unibz

Fakultät für Wirtschaftswissenschaften
Facoltà di Economia
Faculty of Economics and Management

BEMPS –
Bozen Economics & Management
Paper Series

NO 112/ 2025

A nonparametric conditional copula-based imputation method

F. Marta L. Di Lascio, Aurora Gatto

A nonparametric conditional copula-based imputation method

F. Marta L. Di Lascio [*], Aurora Gatto[†]

Abstract

Missing values in multivariate dependent variables may occur during data collection, requiring imputation methods capable of handling complex inter-variable relationships. We propose a nonparametric copula-based method for imputing dependent multivariate missing data, called NPCoImp. By leveraging the conditional empirical beta copula of the missing variables given the observed ones, NPCoImp imputes data while accounting for its distributional shape, particularly radial symmetry, and adjusting the multivariate values used for imputation accordingly. NPCoImp is highly flexible and can handle multivariate missing data with any type of missingness pattern. The performance of the NPCoImp has been evaluated through an extensive Monte Carlo study and compared with classical imputation methods, as well as with its direct competitor, the CoImp algorithm. Our findings indicate that NPCoImp is particularly effective in preserving microdata and dependence structure. The strong performance of the proposed method is further supported by empirical case studies in the agricultural sector. Finally, the NPCoImp algorithm has been implemented in the R package `CoImp`, which is available on CRAN.

Keywords: Asymmetry, conditional copula, empirical copula, imputation methods, multivariate missing data, NPCoImp

JEL Codes: C1, C14, C63, Q1

[*] Faculty of Economics and Management, Free University of Bozen-Bolzano, Italy, e-mail: `marta.dilascio@unibz.it`.

[†] Faculty of Economics and Management, Free University of Bozen-Bolzano, Italy, e-mail: `aurora.gatto@unibz.it`.

1 Introduction

Imputation methods are statistical techniques used to fill in missing data within a dataset (see e.g. Little and Rubin (2019)). Missing data can arise in various fields, such as health care (see e.g. Molenberghs and Kenward (2007)), finance (see e.g. Hüttner et al. (2020), Cont and Kan (2011)), environmental science (see e.g. Chapon et al. (2023)), and social sciences (see e.g. Hammon (2023), Costantini et al. (2023)), due to various reasons, such as data entry errors, non-response in surveys, equipment malfunctions. Handling missing data becomes crucial since incomplete datasets can lead to biased results, reduced statistical power, and inaccurate conclusions. Even in big data analytics (Balusamy et al. (2021)), the inaccuracy caused by missing data can compromise the reliability of results, as the vast and complex nature of these datasets amplifies the impact of any gaps in the data. In this context, imputation methods are useful in (i) maintaining data integrity since big data often comes from various sources, (ii) enhancing model performance since imputation ensures that machine learning models can be trained on the full dataset, (iii) ensuring that the data stream remains consistent and reliable for real-time analytics when data is collected continuously in real-time (see e.g. Enders (2022)). Regardless, there are still some applied contexts where collecting and maintaining big data is challenging, e.g. in clinical and healthcare research as well as in social sciences and humanities, mainly due to data privacy concerns, limited access to digital infrastructure, variability in data formats, and difficulties in standardizing data collection. Therefore, imputation methods prove even more useful as they play a critical role in data preprocessing and ensure the integrity and usefulness of data for subsequent analysis.

In order to decide how to deal with missing data, several factors need to be considered: the missing data rate and the size of the complete data sample, the reason why there are missing values, if possible, and the kind of relationship among the observed variables. Different techniques have been proposed in literature (see e.g. Schafer (1997), Little and Rubin (2019)) and have been applied in various fields including surveys (Chen and Shao (2000), Little (1988)), clinical trials (Rivero et al. (2004)), bioinformatics (Wang et al. (2009)) and agricultural science (Robbins et al. (2013)). Traditional methods are the hot-deck donor (HD hereafter) imputation method (Fuller and Kim (2005)) and the k-nearest neighbour (kNN hereafter) imputation method (Chen and Shao (2000)) with its variants (e.g. Tutz and Ramzan (2015), Hasler and Tillé (2016)). Stochastic imputation methods include, among the others, the regression imputation by the expectation-maximization (EM hereafter) algorithm (Dempster et al. (1977)), the predictive mean matching (PMM hereafter) imputation method (Little (1988)), and more recent methods, such as the imputation method by chained equations (MICE hereafter) (Van Buuren and Groothuis-Oudshoorn (2011)) and the CoImp method (Di Lascio et al. (2015)) based on copulas (Sklar (1959)). All the mentioned methods are developed in the single imputation framework but the MICE; in this work we focus the attention on the single imputation context.

