Calibration Estimation for Semiparametric Copula Models under Missing Data*

Shigeyuki Hamori†  Kaiji Motegi‡  Zheng Zhang§

August 31, 2017

Abstract

This paper investigates the estimation of semiparametric copula models under the presence of missing data. Our models comprise nonparametric marginal distributions and parametric copula functions. The two-step pseudo-likelihood method of Genest, Ghoudi, and Rivest (1995) is infeasible when there exist missing data. Inspired by Chan, Yam, and Zhang (2016), we propose a class of calibration estimators for both marginal distributions and the parameters of interest without imposing additional models on the missing mechanism. We establish consistency and asymptotic normality for our estimators of copula parameters. We also present a natural procedure for consistently estimating the asymptotic variance of our estimators.

Keywords: Calibration weights, Copula, Missing at random, Missing data, Semiparametric model.

---

*This work is supported by JSPS KAKENHI Grant Numbers 17H00983 and 16K17104. The second author is grateful for the financial supports of Kikawada Foundation, Mitsubishi UFJ Trust Scholarship Foundation, and Nomura Foundation.

†Graduate School of Economics, Kobe University. E-mail: hamori@econ.kobe-u.ac.jp
‡Corresponding author. Graduate School of Economics, Kobe University. Address: 2-1 Rokkodai-cho, Nada, Kobe, Hyogo 657-8501 Japan. E-mail: motegi@econ.kobe-u.ac.jp
§Institute of Statistics and Big Data, Renmin University of China. E-mail: zhengzhang@ruc.edu.cn
1 Introduction

Copula is a broadly accepted tool for analyzing complex interdependence among cross-section or time series variables. The concept of copula was originally put forward by Sklar (1959), who proves the existence of a unique copula function that bridges individual marginal distributions and a joint distribution. The copula approach is capable of modelling a wide range of joint distributions with little computational burden (see Trivedi and Zimmer, 2007 for a general overview of copula).

As surveyed in Patton (2009) and Fan and Patton (2014), copulas are extensively used in many fields including economics and finance. Patton (2012, 2013) provide another useful survey that focuses on time series data. For more recent innovations mainly targeting financial variables, see Oh and Patton (2013, 2016, 2017a,b) and Salvatierra and Patton (2015). Also see Aloui, Aissa, and Nguyen (2013) for a financial application and Marra and Wyszynski (2016) for a microeconomic application.


While the copula literature is rather deep and still growing, most of the existing work focus on fully observed data. Little knowledge is available on analyzing general copula models under incomplete observations. Missing data frequently arise in a wide range of empirical research. For instance, in survey analysis, respondents may refuse to report their personal information such as age, education, gender, and salary. In financial econometrics, missing data are a perverse phenomenon since different countries have different time zones,
trading hours, and holidays. There may also be unexpected market closures due to rare
emergency such as terrorism and technical problems.

Wang, Fazayeli, Chatterjee, and Banerjee (2014) studies the estimation of the precision
matrix of Gaussian copula with values missing completely at random (MCAR). Ding and Song
(2016) proposes an EM algorithm for estimating Gaussian copula parameters under miss-
ing at random (MAR); however, they do not include rigorous discussions on theoretical
properties of their estimators. Another practical issue left after those papers is exploring
non-Gaussian copulas for a better fit with data. It is desirable to develop a systematic
inference for general copula models with missing data. The goal of this work is to fill those
gaps.

It is of use to review how missing data are handled outside the copula literature. Existing
approaches can be grossly divided into complete-case analysis, maximum likelihood, and
inverse probability weighting. First, the complete-case analysis simply discards all individ-
uals with incomplete data and performs statistical inference based on individuals with fully
observed data. This method may lead to substantial loss of efficiency in estimation. More
importantly, the consistency of the complete-case analysis requires the data to be MCAR,
otherwise estimators could be severely biased.

The second well-known approach is the maximum likelihood. Under the MAR condi-
tion (cf. Rubin, 1976), the observed data likelihood can be factorized into two parts, and
one can ignore the part that depends solely on the missingness mechanism. While the
missing data mechanism does not have to be modeled, one typically needs to model the co-
variate distribution (see e.g. Little, 1992, Chen, 2004, Ibrahim, Chen, Lipsitz, and Herring,
2005). When the parametric covariate model is misspecified, the estimated parameters are
typically inconsistent. Another likelihood-based method popularly used is full informa-
tion maximum likelihood (FIML), which directly maximizes the likelihood we can calcu-
late from all observed data. Consistency of FIML under the MAR assumption has been
well realized via extensive numerical experiments (cf. Enders, 2001, Enders and Bandalos,
2001); however, its theoretical justification and benefits remain unclear. Furthermore, the computational burden of FIML could be extremely heavy for high-dimensional data (cf. Hirose, Kim, Kano, Imada, Yoshida, and Matsuo, 2016).

Third, one can weight complete observations by the inverse of missing probabilities (cf. Horvitz and Thompson, 1952, Zhao and Lipsitz, 1992). A potential problem of this approach is that the inverse probability weighting is generally inconsistent when the missing data mechanism is misspecified. Another problem is that efficiency loss can be substantial when the selection probability is small (cf. Lawless, Kalbfleisch, and Wild, 1999). Nonparametric approximation of the missing mechanism has been utilized to construct globally efficient estimators (see e.g. Hirano, Imbens, and Ridder, 2003, Imbens, Newey, and Ridder, 2005, Chen, Hong, and Tarozzi, 2008), but their finite sample performance may be poor because the estimated quantities could be extremely sensitive to the estimated missing probabilities.

The present paper takes an alternative approach in order to address missing data. We perform calibration estimation, which is recently established by Chan, Yam, and Zhang (2016) in the literature of average treatment effects. They propose a class of nonparametric calibration estimators by solely balancing the covariates among treated, controlled, and combined groups. There is a conceptual similarity between the missing data phenomenon and the treatment effect study since the potential variables in the latter framework are generally unobserved. It is therefore not surprising that the logic of Chan, Yam, and Zhang (2016) can be extended to the copula estimation with missing data.

A key insight of Chan, Yam, and Zhang (2016) is to interpret the inverse of the propensity score function as a balancing factor for the covariates in different treatment status. Calibration weights are chosen for each individual in a way that they are as consistent as possible with the moments of covariates. Their method does not rely on any additional model or intermediate step for estimating the propensity score. Those features make the calibration estimators globally efficient for average treatment effects.
In our work, we adopt the same logic as Chan, Yam, and Zhang (2016) in order to extend the classical two-step maximum likelihood estimators of Genest, Ghoudi, and Rivest (1995) and Chen and Fan (2005). We apply the calibration method in both the first step of estimating marginal distributions and the second step of maximizing likelihood with respect to copula parameters. We prove that our proposed two-step estimator satisfies consistency and asymptotic normality under the MAR condition and i.i.d. data. We also present a natural procedure for consistently estimating the asymptotic variance of our estimator.\(^1\)

The remainder of this paper is organized as follows. In Section 2 we set up our notations and basic framework. In Section 3 we propose the two-step estimator for semiparametric copula models. In Section 4 we derive the asymptotic properties of our estimator. In Section 5 we present a nonparametric consistent estimator for the asymptotic variance of our estimator. In Section 6 we run Monte Carlo simulations in order to evaluate the finite sample performance of the two-step estimator. In Section 7 we provide some concluding remarks. Proofs of propositions, theorems, and lemmas are collected in Technical Appendices.\(^2\) Omitted simulation results are presented in the supplemental material Hamori, Motegi, and Zhang (2017).

## 2 Notations and Basic Framework

Let \(d \geq 2\) be a fixed positive integer that signifies the dimension of data. Suppose that \(\{Y_i = (Y_{i1}, \ldots, Y_{id})^T\}_{i=1}^N\) are i.i.d. random vectors following the distribution \(F^0(y_1, \ldots, y_d)\). The marginal distributions of \(F^0(y_1, \ldots, y_d)\), denoted by \(\{F^0_j, j = 1, \ldots, d\}\), are assumed to be continuous. Sklar’s (1959) characterization theorem ensures the existence of a unique copula \(C^0\) such that

\[
F^0(y_1, \ldots, y_d) = C^0(F^0_1(y_1), \ldots, F^0_d(y_d)).
\]

\(^1\) The i.i.d. assumption is admittedly a restrictive one that excludes most time series applications. It is left as a future task to extend our theory to non-i.i.d. cases such as martingale difference sequences.

\(^2\) In this online version, the proofs are not disclosed.
Under mild regularity conditions, we have that

\[ f_0(y_1, \ldots, y_d) = c^0(f^0_1(y_1) \cdots f^0_d(y_d)), \quad (2.1) \]

where \( f^0, f^0_j, \) and \( c^0 \) are the density functions of \( F^0, F^0_j, \) and \( C^0, \) respectively.

