(\mathbf{w})$.

A proof of Theorem 1 is presented in Appendix A.

We next establish the asymptotic normality of $\hat{a}(\mathbf{w})$. To handle a potentially non-smooth loss function, the following assumptions are added.

Assumption 5. (i) The loss function $L(v)$ is differentiable almost everywhere; (ii) $E[L'\{g(Y) - a\}c\{F_0(Y), F_1(w_1), \dots, F_d(w_d); \theta_0\}]$ is differentiable with respect to a , and the derivative is nonzero; (iii) $\sum_{i=1}^N T_{0i} \hat{p}_{0K}(\mathbf{X}_i) L'\{g(Y_i) - \hat{a}(\mathbf{w})\} c\{\hat{F}_{0,K}(Y_i), \hat{F}_{1,K}(w_1), \dots, \hat{F}_{d,K}(w_d); \hat{\theta}_K\} = o_p(N^{-1/2})$.

Assumption 6. Let $R(T_{0i}, \mathbf{X}_i, Y_i; a) := T_{0i} \{\pi_0(\mathbf{X}_i)\}^{-1} L'\{g(Y) - a\} c\{F_0(Y), F_1(w_1), \dots, F_d(w_d); \theta_0\}$. Assume that (i) $E[\sup_{a \in \mathcal{A}} |R(T_{0i}, \mathbf{X}_i, Y_i; a)|^{2+\delta}] < \infty$ for some $\delta > 0$; (ii) for any $a \in \mathcal{A}$ and any $\delta > 0$ and some positive constant ψ ,

$$E \left\{ \sup_{\tilde{a}: |\tilde{a}-a|<\delta} |R(T_{0i}, \mathbf{X}_i, Y_i; \tilde{a}) - R(T_{0i}, \mathbf{X}_i, Y_i; a)|^2 \right\}^{\frac{1}{2}} \leq \text{const} \times \delta^\psi;$$

(iii) the function $E\{R(T_{0i}, \mathbf{X}_i, Y_i; a) | \mathbf{X}_i\}$ is Lipschitz continuous in a .

Assumption 5(i) is a mild condition since $L'(v)$ is not required to be continuous. All example loss functions presented in Section 2.1 satisfy Condition (i). Assumption 5(ii) ensures the asymptotic variance to be finite. Assumption 5(iii) is essentially the asymptotic first order condition, similar to that used in Z-estimation. Popular non-smooth loss functions satisfy this first order condition [29]. Assumption 6 is a sufficient condition for stochastic equicontinuity, which is needed for establishing uniform convergence; see Theorem 4 of Andrews [1]. Again, it is satisfied by widely used nonsmooth loss functions.

For any function $f(v_0, v_1, \dots, v_d; \theta)$, define the following derivatives:

$$\begin{aligned} \partial_{\theta} f(v_0, v_1, \dots, v_d; \theta) &:= \frac{\partial}{\partial \theta} f(v_0, v_1, \dots, v_d; \theta), \\ \partial_j f(v_0, v_1, \dots, v_d; \theta) &:= \frac{\partial}{\partial v_j} f(v_0, v_1, \dots, v_d; \theta) \quad \text{for } j \in \{0, 1, \dots, d\}. \end{aligned}$$

The following theorem provides an asymptotic linear representation of $\hat{a}(\mathbf{w})$, which ensures the asymptotic normality of the proposed estimator.

Theorem 2. Impose Assumptions 1–6 and the regularity conditions which ensure Proposition 1. Then,

$$\sqrt{N}\{\hat{a}(\mathbf{w}) - a_0(\mathbf{w})\} = \frac{1}{\sqrt{N}} \sum_{i=1}^N S(\mathbf{T}_i, \mathbf{X}_i, Y_i; \mathbf{w}) + o_p(1),$$

which implies that $\sqrt{N}\{\hat{a}(\mathbf{w}) - a_0(\mathbf{w})\} \xrightarrow{d} \mathcal{N}\{0, \sigma^2(\mathbf{w})\}$, where

$$\sigma^2(\mathbf{w}) = E\left[\{S(\mathbf{T}, \mathbf{X}, Y; \mathbf{w})\}^2\right], \quad S(\mathbf{T}_i, \mathbf{X}_i, Y_i; \mathbf{w}) = \frac{A_{1i}(\mathbf{w}) + A_{2i}(\mathbf{w}) + A_{3i}(\mathbf{w})}{b(\mathbf{w})},$$

$$\begin{aligned}
b(\mathbf{w}) &= -\partial_a E \left[L' \{g(Y) - a\} c \{F_0(Y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \right] \Big|_{a=a_0}, \\
A_{1i}(\mathbf{w}) &= \frac{T_{0i}}{\pi_0(\mathbf{X}_i)} L' \{g(Y_i) - a_0(\mathbf{w})\} c \{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \\
&\quad - \left\{ \frac{T_{0i}}{\pi_0(\mathbf{X}_i)} - 1 \right\} E \left[L' \{g(Y_i) - a_0(\mathbf{w})\} c \{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \mid \mathbf{X}_i \right], \\
A_{2i}(\mathbf{w}) &= \int L' \{g(y) - a_0(\mathbf{w})\} \partial_0 c \{F_0(y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \\
&\quad \times \left[\frac{T_{0i}}{\pi_0(\mathbf{X}_i)} \mathbf{1}(Y_i \leq y) - \left\{ \frac{T_{0i}}{\pi_0(\mathbf{X}_i)} - 1 \right\} F_{Y|\mathbf{X}}(y|\mathbf{X}_i) - F_0(y) \right] dF_0(y), \\
A_{3i}(\mathbf{w}) &= \sum_{j=1}^d E \left[L' \{g(Y) - a_0(\mathbf{w})\} \partial_j c \{F_0(Y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \right] \\
&\quad \times \left[\frac{T_{ji}}{\pi_j(\mathbf{X}_i)} \mathbf{1}(W_{ji} \leq w_j) - \left\{ \frac{T_{ji}}{\pi_j(\mathbf{X}_i)} - 1 \right\} F_{W_j|\mathbf{X}}(w_j|\mathbf{X}_i) - F_j(w_j) \right] \\
&\quad + \xi_i^\top E \left[L' \{g(Y) - a_0(\mathbf{w})\} \partial_\theta c \{F_0(Y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \right],
\end{aligned}$$

and $F_{Y|\mathbf{X}}$ (resp. $F_{W_j|\mathbf{X}}$) denotes the conditional cumulative distribution function of Y given \mathbf{X} (resp. W_j given \mathbf{X}).

A proof of Theorem 2 is presented in Appendix B.

Theorem 2 contains some important results in the existing literature as special cases. Consider the conditional mean regression with complete data, then the influence function $S(\mathbf{T}_i, \mathbf{X}_i, Y_i; \mathbf{w})$ reduces to that of Noh et al. [27].

Corollary 1. Let Theorem 2 hold. Assume further that $\{Y, \mathbf{W}\}$ are all observed, $g(Y) = Y$, and $L(v) = v^2$. Then, $a_0(\mathbf{w}) = E(Y|\mathbf{W} = \mathbf{w})$ and

$$\sqrt{N} \{\hat{a}(\mathbf{w}) - a_0(\mathbf{w})\} = \frac{1}{\sqrt{N}} \sum_{i=1}^N S_{\text{mean}}(\mathbf{T}_i, \mathbf{X}_i, Y_i; \mathbf{w}) + o_p(1),$$

where

$$\begin{aligned}
S_{\text{mean}}(\mathbf{T}_i, \mathbf{X}_i, Y_i; \mathbf{w}) &= \frac{1}{c_{\mathbf{W}} \{F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\}} \\
&\quad \times \left[- \int \{\mathbf{1}(Y_i \leq y) - F_0(y)\} c \{F_0(y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} dy \right. \\
&\quad + \sum_{j=1}^d E \left[\{Y - a_0(\mathbf{w})\} \partial_j c \{F_0(Y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \right] \{\mathbf{1}(W_{ji} \leq w_j) - F_j(w_j)\} \\
&\quad \left. + \xi_i^\top E \left[\{Y - a_0(\mathbf{w})\} \partial_\theta c \{F_0(Y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \right] \right].
\end{aligned}$$

Hence, Theorem 2 reduces to Theorem 2 of [27].

A proof of Corollary 1 is presented in Appendix C.

Consider the conditional quantile regression with complete data, then $S(\mathbf{T}_i, \mathbf{X}_i, Y_i; \mathbf{w})$ reduces to the influence function of Noh et al. [28].

Corollary 2. Let Theorem 2 hold. Assume further that $\{Y, \mathbf{W}\}$ are all observed, $g(Y) = Y$, and $L(v) = v\{\tau - \mathbf{1}(v \leq 0)\}$. Then, $a_0(\mathbf{w}) = F_{Y|\mathbf{W}}^{-1}(\tau|\mathbf{w}) := \inf\{y : F_{Y|\mathbf{W}}(y|\mathbf{w}) \geq \tau\}$ is the τ^{th} conditional quantile of Y given $\mathbf{W} = \mathbf{w}$, and

$$\sqrt{N} \{\hat{a}(\mathbf{w}) - a_0(\mathbf{w})\} = \frac{1}{\sqrt{N}} \sum_{i=1}^N S_{\text{quantile}}(\mathbf{T}_i, \mathbf{X}_i, Y_i; \mathbf{w}) + o_p(1),$$

where

$$\begin{aligned}
S_{\text{quantile}}(\mathbf{T}_i, \mathbf{X}_i, Y_i; \mathbf{w}) &= \frac{1}{c[F_0\{a_0(\mathbf{w})\}, F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0] f_0\{a_0(\mathbf{w})\}} \\
&\quad \times \left[-[\mathbf{1}\{Y_i \leq a_0(\mathbf{w})\} - F_0\{a_0(\mathbf{w})\}] c[F_0\{a_0(\mathbf{w})\}, F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0] \right.
\end{aligned}$$

$$+ \sum_{j=1}^d E \left[\{\tau - \mathbf{1}(Y \leq a_0)\} \partial_j c \{F_0(Y), F_1(w_1), \dots, F_d(w_d); \theta_0\} \right] \{\mathbf{1}(W_{ji} \leq w_j) - F_j(w_j)\} \\ + \xi_i^\top E \left[\{\tau - \mathbf{1}(Y \leq a_0)\} \partial_\theta c \{F_0(Y), F_1(w_1), \dots, F_d(w_d); \theta_0\} \right] \Big].$$

Hence, [Theorem 2](#) reduces to [Theorem 3.1](#) of Noh et al. [28].

A proof of [Corollary 2](#) is presented in [Appendix D](#).

If $u_K(\mathbf{X})$ is a power series, then [Assumptions 2–3](#) require that $K = O(N^\nu)$ for some $r/2s < \nu < 1/6$, where s is the smoothness of approximand and r is the dimension of \mathbf{X} . It implies that $s > 3r$, hence the dimension of \mathbf{X} should be sufficiently small for [Theorems 1–2](#) to hold. We therefore admit that the proposed method is subject to the curse of dimensionality, a common challenge in nonparametric estimation. Hirano et al. [15] impose a similar but stronger undersmoothing condition $r/2(s - 2r) < \nu < 1/9$. Allowing for high-dimensional covariates is beyond the scope of this paper, and will be pursued in the future work.

5. Variance estimation and confidence interval

The asymptotic normality of $\hat{a}(\mathbf{w})$ established in [Theorem 2](#) has a direct implication for constructing the confidence interval of $a_0(\mathbf{w})$. Evidently, the 95% confidence interval for $a_0(\mathbf{w})$ is given by

$$[\hat{a}(\mathbf{w}) - 1.96 \times \widehat{SE}\{\hat{a}(\mathbf{w})\}, \hat{a}(\mathbf{w}) + 1.96 \times \widehat{SE}\{\hat{a}(\mathbf{w})\}], \quad (25)$$

where $\widehat{SE}\{\hat{a}(\mathbf{w})\} = N^{-1/2}\hat{\sigma}(\mathbf{w})$ is the standard error of $\hat{a}(\mathbf{w})$; $\hat{\sigma}(\mathbf{w})$ is a consistent estimator for the asymptotic standard deviation $\sigma(\mathbf{w})$, which is defined in [Theorem 2](#). Constructing the confidence interval (25) essentially requires the consistent variance estimator $\hat{\sigma}^2(\mathbf{w})$.