When the focus of the analysis is on preserving the dependence structure of multivariate data, imputation can be performed by filling in the missing values using draws from the conditional distribution of the missing data given the observed ones. In the so-called fully conditional specification (FCS hereafter) approach (Van Buuren et al. (2006)), which serves as the theoretical framework for the development of MICE, the multivariate model is built through a series of conditional models, each corresponding to an incomplete variable. Two main issues arise from this approach: (*i*) the statistical properties of the implied joint distribution are difficult to establish, and (*ii*) the implied joint distribution may not theoretically exist due to the incompatibility of conditionals (Arnold et al. (1999)). To overcome these issues, the conditional distributions should be derived from the joint multivariate distribution of all the variables of interest. Unfortunately, this is often extremely difficult, especially when the margins are different and/or the data exhibit a complex multivariate dependence structure. In this regard, copula models (see e.g. Nelsen (2006), Durante and Sempi (2016)), which make it possible to flexibly model the multivariate dependence structure of the data generating process (DGP hereafter) while separating it from the univariate margins, have enormous potential.

Copula function in the imputation context

To the best of our knowledge, copulas have been used for imputation purposes only in a few cases. The first work is by Käärik and Käärik (2009) who used Gaussian copulas to impute correlated incomplete data with repeated measurements through the mode of the conditional distributions derived from the joint density function. Few years later, Di Lascio et al. (2015) provided a more flexible solution to the problem of imputing by deriving the conditional probability density function of the available variables given the missing ones through the conditional copula density function. The copula-based imputation algorithm (CoImp hereafter) developed by Di Lascio et al. (2015) and extended in Di Lascio et al. (2014) allows the imputation of multivariate missing values of any missingness pattern by using all the Elliptical and Archimedean copula models. In addition, since the CoImp models margins nonparametrically though local likelihood estimators, it avoids the analytical problems that may arise when deriving conditional densities for certain combinations of copula and margins. Hasler et al. (2018) developed an imputation method based on vine copulas (see e.g. Bedford and Cooke (2002), Czado (2019)) that flexibly builds a joint model by a factorisation of the joint density into a tree of bivariate copulas. The developed method is suitable for multivariate missing data and allows for the use of a broader range of copula families than those considered in CoImp. However, it can only be applied to missing data with a monotone pattern.

In the multiple imputation context, Lun and Khattree (2022) developed an approach for univariate missing pattern that uses the copula only to transform non-normal data to normal one. The idea here is to circumvent the problem of non-symmetric data and allow the use of standard normality-based imputation techniques. More recently, Chapon et al. (2023) developed a method for imputing missing values within a Bayesian framework, specifically designed for extreme missing values. The authors proposed imputing time series of a target site using observations of the same variables from neighboring sites. They modelled the joint distribution of the time series at the target site and its neighboring sites using a D-vine copula with parametric margins and performed imputation by sampling values from the posterior distribution of a missing value, conditional on the observed stations for the given date.

Moving slightly away from the strict context of imputation, several contributions in the field of model estimation in the presence of missing data rely on copula-based methods. Ding and Song (2016) developed the expectation-maximization (EM hereafter) algorithm for the Gaussian copula regression model making it possible to estimate both the marginal parameters and the correlation matrix in presence of missing data. A more applied work that combines functional data analysis with imputation through copulas is in Chen et al. (2019). The authors effectively modelled the inter-sensor relationships in structural strain monitoring systems by using kernel copula density estimators and, inspired by the CoImp approach, used the joint model for imputing structural strains. The authors imputed the marginal distribution of the missing data itself using distribution regression methods instead of estimating the marginal distribution of the missing data directly from non-missing ones. Kertel and Pauly (2022) approached the estimation of the Gaussian copula with an incomplete dataset through the EM algorithm and showed how to model margins through mixture models when no a priori knowledge of their parametric family is available. More recently, Liebscher (2024), assuming a missing completely at random (MCAR hereafter) mechanism, estimated the copula parameter using an approach based on minimizing a linear combination of the Cramér-von Mises divergence measures between the sample copula and a parametric copula, each corresponding to a specific missing data pattern. This method ensures a consistent estimation of the copula parameters while avoiding the inaccuracies introduced by a possible imputation.