Estimation of the copulas is central to econometricians and statisticians, and has been extensively studied in various settings. Genest, Ghoudi, and Rivest (1995) pioneered the study of the estimation of copula in a semiparametric model, in which the copula function is assumed to belong to a parametric family (i.e. \( C^0(F^0_1(y_1), \ldots, F^0_d(y_d)) = C(F^0_1(y_1), \ldots, F^0_d(y_d); \theta^*) \) for some \( \theta^* \in \mathbb{R}^p \)), while the marginal distributions \( \{F^0_j\}_{j=1}^d \) are left unknown.\(^3\) Furthermore, they proposed the well-known \textit{two-step maximum likelihood estimator} for the true parameter \( \theta^*:\)

\[ \hat{\theta} = \arg \max_{\theta \in \Theta} \left\{ \frac{1}{N} \sum_{i=1}^N \log c(\tilde{F}_1(Y_{1i}), \ldots, \tilde{F}_d(Y_{di}); \theta) \right\}, \quad (2.2) \]

where \( c(u_1, \ldots, u_d; \theta) \) is the density of \( C(u_1, \ldots, u_d; \theta), \Theta \subset \mathbb{R}^p \) is the compact domain of the parameter of interest, and \( \tilde{F}_j(y) = (N + 1)^{-1} \sum_{i=1}^N I(Y_{ji} \leq y) \) is a rescaled empirical distribution function.

Most copula literature including Genest, Ghoudi, and Rivest (1995) considers a complete-data framework. Practically, scientists may observe incomplete data for some individuals. Theoretical treatment of copulas gets more involved when there are missing data. It is a challenging question how to keep the two-step estimator (2.2) consistent and asymptotically normal in the presence of missing data. We need to reconsider each of the first step of marginal distribution estimation and the second step of likelihood maximization. To our best knowledge, this paper is the first work addressing that issue.

Let \( T_i = (T_{1i}, \ldots, T_{di})^T \in \{0, 1\}^d \) be a random vector indicating the missing status.

---

\(^3\) See also Oakes (1994), Shih and Louis (1995), Chen and Fan (2005, 2006).
concerning individual $i$:

$$T_{ji} = 0 \quad \text{if } Y_{ji} \text{ is missing;}$$

$$T_{ji} = 1 \quad \text{if } Y_{ji} \text{ is observed.}$$

Let $X_i = (X_{1i}, \ldots, X_{di})^T$ be a vector of covariates that are observable for any individual $i$. We assume that the missing mechanism satisfies the well-known conditional independence condition, which is also known as missing at random (MAR) in Rubin (1976):

**Assumption 2.1 (Missing at Random).** Given $X_i$, $T_i = (T_{1i}, \ldots, T_{di})^T$ is independent of $Y_i = (Y_{1i}, \ldots, Y_{di})^T$:

$$\{T_{1i}, \ldots, T_{di}\} \perp \{Y_{1i}, \ldots, Y_{di}\} | X_i .$$

The MAR condition has been extensively used in econometrics and statistics in order to identify the parameter of interest (cf. Robins and Rotnitzky, 1995, Little and Rubin, 2002, Chen, Hong, and Tarozzi, 2008, Tan, 2011, among others). Note that the MAR condition does not imply the unconditional independence between $T_i$ and $Y_i$. In many applications $T_i$ and $Y_i$ are unconditionally correlated with each other through $X_i$, and that does not violate the MAR condition.

Let the propensity score functions be:

$$\pi_j(x) := \mathbb{P}(T_{ji} = 1 | X_i = x), \ j \in \{1, \ldots, d\} ,$$

$$\eta(x) := \mathbb{P}(T_{1i} = 1, \ldots, T_{di} = 1 | X_i = x) .$$

In words, $\pi_j(x)$ is the conditional probability given $X_i = x$ that $Y_{ji}$ is observed. Similarly, $\eta(x)$ is the conditional probability given $X_i = x$ that all $d$ components of individual $i$ are observed.

We impose the following assumptions in order to ensure desired asymptotic properties of our proposed estimators.
Assumption 2.2. The data \( (Y_i, T_i, X_i)_{i=1}^N \) are i.i.d.

Assumption 2.3. The support of the covariate \( X \), which is denoted by \( \mathcal{X} \), is a Cartesian product of \( r \) compact intervals.

Assumption 2.4. \( \{\pi_j(x)\}_{j=1}^d \) and \( \eta(x) \) are bounded below, i.e. there exist some constant \( \eta_0 \) such that
\[
0 < \eta_0 \leq \pi_j(x), \quad \eta(x) \leq 1 \quad \forall x \in \mathcal{X}, \ j \in \{1, ..., d\}.
\]

Assumption 2.5. \( \{\pi_j(x)\}_{j=1}^d \) and \( \eta(x) \) are \( s \)-times continuously differentiable, where \( s > 13r \).

Assumption 2.6. For any fixed \( y \in \mathbb{R} \), the functions \( \{\mathbb{E}[I(Y_{ji} \leq y)|X_i = x]\}_{j=1}^d \) and \( \mathbb{E}[l_\theta(F_1(Y_{i1}), ..., F_1(Y_{id}); \theta^*) | X_i = x] \) are \( \bar{s} \)-times continuously differentiable in \( x \), where
\[
l_\theta(u_1, ..., u_d) := \frac{\partial}{\partial \theta} \log c(u_1, ..., u_d; \theta), \ \forall (u_1, ..., u_d) \in [0, 1]^d,
\]
and \( \bar{s} > \frac{3}{2}r \).

Assumption 2.7. \( \rho \in C^3(\mathbb{R}) \) is a strictly concave function defined on \( \mathbb{R} \), i.e., \( \rho'' < 0 \), and the range of \( \rho' \) contains \( [1, 1/\eta_0] \) which is a subset of the positive real line.

Assumption 2.8. \( K(N) = O(N^\nu) \), where \( \frac{1}{s/r-2} < \nu < \frac{1}{11} \).

Assumption 2.2 is a standard framework which is also adopted in Genest, Ghoudi, and Rivest (1995) and Chen and Fan (2005). Assumptions 2.3 and 2.4 are used for uniform approximation of \( \{\pi_j(x)^{-1}\}_{j=1}^d \) and \( \eta(x)^{-1} \). Assumption 2.4 is necessary for the nonparametric identification of parameters of interest in the population. Assumptions 2.5 and 2.6 are used to control the approximations error with a given basis function. They are standard assumptions for multivariate smoothing where the order of smoothness required increases with the dimension of \( X \). Assumption 2.7 is a mild assumption on \( \rho \) which is chosen by the statisticians and includes all the important special cases considered in the literature, such as exponential tilting, empirical likelihood, quadratic weighting, inverse logistic (see Section 6.2). Assumption 2.8 is required for controlling the stochastic order of the residual terms. It is a desirable assumption in practice because \( K \) should grow very slowly in comparison with
so that the relatively small number of moment conditions is sufficient for the proposed method to enjoy good performance.

3 Two-Step Estimation with Calibration Weights

In view of (2.1), the true copula parameter $\theta^*$ is identical to the solution of the following maximization problem:

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

By Assumption 2.1 we can identify $\theta^*$ as follows:

$$
\theta^* = \arg \max_{\theta \in \Theta} \mathbb{E} \left[ \frac{I(T_{1i} = 1, ..., T_{di} = 1)}{\eta(X_i)} \log c(F_1^0(Y_{1i}), ..., F_d^0(Y_{di}); \theta) \right] , \quad (3.1)
$$

where $\eta(X_i) := \mathbb{P}(T_{1i} = 1, ..., T_{di} = 1|X_i)$.

In view of (3.1), a natural procedure for estimating $\theta^*$ is as follows.

**Step 1** Estimate the marginal distributions $\{F_j\}$ and the inverse of the propensity score $\eta(x)^{-1}$, denoted by $\{\hat{F}_j\}$ and $q_0(x)$ respectively.

**Step 2** Compute

$$
\hat{\theta} = \arg \max_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} q_0(X_i) I(T_{1i} = 1, ..., T_{di} = 1) \log c(\hat{F}_1(Y_{1i}), ..., \hat{F}_d(Y_{di}); \theta) .
$$

Step 1 is elaborated in Section 3.1, where we present a class of calibration estimators for the marginal distributions $\{F_j\}$. Step 2 is elaborated in Section 3.2, where we propose the maximum likelihood estimation with calibration weights.
3.1 Calibration Estimators for Marginal Distributions

3.1.1 Existing Approaches

Under Assumption 2.1, for $j \in \{1, ..., d\}$, the marginal distribution $F^0_j$ can be represented by

$$F^0_j(y) = \mathbb{E}[I(Y_{ji} \leq y)] = \mathbb{E} \left[ \mathbb{E}[I(Y_{ji} \leq y) | X_i] \right]$$

$$= \mathbb{E} \left[ \frac{T_{ji}}{\pi_j(X_i)} I(Y_{ji} \leq y) \right],$$

where $\pi_j(x) = \mathbb{P}(T_{ji} = 1 | X_i = x)$. Expression (3.2) suggests that, if the propensity score $\pi_j(x)$ were known, then the inverse probability weighting (IPW) estimator for $F_j$ would operate without any issue:

$$\tilde{F}_j(y) := \frac{1}{N} \sum_{i=1}^{N} \frac{T_{ji}}{\pi_j(X_i)} I(Y_{ji} \leq y)$$

(see Horvitz and Thompson, 1952). $\pi_j(x)$ is, however, usually unknown and need to be estimated in practice. The existing methods for estimating $\pi_j(x)$ can roughly be divided into parametric and nonparametric approaches. For the parametric approach, Zhao and Lipsitz (1992) propose ML estimators based on a logistic model (see e.g. Robins, Rotnitzky, and Zhao, 1994; Bang and Robins, 2005). Parametric estimators are easy to compute, but they are haunted by misspecification problems. If the missing data mechanism is misspecified, the estimator is severely biased in general (cf. Lawless, Kalbfleisch, and Wild, 1999).