$\hat{\sigma}^2(\mathbf{w})$ can broadly be constructed in three ways: plug-in, jackknife, and bootstrap. First, a conceptually straightforward way to estimate $\sigma^2(\mathbf{w})$ is the plug-in method. A practical problem, however, is that $\sigma^2(\mathbf{w})$ has an extremely complicated expression, and many unknown quantities need to be approximated nonparametrically (see [Theorem 2](#)). Hence, the plug-in method is not recommended from a practical point of view. (The formula of the plug-in variance estimator under the conditional mean regression is omitted to save space, but is available upon request.)

Second, we consider the jackknife method [7]. The i th jackknife sample is constructed by deleting the i th observation from the dataset:

$$\mathbf{J}_{[-i]} := \{\mathbf{T}_j, \mathbf{W}_j, \mathbf{X}_j, Y_j : j \in \{1, 2, \dots, i-1, i+1, \dots, N\}\}.$$

The i th jackknife replicate, denoted as $\hat{a}_{[-i]}(\mathbf{w})$, is defined as the point estimator for $a_0(\mathbf{w})$ computed on the i th jackknife sample $\mathbf{J}_{[-i]}$. The jackknife-based standard error is given by

$$\widehat{SE}_{jack}\{\hat{a}(\mathbf{w})\} = \left[\frac{N-1}{N} \sum_{i=1}^N \{\hat{a}_{[-i]}(\mathbf{w}) - \hat{a}_{[-\cdot]}(\mathbf{w})\}^2 \right]^{\frac{1}{2}}, \quad (26)$$

where $\hat{a}_{[-\cdot]}(\mathbf{w}) = N^{-1} \sum_{i=1}^N \hat{a}_{[-i]}(\mathbf{w})$. Substitute (26) into (25) to compute the confidence interval.

Third, we consider the bootstrap method [7]. The b th bootstrap sample $\{\mathbf{T}_i^{[b]}, \mathbf{W}_i^{[b]}, \mathbf{X}_i^{[b]}, Y_i^{[b]}\}_{i=1}^N$ is resampled with replacement from the original sample $\{\mathbf{T}_i, \mathbf{W}_i, \mathbf{X}_i, Y_i\}_{i=1}^N$ with the uniform probability. The b th bootstrap replicate, denoted as $\hat{a}^{[b]}(\mathbf{w})$, is defined as the point estimator for $a_0(\mathbf{w})$ computed on the b th bootstrap sample. Repeat B times to get $\{\hat{a}^{[b]}(\mathbf{w})\}_{b=1}^B$. The bootstrapped standard error is given by

$$\widehat{SE}_{boot}\{\hat{a}(\mathbf{w})\} = \left[\frac{1}{B} \sum_{b=1}^B \{\hat{a}^{[b]}(\mathbf{w}) - \hat{a}^{[-\cdot]}(\mathbf{w})\}^2 \right]^{\frac{1}{2}}, \quad (27)$$

where $\hat{a}^{[-\cdot]}(\mathbf{w}) = B^{-1} \sum_{b=1}^B \hat{a}^{[b]}(\mathbf{w})$. Substitute (27) into (25) to compute the confidence interval.

Bootstrapping provides another way to construct a confidence interval. Sort the B bootstrap replicates from the smallest to the largest, and relabel them as $\hat{a}^{(1)}(\mathbf{w}) \leq \dots \leq \hat{a}^{(B)}(\mathbf{w})$. The 95% bootstrapped confidence interval is given as

$$[\hat{a}^{(0.025B)}(\mathbf{w}), \hat{a}^{(0.975B)}(\mathbf{w})]. \quad (28)$$

The entire confidence interval (25) can be replaced with (28). The former bootstrap approach (27) relies on the asymptotic normality result, while the latter approach (28) does not. We distinguish them hereafter, calling the former bootstrap method I and the latter bootstrap method II.

6. Data-driven selection of tuning parameters

The calibration approach requires a selection of the tuning parameter K . More generally, the calibration weights can be computed with distinct tuning parameters as $\{\hat{p}_{0,K_0}(\mathbf{X}), \dots, \hat{p}_{d,K_d}(\mathbf{X}), \hat{q}_{K_\eta}(\mathbf{X})\}$. The asymptotic theory of the proposed estimator permits various values for $\mathbf{K} = (K_0, \dots, K_d, K_\eta)^\top$. This poses a dilemma for applied researchers who have only one finite sample and would like some guidance on the selection of \mathbf{K} . Several data-driven selection methods are discussed in [20,21] among others. Based on this background, we present two well-known approaches: covariate balancing and cross validation.

6.1. Covariate balancing approach

The first approach is to select the tuning parameters that achieve the best balance of the covariate distributions between the observed and whole groups. Note that

$$E\{T_{ji}\pi_j^{-1}(\mathbf{X}_i)\mathbf{1}(\mathbf{X}_i \leq \mathbf{x})\} = E\{\mathbf{1}(\mathbf{X}_i \leq \mathbf{x})\}, \quad j \in \{0, 1, \dots, d\},$$

$$E\{\mathbf{1}(T_{0i} = \dots = T_{di} = 1)\eta^{-1}(\mathbf{X}_i)\mathbf{1}(\mathbf{X}_i \leq \mathbf{x})\} = E\{\mathbf{1}(\mathbf{X}_i \leq \mathbf{x})\}.$$

Each tuning parameter is selected so that it minimizes the integrated distance function:

$$\hat{K}_j := \arg \min_{K_j \in \{1, \dots, \bar{K}\}} \sum_{i=1}^N \left\{ \hat{F}_{j,K_j}(\mathbf{X}_i) - \hat{F}_X(\mathbf{X}_i) \right\}^2, \quad j \in \{0, 1, \dots, d\},$$

$$\hat{K}_\eta := \arg \min_{K \in \{1, \dots, \bar{K}\}} \sum_{i=1}^N \left\{ \hat{F}_{\eta,K}(\mathbf{X}_i) - \hat{F}_X(\mathbf{X}_i) \right\}^2,$$

where $\bar{K} \in \mathbb{N}$ is a prespecified upper bound, $\hat{F}_X(\mathbf{x}) := (N+1)^{-1} \sum_{i=1}^N \mathbf{1}(\mathbf{X}_i \leq \mathbf{x})$ is the empirical distribution of \mathbf{X} based on the whole group, and

$$\hat{F}_{j,K_j}(\mathbf{x}) := \sum_{i=1}^N T_{ji} \hat{p}_{jK}(\mathbf{X}_i) \mathbf{1}(\mathbf{X}_i \leq \mathbf{x}),$$

$$\hat{F}_{\eta,K}(\mathbf{x}) := \sum_{i=1}^N \mathbf{1}(T_{0i} = \dots = T_{di} = 1) \hat{q}_K(\mathbf{X}_i) \mathbf{1}(\mathbf{X}_i \leq \mathbf{x}).$$

An advantage of this approach is that its performance is not affected by the dimension of regressors d , since \mathbf{W} does not play any role here. The performance of this approach, however, is affected by the dimension of covariates r . Another potential issue is that this approach exploits the information contained in (\mathbf{T}, \mathbf{X}) but not in (Y, \mathbf{W}) .

6.2. M-Fold cross validation

The second approach is the M -fold cross validation (CV). Note that $E[L\{g(Y) - a_0(\mathbf{W})\}] \leq E[L\{g(Y) - a(\mathbf{W})\}]$ holds for any $\sigma(\mathbf{W})$ -measurable function $a(\mathbf{W})$, and that $E[L\{g(Y_i) - a(\mathbf{W}_i)\}] = E[\mathbf{1}(T_{0i} = T_{1i} = \dots = T_{di} = 1)\eta^{-1}(\mathbf{X}_i)L\{g(Y_i) - a(\mathbf{W}_i)\}]$. Based on these observations, the M -fold CV proceeds as follows:

Step 1 Divide N individuals into M groups (e.g., $M = 5$ or 10), and let $n = N/M$. Assume for simplicity that n is a natural number. Data in the m th group are denoted as $S_m \equiv \{\mathbf{X}_i^{(m)}, \mathbf{W}_i^{(m)}, \mathbf{T}_i^{(m)}, Y_i^{(m)}\}_{i=1}^n$ for $m \in \{1, \dots, M\}$.

Step 2 For each $m \in \{1, \dots, M\}$, define $S_{(-m)} = \{\mathbf{X}_i, \mathbf{W}_i, \mathbf{T}_i, Y_i\}_{i=1}^N \setminus S_m$. Compute calibration weights on $S_{(-m)}$:

$$\begin{aligned} \hat{\lambda}_{jK_j}^{(-m)} &= \arg \max_{\lambda \in \mathbb{R}^{K_j}} \hat{G}_{jK_j}^{(-m)}(\lambda), \quad j \in \{0, 1, \dots, d\}, \quad \hat{\beta}_{K_\eta}^{(-m)} = \arg \max_{\beta \in \mathbb{R}^{K_\eta}} \hat{H}_{K_\eta}^{(-m)}(\beta), \\ \hat{G}_{jK_j}^{(-m)}(\lambda) &= \frac{1}{N-n} \sum_{i \in S_{(-m)}} T_{ji} \rho \left\{ \lambda^\top u_{K_j}(\mathbf{X}_i) \right\} - \frac{1}{N-n} \sum_{i \in S_{(-m)}} \lambda^\top u_{K_j}(\mathbf{X}_i), \\ \hat{H}_{K_\eta}^{(-m)}(\beta) &= \frac{1}{N-n} \sum_{i \in S_{(-m)}} \mathbf{1}(T_{0i} = \dots = T_{di} = 1) \rho \left\{ \beta^\top u_{K_\eta}(\mathbf{X}_i) \right\} - \frac{1}{N-n} \sum_{i \in S_{(-m)}} \beta^\top u_{K_\eta}(\mathbf{X}_i), \\ \hat{p}_{jK_j}^{(-m)}(\mathbf{X}) &= \frac{\rho' \left[\left\{ \hat{\lambda}_{jK_j}^{(-m)} \right\}^\top u_{K_j}(\mathbf{X}) \right]}{N-n}, \quad \hat{q}_{K_\eta}^{(-m)}(\mathbf{X}) = \frac{\rho' \left[\left\{ \hat{\beta}_{K_\eta}^{(-m)} \right\}^\top u_{K_\eta}(\mathbf{X}) \right]}{N-n}. \end{aligned}$$

Estimate marginal and joint distributions using the calibration weights:

$$\begin{aligned}\hat{F}_{0,K_0}^{(-m)}(y) &= \sum_{i \in S_{(-m)}} T_{0i} \hat{p}_{0K_0}^{(-m)}(\mathbf{X}_i) \mathbf{1}(Y_i \leq y), \quad \hat{F}_{j,K_j}^{(-m)}(w) = \sum_{i \in S_{(-m)}} T_{ji} \hat{p}_{jK_j}^{(-m)}(\mathbf{X}_i) \mathbf{1}(W_{ji} \leq w), \quad j \in \{1, \dots, d\}, \\ \hat{\theta}_K^{(-m)} &= \arg \max_{\theta \in \Theta} \sum_{i \in S_{(-m)}} \mathbf{1}(T_{0i} = \dots = T_{di} = 1) \hat{q}_{K_\eta}^{(-m)}(\mathbf{X}_i) \ln c \left\{ \hat{F}_{0,K_0}^{(-m)}(Y_i), \hat{F}_{1,K_1}^{(-m)}(W_{1i}), \dots, \hat{F}_{d,K_d}^{(-m)}(W_{di}); \theta \right\}.\end{aligned}$$

Finally, estimate the regression curve:

$$\hat{a}_K^{(-m)}(w) = \arg \min_{a \in \mathbb{R}} \sum_{i \in S_{(-m)}} T_{0i} \hat{p}_{0K_0}^{(-m)}(\mathbf{X}_i) L\{g(Y_i) - a\} c \left\{ \hat{F}_{0,K_0}^{(-m)}(Y_i), \hat{F}_{1,K_1}^{(-m)}(w_1), \dots, \hat{F}_{d,K_d}^{(-m)}(w_d); \hat{\theta}_K^{(-m)} \right\}.$$

Step 3 Choose optimal $\mathbf{K} = (K_0, \dots, K_d, K_\eta)^\top$ that minimizes the CV criterion:

$$CV(\mathbf{K}) = \sum_{m=1}^M \left[\sum_{i \in S_m} \mathbf{1}(T_{0i} = \dots = T_{di} = 1) \hat{q}_{K_\eta}^{(-m)}(\mathbf{X}_i) L \left\{ g(Y_i) - \hat{a}_K^{(-m)}(\mathbf{W}_i) \right\} \right].$$

When $M = N$, this approach coincides with the leave-out CV. Li [20] shows that the M -fold CV is asymptotically optimal in the sense of minimizing a weighted loss function for regression.