In the face of a now extensive literature of imputation methods, it is inevitable to ask what is the best method to use. In the context of single imputation using stochastic methods, the CoImp algorithm appears to be an attractive approach both theoretically and empirically, particularly when the primary goal is to preserve the complex multivariate dependence structure of the DGP and the missing data follow any pattern. Aissia et al. (2017) compared the performance of several imputation methods and applied them to hydrological data finding that the CoImp approach showed the best performance. Shiau and Lien (2021) imputed daily suspended sediment loads, demonstrating the usefulness of CoImp in terms of various performance measures, such as the root mean squared error. Nevertheless, some authors (Kim et al. (2017), Hasler et al. (2018), Hüttner et al. (2020)) compared CoImp with other imputation methods, highlighting some of its weaknesses, such as slow imputation due to the computational cost of the Hit-or-Miss Monte Carlo method and the limited families of copula models that are used to derive the conditional functions. Specifically, Kim et al. (2017) compared different imputation methods to assess their impact when re-sampling techniques, such as the jackknife and bootstrap, are used to estimate the variance of parameters in complex sampling designs. The CoImp showed a good performance in cluster sampling design where the data are skewed to right, but required more computational time than the other methods considered. This latter feature of the CoImp has also been confirmed by Hasler et al. (2018) who developed a D-vine imputation method for MCAR data with monotone non-response pattern. The method proposed appeared to have a better performance than the CoImp in many of the considered cases, even though it did not overcome all the considered competitors. A more applied work considering the performance of the CoImp is by Hüttner et al. (2020) who interpolated missing values in dependent credit spreads data using the kriging technique. The authors found that the CoImp does not perform very well with this kind of data, likely due to the fact that the imputation carried out through a randomly drawn imputed value is not the optimal point forecast.

Based on the discussed literature, there remains a need to develop a flexible and powerful copula-based imputation method that: (*i*) enables reliable multivariate missing data imputation while preserving the multivariate dependence structure of the DGP, (*ii*) is data-driven and distribution-free to maximize flexibility, (*iii*) has a competitive computational burden, and (*iv*) does not present theoretical issues in using conditional distributions, as these are derived from the joint distribution of missing and available values.

Summary

The main purpose of this paper is to develop a copula-based imputation method for multivariate dependent missing data of any pattern that: (*i*) overcomes the weaknesses of the CoImp method, primarily the limited set of copula families that can be used and its computational burden, (*ii*) avoids the problem of the incompatibility of conditionals found in other well-known imputation methods based on the FCS approach, and (*iii*) provides accurate imputations to preserve the dependence structure despite using a single guess to replace each missing value. To achieve this, we propose a nonparametric imputation method that is highly flexible, eliminates the risk of model misspecification, avoids the use of the Hit-or-Miss approach for generating missing data, and prevents theoretical issues in deriving conditional distributions.

The paper is organized as follows. In Section 2 we present in detail the nonparametric copula-based imputation method (NPCoImp hereafter) after describing its mathematical framework and setting the notation. In Section 3 we investigate the performance of the NPCoImp in a large Monte Carlo study comparing it with the CoImp and other competing methods. Section 4 presents empirical applications to two case studies concerning the agricultural sector. Finally, Section 5 outlines conclusions and discusses proposals for further research.

2 The NPCoImp method

Here we present the nonparametric copula-based imputation method, called NPCoImp, which imputes multivariate missing data of any pattern and dimension and leverages the conditional empirical copula of the missing variables given the observed ones.