Nonparametric estimators provide valid inference in large samples without relying on parametric assumptions. Hahn (1998), Hirano, Imbens, and Ridder (2003), Imbens, Newey, and Ridder (2005), and Chen, Hong, and Tarozzi (2008) present various nonparametric estimators for propensity scores in the context of average treatment effects. A well-known shortcoming of the nonparametric approach, called the curse of dimensionality, is that it
tends to have poor performance in small sample.

### 3.1.2 Proposed Approach

We bypass a direct estimation of the unknown propensity score functions by using a class of calibration weights. They are constructed by balancing the moments of observed covariates between the non-missing group and the whole group. This approach was first put forward by Chan, Yam, and Zhang (2016) in the context of average treatment effects.

For any integrable function $u(\cdot)$, the propensity score $\pi_j$ balances the covariates between the non-missing group and the whole group:

$$
E \left[ \frac{T_{ji}}{\pi_j(X_i)} u(X_i) \right] = E[u(X_i)], \quad j \in \{1, \ldots, d\}.
$$

Let $D(v, v_0)$ be a distance measure that is continuously differentiable in $v \in \mathbb{R}$, non-negative, strictly convex in $v$, and $D(v_0, v_0) = 0$. The core idea of calibration as in Deville and Särndal (1992) is to minimize the aggregate distance between the final weights $w = (w_1, \ldots, w_N)$ and a pre-specified vector of design weights $d = (d_1, \ldots, d_N)$ subject to certain moment constraints. In survey applications, the design weights are known inverse probability weights. In our missing data setting, for some fixed $j \in \{1, \ldots, d\}$, the inverse probability weights are $d_i = \pi_j(X_i)$ if $T_{ji} = 1$ for $i = 1, \ldots, N$, which are unknown and need to be estimated.

To avoid direct estimation of $\pi_j$, we follow the spirit of Chan, Yam, and Zhang (2016) to construct calibration weights by solving the following minimization problem subject to a sample counterpart of (3.3):

Minimize $\sum_{i=1}^{N} T_{ji} D(Np_{ji}, 1)$ subject to $\sum_{i=1}^{N} T_{ji}p_{ji}u_K(X_i) = \frac{1}{N} \sum_{i=1}^{N} u_K(X_i)$, \quad (3.4)

where $u_K(X)$ is a $K(N)$-dimensional function of $X$, whose components form a set of orthonormal polynomials. We assume that $K(N) \rightarrow \infty$ as $N \rightarrow \infty$ yet with $K(N) = o(N)$. Assume further that all these $u_K$’s form a basis on $L^\infty$ as $K(N) \rightarrow \infty$. 

11
The choice of uniform design weights in (3.4) is justified by a few observations. First, if there are no missing data, then we can estimate $F^0_j(y)$ by the empirical distribution $N^{-1} \sum_{i=1}^N I(Y_{ji} \leq y)$, which assigns equal weights for each individual. Second, there is no need to estimate $\tau_j(x)$ when the uniform design weights are used. Third, by minimizing the aggregate distance from constant weights, the dispersion of the resulting weights is controlled and extreme weights are avoided. It is well known that extreme weights cause instability in the Horvitz-Thompson estimators under model misspecification.

A unique challenge posed by the use of uniform design weights is how to choose the number of matching conditions $K(N)$. Hellerstein and Imbens (1999) shows that, when the number of matching conditions is fixed, an empirical likelihood calibration estimator with misspecified design weights is inconsistent in general. To circumvent this theoretical difficulty, we use $u_K$ with $K(N) \to \infty$ and $K(N) = o(N)$.

To gain computational efficiency, we consider the dual problem of (3.4). The primal problem (3.4) is a convex separable programming with linear constraints. The dual problem, by contrast, is an unconstrained convex maximization problem. The latter enhances the speed and stability of numerical optimization algorithms (cf. Tseng and Bertsekas, 1987).

With a slight abuse of notation, let $D(v) = D(v, 1)$, $f(v) = D(1 - v)$, and $f'(v) = \partial f(v)/\partial v$. When $T_{ji} = 1$, the dual solution of (3.4) is given by

$$
\hat{p}_{jK}(X_i) := \frac{1}{N} \rho'(\hat{\lambda}^T_j u_K(X_i)) ,
$$

where $\rho'$ is the first derivative of a strictly concave function

$$
\rho(v) = f((f')^{-1}(v)) + v - v(f')^{-1}(v)
$$

and $\hat{\lambda}_{jK} \in \mathbb{R}^K$ maximizes the following concave objective function

$$
\hat{G}_{jK}(\lambda) := \frac{1}{N} \sum_{i=1}^N [T_{ji} \rho(\lambda^T u_K(X_i)) - \lambda^T u_K(X_i)] .
$$
In view of the first-order conditions with respect to the objective function (3.6), it is straightforward to verify that the linear constraints in primal problem (3.4) are satisfied. From Lemmas A.1 and A.2 in Technical Appendix A, we can find that our calibration weights $N\hat{p}_{jK}$ defined in (3.5) uniformly approximate the inverse of the propensity score function $\pi_j^{-1}$ up to a certain rate:

$$\sup_{x \in \mathcal{X}} |N\hat{p}_{jK}(x) - \pi_j(x)^{-1}| = O_p \left( K^{-\frac{2}{3}} + \sqrt{\frac{K^3}{N}} \right).$$

Finally, we define the empirical calibration estimator of $F^0_j$ by

$$\hat{F}_j(y) := \sum_{i=1}^N T_{ji}\hat{p}_{jK}(X_i)I(Y_{ji} \leq y). \tag{3.7}$$

### 3.2 Calibration Estimators for Copula Parameter

Note that $\eta(x)$ balances the covariates between the non-missing group and the whole group:

$$\mathbb{E} \left[ \frac{I(T_{1i} = 1, \ldots, T_{di} = 1)}{\eta(X_i)} u(X_i) \right] = \mathbb{E}[u(X_i)] \tag{3.8}.$$

Applying a similar logic in (3.5) to the moment condition (3.8), we can define calibration weights for the likelihood maximization step by

$$\hat{q}_K(X_i) = \frac{1}{N} \rho' \left( \hat{\beta}^T u_K(X_i) \right) \text{ for } i \text{ such that } T_{1i} = \cdots = T_{di} = 1, \tag{3.9}$$

where $\hat{\beta}_K$ maximizes the following concave objective function:

$$\hat{H}_K(\beta) = \frac{1}{N} \sum_{i=1}^N I(T_{1i} = 1, \ldots, T_{di} = 1) \rho' \left( \beta^T u_K(X_i) \right) - \frac{1}{N} \sum_{i=1}^N \beta^T u_K(X_i). \tag{3.10}$$

Lemmas A.1 and A.2 in Technical Appendix A imply that calibration weights $N\hat{q}_K$ defined in (3.9) uniformly approximate the inverse of the propensity score function $\eta^{-1}$ up to a
certain rate:
\[
\sup_{x \in \mathcal{X}} |N \hat{q}_K(x) - \eta(x)^{-1}| = O_p \left( K^{-\frac{3}{2}} + \sqrt{\frac{K^3}{N}} \right).
\]
In light of (3.1), the calibrated maximum likelihood estimator of \( \theta^* \) is defined by
\[
\hat{\theta} = \arg \max_{\theta \in \Theta} \left\{ \sum_{i=1}^{N} I(T_{1i} = 1, \ldots, T_{di} = 1) \hat{q}_K(X_i) \log c(\hat{F}_1(Y_{1i}), \ldots, \hat{F}_d(Y_{di}); \theta) \right\}.
\]

(3.11)

4 Asymptotic Properties of Calibration Estimators

We first demonstrate that the calibration estimator \( \{ \hat{F}_j \} \) is \( \sqrt{N} \)-consistent for the marginal distribution \( \{ F_j^0 \} \). More importantly, it gives an equivalent asymptotic representation of \( \sqrt{N}\{ \hat{F}_j - F_j^0 \} \) in which the summands are \( i.i.d. \).