7. Monte Carlo simulation

In this section, we perform Monte Carlo simulations in order to evaluate the finite sample performance of the calibration approach. Here we focus on a benchmark scenario which has one regressor ($d = 1$) and one covariate ($r = 1$) in order to conserve space. In Sections 6.2 and 6.3 of the Online Supplement [14], we discuss two extended scenarios for completeness: the extended scenario I which has one regressor ($d = 1$) and two covariates ($r = 2$) and the extended scenario II which has two regressors ($d = 2$) and one covariate ($r = 1$).

Let $\mathbf{Z}_i = (Z_{1i}, Z_{2i}, Z_{3i})^\top$, and draw \mathbf{Z}_i independently and identically from either the trivariate Clayton copula with parameter $\theta_0 = 1.333$ or the trivariate Gumbel copula with $\theta_0 = 1.667$. For both cases, the implied Kendall's tau is $\tau = 0.4$, a moderate level of association. Define the regressand $Y_i = \Phi^{-1}(Z_{1i})$, regressor $W_i = \Phi^{-1}(Z_{2i})$, and covariate $X_i = \Phi^{-1}(Z_{3i})$, where $\Phi^{-1}(\cdot)$ is the inverse distribution function of $\mathcal{N}(0, 1)$. Assume that W_i and X_i are observed for all $i \in \{1, \dots, N\}$, where the sample size is $N \in \{250, 500, 750\}$. Let T_i be a binary indicator which equals 1 if Y_i is observed and 0 if Y_i is missing. The regressand Y_i may be missing with the propensity score function:

$$\Pr(T_i = 1 | X_i = x) = \frac{1}{1 + \exp(b_0 + b_1 x)}. \quad (29)$$

The logistic function is commonly used to specify the propensity score function in the literature of missing data analysis. Suppose that $(b_0, b_1) = (-0.57, 1.5)$, in which case $E(T_i) = 0.6$ and Y_i is MAR.

Following Noh et al. [27], the conditional mean regression is considered here. In this case, the estimated regression curve $\hat{a}(w)$ has the closed-form solution (see Remark 3):

$$\hat{a}(w) = \frac{\sum_{i=1}^N T_i \hat{p}_K(X_i) Y_i c \left\{ \hat{F}_{0,K}(Y_i), \hat{F}_{1,K}(w); \hat{\theta}_K \right\}}{\sum_{i=1}^N T_i \hat{p}_K(X_i) c \left\{ \hat{F}_{0,K}(Y_i), \hat{F}_{1,K}(w); \hat{\theta}_K \right\}}.$$

Hence, the computation of $\hat{a}(w)$ is straightforward once the calibration weight $\hat{p}_K(X_i)$ is computed.

The calibration weight $\hat{p}_K(X_i)$ is computed with two alternative specifications for $\rho(\cdot)$. The first one is the exponential tilting (ET): $\rho(v) = -\exp(-v)$. The second one is the inverse logistic (IL): $\rho(v) = v - \exp(-v)$. A practical advantage of these specifications is that $\rho(v)$ is a well-defined increasing, strictly concave function for any $v \in \mathbb{R}$. This property stabilizes numerical optimization especially in small samples, since the calibration weight must be positive by definition and its formula is indeed $\hat{p}_K(X) = N^{-1} \rho'(\hat{\lambda}_K^\top u_K(X))$. See Section 6.1 of the Online Supplement [14] for other possible but less practical specifications for $\rho(\cdot)$.

To compute the calibration weight, the sieve basis function is specified as $u_K(X) = (1, X, \dots, X^{K-1})^\top$. The tuning parameter K is either fixed at $K \in \{2, 3, 4\}$, or automatically selected from the choice set $\{2, 3, 4\}$ via the balancing principle of covariate distributions (CB) or the M -fold cross validation (CV) with $M \in \{5, 10\}$. See Sections 6.1 and 6.2 for the procedures of CB and CV, respectively.

Since W is assumed to be always observed, the single calibration weight $\hat{p}_K(X)$ suffices in this study. The calibration weight with respect to the joint distribution of (Y, W) , namely $\hat{q}_K(X)$, is identical to $\hat{p}_K(X)$ by construction. In the present simulation, a model misspecification is not discussed. When the underlying copula is Clayton (Gumbel), we fit the Clayton (Gumbel) copula to estimate θ_0 and $a_0(w)$.

For comparison, the equal weight $\hat{p}_K(X) = 1/\sum_{i=1}^N T_i$ is also considered. The equal-weight approach, which is essentially equivalent to the list-wise deletion, should fail under the MAR mechanism (29), since by construction it ignores the impact of X on the propensity score.

Table 1

IRMSE of $\hat{a}(w)$ based on the calibration approach. This table reports the integrated root mean squared errors (IRMSEs) of the estimated regression curve $\hat{a}(w)$ under the benchmark scenario. Y is MAR and the sample size is $N \in \{250, 500, 750\}$. The IRMSE is computed across the grid $w \in \{-3.00, -2.95, \dots, 2.95, 3.00\}$ and $J = 1000$ Monte Carlo samples. The calibration approach is taken, and two alternative specifications are used for $\rho(\cdot)$: exponential tilting (ET) and inverse logistic (IL). The sieve basis function is specified as $u_K(X) = (1, X, \dots, X^{K-1})^T$. The tuning parameter K is either fixed at $K \in \{2, 3, 4\}$, or automatically selected from the choice set $\{2, 3, 4\}$ via covariate balancing (CB) or M -fold cross validation (CV) with $M \in \{5, 10\}$. For comparison, the IRMSEs based on the equal-weight approach are also reported.

Clayton copula, $N = 250$							Gumbel copula, $N = 250$						
K	2	3	4	CB	CV5	CV10	K	2	3	4	CB	CV5	CV10
ET	0.174	0.160	0.170	0.162	0.169	0.173	ET	0.291	0.258	0.282	0.253	0.288	0.289
IL	0.147	0.151	0.194	0.157	0.146	0.144	IL	0.255	0.262	0.350	0.261	0.259	0.256
IRMSE of the equal-weight approach = 0.270							IRMSE of the equal-weight approach = 0.427						
Clayton copula, $N = 500$							Gumbel copula, $N = 500$						
K	2	3	4	CB	CV5	CV10	K	2	3	4	CB	CV5	CV10
ET	0.133	0.123	0.113	0.120	0.127	0.126	ET	0.226	0.177	0.186	0.173	0.224	0.220
IL	0.104	0.109	0.125	0.104	0.106	0.103	IL	0.181	0.180	0.209	0.184	0.183	0.179
IRMSE of the equal-weight approach = 0.249							IRMSE of the equal-weight approach = 0.384						
Clayton copula, $N = 750$							Gumbel copula, $N = 750$						
K	2	3	4	CB	CV5	CV10	K	2	3	4	CB	CV5	CV10
ET	0.094	0.105	0.088	0.103	0.096	0.096	ET	0.236	0.145	0.151	0.142	0.215	0.212
IL	0.085	0.090	0.097	0.085	0.088	0.087	IL	0.150	0.144	0.162	0.150	0.152	0.148
IRMSE of the equal-weight approach = 0.241							IRMSE of the equal-weight approach = 0.370						

7.1. Performance of point estimation

To evaluate the finite sample performance of the point estimation for $a_0(w)$, draw $J = 1000$ Monte Carlo samples and compute integrated root mean squared errors (IRMSEs) as follows: First, RMSE in the j th sample is defined as

$$RMSE_j = \sqrt{\frac{1}{\#\mathcal{W}} \sum_{w \in \mathcal{W}} \{\hat{a}_j(w) - a_0(w)\}^2}, \quad (30)$$

where $\hat{a}_j(w)$ is the estimated target function; \mathcal{W} is the set of w 's considered; $\#\mathcal{W}$ is the number of w 's considered. Since the marginal distribution of the regressor W_i is $\mathcal{N}(0, 1)$, the range $w \in [-3, 3]$ should be covered in order to properly evaluate the performance of each estimator. Dividing this interval more finely would make the evaluation more accurate, but it would raise computational burden. To balance the trade-off between evaluation accuracy and computational speed, we use $\mathcal{W} = \{-3.00, -2.95, \dots, 2.95, 3.00\}$ and hence $\#\mathcal{W} = 121$.

Eq. (30) contains the true value $a_0(w)$. Recall that $a_0(w) = \arg \min_{a \in \mathbb{R}} E[L\{g(Y) - a\}c\{F_0(Y), F_1(w); \theta_0\}]$, which can be approximated by $a_0(w) \simeq \arg \min_{a \in \mathbb{R}} (1/N) \sum_{i=1}^N L\{g(Y_i) - a\}c\{F_0(Y_i), F_1(w); \theta_0\}$. Since $L\{g(Y) - a\} = (Y - a)^2$ in the present study, $a_0(w)$ is simply given by

$$a_0(w) = \frac{\sum_{i=1}^N Y_i c\{F_0(Y_i), F_1(w); \theta_0\}}{\sum_{i=1}^N c\{F_0(Y_i), F_1(w); \theta_0\}}. \quad (31)$$

Substitute (31) into (30) to compute $RMSE_j$. Finally, the IRMSE is defined as $IRMSE = (1/J) \sum_{j=1}^J RMSE_j$ with $J = 1000$. The smaller value of IRMSE implies the higher precision in the point estimation of $a_0(w)$.

See Table 1 for the resulting IRMSEs. For each copula and sample size, there are $2 \times 6 = 12$ versions of the calibration approach, since $\rho(\cdot)$ is specified in 2 ways and K is selected in 6 ways. For any copula and sample size, the IRMSE of any version of the calibration approach is always smaller than the IRMSE of the equal-weight approach. This is overwhelming evidence that the calibration approach strictly dominates the equal-weight approach when Y is MAR. Focus on the Clayton copula with $N = 250$, for example. The IRMSE of the calibration approach takes the smallest value of 0.144 when the inverse logistic (IL) ρ -function is used and K is selected via the 10-fold cross validation (CV10); it takes the largest value of 0.194 when the IL ρ -function is used and K is fixed at 4. The IRMSE of the equal-weight approach is 0.270, which is clearly worse than the worst case of the calibration approach.

For any of the 12 versions of the calibration approach, the IRMSE shrinks as the sample size grows, which confirms the asymptotic validity of the calibration approach. See the pair (IL, CV10) under the Clayton copula, for example. The IRMSEs are $\{0.144, 0.103, 0.087\}$ for $N \in \{250, 500, 750\}$, respectively.