2.1 Mathematical framework and notation

Given a p-dimensional real random row vector $\boldsymbol{X} = (X_1, \ldots, X_j, \ldots, X_p)$ and its realization $\boldsymbol{x} = (x_1, \ldots, x_j, \ldots, x_p)$, let $F(\cdot)$ be a p-dimensional cumulative distribution function with univariate marginal distribution functions $F_j(\cdot)$, for $j = 1, \ldots, p$. Then, following Sklar (1959), there exists a copula $C : [0,1]^p \to [0,1]$ such that:

$$F(\boldsymbol{x}) = F(x_1, \ldots, x_j, \ldots, x_p) = C\left(F_1(x_1), \ldots, F_j(x_j), \ldots, F_p(x_p)\right).$$

If $F_j(\cdot)$, $j = 1, \dots, p$ are continuous, then the copula $C(\cdot)$ associated with distribution function $F(\cdot)$ is unique and is given by:

$$\begin{aligned} C(\boldsymbol{u}) = C(u_1, \dots, u_j, \dots, u_p) &= P\left(U_1 \leq u_1, \dots, U_j \leq u_j, \dots, U_p \leq u_p\right) \\ &= F\left(F_1^{-1}(u_1), \dots, F_j^{-1}(u_j), \dots, F_p^{-1}(u_p)\right) \\ &= F(x_1, \dots, x_j, \dots, x_p), \qquad (1) \end{aligned}$$

where $\boldsymbol{U} = (U_1, \dots, U_j, \dots, U_p)$ is the vector of the probability integral transforms, defined as $U_j = F_j(X_j)$ for $j = 1, \dots, p$, such that $U_j \sim \mathcal{U}(0,1)$, $\forall j$, and $F_1^{-1}(\cdot), \dots, F_j^{-1}(\cdot), \dots, F_p^{-1}(\cdot)$ are the quantile functions (i.e. the inverse of marginal distribution functions). Therefore, a p–dimensional copula $C(\cdot)$ (with $p \geq 2$) is a p–dimensional distribution function with univariate uniform margins.

Let $\boldsymbol{X}_\nu = (X_j)_{j \in \nu} \in \mathbb{R}^q$ be the subvector of $\boldsymbol{X}$ corresponding to the indices $\nu \subset \{1, \dots, p\}$, where $|\nu| = q < p$, containing missing data. Let $\boldsymbol{U}_\nu$ be the subvector of $\boldsymbol{U}$ corresponding to $\boldsymbol{X}_\nu$. Furthermore, let $\boldsymbol{x}_\nu$ and $\boldsymbol{u}_\nu$ be realizations of $\boldsymbol{X}_\nu$ and $\boldsymbol{U}_\nu$, respectively. Let $\boldsymbol{X}_{\bar{\nu}} = (X_j)_{j \in \bar{\nu}} \in \mathbb{R}^{(p-q)}$ be the subvector of $\boldsymbol{X}$ corresponding to the indices $\bar{\nu} \subset \{1, \dots, p\} \setminus \nu$, where $|\bar{\nu}| = p - q$, containing no missing data. Let $\boldsymbol{U}_{\bar{\nu}}$ be the vector of uniform variables corresponding to $\boldsymbol{X}_{\bar{\nu}}$, i.e. a subvector of $\boldsymbol{U}$. Hence, $\boldsymbol{x}_{\bar{\nu}}$ and $\boldsymbol{u}_{\bar{\nu}}$ are realizations of $\boldsymbol{X}_{\bar{\nu}}$ and $\boldsymbol{U}_{\bar{\nu}}$, respectively. The conditional copula function $C_{\boldsymbol{U}_\nu|\boldsymbol{U}_{\bar{\nu}}}(\boldsymbol{u})$ used to impute the missing values $\boldsymbol{u}_\nu$ is defined through the Bayes' rule as follows:

$$C_{\boldsymbol{U}_\nu|\boldsymbol{U}_{\bar{\nu}}}(\boldsymbol{u}) = \frac{C_{\boldsymbol{U}}(\boldsymbol{u})}{C_{\boldsymbol{U}_{\bar{\nu}}}(\boldsymbol{u}_{\bar{\nu}})}. \qquad (2)$$