Proposition 4.1. Under Assumptions 2.1-2.8, we have for each \( j \in \{ 1, \ldots, d \} \) that
\[
\sqrt{N}\{ \hat{F}_j(y) - F_j^0(y) \} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi_j(Y_i, X_i, T_{ji}; y) + o_p(1), \ \forall y \in \mathbb{R},
\]
where \( \hat{F}_j \) is defined in (3.7) and
\[
\psi_j(Y_i, X_i, T_{ji}; y) := \frac{T_{ji}}{\pi_j(X_i)} I(Y_{ji} \leq y) - \frac{T_{ji}}{\pi_j(X_i)} \mathbb{E}[I(Y_{ji} \leq y) | X_i] + \mathbb{E}[I(Y_{ji} \leq y) | X_i] - F_j^0(y).
\]

Proof. See Technical Appendix B.

We next prove the consistency and asymptotic normality of \( \hat{\theta} \) appearing in (3.11). See Section 4.1 for consistency and Section 4.2 for asymptotic normality.
4.1 Consistency

Let $\Theta \subset \mathbb{R}^p$ be the parameter space. For $\theta, \theta^* \in \Theta$, let $\|\theta - \theta^*\|$ be the usual Euclidean metric. In addition, for $j \in \{1, \ldots, d\}$, define the following notations:

$$U_i := (U_{1i}, \ldots, U_{di})^T, \quad U_{ji} := F_j^0(Y_{ji}),$$

$$l(v_1, \ldots, v_d; \theta) := \log c(v_1, \ldots, v_d; \theta), \quad l_\theta(v_1, \ldots, v_d; \theta) := \frac{\partial}{\partial \theta} l(v_1, \ldots, v_d; \theta),$$

$$l_j(v_1, \ldots, v_d; \theta) := \frac{\partial}{\partial v_j} l(v_1, \ldots, v_d; \theta), \quad l_{\theta\theta}(v_1, \ldots, v_d; \theta) := \frac{\partial^2}{\partial \theta \partial \theta} l(v_1, \ldots, v_d; \theta),$$

$$l_{\theta j}(v_1, \ldots, v_d; \theta) := \frac{\partial^2}{\partial \theta \partial v_j} l(v_1, \ldots, v_d; \theta).$$

We impose the following standard assumptions in order to ensure the consistency of $\hat{\theta}$. They are also imposed in Chen and Fan (2005).

1. **(C1)** $\mathbb{E}[l(U_{1i}, \ldots, U_{di}; \theta)]$ has a unique maximum $\theta^*$ in $\Theta$, where $\Theta$ is a compact subset of $\mathbb{R}^p$.

2. **(C2)** $\{Y_i = (Y_{1i}, \ldots, Y_{di})^T\}_{i=1}^N$ are i.i.d. samples from unknown distribution $F_0(y_1, \ldots, y_d)$ with continuous marginal distributions $F_1^0, \ldots, F_d^0$.

3. **(C3)** The true (unknown) copula function $C^0(u_1, \ldots, u_d)$ has continuous partial derivatives.

4. **(C4)** (i) For any $u \in (0, 1)^d$, $l(u; \theta)$ is a continuous function of $\theta$. (ii) $\mathbb{E}[\sup_{\theta \in \Theta} |l(U_{1i}, \ldots, U_{di}; \theta)|] < \infty$.

**Theorem 4.2.** Under Assumptions 2.1-2.8 and (C1)-(C4), we have that

- $\|\hat{\theta} - \theta^*\| = o_p(1),$
- $\sum_{i=1}^N I(T_{1i} = 1, \ldots, T_{di} = 1)\hat{q}_K(X_i)l(\hat{F}_1(Y_{1i}), \ldots, \hat{F}_d(Y_{di}); \hat{\theta}) = \mathbb{E}[l(F_1^0(Y_{1i}), \ldots, F_d^0(Y_{di}); \theta^*)] + o_p(1),$

where $\hat{q}_K$ and $\hat{\theta}$ are defined in (3.9) and (3.11), respectively.
Proof. See Technical Appendix C.

4.2 Asymptotic Normality

For \( j \in \{1, ..., d\} \), we define the following notations:

\[
\varphi(T_i, X_i, U_i; \theta^*) := \frac{I(T_{i1} = 1, ..., T_{di} = 1)}{\eta(X_i)} \left\{ I(\eta(U_{1i}, ..., U_{di}; \theta^*)) - \mathbb{E}[I(\eta(U_{1i}, ..., U_{di}; \theta^*)) | X_i] \right\} \\
+ \mathbb{E}[I(\eta(U_{1i}, ..., U_{di}; \theta^*)) | X_i] - \mathbb{E}[I(\eta(U_{1i}, ..., U_{di}; \theta^*))] ,
\]

\[
\phi_j(T_{ji}, X_i, U_{ji}; v) := \frac{T_{ji}}{\pi_j(X_i)} I(U_{ji} \leq v) - \frac{T_{ji}}{\pi_j(X_i)} \mathbb{E}[I(U_{ji} \leq v) | X_i] \\
+ \mathbb{E}[I(U_{ji} \leq v) | X_i] , \; v \in [0, 1] ,
\]

\[
W_j(T_{ji}, X_i, U_{ji}; \theta^*) := \mathbb{E} [I(\eta_j(U_{1s}, ..., U_{ds}; \theta^*)) \{ \phi_j(T_{ji}, X_i, U_{ji}; U_{js}) - U_{js} \} | U_{ji}, X_i, T_{ji}] \; (s \neq i) .
\]

The following standard conditions will be used to establish the \( \sqrt{N} \)-asymptotic normality of \( \hat{\theta} \). They are analogous to the assumptions imposed in Chen and Fan (2005).

(A1) (i) (C1) is valid with \( \theta^* \in \text{int}(\Theta) \); (ii) The following matrices are finite and positive definite:

\[
B := -\mathbb{E} [I(\eta(U_{1i}, ..., U_{di}; \theta^*))] .
\]

\[
\Sigma := \text{Var} \left( \varphi(T_i, X_i, U_i; \theta^*) + \sum_{j=1}^d W_j(T_{ji}, X_i, U_{ji}; \theta^*) \right) .
\]

(A2) For \( j \in \{1, ..., d\} \), \( I(\eta_j(u_1, ..., u_d; \theta^*)) \) is well defined and continuous in \( (u_1, ..., u_d) \in (0, 1)^d \).
(A3) (i) \( ||l_{\theta}(u_1, ..., u_d; \theta^*)|| \leq \text{constant} \times \prod_{j=1}^{d} \{v_j(1 - v_j)\}^{-a_j} \) for some \( a_j \geq 0 \) such that

\[
E \left[ \prod_{j=1}^{d} \{U_{ji}(1 - U_{ji})\}^{-2a_j} \right] < \infty ;
\]

(ii) \( ||l_{\theta_k}(u_1, ..., u_d; \theta^*)|| \leq \text{constant} \times \{v_k(1 - v_k)\}^{-b_k} \prod_{j=1, j \neq k}^{d} \{v_j(1 - v_j)\}^{-a_j} \) for some \( b_k > a_k \) such that

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

for some \( \xi_k \in (0, 1/2) \).

(A4) (i) For every \((u_1, ..., u_d) \in (0, 1)^d\), the function \( l_{\theta}(u_1, ..., u_d; \theta) \) is continuous with respect to \( \theta \) in a neighborhood of \( \theta^* \); (ii) \( E \left[ \sup_{\theta \in \Theta: ||\theta - \theta^*|| = o(1)} ||l_{\theta}(U_{1i}, ..., U_{di}; \theta)|| \right] < \infty \).

Condition (A3) allows the score function and its partial derivatives with respect to the first \( d \) arguments to blow up at the boundaries, which occurs for many popular copula functions such as Gaussian, Clayton, and \( t \)-copulas.

**Theorem 4.3.** Under Assumptions 2.1-2.8, (C1)-(C4), and (A1)-(A4), we have that

\[
\sqrt{N}(\hat{\theta} - \theta^*) \overset{d}{\to} N(0, V_{\theta^*}) ,
\]

where \( V_{\theta^*} = B^{-1}\Sigma B^{-1} \) and \( \hat{\theta}, B, \) and \( \Sigma \) are defined in (3.11), (4.4), and (4.5) respectively.

**Proof.** See Technical Appendix D.

Asymptotic variance \( V_{\theta^*} \) differs from that in Chen and Fan (2005) due to the presence of missing data. The extra term \( \sum_{j=1}^{d} W_j(T_{ji}, X_i, U_{ji}; \theta^*) \) appears in the expression of \( \Sigma \) since we need to estimate the marginal distributions \( \{F_j^0\}_{j=1}^{d} \) under missing data. Indeed, the term \( \phi_j(T_{ji}, X_i, U_{ji}; U_{js}) - U_{js} \) in \( W_j(T_{ji}, X_i, U_{ji}; \theta^*) \) is the influence function of \( \sqrt{N}(\hat{F}_j(U_{js}) - F_j^0(U_{js})) \) as shown in Proposition 4.1.
5 Nonparametric Variance Estimator

As shown in Theorem 4.3, the asymptotic variance of \( \hat{\theta} \) is given by \( V_{\theta^*} = B^{-1}\Sigma B^{-1} \). In this section we describe how to consistently estimate each of \( B \) and \( \Sigma \) in order to compute a consistent estimator for \( V_{\theta^*} \).