We next discuss the relative performance between the exponential tilting (ET) and IL ρ -functions. There are $2 \times 3 \times 6 = 36$ cases where they can be compared, since 2 copulas, 3 sample sizes, and 6 ways to select K are considered. In 25 out of the 36 cases (69.4%), IL leads to the smaller IRMSE than ET. In particular, IL always outperforms ET whenever the cross

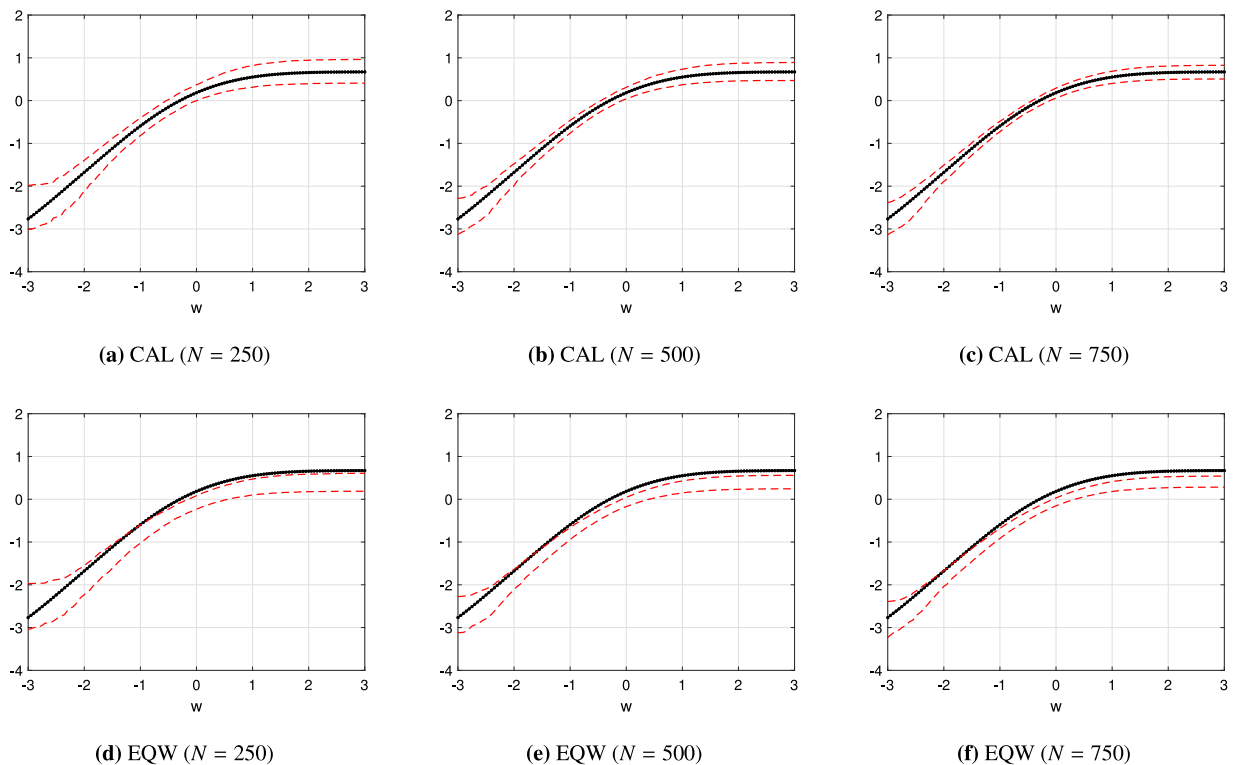


Fig. 1. True $a_0(w)$ and the upper and lower 2.5-percentiles of $\{\hat{a}_j(w)\}_{j=1}^J$ (Clayton copula). This figure plots the true regression curve $a_0(w)$ in black, solid lines and the upper and lower 2.5-percentiles of estimated regression curves $\{\hat{a}_j(w)\}_{j=1}^J$ across $J = 1000$ Monte Carlo samples in red, dashed lines. The benchmark scenario with the Clayton copula is considered, and the sample sizes are $N \in \{250, 500, 750\}$. The regression curve is estimated via the calibration approach (CAL) or the equal-weight approach (EQW). For CAL, the inverse logistic $\rho(\cdot)$ and the data-driven selection of K based on the 10-fold cross validation are used, where the choice set is $\{2, 3, 4\}$.

validations are used to select K . In the extended scenarios of Sections 6.2 and 6.3 of the Online Supplement [14], we shall observe similar or even stronger evidence for IL. Hence, applied researchers are advised to use IL instead of ET in general.

Focusing on IL, we investigate the relative performance among the 6 selection mechanisms for K . For each copula and sample size, the IRMSE is (nearly) minimized when CV10 is used. CV10 attains the smallest IRMSE in 3 out of the 6 cases, and almost the smallest IRMSE in the other 3 cases. Thus, applied researchers are advised to use the pair (IL, CV10) in order to maximize the accuracy of the point estimation of the regression curve $a_0(w)$.

We plot the true regression curve $a_0(w)$ and the upper and lower 2.5-percentiles of estimated regression curves $\{\hat{a}_j(w)\}_{j=1}^J$ across $J = 1000$ Monte Carlo samples. See Figs. 1–2 for the plots under the Clayton and Gumbel copulas, respectively. The calibration approach is implemented with (IL, CV10) as concluded above. For comparison, the conventional equal-weight approach is also considered.

Under the Clayton copula, the upper and lower 2.5-percentiles of $\{\hat{a}_j(w)\}_{j=1}^J$ based on the calibration approach are almost identical to $a_0(w)$ for any $w \in [-3, 3]$ (Panels (a)–(c), Fig. 1). The distance between the two percentiles is sufficiently small even for $N = 250$, and it further shrinks toward $a_0(w)$ as N increases. It indicates that the calibration approach performs strikingly well.

For the equal-weight approach, the upper 2.5-percentile is smaller than $a_0(w)$ for most $w \in [-3, 3]$, indicating the presence of negative bias (Panels (d)–(f), Fig. 1). This bias is a major source of the large IRMSE observed in Table 1. The bias does not vanish as N increases, which implies that the equal-weight approach fails when Y is MAR.

Under the Gumbel copula, the upper and lower 2.5-percentiles of $\{\hat{a}_j(w)\}_{j=1}^J$ based on the calibration approach are almost identical to $a_0(w)$ for $w \leq 1$ (Panels (a)–(c), Fig. 2). The distance between the two percentiles is relatively large for $w > 1$, suggesting that the upper tail dependence of the Gumbel copula combined with the logistic-type MAR mechanism (29) has a large adverse effect on point estimation. Nevertheless, the two percentiles contain $a_0(w)$ especially when $N = 750$, indicating that reasonably sharp estimation at the upper tail is guaranteed in large samples.

The equal-weight approach utterly fails under the Gumbel copula (Panels (d)–(f), Fig. 2). There exists extremely large negative bias for $w > 0$ regardless of the sample size. This result highlights the fatal shortcoming of the equal-weight approach and the remarkable use of the calibration approach under the MAR mechanism.

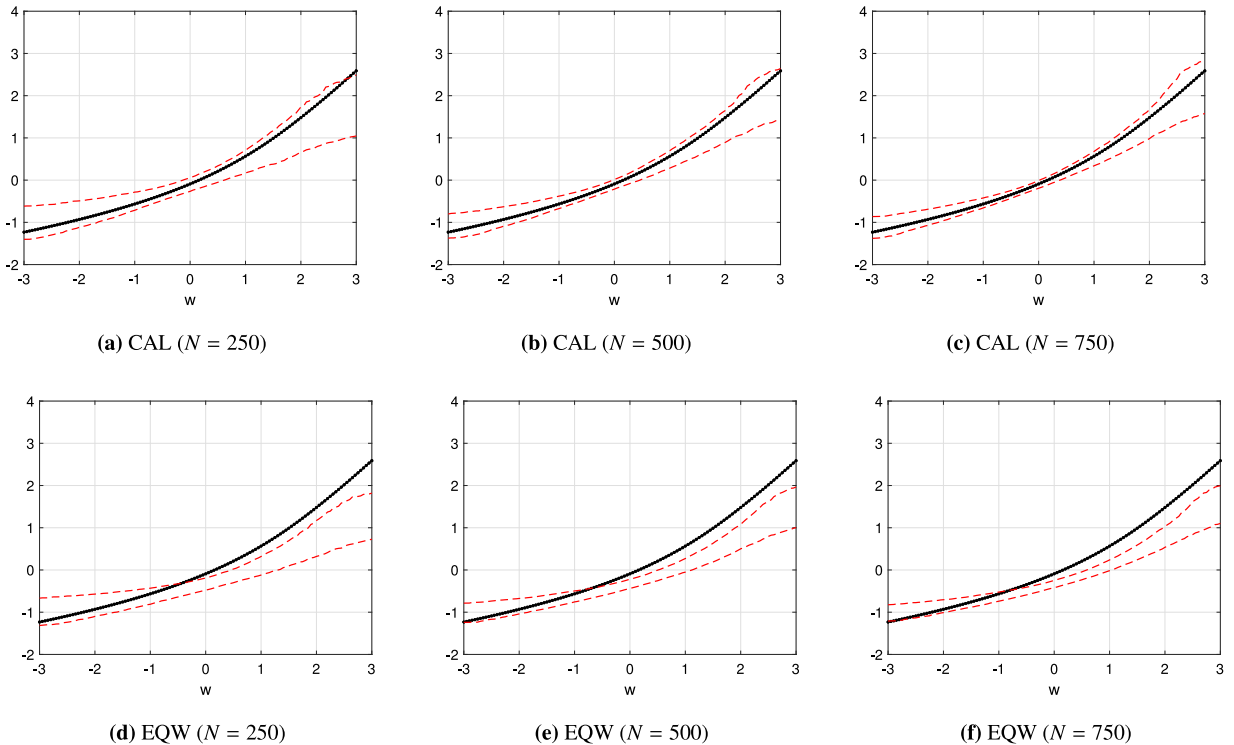


Fig. 2. True $a_0(w)$ and the upper and lower 2.5-percentiles of $\{\hat{a}_j(w)\}_{j=1}^J$ (Gumbel copula). This figure plots the true regression curve $a_0(w)$ in black, solid lines and the upper and lower 2.5-percentiles of estimated regression curves $\{\hat{a}_j(w)\}_{j=1}^J$ across $J = 1000$ Monte Carlo samples in red, dashed lines. The benchmark scenario with the Gumbel copula is considered, and the sample sizes are $N \in \{250, 500, 750\}$. The regression curve is estimated via the calibration approach (CAL) or the equal-weight approach (EQW). For CAL, the inverse logistic $\rho(\cdot)$ and the data-driven selection of K based on the 10-fold cross validation are used, where the choice set is $\{2, 3, 4\}$.

7.2. Performance of variance estimation

In this section, we investigate the finite sample performance of the variance estimation associated with the calibration approach. The inverse logistic $\rho(v) = v - \exp(-v)$ is used to compute calibration weights, since we know from the previous section that it outperforms the exponential tilting $\rho(\cdot)$ in point estimation. We compute 95% confidence intervals for $a_0(w)$, using the bootstrap method I (recall Section 5). In this method, the confidence interval is constructed in accordance with the asymptotic normality result, where the asymptotic variance is approximated via bootstrapping with $B = 500$ iterations (see (25) and (27)).

In Section 6.1.2 of the Online Supplement [14], we also consider the jackknife method and the bootstrap method II. The three methods perform as well as each other in many cases, but the bootstrap method I performs best in some challenging cases such as the upper tail of the Gumbel distribution. Hence, in the present section, we focus on the bootstrap method I in order to conserve space.

The grid is simplified as $w \in \{-2.5, -1.5, -0.5, 0.5, 1.5, 2.5\}$. Another simplification is that the tuning parameter is fixed at $K \in \{2, 3, 4\}$ and the data-driven selection of K is not considered. Each K is used for computing a point estimator $\hat{a}(w)$, and the same K is used for computing a confidence interval. These simplifications partly stem from the heavy computational burden of bootstrapping. Another reason is that it is well known that a data-driven selection of K in variance estimation is rather challenging. Nevertheless, the present simple set-up is informative enough to evaluate the finite sample performance of the bootstrap method I.

For each of $J = 1000$ Monte Carlo samples, compute a 95% confidence interval $CI(w) = [b_\ell(w), b_u(w)]$, resulting in $\{CI^{(1)}(w), \dots, CI^{(J)}(w)\}$. Compute the coverage probability of the 95% confidence intervals as $CP(w) = (1/J) \sum_{j=1}^J \mathbf{1}_{\{a_0(w) \in CI^{(j)}(w)\}}$ and the mean length of the 95% confidence intervals as $\bar{L}(w) = (1/J) \sum_{j=1}^J \{b_u^{(j)}(w) - b_\ell^{(j)}(w)\}$, where $CI^{(j)}(w) = [b_\ell^{(j)}(w), b_u^{(j)}(w)]$. Both $CP(w)$ and $\bar{L}(w)$ should be inspected when evaluating the finite sample performance of the confidence interval. There is often a trade-off between the two measures, and an accurate confidence interval is supposed to achieve both $CP(w)$ which is close enough to 0.95 and small enough $\bar{L}(w)$.