Specifically, in the method we propose we use the empirical version of the conditional copula function in Eq. (2) based on the following empirical beta copula (Segers et al. (2017)):

$$C_{\boldsymbol{U}}^{\mathcal{B}}(\boldsymbol{u}) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{p} F_{n,R_{i,j}^{(n)}}(u_j), \quad u_j \in [0,1]\,, j = 1, \dots, p, \qquad (3)$$

where

$$R_{i,j}^{(n)} = \sum_{k=1}^{n} \mathbb{1}\{X_{k,j} \leq X_{i,j}\}$$

is the rank of $X_{i,j}$ among $(X_{1,j}, \dots, X_{n,j})$, and, for $u \in [0,1]$ and $R_{i,j}^{(n)} = r \in \{1, \dots, n\}$,

$$F_{n,r}(u) = P(U_{r:n} \leq u) = \sum_{s=r}^{n} \binom{n}{s} u^s (1-u)^{n-s}$$

is the cumulative distribution function of a beta probabilistic model $\mathcal{B}(r, n+1-r)$ and $U_{1:n} < \cdots < U_{n:n}$ generically denote the order statistics based on n independent random variables $U_1, \ldots, U_n$, uniformly distributed on $[0,1]$. The empirical beta copula is then defined as a smoothed version of the empirical copula and it is a genuine copula that does not require the choice of a smoothing parameter (see e.g. Segers et al. (2017)). Hence, the conditional copula in Eq. (2) can be estimated through its empirical beta version by exploiting the Eq. (3):

$$C^{\mathcal{B}}_{\boldsymbol{U}_{\nu}|\boldsymbol{U}_{\bar{\nu}}}(\boldsymbol{u}) = \frac{C^{\mathcal{B}}_{\boldsymbol{U}}(\boldsymbol{u})}{C^{\mathcal{B}}_{\boldsymbol{U}_{\bar{\nu}}}(\boldsymbol{u}_{\bar{\nu}})}. \tag{4}$$

As an example, in case of missing bivariate variables, say X_j and $X_{j'}$ so that $\nu = (j, j')$ and $\bar{\nu} = (1, \ldots, j-1, j+1, \ldots, j'-1, j'+1, \ldots, p)$, the conditional empirical beta copula $C^{\mathcal{B}}_{\boldsymbol{U}_{\nu}|\boldsymbol{U}_{\bar{\nu}}}(\boldsymbol{u})$ in Eq. (4), where $\boldsymbol{U}_{\nu} = (U_j, U_{j'})$ and $\boldsymbol{U}_{\bar{\nu}} = (U_1, \ldots, U_{j-1}, U_{j+1}, \ldots, U_{j'-1}, U_{j'+1}, \ldots, U_p)$, is as follows:

$$\begin{aligned} C^{\mathcal{B}}_{\boldsymbol{U}_{\nu}|\boldsymbol{U}_{\bar{\nu}}}(\boldsymbol{u}) &= \frac{C^{\mathcal{B}}_{\boldsymbol{U}}(u_1, \ldots, u_{j-1}, u_j, u_{j+1}, \ldots, u_{j'-1}, u_{j'}, u_{j'+1}, \ldots, u_p)}{C^{\mathcal{B}}_{\boldsymbol{U}_{\bar{\nu}}}(u_1, \ldots, u_{j-1}, u_{j+1}, \ldots, u_{j'-1}, u_{j'+1}, \ldots, u_p)} \\ &= \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\prod_{j=1}^{p} F_{n,R^{(n)}_{i,j}}(u_j)}{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\prod_{j\in\bar{\nu}} F_{n,R^{(n)}_{i,j}}(u_j)}. \end{aligned}$$

The NPCoImp method we propose requires the use of an indicator of the symmetry/direction of the asymmetry of the copula $C^{\mathcal{B}}_{\boldsymbol{U}_{\nu}|\boldsymbol{U}_{\bar{\nu}}}(\cdot)$. We exploit the concept of radially symmetric copula (see e.g. Nelsen (1993), Nelsen (2006), and Durante and Sempi (2016)) that involves the survival copula $\widetilde{C}(\cdot)$ associated with a copula $C(\cdot)$ (see Eq. (1)) defined as follows:

$$\begin{aligned} \widetilde{C}(\boldsymbol{u}) &= = P\left(U_1 \geq 1-u_1, \ldots, U_j \geq 1-u_j, \ldots, U_p \geq 1-u_p\right) \\ &= 1 + \sum_{s=1}^{p}(-1)^s \sum_{1\leq j_1<\cdots<j_s\leq n} C_{j_1\ldots j_s}(1-u_{j_1}, \ldots, 1-u_{j_s}) \end{aligned}$$

where $C_{j_1\ldots j_s}(\cdot)$ denoting the marginal of $C(\cdot)$ related to $(j_1, \ldots, j_s)$. Consequently, the conditional empirical beta survival copula can be defined as follows:

$$\widetilde{C}^{\mathcal{B}}_{\boldsymbol{U}_{\nu}|\boldsymbol{U}_{\bar{\nu}}}(\boldsymbol{u}) = \frac{\widetilde{C}^{\mathcal{B}}_{\boldsymbol{U}}(\boldsymbol{u})}{\widetilde{C}^{\mathcal{B}}_{\boldsymbol{U}_{\bar{\nu}}}(\boldsymbol{u}_{\bar{\nu}})}. \tag{5}$$