5.1 Estimation of \( B \)

Using the missing at random condition (Assumption 2.1), \( B \) can be rewritten as

\[
B = -\mathbb{E} \left[ I(T_{i1} = 1, ..., T_{di} = 1) l_{\theta}(U_{1i}, ..., U_{di}; \theta^*) / \eta(X_i) \right].
\]

By Lemmas A.1 and A.2 in Technical Appendix A, we know that the calibration weight \( N\hat{q}_K(x) \) is consistent for \( \eta^{-1}(x) \). Hence we define the estimator of \( B \) as

\[
\hat{B} = -\sum_{i=1}^{N} I(T_{i1} = 1, ..., T_{di} = 1) \hat{q}_K(X_i) l_{\theta}(\hat{U}_{1i}, ..., \hat{U}_{di}; \hat{\theta}) \tag{5.1}
\]

where

\[
\hat{U}_{ji} := \hat{F}_j(Y_{ji}) = \sum_{s=1}^{N} T_{js}\hat{p}_jK(X_s)I(Y_{js} \leq Y_{ji})
\]

**Theorem 5.1.** Under Assumptions 2.1-2.8, (C1)-(C4), and (A1)-(A4), \( \| \hat{B} - B \| \overset{p}{\to} 0 \).

**Proof.** See Technical Appendix E.

5.2 Estimation of \( \Sigma \)

Note that

\[
\mathbb{E} \left[ \varphi(T_i, X_i, U_i; \theta^*) + \sum_{j=1}^{d} W_j(T_{ji}, X_i, U_{ji}; \theta^*) \right] = 0.
\]
Then use Assumption 2.1 to rewrite $\Sigma$ as

$$
\Sigma = \text{Var} \left( \varphi(T, X_i, U_i; \theta^*) + \sum_{j=1}^{d} W_j(T_{ji}, X_i, U_{ji}; \theta^*) \right)
$$

$$
= \mathbb{E} \left[ \left( \varphi(T, X_i, U_i; \theta^*) + \sum_{j=1}^{d} W_j(T_{ji}, X_i, U_{ji}; \theta^*) \right)^2 \right]
$$

$$
= \mathbb{E} \left[ \frac{I(T_i = 1, ..., T_{di} = 1)}{\eta(X_i)} \left( \varphi(T, X_i, U_i; \theta^*) + \sum_{j=1}^{d} W_j(T_{ji}, X_i, U_{ji}; \theta^*) \right)^2 \right]. \quad (5.2)
$$

In view of (5.2), it is sufficient to estimate $\varphi(T, X_i, U_i; \theta^*)$ and $W_j(T_{ji}, X_i, U_{ji}; \theta^*)$ in order to estimate $\Sigma$.

We first estimate $\varphi(T, X_i, U_i; \theta^*)$, which is defined in (4.1). In view of (4.1), it is sufficient to estimate $\mathbb{E}[l_\theta(U_{1i}, ..., U_{di}; \theta^*)|X_i]$ and $\mathbb{E}[l_\theta(U_{1i}, ..., U_{di}; \theta^*)]$ in order to estimate $\varphi(T, X_i, U_i; \theta^*)$.

Assumption 2.1 implies that

$$
\mathbb{E}[l_\theta(U_{1i}, ..., U_{di}; \theta^*)|X_i] = \mathbb{E}[l_\theta(U_{1i}, ..., U_{di}; \theta^*)|X_i, T_{1i} = \cdots = T_{di} = 1], \quad (5.3)
$$

$$
\mathbb{E}[l_\theta(U_{1i}, ..., U_{di}; \theta^*)] = \mathbb{E} \left[ \frac{I(T_i = 1, ..., T_{di} = 1)}{\eta(X_i)} l_\theta(U_{1i}, ..., U_{di}; \theta^*) \right]. \quad (5.4)
$$

Based on (5.3), $\mathbb{E}[l_\theta(U_{1i}, ..., U_{di}; \theta^*)|X_i]$ can be estimated by projecting $l_\theta(\hat{U}_{1i}, ..., \hat{U}_{di}; \hat{\theta})$ into the space linearly spanned by $\{u_K(X_i)\}$:

$$
\hat{\mathbb{E}}[l_\theta(U_{1i}, ..., U_{di}; \theta^*)|X_i] := \left\{ \sum_{i=1}^{N} I(T_{1i} = 1, ..., T_{di} = 1) l_\theta(\hat{U}_{1i}, ..., \hat{U}_{di}; \hat{\theta}) u_K(X_i) \right\}^T
$$

$$
\times \left\{ \sum_{i=1}^{N} I(T_{1i} = 1, ..., T_{di} = 1) u_K(X_i) u_K^T(X_i) \right\}^{-1} u_K(X_i).
$$

Based on (5.4) and the fact that $\sup_{x \in \mathcal{X}} |N \hat{q}_K(x) - \eta(x)^{-1}| = O_p \left( K^{-\frac{3}{2}+1} + \sqrt{\frac{K^3}{N}} \right)$ (cf.
Lemmas A.1 and A.2 in Technical Appendix A), \( \mathbb{E}[l_\theta(U_{1i}, \ldots, U_{di}; \theta^*)] \) can be estimated by

\[
\widehat{\mathbb{E}}[l_\theta(U_{1i}, \ldots, U_{di}; \theta^*)] := \sum_{i=1}^{N} I(T_{1i} = 1, \ldots, T_{di} = 1) \hat{q}_K(X_i) l_\theta(\hat{U}_{1i}, \ldots, \hat{U}_{di}; \hat{\theta}) .
\]

Substitute those results into (4.1) to estimate \( \varphi(T_i, X_i, U_i; \theta^*) \):

\[
\widehat{\varphi}(T_i, X_i, U_i; \theta^*) := I(T_{1i} = 1, \ldots, T_{di} = 1) N \hat{q}_K(X_i) \left\{ l_\theta(\hat{U}_{1i}, \ldots, \hat{U}_{di}; \hat{\theta}) - \mathbb{E}[l_\theta(U_{1i}, \ldots, U_{di}; \theta^*)|X_i] \right\}
\]

\[+ \mathbb{E}[l_\theta(U_{1i}, \ldots, U_{di}; \theta^*)|X_i] - \mathbb{E}[l_\theta(U_{1i}, \ldots, U_{di}; \theta^*)] . \tag{5.5}\]

We next estimate \( W_j(T_{ji}, X_i, U_{ji}; \theta^*) \), which is defined in (4.3). Using Assumption 2.1, (4.3) can be rewritten as

\[
W_j(T_{ji}, X_i, U_{ji}; \theta^*) = \mathbb{E}\left[ \frac{I(T_{1s} = 1, \ldots, T_{ds} = 1)}{\eta(X_s)} l_{\theta_j}(U_{1s}, \ldots, U_{ds}; \theta^*) \right. \\
\times \left. \{ \phi_j(T_{ji}, X_i, U_{ji}; U_{js}) - U_{js}\} \right| U_{ji}, X_i, T_{ji} \right]. \tag{5.6}\]

It is thus sufficient to estimate \( \phi_j(T_{ji}, X_i, U_{ji}; u) \), which is defined in (4.2). Using Assumption 2.1, (4.2) can be rewritten as

\[
\phi_j(T_{ji}, X_i, U_{ji}; u) = \frac{T_{ji}}{\pi_j(X_i)} I(U_{ji} \leq v) - \frac{T_{ji}}{\pi_j(X_i)} \mathbb{E}[I(U_{ji} \leq v)|X_i, T_{ji} = 1] \\
+ \mathbb{E}[I(U_{ji} \leq v)|X_i, T_{ji} = 1] . \tag{5.7}\]