See Table 2 for simulation results under the Clayton copula. For almost all values of $N \in \{250, 500, 750\}$, $K \in \{2, 3, 4\}$, and $w \in \{-2.5, \dots, 2.5\}$, the coverage probability is close enough to 0.95 as desired. It exceeds 0.9 in 48 out of the

Table 2

Performance of confidence intervals based on the bootstrap method I (Clayton copula). In this table, the benchmark scenario with the Clayton copula is considered. Y is MAR and the sample size is $N \in \{250, 500, 750\}$. The regression curve $a_0(w)$ is estimated via the calibration approach with inverse logistic $\rho(v) = v - \exp(-v)$, where $w \in \{-2.5, -1.5, \dots, 1.5, 2.5\}$. The sieve basis function is specified as $u_K(X) = (1, X, \dots, X^{K-1})^T$ with $K \in \{2, 3, 4\}$. The 95% confidence interval (CI) for $a_0(w)$ is computed in accordance with the asymptotic normality result, where the asymptotic variance is approximated by bootstrapping. The same K as in the point estimation is used for $B = 500$ bootstrap iterations. This table reports the coverage probability and the mean length of the CIs across $J = 1000$ Monte Carlo samples.

Coverage probability of 95% CI ($N = 250$)							Mean length of 95% CI ($N = 250$)						
$K \backslash w$	-2.5	-1.5	-0.5	0.5	1.5	2.5	$K \backslash w$	-2.5	-1.5	-0.5	0.5	1.5	2.5
2	0.900	0.934	0.941	0.933	0.924	0.924	2	0.748	0.519	0.348	0.420	0.507	0.528
3	0.888	0.939	0.947	0.938	0.943	0.942	3	0.768	0.564	0.357	0.459	0.555	0.578
4	0.908	0.959	0.954	0.939	0.938	0.937	4	0.841	0.576	0.430	0.632	0.761	0.789

Coverage probability of 95% CI ($N = 500$)							Mean length of 95% CI ($N = 500$)						
$K \backslash w$	-2.5	-1.5	-0.5	0.5	1.5	2.5	$K \backslash w$	-2.5	-1.5	-0.5	0.5	1.5	2.5
2	0.889	0.950	0.956	0.948	0.935	0.935	2	0.613	0.359	0.244	0.294	0.356	0.371
3	0.855	0.949	0.947	0.932	0.931	0.932	3	0.639	0.375	0.249	0.314	0.381	0.398
4	0.893	0.955	0.954	0.954	0.951	0.948	4	0.668	0.392	0.295	0.439	0.536	0.557

Coverage probability of 95% CI ($N = 750$)							Mean length of 95% CI ($N = 750$)						
$K \backslash w$	-2.5	-1.5	-0.5	0.5	1.5	2.5	$K \backslash w$	-2.5	-1.5	-0.5	0.5	1.5	2.5
2	0.897	0.937	0.935	0.942	0.926	0.931	2	0.564	0.292	0.200	0.244	0.295	0.308
3	0.897	0.934	0.947	0.939	0.943	0.943	3	0.583	0.300	0.202	0.254	0.308	0.322
4	0.909	0.952	0.957	0.956	0.963	0.960	4	0.579	0.307	0.235	0.344	0.419	0.436

$3 \times 3 \times 6 = 54$ cases. In the remaining 6 cases, the coverage probability is always above 0.85. The mean length of the confidence intervals diminishes as N increases for all cases, indicating the asymptotic validity of the bootstrap method I. In all cases, the mean length is sufficiently small; it ranges between 0.200 and 0.841. Thus, the bootstrap method I leads to strikingly accurate confidence intervals under the Clayton copula.

See Table 3 for simulation results under the Gumbel copula. The coverage probability is sufficiently close to 0.95 for $w \leq 0.5$, but sometimes well below 0.95 for $w \geq 1.5$. See the case with $N = 250$ and $K = 3$, for example. The coverage probabilities are $\{.867, .907, .944, .871, .764, .467\}$ for $w \in \{-2.5, -1.5, -0.5, 0.5, 1.5, 2.5\}$, respectively. Similarly, the mean length of the confidence intervals is sufficiently small for $w \leq 0.5$, but sometimes large for $w \geq 1.5$. Keeping the same example, the mean length is $\{0.752, 0.553, 0.367, 0.390, 0.716, 1.033\}$. These are not surprising results, since the point estimation itself exhibits some distortions at the upper tail of the Gumbel copula (recall Fig. 2). We observe that the bootstrap is asymptotically valid for any $w \in \{-2.5, \dots, 2.5\}$, since both coverage probability and mean length improve as the sample size increases for each $K \in \{2, 3, 4\}$.

In summary, the bootstrap method I delivers remarkably accurate confidence intervals except for the upper tail of Gumbel. As the sample size grows, the performance at the upper tail of the Gumbel copula improves steadily.

8. Empirical application

In this section, we analyze the relationship between research and development (R&D) expenses and revenues of German manufacturing firms in 2017. R&D plays a key role in the manufacturing industry, and we use revenues as a proxy of the firm size. Intuitively, the larger manufacturing firm should have the larger R&D expense, hence the two variables should be positively correlated. A primary goal of this study is to check if that is indeed the case in Germany, with an explicit attention to missing data.

8.1. Data and methodology

Let the regressor W_i be the log of operating revenue (turnover) of firm i . Let the regressand Y_i be the log of R&D expense of firm i . (Both revenue and R&D are measured in EUR.) In practice, there are many firms which report their operating revenues but not R&D, while there are few firms which report their R&D but not revenues. It is therefore reasonable to assume that the revenue W is always observed while the R&D expense Y is possibly missing.

Assume that the covariate is identical to the regressor: $X_i = W_i$. Intuitively, whether a firm reports its R&D expense should depend on the stringency of the accounting requirement that the firm is facing. The larger firm should meet the more stringent accounting rule in order to keep its social credibility. Hence, the magnitude of revenues is supposed to have a positive impact on the probability of reporting R&D.

All data used in this study are retrieved on Orbis maintained by Bureau van Dijk. On Orbis, there are 14836 firms whose (i) country ISO code is DE (i.e., Germany), (ii) NACE Revision 2 code is "C" (i.e., Manufacturing), and (iii) operating

Table 3

Performance of confidence intervals based on the bootstrap method I (Gumbel copula). In this table, the benchmark scenario with the Gumbel copula is considered. Y is MAR and the sample size is $N \in \{250, 500, 750\}$. The regression curve $a_0(w)$ is estimated via the calibration approach with inverse logistic $\rho(v) = v - \exp(-v)$, where $w \in \{-2.5, -1.5, \dots, 1.5, 2.5\}$. The sieve basis function is specified as $u_K(X) = (1, X, \dots, X^{K-1})^T$ with $K \in \{2, 3, 4\}$. The 95% confidence interval (CI) for $a_0(w)$ is computed in accordance with the asymptotic normality result, where the asymptotic variance is approximated by bootstrapping. The same K as in the point estimation is used for $B = 500$ bootstrap iterations. This table reports the coverage probability and the mean length of the CIs across $J = 1000$ Monte Carlo samples.

Coverage probability of 95% CI ($N = 250$)							Mean length of 95% CI ($N = 250$)						
$K \backslash w$	-2.5	-1.5	-0.5	0.5	1.5	2.5	$K \backslash w$	-2.5	-1.5	-0.5	0.5	1.5	2.5
2	0.848	0.898	0.956	0.840	0.726	0.464	2	0.705	0.519	0.345	0.364	0.680	1.035
3	0.867	0.907	0.944	0.871	0.764	0.467	3	0.752	0.553	0.367	0.390	0.716	1.033
4	0.912	0.945	0.966	0.904	0.826	0.574	4	1.054	0.766	0.482	0.553	1.058	1.432

Coverage probability of 95% CI ($N = 500$)							Mean length of 95% CI ($N = 500$)						
$K \backslash w$	-2.5	-1.5	-0.5	0.5	1.5	2.5	$K \backslash w$	-2.5	-1.5	-0.5	0.5	1.5	2.5
2	0.849	0.883	0.935	0.900	0.783	0.567	2	0.513	0.361	0.235	0.265	0.501	0.889
3	0.879	0.914	0.950	0.911	0.813	0.574	3	0.528	0.370	0.241	0.278	0.520	0.861
4	0.934	0.952	0.979	0.933	0.890	0.697	4	0.739	0.513	0.310	0.381	0.785	1.155

Coverage probability of 95% CI ($N = 750$)							Mean length of 95% CI ($N = 750$)						
$K \backslash w$	-2.5	-1.5	-0.5	0.5	1.5	2.5	$K \backslash w$	-2.5	-1.5	-0.5	0.5	1.5	2.5
2	0.883	0.907	0.946	0.919	0.818	0.615	2	0.430	0.296	0.192	0.224	0.428	0.801
3	0.893	0.915	0.943	0.914	0.833	0.633	3	0.438	0.301	0.195	0.229	0.432	0.803
4	0.939	0.956	0.964	0.936	0.917	0.739	4	0.573	0.390	0.232	0.290	0.609	0.970

revenues in 2017 are observed. Sort the 14836 firms in revenues, and keep top $N = 500$ firms for analysis. R&D expenses are observed for 125 out of the 500 firms and missing for the other 375. The missing probability of R&D is therefore 75%, strongly motivating the use of the calibration estimation.

Perform the conditional mean regression of Noh et al. [27] with the Gumbel copula. Under the conditional mean regression, the regression curve $a_0(w)$ can be estimated via (24). The grid is specified as $w \in \{20.00, 20.01, \dots, 27.00\}$, which tightly contains the minimum and maximum of the revenues.

The calibration estimation for $a_0(w)$ is performed with the inverse logistic $\rho(v) = v - \exp(-v)$ and the sieve basis function $u_K(X) = (1, X, \dots, X^{K-1})^T$. The tuning parameter K is chosen from $\{2, 3, 4, 5\}$ via the 10-fold cross validation. These options are proven to perform best in the simulation study (Section 7.1). In the present case, $K^* = 4$ is chosen as the optimal value that minimizes the cross-validation criterion.

The 95% confidence interval for $\hat{a}(w)$ is computed in accordance with the asymptotic normality result, where the standard error is approximated by bootstrapping with $B = 1000$ iterations. This method, which is called the bootstrap method I in this paper, is proven to perform best in the simulation study (Section 7.2). The tuning parameter is fixed at the optimal value $K^* = 4$ in all bootstrap samples.

For comparison, the benchmark equal-weight approach, which is equivalent to the list-wise deletion, is also performed. The bootstrap method I with $B = 1000$ iterations is used to construct the 95% confidence interval for $\hat{a}(w)$ associated with the equal-weight approach.

8.2. Empirical result

See Fig. 3 for a summary of empirical results. The blue asterisks represent the 125 firms whose revenues and R&D are both observed. The black, solid line represents the estimated regression curve $\hat{a}(w)$. The green circles represent the 375 firms with unobserved R&D expenses, and those values are imputed with the conditional expectation given W_i . By construction, all green circles lie on $\hat{a}(w)$. The 95% confidence interval based on the bootstrap method I is plotted with the red, dashed lines.

Focus on Panel (a) of Fig. 3, where the result of the calibration approach (CAL) is summarized. The estimated copula parameter is $\hat{\theta} = 1.736$ and the associated Kendall's tau is $\hat{\tau} = 0.424$. The estimated regression curve $\hat{a}_{CAL}(w)$ is positively sloped, suggesting a positive correlation between revenues and R&D. The vast majority of the 375 firms with unobserved R&D have relatively small revenues $w \in [20, 22]$. The confidence interval becomes wider as the revenue w increases, reflecting the fact that there are fewer observations on the upper tail. These results are all as expected, hence we can conclude that the calibration approach produces intuitively reasonable results.