Hence we propose the following criterion to assess the radial symmetry (about $\mathbf{1}_q 0.5$ where $\mathbf{1}_q$ is the all-one vector of dimension $(1 \times q)$) of the copula $\widetilde{C}^{\mathcal{B}}_{\boldsymbol{U}_\nu|\boldsymbol{U}_{\bar{\nu}}}(\cdot)$. Given a vector $\boldsymbol{\Psi} = (\Psi_1, \ldots, \Psi_a, \ldots, \Psi_A)$ of probabilities in $]0, 0.5[$, we define:

$$D_a = C^{\mathcal{B}}_{\boldsymbol{U}_\nu|\boldsymbol{U}_{\bar{\nu}}}\left(\boldsymbol{u}^{\Psi_a^-}\right) - \left(\widetilde{C}^{\mathcal{B}}_{\boldsymbol{U}_\nu|\boldsymbol{U}_{\bar{\nu}}}\left(\boldsymbol{u}^{\Psi_a^+}\right)\right) \tag{6}$$

where $\boldsymbol{u}^{\Psi_a^+}$ $\left(\boldsymbol{u}^{\Psi_a^-}\right)$ is the vector $\boldsymbol{u}$ in which each missing uniform number has been replaced by $0.5 + \Psi_a$ $(0.5 - \Psi_a)$, and we evaluate the radial symmetry of $C^{\mathcal{B}}_{\boldsymbol{U}_\nu|\boldsymbol{U}_{\bar{\nu}}}(\cdot)$ through the following:

$$\sum_{a=1}^{A} D_a \begin{cases} > 0 & \Rightarrow \quad C^{\mathcal{B}}_{\boldsymbol{U}_\nu|\boldsymbol{U}_{\bar{\nu}}}(\cdot) \quad \text{is assumed to be negative asymmetric} \\ = 0 & \Rightarrow \quad C^{\mathcal{B}}_{\boldsymbol{U}_\nu|\boldsymbol{U}_{\bar{\nu}}}(\cdot) \quad \text{is assumed to be symmetric (about } \mathbf{1}_q 0.5) \\ < 0 & \Rightarrow \quad C^{\mathcal{B}}_{\boldsymbol{U}_\nu|\boldsymbol{U}_{\bar{\nu}}}(\cdot) \quad \text{is assumed to be positive asymmetric.} \end{cases} \tag{7}$$

2.2 The NPCoImp algorithm

Here we describe in detail the procedure of the NPCoImp by making use of the preliminaries provided in Section 2.1. Suppose you have a $(n \times p)$-dimensional data matrix $\mathbb{X}$:

$$\mathbb{X} = \begin{pmatrix} x_{11} & \ldots & x_{1j} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots & \ldots & \vdots \\ x_{i1} & \ldots & x_{ij} & \ldots & x_{ip} \\ \vdots & \ldots & \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{nj} & \ldots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_i \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix}$$

whose corresponding pseudo-observations matrix is the following $(n \times p)$-dimensional matrix $\mathbb{U}$:

$$\mathbb{U} = \begin{pmatrix} u_{11} & \ldots & u_{1j} & \ldots & u_{1p} \\ \vdots & \ddots & \vdots & \ldots & \vdots \\ u_{i1} & \ldots & u_{ij} & \ldots & u_{ip} \\ \vdots & \ldots & \vdots & \ddots & \vdots \\ u_{n1} & \ldots & u_{nj} & \ldots & u_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_i \\ \vdots \\ \boldsymbol{u}_n \end{pmatrix}.$$