Note that \( \mathbb{E}[I(U_{ji} \leq v)|X_i, T_{ii} = 1] \) can be estimate by the following projection:

\[
\widehat{\mathbb{E}}[I(U_{ji} \leq v)|X_i, T_{ji} = 1] = \left\{ \sum_{i=1}^{N} I(T_{ji} = 1) I(U_{ji} \leq v) u_K(X_i) \right\}^T \\
\times \left\{ \sum_{i=1}^{N} I(T_{ji} = 1) u_K(X_i) u_K^T(X_i) \right\}^{-1} u_K(X_i) . \tag{5.8}\]
Based on (5.7), (5.8), and the fact that \( \sup_{x \in X} |\tilde{N}(\hat{p}_j K(x) - \pi_j(x))^{-1}| = O_p \left( K^{-\frac{p}{2} + 1} + \sqrt{\frac{K^r}{N}} \right) \) (cf. Lemmas A.1 and A.2 in Technical Appendix A), \( \phi_j(T_{ji}, X_i, U_{ji}; v) \) can be estimated by

\[
\hat{\phi}_j(T_{ji}, X_i, U_{ji}; v) := \tilde{N}(\tilde{p}_j K(X_i)) I(\tilde{U}_{ji} \leq v) - T_{ji} \{ N(\tilde{p}_j K(X_i)) \} \cdot \tilde{E}[I(\tilde{U}_{ji} \leq v) | X_i, T_{ji} = 1] + \tilde{E}[I(\tilde{U}_{ji} \leq v) | X_i, T_{ji} = 1]. 
\] (5.9)

In light of (5.6) and (5.9), the estimator of \( W_j(T_{ji}, X_i, U_{ji}; \theta^*) \) can be defined by

\[
\tilde{W}_j(T_{ji}, X_i, U_{ji}; \theta^*) := \sum_{s=1}^N I(T_{1s} = 1, ..., T_{ds} = 1) \tilde{q}_K(X_s) 
\times \tilde{q}_j(\tilde{U}_1, ..., \tilde{U}_d; \hat{\theta}) \{ \hat{\phi}_j(T_{ji}, X_i, U_{ji}; \hat{\theta}) - \tilde{U}_{js} \}. 
\] (5.10)

Finally, substitute (5.5) and (5.10) into the sample counterpart of (5.2) to obtain

\[
\tilde{\Sigma} := \sum_{i=1}^N I(T_{1i} = 1, ..., T_{di} = 1) \tilde{q}_K(X_i) \left( \tilde{\varphi}(T_i, X_i, U_i; \theta^*) + \sum_{j=1}^d \tilde{W}_j(T_{ji}, X_i, U_{ji}; \theta^*) \right)^2 . 
\] (5.11)

**Theorem 5.2.** Under Assumptions 2.1-2.8, (C1)-(C4), and (A1)-(A4), \( \| \tilde{\Sigma} - \Sigma \| \overset{p}{\to} 0. \)

**Proof.** The proof of Theorem 5.2 is omitted since it is analogous to that of Theorem 5.1.

Substitute (5.1) and (5.11) into \( V_{\theta^*} = B^{-1} \Sigma B^{-1} \) to get \( \hat{V} := \hat{B}^{-1} \hat{\Sigma} \hat{B}^{-1} \).

**Theorem 5.3.** Under Assumptions 2.1-2.8, (C1)-(C4), and (A1)-(A4), \( \| \hat{V} - V_{\theta^*} \| \overset{p}{\to} 0. \)

**Proof.** Theorem 5.3 is a straightforward implication of Theorems 5.1 and 5.2.

**6 Monte Carlo Simulations**

We run Monte Carlo simulations in order to evaluate the performance of calibration estimators. We describe our entire simulation design below, but report only some representative
results in order to save space. See the supplemental material Hamori, Motegi, and Zhang (2017) for complete results.

6.1 Data Generating Process

Consider a trivariate case (i.e. \( d = 3 \)). The number of individuals is \( N \in \{100, 500\} \) and we draw \( J = 1000 \) Monte Carlo samples. We use a scalar covariate \( X_i \) that is independently drawn from the uniform distribution in \([0, 1]\): \( X_1, \ldots, X_N \overset{	ext{i.i.d.}}{\sim} U(0, 1) \).

Define an idiosyncratic shock \( \epsilon_i \overset{	ext{i.i.d.}}{\sim} N(0, \Sigma_\epsilon) \), where

\[
\Sigma_\epsilon = \begin{bmatrix}
1 & r_\epsilon & r_\epsilon \\
 r_\epsilon & 1 & r_\epsilon \\
 r_\epsilon & r_\epsilon & 1
\end{bmatrix}
\]

with \( r_\epsilon \in \{0, 0.5\} \).

Let \( Z_i = \Phi^{-1}(X_i) \), where \( \Phi^{-1}(\cdot) \) is the inverse cumulative distribution function of \( N(0, 1) \). Assume that a data generating process (DGP) for \( Y_i = [Y_{1i}, Y_{2i}, Y_{3i}]^T \) is given by \( Y_i = a + b \times Z_i + \epsilon_i \). We fix \( a = [1, 0, 0]^T \) and consider two cases \( b = [0, 0, 0, 0]^T \) and \( b = [0.7, 1.4, 2.1]^T \).

Trivially, we have that \( Y_i | X_i \overset{\text{i.i.d.}}{\sim} N(a + b \times Z_i, \Sigma_\epsilon) \) and \( Y_i \overset{\text{i.i.d.}}{\sim} N(a, S) \), where \( S = b \times b^T + \Sigma_\epsilon \). The correlation matrix of \( Y_i \) is given by \( \text{Corr}[Y_i] = (\text{diag}[S])^{-1/2} \times S \times (\text{diag}[S])^{-1/2} = [r_{nm}] \) with \( r_{mm} = 1; r_{mn} = r_{nm}; m, n \in \{1, 2, 3\} \).

Recall that we try two values for \( r_\epsilon \) and two values for \( b \). In each of the four cases, the key parameters \( \{r_{12}, r_{13}, r_{23}\} \) are calculated as follows.

**Case #1.** \( r_\epsilon = 0.0 \) and \( b = [0.0, 0.0, 0.0]^T \). In this case \( (r_{12}, r_{13}, r_{23}) = (0.000, 0.000, 0.000) \), no correlation in \( Y_i \) due to no correlation in \( \epsilon_i \) and no impact of \( Z_i \).

**Case #2.** \( r_\epsilon = 0.0 \) and \( b = [0.7, 1.4, 2.1]^T \). In this case \( (r_{12}, r_{13}, r_{23}) = (0.467, 0.518, 0.735) \), a moderately large correlation in \( Y_i \) due to a common impact of \( Z_i \).
Case #3. \( r_\epsilon = 0.5 \) and \( b = [0.0, 0.0, 0.0]^T \). In this case \( (r_{12}, r_{13}, r_{23}) = (0.500, 0.500, 0.500) \), a moderately large correlation in \( Y_i \) due to a positive correlation in \( \epsilon_i \).

Case #4. \( r_\epsilon = 0.5 \) and \( b = [0.7, 1.4, 2.1]^T \). In this case \( (r_{12}, r_{13}, r_{23}) = (0.705, 0.694, 0.860) \), a large correlation in \( Y_i \) due to a positive correlation in \( \epsilon_i \) combined with a common impact of \( Z_i \).

The propensity score is written as \( \pi_j(x_i) = P(T_{ji} = 1|X_i = x_i) \) for \( j = 1, 2, 3 \). We assume that \( \pi_1(x_i) = \pi_2(x_i) = \pi_3(x_i) = l + (1 - l)x_i \) with \( l \in \{0.5, 0.8\} \). This implies that \( \pi_j(X_i) \sim U(l, 1) \) for \( j \in \{1, 2, 3\} \).

The conditional probability given \( X_i = x_i \) that we observe all components of individual \( i \) is written as \( \eta(x_i) = P(T_{1i} = T_{2i} = T_{3i} = 1|X_i = x_i) \). We assume that, conditional on \( X_i = x_i \), \( T_{1i}, T_{2i}, \) and \( T_{3i} \) are independent of each other: \( \eta(x_i) = \pi_1(x_i)\pi_2(x_i)\pi_3(x_i) \). The smaller value of \( l \) leads to more missing observations for each component on average.

### 6.2 Estimation

To compute marginal distributions with calibration weights, we need to specify a functional form of \( \rho(v) \). In theory \( \rho(v) \) can be any strictly concave function. Following Chan, Yam, and Zhang (2016) and their supplemental material Chan, Yam, and Zhang (2015), we consider four specifications for \( \rho(v) \):

**Exponential Tilting (ET).** \( \rho(v) = -\exp(-v) \).

**Empirical Likelihood (EL).** \( \rho(v) = \log(1 + v) \).

**Quadratic (QR).** \( \rho(v) = -0.5(1 - v)^2 \).

**Inverse Logistic (IL).** \( \rho(v) = v - \exp(-v) \).

For the \( K \)-dimensional polynomial function \( u_K \), we consider \( K = 3 \) and \( u_3(x_i) = [1, x_i, x_i^2]^T \). In theory, we need \( K(N) \to \infty \) and \( K(N) = o(N) \) in order to derive the consistency and asymptotic normality. In practice, a large value of \( K \) leads to poor performance.
in finite sample. We are restricting $K$ to take a small value based on the suggestions of Hirano, Imbens, and Ridder (2003), Imbens, Newey, and Ridder (2005), and Chan, Yam, and Zhang (2015).\footnote{In extra simulations not reported, we tried $K = 2$ (i.e. $u_2(x_i) = [1, x_i]^T$). Results are almost identical to those with $K = 3$.}

We now execute the first step to compute $\hat{F}_j(y)$, using (3.5), (3.6), and (3.7). We then execute the second step to maximize the log-likelihood function with calibration weights, using (3.9), (3.10), and (3.11).