Now focus on Panel (b), where the result of the benchmark equal-weight approach (EQW) is summarized. The estimated copula parameter is $\hat{\theta} = 1.465$ and the associated Kendall's tau is $\hat{\tau} = 0.317$. $\hat{a}_{EQW}(w)$ is positively sloped, but note that $\hat{a}_{CAL}(w) < \hat{a}_{EQW}(w)$ for any $w \in [20, 27]$. Taking $w = 22$ as an example, $\hat{a}_{CAL}(w) = 18.10$ with the associated

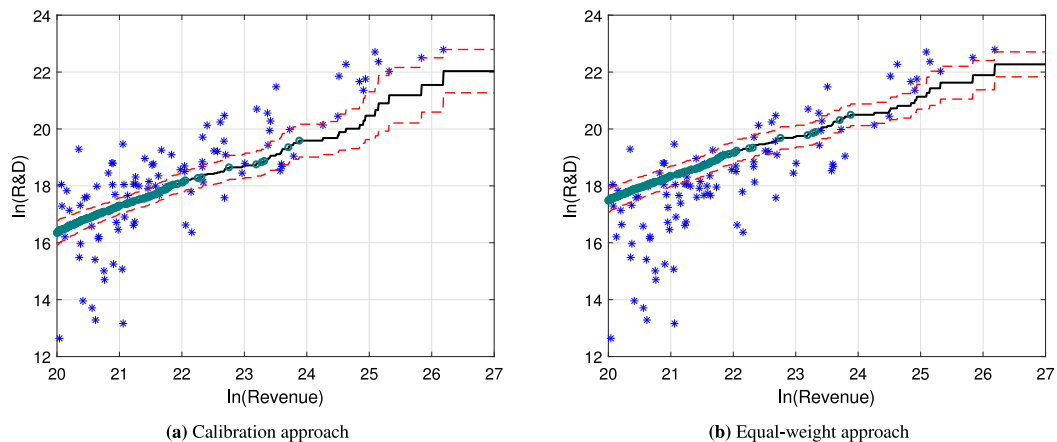


Fig. 3. The estimated regression curve $\hat{a}(w)$ and the 95% confidence interval. This figure summarizes the empirical results on $N = 500$ German manufacturing firms in 2017. The regressand Y is the log of R&D expense, and it is missing for 375 firms. The regressor W is the log of operating revenue, and it is observed for all firms. The covariate X is identical to W . The blue asterisks represent the 125 firms whose revenues and R&D are both observed. The conditional mean regression is performed via the calibration or equal-weight approach. The estimated regression curve $\hat{a}(w)$ is plotted with the black, solid line. The green circles represent the 375 firms with unobserved R&D, and they are imputed with the conditional expectation given W_i . The 95% confidence interval, plotted with the red, dashed lines, is computed in accordance with the asymptotic normality result, where the standard error is approximated by bootstrapping.

confidence interval being $[17.77, 18.43]$; $\hat{a}_{EQW}(w) = 19.14$ with the associated confidence interval being $[18.76, 19.52]$. Clearly, $\hat{a}_{CAL}(w) < \hat{a}_{EQW}(w)$ and their confidence intervals are disjoint. Similar results appear at any other level of revenue. It implies that the predicted value of R&D given the level of revenue is always smaller under the calibration approach than the equal-weight approach.

There is a logical reason why $\hat{a}_{CAL}(w) < \hat{a}_{EQW}(w)$. Recall the key fact that most firms with missing R&D concentrate on the lower tail of revenue, as shown in Fig. 3. The calibration approach takes it into account, and puts a sufficiently large emphasis on the fit of the regression curve at the lower tail. This is why $\hat{a}_{CAL}(w)$ passes through roughly the middle of the observations at the lower tail (Panel (a)).

The equal-weight approach, by contrast, ignores the 375 firms with missing R&D and assigns the uniform weight on the remaining 125 firms. Consequently, the weights assigned on the observed firms at the lower tail are smaller than what they should be. This is why $\hat{a}_{EQW}(w)$ passes through nearly the upper bound of the lower-tail observations, and passes through roughly the middle of the observations at the medium level $w \in [23, 24]$ (Panel (b)). Thus, ignoring the missing mechanism of R&D results in the positively biased regression curve.

In summary, the calibration approach delivers the plausible empirical result by assigning large enough weights on small firms, while the benchmark equal-weight approach delivers the biased result by assigning the uniform weight on all observed firms. This stark contrast highlights how the calibration approach ensures valid inference under the MAR mechanism.

9. Conclusion

The existing literature of copula-based regression models assumes that complete data are available. This assumption is violated in many real applications such as applied microeconomics, corporate finance, and survey sampling. In the present paper, the regressand and regressors are allowed to be missing at random. We formulate the generalized regression model which unifies many prominent cases such as the conditional mean and quantile regressions. The calibration estimator of the regression curve is proposed, and its consistency and asymptotic normality are proved.

In the Monte Carlo simulation, the proposed approach exhibits sharp finite sample performance in both point and variance estimation. Besides, the data-driven selection of the tuning parameter K based on the cross validation operates well. The benchmark equal-weight approach, by contrast, fails with substantial bias under MAR.

In the empirical application, we regress the R&D expenses of German manufacturing firms onto their revenues. The revenue is observed for all 500 firms, while R&D is observed for only 125 firms. The vast majority of the 375 firms with missing R&D have small revenues. The calibration approach delivers a plausible result by assigning sufficiently large weights on firms with small revenues. The equal-weight approach, by contrast, delivers a misleading result by discarding the 375 firms with missing R&D and assigning the uniform weight on the 125 firms left. This contrast highlights the practical use of the calibration approach.

CRedit authorship contribution statement

Shigeyuki Hamori: Validation, Formal analysis, Writing - review & editing, Supervision, Project administration. **Kaiji Motegi:** Software, Formal analysis, Investigation, Visualization, Writing - review & editing. **Zheng Zhang:** Conceptualization, Methodology, Investigation, Writing - original draft, Writing - review & editing.

Acknowledgments

We are grateful to the Editor-in-Chief, Dietrich von Rosen, an anonymous Associate Editor, and two anonymous referees for their insightful comments and suggestions. We also thank Marcus Chambers, Daisuke Nagakura, Teruo Nakatsuma, Tatsuyoshi Okimoto, and Naoya Sueishi, seminar participants at Kobe University, the University of Essex, and Keio University, and conference participants at the 2019 Japanese Joint Statistical Meeting for their helpful comments. The first author, Shigeyuki Hamori, is grateful for the financial support of JSPS, Japan KAKENHI Grant No. (A) 17H00983 and the Organization for Advanced and Integrated Research (OAIR), Kobe University. The second author, Kaiji Motegi, is grateful for the financial support of JSPS, Japan KAKENHI Grant No. 19K13670, Japan Center for Economic Research, and OAIR. The last author, Zheng Zhang, acknowledges 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

Note that

$$\begin{aligned} & \sup_{a \in \mathcal{A}} \left| \sum_{i=1}^N T_{0i} \hat{p}_{0K}(\mathbf{X}_i) L\{g(Y_i) - a\} c \left\{ \hat{F}_{0,K}(Y_i), \hat{F}_{1,K}(w_1), \dots, \hat{F}_{d,K}(w_d); \hat{\boldsymbol{\theta}}_K \right\} \right. \\ & \quad \left. - E \left[\frac{T_{0i}}{\pi_0(\mathbf{X}_i)} L\{g(Y_i) - a\} c \{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \right] \right| \\ & \leq \sup_{a \in \mathcal{A}} \left| \sum_{i=1}^N T_{0i} \hat{p}_{0K}(\mathbf{X}_i) L\{g(Y_i) - a\} c \left\{ \hat{F}_{0,K}(Y_i), \hat{F}_{1,K}(w_1), \dots, \hat{F}_{d,K}(w_d); \hat{\boldsymbol{\theta}}_K \right\} \right. \\ & \quad \left. - \frac{1}{N} \sum_{i=1}^N \frac{T_{0i}}{\pi_0(\mathbf{X}_i)} L\{g(Y_i) - a\} c \{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \right| \end{aligned} \quad (\text{A.1})$$

$$\begin{aligned} & + \sup_{a \in \mathcal{A}} \left| \frac{1}{N} \sum_{i=1}^N \frac{T_{0i}}{\pi_0(\mathbf{X}_i)} L\{g(Y_i) - a\} c \{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \right. \\ & \quad \left. - E \left[\frac{T_{0i}}{\pi_0(\mathbf{X}_i)} L\{g(Y_i) - a\} c \{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \right] \right|. \end{aligned} \quad (\text{A.2})$$

To show $\hat{a}(\mathbf{w}) \xrightarrow{p} a_0(\mathbf{w})$, it is sufficient to show both (A.1) and (A.2) are of $o_p(1)$. By Assumption 4 and the results that $\hat{\boldsymbol{\theta}}_K \xrightarrow{p} \boldsymbol{\theta}_0$ and $\sup_{x \in \mathcal{X}} |\hat{p}_{0K}(x) - \pi_0^{-1}(x)| = o_p(1)$, the term (A.1) is trivially of $o_p(1)$; see Theorem 1 of Hamori et al. [13]. For the term (A.2), note that for fixed $a \in \mathcal{A}$,

$$\left| \frac{1}{N} \sum_{i=1}^N \frac{T_{0i}}{\pi_0(\mathbf{X}_i)} L\{g(Y_i) - a\} c \{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} - E \left[\frac{T_{0i}}{\pi_0(\mathbf{X}_i)} L\{g(Y_i) - a\} c \{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \right] \right| = o_p(1).$$

Moreover, the dominating function is integrable:

$$\begin{aligned} & E \left[\sup_{a \in \mathcal{A}} \left| \frac{T_{0i}}{\pi_0(\mathbf{X}_i)} L\{g(Y_i) - a\} c \{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \right| \right] \\ & \leq O(1) \times E \left[\sup_{a \in \mathcal{A}} L\{g(Y_i) - a\} c \{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \right] \\ & = O(1) \times \frac{1}{c_{\mathbf{W}} \{F_1(w_1), \dots, F_d(w_d)\}} \times E \left[\sup_{a \in \mathcal{A}} L\{g(Y_i) - a\} \mid \mathbf{W}_i = \mathbf{w} \right] < \infty, \end{aligned}$$

where the first inequality holds because $\pi_0(x)$ is uniformly bounded away from zero; see Assumption 4 of the Online Supplement [14]. Hence, by the uniform law of large numbers, the term (A.2) is of $o_p(1)$.

Appendix B. Proof of Theorem 2

To prove Theorem 2, the following lemma will be of use.

Lemma A. Under the regularity conditions imposed in [13], for all $j \in \{0, 1, \dots, d\}$ and any square integrable function $\phi(Y, \mathbf{W}, \mathbf{X})$, we have

$$\begin{aligned} & \sqrt{N} \sum_{i=1}^N [T_{ji} \hat{p}_{j,K}(\mathbf{X}_i) \phi(Y_i, \mathbf{W}_i, \mathbf{X}_i) - E\{\phi(Y_i, \mathbf{W}_i, \mathbf{X}_i)\}] \\ &= \frac{1}{\sqrt{N}} \sum_{i=1}^N \left[\frac{T_{ji}}{\pi_j(\mathbf{X}_i)} \phi(Y_i, \mathbf{W}_i, \mathbf{X}_i) - E\{\phi(Y_i, \mathbf{W}_i, \mathbf{X}_i)\} \right] - \frac{1}{\sqrt{N}} \sum_{i=1}^N \left\{ \frac{T_{ji}}{\pi_j(\mathbf{X}_i)} - 1 \right\} E\{\phi(Y_i, \mathbf{W}_i, \mathbf{X}_i) | \mathbf{X}_i\} + o_p(1). \end{aligned}$$

A proof of Lemma A is omitted since it is similar to the proof of Theorem 2 of Hamori et al. [13].