Recall that our target quantity is $\{r_{12}, r_{13}, r_{23}\}$, the correlation coefficients among $Y_{1i}$, $Y_{2i}$, and $Y_{3i}$. We estimate $\theta = [r_{12}, r_{13}, r_{23}]^T$ based on the Gaussian copula

$$c(u_1, u_2, u_3; \theta) = \frac{1}{|R|} \exp \left[ -\frac{1}{2} \phi(u_1, u_2, u_3)^T (R^{-1} - I) \phi(u_1, u_2, u_3) \right],$$

where

$$R = \begin{bmatrix}
1 & r_{12} & r_{13} \\
r_{12} & 1 & r_{23} \\
r_{13} & r_{23} & 1
\end{bmatrix}, \quad \phi(u_1, u_2, u_3) = [\Phi^{-1}(u_1), \Phi^{-1}(u_2), \Phi^{-1}(u_3)]^T,$$

and $\Phi^{-1}(\cdot)$ is the inverse c.d.f. of $N(0, 1)$.

Recall that, in the first step, we compute the calibration weights $\hat{p}_{jK}(x_i)$ with ET, EL, QR, and IL. Besides those cases, we try naïve equal weights $\hat{p}_{jK}(x_i) = 1/N$ (called "EQ"). The latter is included as a benchmark for evaluating the performance of calibration estimators.

The same goes for the second step, where we compute $\hat{q}_K(x_i)$. Besides ET, EL, QR, and IL, we try naïve equal weights $\hat{q}_K(x_i) = 1/N$ ("EQ"). Thus, we have $5 \times 5 = 25$ ways of computing the pair $(\hat{p}_{jK}(x_i), \hat{q}_K(x_i))$ in total.
6.3 Simulation Results

In Table 1 we report bias, variance, and mean squared error (MSE) for each case. To save space, we only report simulation results with \( N = 100 \); Case #1 (\( r_\epsilon = 0 \) and \( b = 0 \); \( l = 0.5 \)). See the supplemental material Hamori, Motegi, and Zhang (2017) for complete results.

Insert Table 1 here.

In view of Table 1, \( \hat{r}_{12}, \hat{r}_{13} \), and \( \hat{r}_{23} \) are biased if and only if the naïve equal weight (EQ) is used for the first step. If EQ is used for the first step, there is a substantial bias of around 0.37 for each of \( \{\hat{r}_{12}, \hat{r}_{13}, \hat{r}_{23}\} \). In Case #4 (\( r_\epsilon \neq 0 \) and \( b \neq 0 \)), the bias decreases dramatically to around 0.10 for \( \{\hat{r}_{12}, \hat{r}_{13}\} \) and 0.03 for \( \hat{r}_{23} \) (cf. supplemental material).

There is a logical reason why the bias is larger in Case #1 than in Case #4. Recall that \( (r_{12}, r_{13}, r_{23}) = (0.000, 0.000, 0.000) \) in Case #1 and \( (r_{12}, r_{13}, r_{23}) = (0.705, 0.694, 0.860) \) in Case #4. Since we search optimal \( (\hat{r}_{12}, \hat{r}_{13}, \hat{r}_{23}) \) between -1 and +1, large upward bias cannot occur in Case #4.

Table 1 is concerned of the relatively large missing probability with \( l = 0.5 \). When \( l = 0.8 \), the bias caused by EQ in the first step decreases to about 0.10 (cf. supplemental material). It is expected that the smaller missing probability implies less bias since more information is available to the researcher.

Unless EQ is used for the first step, all estimators have virtually no bias and similar MSE. Different calibration weights lead to similar MSE, suggesting that a choice of calibration weights is not crucial. MSE is around 0.003 for each of \( \{\hat{r}_{12}, \hat{r}_{13}, \hat{r}_{23}\} \), indicating that the correlation coefficients are very accurately estimated in spite of the small sample size \( N = 100 \). This sharp performance arguably stems from the simplicity of the Gaussian DGP with Gaussian copula model. MSE gets even smaller when sample size grows up to \( N = 500 \) (cf. supplemental material).\(^5\)

\(^5\) As a future task, it is of interest to consider a more involved simulation design (e.g. larger...
7 Conclusions

Copula is an increasingly popular tool for analyzing complex interdependence among cross-section or time series variables. It is used in various research fields including economics and finance. Missing data are a challenging problem that appears in virtually any field.

The present work is one of the earliest contributions to bridge copula and missing data. We investigate the estimation of semiparametric copula models under the presence of missing data. Our models consist of nonparametric marginal distributions and parametric copula functions.

We handle missing data with calibration estimators, which are recently put forward by Chan, Yam, and Zhang (2016) in the literature of average treatment effects. An innovative aspect of Chan, Yam, and Zhang’s (2016) calibration estimators is that they do not require a direct estimation of propensity score functions. Their main insight is to interpret a propensity score function as a balancing factor between observable covariates of the non-missing group and those of the whole group. Calibration weights are chosen for each individual in a way that they are as consistent as possible with the moments of covariates.

We follow the same logic as Chan, Yam, and Zhang (2016) in order to extend Genest, Ghoudi, and Rivest’s (1995) two-step pseudo-likelihood estimator to a missing data framework. The calibration weights are used for both the first step of estimating marginal distributions and the second step of likelihood maximization with respect to copula parameters.

We prove the consistency and asymptotic normality of our estimators for copula parameters. We also present a natural procedure for consistently estimating the asymptotic variance of our estimators.

dimension $d$, non-Gaussian DGP, non-Gaussian copula model, different missing mechanism, comparison with other existing methods like imputations, etc.). In particular, it is of interest to compare the calibration estimators with FIML and imputations in such a realistic set-up. The current simulation indicates that the calibration estimators perform well in a simple setting at least.
References


Technical Appendices

A Lemmas on Calibration Weights

Recall (3.5) and (3.6), where the calibration weights in the first step are presented. For \(j \in \{1, \ldots, d\}\), define the population counterparts of \(\hat{G}_{jK}, \hat{\lambda}_{jK}, \hat{p}_{jK}(x)\):

\[
G_{jK}^*(\lambda) := \mathbb{E}[\hat{G}_{jK}(\lambda)] = \mathbb{E}[\pi_j(X_i)\rho(\lambda^T X_i) - \lambda^T u_K(X_i)],
\]

\[
\lambda_{jK}^* := \arg \max_{\lambda \in \mathbb{R}^K} G_{jK}^*(\lambda) \quad \text{and} \quad p_{jK}^*(x) := \frac{1}{N} \rho'((\lambda_{jK}^*)^T u_K(x)).
\]

Similarly, recall (3.9) for the calibration weights in the second step. Define the population counterparts of \(\hat{H}_K, \beta_K, \hat{q}_K(x)\):

\[
H_K^*(\beta) := \mathbb{E}[\hat{H}_K(\beta)] = \mathbb{E}[\eta(X_i)\rho(\beta^T X_i) - \beta^T u_K(X_i)],
\]

\[
\beta_K^* := \arg \max_{\beta \in \mathbb{R}^K} H_K^*(\beta) \quad \text{and} \quad q_K^*(x) := \frac{1}{N} \rho'((\beta_K^*)^T u_K(x)).
\]

The following lemma establishes the approximation result of the function \(\pi_j(x)^{-1}\) (resp. \(\eta(x)^{-1}\)) by \(Np_{jK}^*(x)\) (resp. \(Nq_K^*(x)\)):

**Lemma A.1.** Under Assumptions 2.2-2.5, we have

\[
\sup_{x \in \mathcal{X}} |Np_{jK}^*(x) - \pi_j(x)^{-1}| = O_p \left(K^{-\frac{1}{p} + 1}\right)
\]

and

\[
\sup_{x \in \mathcal{X}} |Nq_K^*(x) - \eta(x)^{-1}| = O_p \left(K^{-\frac{1}{p} + 1}\right).
\]

**Proof.** Currently not disclosed. \(\square\)

The other lemma is about the performance of the approximation of \(Np_{jK}^*(x)\) (resp. \(Nq_K^*(x)\)) by \(N\hat{p}_{jK}(x)\) (resp. \(N\hat{q}_K(x)\)):

**Lemma A.2.** Under Assumptions 2.2-2.5, we have

\[
\sup_{x \in \mathcal{X}} |N\hat{p}_{jK}(x) - Np_{jK}^*(x)| = O_p \left(\sqrt{\frac{K^3}{N}}\right)
\]

and

\[
\sup_{x \in \mathcal{X}} |N\hat{q}_K(x) - Nq_K^*(x)| = O_p \left(\sqrt{\frac{K^3}{N}}\right).
\]

**Proof.** Currently not disclosed. \(\square\)
B Proof of Proposition 4.1
Currently not disclosed.

C Proof of Theorem 4.2
Currently not disclosed.