By Assumption 5(iii), we have $\sum_{i=1}^N (T_{0i} \hat{p}_{0K}(\mathbf{X}_i) L'\{g(Y_i) - \hat{a}\} c\{\hat{F}_{0,K}(Y_i), \hat{F}_{1,K}(w_1), \dots, \hat{F}_{d,K}(w_d); \hat{\theta}_K\} = o_p(N^{-1/2})$. Since the loss function $L(\cdot)$ may not be twice differentiable, we cannot directly apply the Taylor's expansion to obtain the expression for $\sqrt{N}(\hat{a} - a_0)$. Note that

$$R(T_0, \mathbf{X}, Y; a) = \frac{T_0}{\pi_0(\mathbf{X})} L'\{g(Y) - a\} c\{F_0(Y), F_1(w_1), \dots, F_d(w_d); \theta_0\},$$

and $f(a) := E\{R(T_0, \mathbf{X}, Y; a)\}$ is differentiable with respect to a . Since $f(a_0) = 0$ holds by definition, we have by the mean value theorem that $0 = \sqrt{N}f(a_0) = \sqrt{N}f(\hat{a}) - f'(\tilde{a})\sqrt{N}(\hat{a} - a_0)$, where \tilde{a} lies between a_0 and \hat{a} . Using Assumption 5(iii) and the facts that $f'(a)$ is a continuous function of a and $\hat{a} \xrightarrow{p} a_0$, we have

$$\begin{aligned} & \sqrt{N}(\hat{a} - a_0) = f'(a_0)^{-1} \sqrt{N}f(\hat{a}) + o_p(1) \\ &= -f'(a_0)^{-1} \sqrt{N} \sum_{i=1}^N \left[T_{0i} \hat{p}_{0K}(\mathbf{X}_i) L'\{g(Y_i) - \hat{a}\} c\{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \theta_0\} - f(\hat{a}) \right] \\ & \quad + f'(a_0)^{-1} \sqrt{N} \sum_{i=1}^N T_{0i} \hat{p}_{0K}(\mathbf{X}_i) L'\{g(Y_i) - \hat{a}\} c\{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \theta_0\} \\ & \quad - f'(a_0)^{-1} \sqrt{N} \sum_{i=1}^N T_{0i} \hat{p}_{0K}(\mathbf{X}_i) L'\{g(Y_i) - \hat{a}\} c\{\hat{F}_{0,K}(Y_i), \hat{F}_{1,K}(w_1), \dots, \hat{F}_{d,K}(w_d); \hat{\theta}_K\} + o_p(1) \\ &= -f'(a_0)^{-1} \sqrt{N} \sum_{i=1}^N \left[T_{0i} \hat{p}_{0K}(\mathbf{X}_i) L'\{g(Y_i) - \hat{a}\} c\{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \theta_0\} - f(\hat{a}) \right] \\ & \quad - f'(a_0)^{-1} \sum_{i=1}^N T_{0i} \hat{p}_{0K}(\mathbf{X}_i) L'\{g(Y_i) - \hat{a}\} \partial_0 c\{\tilde{F}_{0,K}(Y_i), \tilde{F}_{1,K}(w_1), \dots, \tilde{F}_{d,K}(w_d); \tilde{\theta}_K\} \times \sqrt{N}\{\hat{F}_{0,K}(Y_i) - F_0(Y_i)\} \\ & \quad - f'(a_0)^{-1} \sum_{i=1}^N \sum_{j=1}^d T_{0i} \hat{p}_{0K}(\mathbf{X}_i) L'\{g(Y_i) - \hat{a}\} \partial_j c\{\tilde{F}_{0,K}(Y_i), \tilde{F}_{1,K}(w_1), \dots, \tilde{F}_{d,K}(w_d); \tilde{\theta}_K\} \times \sqrt{N}\{\hat{F}_{j,K}(w_j) - F_j(w_j)\} \\ & \quad - f'(a_0)^{-1} \sqrt{N} \sum_{i=1}^N T_{0i} \hat{p}_{0K}(\mathbf{X}_i) L'\{g(Y_i) - \hat{a}\} \partial_\theta c\{\tilde{F}_{0,K}(Y_i), \tilde{F}_{1,K}(w_1), \dots, \tilde{F}_{d,K}(w_d); \tilde{\theta}_K\} \times \sqrt{N}(\hat{\theta}_K - \theta_0) + o_p(1), \end{aligned} \quad (\text{B.1})$$

where $\{\tilde{F}_{0,K}(Y_i), \tilde{F}_{1,K}(w_1), \dots, \tilde{F}_{d,K}(w_d)\}$ lie on the line joining from $\{\hat{F}_{0,K}(Y_i), \hat{F}_{1,K}(w_1), \dots, \hat{F}_{d,K}(w_d)\}$ to $\{F_0(Y_i), F_1(w_1), \dots, F_d(w_d)\}$, and $\tilde{\theta}_K$ lies on the line joining $\hat{\theta}_K$ and θ_0 .

By Lemma A, we can deduce the following identities:

$$\begin{aligned} & \sqrt{N} \sum_{i=1}^N \left[T_{0i} \hat{p}_{0K}(\mathbf{X}_i) L'\{g(Y_i) - \hat{a}\} c\{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \theta_0\} - f(\hat{a}) \right] \\ &= \frac{1}{\sqrt{N}} \sum_{i=1}^N \left[\frac{T_{0i}}{\pi_0(\mathbf{X}_i)} L'\{g(Y_i) - \hat{a}\} c\{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \theta_0\} - f(\hat{a}) \right. \\ & \quad \left. - \left\{ \frac{T_{0i}}{\pi_0(\mathbf{X}_i)} - 1 \right\} E\{L'\{g(Y_i) - \hat{a}\} c\{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \theta_0\} | \mathbf{X}_i\} \right] + o_p(1), \end{aligned}$$

$$= \frac{1}{\sqrt{N}} \sum_{i=1}^N \left[\frac{T_{0i}}{\pi_0(\mathbf{X}_i)} L' \{g(Y_i) - a_0\} c \{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \right. \\ \left. - \left\{ \frac{T_{0i}}{\pi_0(\mathbf{X}_i)} - 1 \right\} E \left[L' \{g(Y_i) - a_0\} c \{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} | \mathbf{X}_i \right] \right] + o_p(1), \quad (\text{B.2})$$

where the second equality holds from the facts that $\hat{a} \xrightarrow{p} a_0$, $E\{R(T_{0i}, \mathbf{X}_i, Y_i; a)\} = 0$, and that [Assumption 6](#) and [Theorem 4](#) of [Andrews \[1\]](#) imply the empirical processes

$$\left[\frac{1}{\sqrt{N}} \sum_{i=1}^N [R(T_{0i}, \mathbf{X}_i, Y_i; a) - E\{R(T_{0i}, \mathbf{X}_i, Y_i; a)\}] : a \in \mathcal{A} \right]$$

and

$$\left[\frac{1}{\sqrt{N}} \sum_{i=1}^N \left\{ \frac{T_{0i}}{\pi_0(\mathbf{X}_i)} - 1 \right\} E \left[L' \{g(Y_i) - a\} c \{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} | \mathbf{X}_i \right] : a \in \mathcal{A} \right]$$

are stochastically equicontinuous.

Note that

$$\begin{aligned} & \sqrt{N} \sum_{i=1}^N T_{0i} \hat{p}_{0K}(\mathbf{X}_i) L' \{g(Y_i) - \hat{a}\} \partial_0 c \left\{ \tilde{F}_{0,K}(Y_i), \tilde{F}_1(w_1), \dots, \tilde{F}_d(w_d); \tilde{\boldsymbol{\theta}}_K \right\} \{\hat{F}_{0,K}(Y_i) - F_0(Y_i)\} \\ &= o_p(1) + \int L' \{g(y) - a_0\} \partial_0 c \{F_0(y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \sqrt{N} \{\hat{F}_{0,K}(y) - F_0(y)\} dF_0(y) \\ &= o_p(1) + \frac{1}{\sqrt{N}} \sum_{i=1}^N \int L' \{g(y) - a_0\} \partial_0 c \{F_0(y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \\ & \quad \times \left[\frac{T_{0i}}{\pi_0(\mathbf{X}_i)} \mathbf{1}(Y_i \leq y) - \left\{ \frac{T_{0i}}{\pi_0(\mathbf{X}_i)} - 1 \right\} F_{Y|\mathbf{X}}(y|\mathbf{x}) - F_0(y) \right] dF_0(y), \end{aligned} \quad (\text{B.3})$$

and

$$\begin{aligned} & \sqrt{N} \sum_{i=1}^N T_{0i} \hat{p}_{0K}(\mathbf{X}_i) L' \{g(Y_i) - \hat{a}\} \sum_{j=1}^d \partial_j c \left\{ \tilde{F}_{0,K}(Y_i), \tilde{F}_{1,K}(w_1), \dots, \tilde{F}_{d,K}(w_d); \tilde{\boldsymbol{\theta}}_K \right\} \{\hat{F}_j(w_j) - F_j(w_j)\} \\ &= o_p(1) + \frac{1}{\sqrt{N}} \sum_{i=1}^N \sum_{j=1}^d E \left[L' \{g(Y) - a_0\} \partial_j c \{F_0(Y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \right] \\ & \quad \times \left[\frac{T_{ji}}{\pi_j(\mathbf{X}_i)} \mathbf{1}(W_{ji} \leq w_j) - \left\{ \frac{T_{ji}}{\pi_j(\mathbf{X}_i)} - 1 \right\} F_{W_j|\mathbf{X}}(w_j|\mathbf{x}) - F_j(w_j) \right]. \end{aligned} \quad (\text{B.4})$$

and

$$\begin{aligned} & \sqrt{N} \sum_{i=1}^N T_{0i} \hat{p}_{0K}(\mathbf{X}_i) L' \{g(Y_i) - \hat{a}\} \partial_{\boldsymbol{\theta}} c \left\{ \tilde{F}_{0,K}(Y_i), \tilde{F}_{1,K}(w_1), \dots, \tilde{F}_{d,K}(w_d); \tilde{\boldsymbol{\theta}}_K \right\} (\hat{\boldsymbol{\theta}}_K - \boldsymbol{\theta}_0) \\ &= o_p(1) + E \left[L' \{g(Y) - a_0\} \partial_{\boldsymbol{\theta}} c \{F_0(Y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \right] \times \frac{1}{\sqrt{N}} \sum_{i=1}^N \boldsymbol{\xi}_i. \end{aligned} \quad (\text{B.5})$$

Combine [\(B.1\)–\(B.5\)](#) to obtain the desired result $\sqrt{N} \{\hat{a}(\mathbf{w}) - a_0(\mathbf{w})\} = N^{-1/2} \sum_{i=1}^N S(T_i, \mathbf{X}_i, Y_i; \mathbf{w}) + o_p(1)$.

Appendix C. Proof of [Corollary 1](#)

Suppose that $\Pr(T_{ji} = 1) = 1$ for all $j \in \{0, 1, \dots, d\}$ and $i \in \{1, \dots, N\}$ (i.e., complete data), $L(v) = v^2$, and $a_0(\mathbf{w}) = E(Y|\mathbf{W} = \mathbf{w})$. Then, the key quantities in [Theorem 2](#) are simplified as follows:

$$b(\mathbf{w}) = 2E[c \{F_0(Y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\}] = 2c_{\mathbf{W}} \{F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\},$$

$$A_{1i}(\mathbf{w}) = 2\{Y_i - a_0(\mathbf{w})\} c \{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\},$$

$$\begin{aligned} A_{2i}(\mathbf{w}) &= 2 \int \{y - a_0(\mathbf{w})\} \{\mathbf{1}(Y_i \leq y) - F_0(y)\} dc \{F_0(y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \\ &= 2 \int_{Y_i}^{\infty} \{y - a_0(\mathbf{w})\} dc \{F_0(y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \end{aligned}$$

$$\begin{aligned}
& -2 \int_{-\infty}^{\infty} \{y - a_0(\mathbf{w})\} F_0(y) dc \{F_0(y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \\
& = -2 \{Y_i - a_0(\mathbf{w})\} c \{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \\
& \quad - 2 \int_{Y_i}^{\infty} c \{F_0(y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} dy \\
& \quad + 2 \int_{-\infty}^{\infty} c \{F_0(y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} F_0(y) dy \\
& \quad + 2 \int_{-\infty}^{\infty} c \{F_0(y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \{y - a_0(\mathbf{w})\} dF_0(y) \\
& = 2 \int_{-\infty}^{\infty} c \{F_0(y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \{y - a_0(\mathbf{w})\} dF_0(y) \\
& \quad - 2 \{Y_i - a_0(\mathbf{w})\} c \{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \\
& \quad - 2 \int \{\mathbf{1}(Y_i \leq y) - F_0(y)\} c \{F_0(y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} dy, \\
A_{3i}(\mathbf{w}) & = 2 \sum_{j=1}^d E \left[\{Y - a_0(\mathbf{w})\} \partial_j c \{F_0(Y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \right] \{\mathbf{1}(W_{ji} \leq w_j) - F_j(w_j)\} \\
& \quad + 2 \xi_i^\top E \left[\{Y - a_0(\mathbf{w})\} \partial_\theta c \{F_0(Y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \right].
\end{aligned}$$

Hence, in light of

$$\frac{1}{c_{\mathbf{w}}\{F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\}} \left[\int_{-\infty}^{\infty} c \{F_0(y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \{y - a_0(\mathbf{w})\} dF_0(y) \right] = 0,$$

then the influence function $S(T_i, \mathbf{X}_i, Y_i; \mathbf{w})$ in Theorem 2 can be simplified as follows:

$$\begin{aligned}
S_{\text{mean}} & = \frac{A_{1i}(\mathbf{w}) + A_{2i}(\mathbf{w}) + A_{3i}(\mathbf{w})}{b(\mathbf{w})} \\
& = \frac{1}{c_{\mathbf{w}}\{F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\}} \times \left[- \int \{\mathbf{1}(Y_i \leq y) - F_0(y)\} c \{F_0(y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} dy \right. \\
& \quad + \sum_{j=1}^d E \left[\{Y - a_0(\mathbf{w})\} \partial_j c \{F_0(Y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \right] \{\mathbf{1}(W_{ji} \leq w_j) - F_j(w_j)\} \\
& \quad \left. + \xi_i^\top E \left[\{Y - a_0(\mathbf{w})\} \partial_\theta c \{F_0(Y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \right] \right].
\end{aligned}$$

Hence, our influence function reduces to that of Noh et al. [27].