D Proof of Theorem 4.3
Currently not disclosed.

E Proof of Theorem 5.1
Currently not disclosed.
Table 1: Simulation Results

<table>
<thead>
<tr>
<th>$(p, q)$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r_{12}$</td>
<td></td>
<td></td>
<td>$r_{13}$</td>
<td></td>
<td></td>
<td>$r_{23}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bias</td>
<td>Variance</td>
<td>MSE</td>
<td>Bias</td>
<td>Variance</td>
<td>MSE</td>
<td>Bias</td>
<td>Variance</td>
<td>MSE</td>
</tr>
<tr>
<td>$(EQ, EQ)$</td>
<td>0.3764</td>
<td>0.0196</td>
<td>0.1613</td>
<td>0.3744</td>
<td>0.0194</td>
<td>0.1596</td>
<td>0.3749</td>
<td>0.0182</td>
<td>0.1587</td>
</tr>
<tr>
<td>$(EQ, ET)$</td>
<td>0.3705</td>
<td>0.0291</td>
<td>0.1664</td>
<td>0.3727</td>
<td>0.0292</td>
<td>0.1682</td>
<td>0.3669</td>
<td>0.0289</td>
<td>0.1635</td>
</tr>
<tr>
<td>$(EQ, EL)$</td>
<td>0.3718</td>
<td>0.0321</td>
<td>0.1704</td>
<td>0.3702</td>
<td>0.0300</td>
<td>0.1670</td>
<td>0.3761</td>
<td>0.0307</td>
<td>0.1721</td>
</tr>
<tr>
<td>$(EQ, QR)$</td>
<td>0.3687</td>
<td>0.0288</td>
<td>0.1647</td>
<td>0.3761</td>
<td>0.0261</td>
<td>0.1675</td>
<td>0.3645</td>
<td>0.0298</td>
<td>0.1627</td>
</tr>
<tr>
<td>$(EQ, IL)$</td>
<td>0.3749</td>
<td>0.0298</td>
<td>0.1704</td>
<td>0.3795</td>
<td>0.0266</td>
<td>0.1707</td>
<td>0.3753</td>
<td>0.0295</td>
<td>0.1703</td>
</tr>
<tr>
<td>$(ET, EQ)$</td>
<td>-0.0015</td>
<td>0.0238</td>
<td>0.0238</td>
<td>-0.0022</td>
<td>0.0248</td>
<td>0.0248</td>
<td>0.0078</td>
<td>0.0232</td>
<td>0.0232</td>
</tr>
<tr>
<td>$(ET, ET)$</td>
<td>0.0018</td>
<td>0.0326</td>
<td>0.0326</td>
<td>-0.0060</td>
<td>0.0359</td>
<td>0.0360</td>
<td>0.0036</td>
<td>0.0345</td>
<td>0.0345</td>
</tr>
<tr>
<td>$(ET, EL)$</td>
<td>0.0041</td>
<td>0.0396</td>
<td>0.0396</td>
<td>-0.0010</td>
<td>0.0400</td>
<td>0.0400</td>
<td>0.0060</td>
<td>0.0361</td>
<td>0.0361</td>
</tr>
<tr>
<td>$(ET, QR)$</td>
<td>-0.0020</td>
<td>0.0339</td>
<td>0.0339</td>
<td>0.0036</td>
<td>0.0327</td>
<td>0.0327</td>
<td>0.0127</td>
<td>0.0339</td>
<td>0.0340</td>
</tr>
<tr>
<td>$(ET, IL)$</td>
<td>0.0118</td>
<td>0.0346</td>
<td>0.0347</td>
<td>-0.0056</td>
<td>0.0367</td>
<td>0.0367</td>
<td>0.0025</td>
<td>0.0347</td>
<td>0.0347</td>
</tr>
<tr>
<td>$(EL, EQ)$</td>
<td>-0.0043</td>
<td>0.0238</td>
<td>0.0238</td>
<td>-0.0023</td>
<td>0.0221</td>
<td>0.0221</td>
<td>0.0038</td>
<td>0.0241</td>
<td>0.0241</td>
</tr>
<tr>
<td>$(EL, ET)$</td>
<td>-0.0069</td>
<td>0.0382</td>
<td>0.0382</td>
<td>0.0084</td>
<td>0.0348</td>
<td>0.0349</td>
<td>0.0072</td>
<td>0.0345</td>
<td>0.0346</td>
</tr>
<tr>
<td>$(EL, EL)$</td>
<td>-0.0007</td>
<td>0.0390</td>
<td>0.0390</td>
<td>0.0016</td>
<td>0.0385</td>
<td>0.0385</td>
<td>-0.0021</td>
<td>0.0333</td>
<td>0.0333</td>
</tr>
<tr>
<td>$(EL, QR)$</td>
<td>0.0108</td>
<td>0.0378</td>
<td>0.0379</td>
<td>0.0074</td>
<td>0.0366</td>
<td>0.0366</td>
<td>0.0054</td>
<td>0.0367</td>
<td>0.0368</td>
</tr>
<tr>
<td>$(EL, IL)$</td>
<td>0.0053</td>
<td>0.0359</td>
<td>0.0359</td>
<td>-0.0073</td>
<td>0.0362</td>
<td>0.0363</td>
<td>-0.0041</td>
<td>0.0357</td>
<td>0.0357</td>
</tr>
<tr>
<td>$(QR, EQ)$</td>
<td>-0.0019</td>
<td>0.0240</td>
<td>0.0240</td>
<td>0.0025</td>
<td>0.0250</td>
<td>0.0250</td>
<td>0.0017</td>
<td>0.0221</td>
<td>0.0221</td>
</tr>
<tr>
<td>$(QR, ET)$</td>
<td>0.0004</td>
<td>0.0370</td>
<td>0.0370</td>
<td>0.0041</td>
<td>0.0361</td>
<td>0.0361</td>
<td>0.0090</td>
<td>0.0353</td>
<td>0.0354</td>
</tr>
<tr>
<td>$(QR, EL)$</td>
<td>-0.0081</td>
<td>0.0401</td>
<td>0.0401</td>
<td>0.0017</td>
<td>0.0378</td>
<td>0.0378</td>
<td>-0.0017</td>
<td>0.0376</td>
<td>0.0376</td>
</tr>
<tr>
<td>$(QR, QR)$</td>
<td>0.0003</td>
<td>0.0376</td>
<td>0.0376</td>
<td>-0.0041</td>
<td>0.0375</td>
<td>0.0375</td>
<td>0.0084</td>
<td>0.0352</td>
<td>0.0353</td>
</tr>
<tr>
<td>$(QR, IL)$</td>
<td>-0.0001</td>
<td>0.0359</td>
<td>0.0359</td>
<td>0.0012</td>
<td>0.0341</td>
<td>0.0341</td>
<td>-0.0028</td>
<td>0.0350</td>
<td>0.0350</td>
</tr>
<tr>
<td>$(IL, EQ)$</td>
<td>0.0078</td>
<td>0.0227</td>
<td>0.0228</td>
<td>-0.0052</td>
<td>0.0255</td>
<td>0.0256</td>
<td>0.0111</td>
<td>0.0226</td>
<td>0.0227</td>
</tr>
<tr>
<td>$(IL, ET)$</td>
<td>0.0018</td>
<td>0.0373</td>
<td>0.0373</td>
<td>0.0034</td>
<td>0.0377</td>
<td>0.0377</td>
<td>-0.0056</td>
<td>0.0347</td>
<td>0.0347</td>
</tr>
<tr>
<td>$(IL, EL)$</td>
<td>0.0136</td>
<td>0.0392</td>
<td>0.0394</td>
<td>0.0012</td>
<td>0.0405</td>
<td>0.0405</td>
<td>0.0081</td>
<td>0.0384</td>
<td>0.0384</td>
</tr>
<tr>
<td>$(IL, QR)$</td>
<td>0.0022</td>
<td>0.0338</td>
<td>0.0338</td>
<td>0.0054</td>
<td>0.0381</td>
<td>0.0381</td>
<td>-0.0021</td>
<td>0.0346</td>
<td>0.0346</td>
</tr>
<tr>
<td>$(IL, IL)$</td>
<td>0.0054</td>
<td>0.0361</td>
<td>0.0361</td>
<td>0.0131</td>
<td>0.0379</td>
<td>0.0381</td>
<td>0.0037</td>
<td>0.0358</td>
<td>0.0358</td>
</tr>
</tbody>
</table>

EQ: equal weights. ET: exponential tilting. EL: empirical likelihood. QR: quadratic. IL: inverse logistic. $(EQ, ET)$, for example, means that EQ is used for the first step of estimating marginal distributions while ET is used for the second step of maximizing likelihood. Sample size is $N = 100$; Case #1 ($r_c = 0$ and $b = 0$) is considered; there is relatively large missing probability ($l = 0.5$).