Appendix D. Proof of Corollary 2

Suppose that $\Pr(T_{ji} = 1) = 1$ for all $j \in \{0, 1, \dots, d\}$ and $i \in \{1, \dots, N\}$ (i.e., complete data) and $L(v) = v \{\tau - \mathbf{1}(v \leq 0)\}$, which implies that $L'(v) = \tau - \mathbf{1}(v \leq 0)$. Then, the key quantities in Theorem 2 are simplified as follows:

$$\begin{aligned}
b(\mathbf{w}) & = \partial_a \int_{-\infty}^a c \{F_0(y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} f_0(y) dy \Big|_{a=a_0(\mathbf{w})} = c \{F_0\{a_0(\mathbf{w})\}, F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} f_0\{a_0(\mathbf{w})\}, \\
A_{1i}(\mathbf{w}) & = [\tau - \mathbf{1}\{Y_i \leq a_0(\mathbf{w})\}] c \{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\}, \\
A_{2i}(\mathbf{w}) & = \int [\tau - \mathbf{1}(y \leq a_0)] \{\mathbf{1}(Y_i \leq y) - F_0(y)\} dc \{F_0(y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \\
& = \tau \left[\int_{Y_i}^{\infty} dc \{F_0(y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} - \int_{-\infty}^{\infty} F_0(y) dc \{F_0(y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \right] \\
& \quad - \int_{Y_i}^{a_0} dc \{F_0(y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} + \int_{-\infty}^{a_0} F_0(y) dc \{F_0(y), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \\
& = -\tau c \{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} + \tau c_{\mathbf{w}}\{F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \\
& \quad - \mathbf{1}(Y_i \leq a_0) c \{F_0\{a_0(\mathbf{w})\}, F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \\
& \quad + \mathbf{1}(Y_i \leq a_0) c \{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\} \\
& \quad + F_0\{a_0(\mathbf{w})\} c \{F_0\{a_0(\mathbf{w})\}, F_1(w_1), \dots, F_d(w_d); \boldsymbol{\theta}_0\}
\end{aligned}$$

$$\begin{aligned}
& - \int_{-\infty}^{a_0} c \{F_0(y), F_1(w_1), \dots, F_d(w_d); \theta_0\} dF_0(y) \\
& = -\tau c \{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \theta_0\} \\
& \quad - \mathbf{1}\{Y_i \leq a_0(\mathbf{w})\} c \{F_0\{a_0(\mathbf{w})\}, F_1(w_1), \dots, F_d(w_d); \theta_0\} \\
& \quad + \mathbf{1}\{Y_i \leq a_0(\mathbf{w})\} c \{F_0(Y_i), F_1(w_1), \dots, F_d(w_d); \theta_0\} \\
& \quad + F_0\{a_0(\mathbf{w})\} c \{F_0\{a_0(\mathbf{w})\}, F_1(w_1), \dots, F_d(w_d); \theta_0\},
\end{aligned}$$

where the last equality holds because

$$\begin{aligned}
& \tau c_{\mathbf{W}}\{F_1(w_1), \dots, F_d(w_d); \theta_0\} - \int_{-\infty}^{a_0} c \{F_0(y), F_1(w_1), \dots, F_d(w_d); \theta_0\} dF_0(y) \\
& = E[L'(Y - a_0) c \{F_0(y), F_1(w_1), \dots, F_d(w_d); \theta_0\}] = o_p(N^{-1/2}).
\end{aligned}$$

Finally,

$$\begin{aligned}
A_{3i}(\mathbf{w}) &= \sum_{j=1}^d E \left[\{\tau - \mathbf{1}(Y \leq a_0)\} \partial_j c \{F_0(Y), F_1(w_1), \dots, F_d(w_d); \theta_0\} \right] \{\mathbf{1}(W_{ji} \leq w_j) - F_j(w_j)\} \\
& \quad + \xi_i^\top E \left[\{\tau - \mathbf{1}(Y \leq a_0)\} \partial_\theta c \{F_0(Y), F_1(w_1), \dots, F_d(w_d); \theta_0\} \right].
\end{aligned}$$

Hence, the influence function $S(T_i, \mathbf{X}_i, Y_i; \mathbf{w})$ in Theorem 2 is simplified as follows:

$$\begin{aligned}
S_{\text{quantile}} &= \frac{A_{1i}(\mathbf{w}) + A_{2i}(\mathbf{w}) + A_{3i}(\mathbf{w})}{b(\mathbf{w})} = \frac{1}{c \{F_0\{a_0(\mathbf{w})\}, F_1(w_1), \dots, F_d(w_d); \theta_0\} f_0\{a_0(\mathbf{w})\}} \\
& \times \left[-[\mathbf{1}\{Y_i \leq a_0(\mathbf{w})\} - F_0\{a_0(\mathbf{w})\}] c \{F_0\{a_0(\mathbf{w})\}, F_1(w_1), \dots, F_d(w_d); \theta_0\} \right. \\
& \quad + \sum_{j=1}^d E \left[\{\tau - \mathbf{1}(Y \leq a_0)\} \partial_j c \{F_0(Y), F_1(w_1), \dots, F_d(w_d); \theta_0\} \right] \{\mathbf{1}(W_{ji} \leq w_j) - F_j(w_j)\} \\
& \quad \left. + \xi_i^\top E \left[\{\tau - \mathbf{1}(Y \leq a_0)\} \partial_\theta c \{F_0(Y), F_1(w_1), \dots, F_d(w_d); \theta_0\} \right] \right],
\end{aligned}$$

which coincides with the influence function of Noh et al. [28].

Appendix E. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.jmva.2020.104654>.

References

- [1] D.W. Andrews, Empirical process methods in econometrics, *Handb. Econom.* 4 (1994) 2247–2294.
- [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] B. Chang, H. Joe, Prediction based on conditional distributions of vine copulas, *Comput. Statist. Data Anal.* 139 (2019) 45–63.
- [4] M. De Backer, A. El Ghouch, I. Van Keilegom, Semiparametric copula quantile regression for complete or censored data, *Electron. J. Stat.* 11 (2017) 1660–1698.
- [5] H. Dette, R. Van Hecke, S. Volgushev, Some comments on copula-based regression, *J. Amer. Statist. Assoc.* 109 (2014) 1319–1324.
- [6] W. Ding, P.X.-K. Song, EM algorithm in Gaussian copula with missing data, *Comput. Statist. Data Anal.* 101 (2016) 1–11.
- [7] B. Efron, R.J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall/CRC, 1994.
- [8] T. Emura, C.-W. Lin, W. Wang, A goodness-of-fit test for Archimedean copula models in the presence of right censoring, *Comput. Statist. Data Anal.* 54 (12) (2010) 3033–3043.
- [9] T. Emura, W. Wang, Testing quasi-independence for truncation data, *J. Multivariate Anal.* 101 (1) (2010) 223–239.
- [10] T. Emura, W. Wang, Nonparametric maximum likelihood estimation for dependent truncation data based on copulas, *J. Multivariate Anal.* 110 (2012) 171–188.
- [11] C. Genest, K. Ghoudi, L.-P. Rivest, A semiparametric estimation procedure of dependence parameters in multivariate families of distributions, *Biometrika* 82 (1995) 543–552.
- [12] F. Guo, W. Ma, L. Wang, Semiparametric estimation of copula models with nonignorable missing data, *J. Nonparametr. Stat.* 32 (2020) 109–130.
- [13] S. Hamori, K. Motegi, Z. Zhang, Calibration estimation of semiparametric copula models with data missing at random, *J. Multivariate Anal.* 173 (2019) 85–109.
- [14] S. Hamori, K. Motegi, Z. Zhang, Online supplement for “Copula-based regression models with data missing at random”, 2020, Kobe University and Renmin University of China.
- [15] K. Hirano, G.W. Imbens, G. Ridder, Efficient estimation of average treatment effects using the estimated propensity score, *Econometrica* 71 (2003) 1161–1189.
- [16] D.G. Horvitz, D.J. Thompson, A generalization of sampling without replacement from a finite universe, *J. Amer. Statist. Assoc.* 47 (1952) 663–685.
- [17] H. Joe, D. Kurowicka, *Dependence Modeling: Vine Copula Handbook*, World Scientific, 2011.

- [18] J.D.Y. Kang, J.L. Schafer, Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data, *Statist. Sci.* 22 (4) (2007) 523–539.
- [19] D. Kraus, C. Czado, D-vine copula based quantile regression, *Comput. Statist. Data Anal.* 110 (2017) 1–18.
- [20] K.-C. Li, Asymptotic optimality for C_p , C_L , cross-validation and generalized cross-validation: Discrete index set, *Ann. Statist.* 15 (3) (1987) 958–975.
- [21] Q. Li, J.S. Racine, *Nonparametric Econometrics: Theory and Practice*, Princeton University Press, 2007.
- [22] R.J.A. Little, D.B. Rubin, *Statistical Analysis with Missing Data*, second ed., Wiley-Interscience, 2002.
- [23] T. Nagler, T. Vatter, *Solving estimating equations with copulas*, 2018, arXiv:1801.10576.
- [24] R.B. Nelsen, *An Introduction to Copulas*, Springer Science & Business Media, 2007.
- [25] W.K. Newey, Convergence rates and asymptotic normality for series estimators, *J. Econometrics* 79 (1) (1997) 147–168.
- [26] W.K. Newey, J.L. Powell, Asymmetric least squares estimation and testing, *Econometrica* 55 (1987) 819–847.
- [27] H. Noh, A. El Ghouch, T. Bouezmarni, Copula-based regression estimation and inference, *J. Amer. Statist. Assoc.* 108 (502) (2013) 676–688.
- [28] H. Noh, A. El Ghouch, I. Van Keilegom, Semiparametric conditional quantile estimation through copula-based multivariate models, *J. Bus. Econom. Statist.* 33 (2) (2015) 167–178.
- [29] A. Pakes, D. Pollard, Simulation and the asymptotics of optimization estimators, *Econometrica* 57 (1989) 1027–1057.
- [30] B. Rémillard, B. Nasri, T. Bouezmarni, On copula-based conditional quantile estimators, *Statist. Probab. Lett.* 128 (2017) 14–20.
- [31] D.B. Rubin, Inference and missing data, *Biometrika* 63 (1976) 581–592.
- [32] A. Sklar, Fonctions de répartition à n dimensions et leurs marges, *Publ. Inst. Statist. Univ. Paris* 8 (1959) 229–231.
- [33] H. Tsukahara, Semiparametric estimation in copula models, *Canad. J. Statist.* 33 (2005) 357–375.