We define $\mathbb{X}^c$ as an $(n_c \times p)$-dimensional submatrix of $\mathbb{X}$ that contains only the complete observations, where $n_c < n$ and n_c represents the number of complete rows. $\mathbb{U}^c$ is the corresponding submatrix of $\mathbb{U}$. Given a vector of observations $\boldsymbol{x}_i$, where the subvector $\boldsymbol{x}_{i\nu} = (x_{ij})_{j\in\nu} \in \mathbb{R}^q$ consists of missing values, with $\nu \subset \{1, \dots, p\}$ (where $|\nu| = q < p$) denoting the set of indices corresponding to the variables in $\boldsymbol{X}$ that contain missing values in $\mathbb{X}$. We impute $\boldsymbol{x}_{i\nu}$ by using the corresponding vector of pseudo-observations $\boldsymbol{u}_{i\nu} = (u_{ij})_{j\in\nu} \in \mathbb{R}^q$ given $\boldsymbol{u}_{i\bar{\nu}} = (u_{ij})_{j\in\bar{\nu}} \in \mathbb{R}^{(p-q)}$ (where $\bar{\nu} \subset \{1, \dots, p\} \setminus \nu$ and $|\bar{\nu}| = p - q$) through the following procedure.

Given a row vector $\boldsymbol{x}_i$ with missing values at positions ν (where ν has length q) and its corresponding pseudo-observations vector $\boldsymbol{u}_i$, we

1. estimate nonparametrically the conditional copula of the missing variables given the available ones through the conditional empirical beta copula $C^{\mathcal{B}}_{\boldsymbol{U}_\nu|\boldsymbol{U}_{\bar{\nu}}}(\cdot)$ in Eq. (4) and its survival version $\widetilde{C}^{\mathcal{B}}_{\boldsymbol{U}_\nu|\boldsymbol{U}_{\bar{\nu}}}(\cdot)$ in Eq. (5) using the complete pseudo-observations in $\mathbb{U}^c$;

2. evaluate the radial symmetry/asymmetry of $C^{\mathcal{B}}_{\boldsymbol{U}_\nu|\boldsymbol{U}_{\bar{\nu}}}(\cdot)$ by exploiting Eqs. (6)-(7);

3. impute $\boldsymbol{u}_{i\nu}$ through $\boldsymbol{u}^\star_{i\nu} \in [0, 1]^q$ whose values depend on the radial (a)symmetry of the conditional empirical beta copula; specifically:

   - if $C^{\mathcal{B}}_{\boldsymbol{U}_\nu|\boldsymbol{U}_{\bar{\nu}}}(\cdot)$ is assumed to be symmetric about $\mathbf{1}_q 0.5$, then $\boldsymbol{u}^\star_{i\nu} = \mathbf{1}_q 0.5$
   - if $C^{\mathcal{B}}_{\boldsymbol{U}_\nu|\boldsymbol{U}_{\bar{\nu}}}(\cdot)$ is assumed to be negative asymmetric, then $\boldsymbol{u}^\star_{i\nu} = \mathbf{1}_q \left(0.5 - \arg\max_{\Psi_a} D_a\right)$
   - if $C^{\mathcal{B}}_{\boldsymbol{U}_\nu|\boldsymbol{U}_{\bar{\nu}}}(\cdot)$ is assumed to be positive asymmetric, then $\boldsymbol{u}^\star_{i\nu} = \mathbf{1}_q \left(0.5 + \arg\max_{\Psi_a} D_a\right)$

   where $\mathbf{1}_q$ is the all-one vector of dimension $(1 \times q)$;

4. obtain the imputed vector $\hat{\boldsymbol{u}}_i$ by merging $\boldsymbol{u}_{i\bar{\nu}}$ and $\boldsymbol{u}^\star_{i\nu}$;

5. compute the dissimilarity between $\hat{\boldsymbol{u}}_i$ and each row of the complete data matrix $\mathbb{U}^c$ through an appropriate dissimilarity measures $d_{(\cdot,\cdot)}$, e.g. Kendall-based correlation measures, Gower's index, Euclidean, Manhattan and Canberra distances; then, select the K rows of pseudo-observations that minimize the computed dissimilarities, say $(\boldsymbol{u}^c_1, \dots, \boldsymbol{u}^c_k, \dots, \boldsymbol{u}^c_K)^T$, and the corresponding complete observations $(\boldsymbol{x}^c_1, \dots, \boldsymbol{x}^c_k, \dots, \boldsymbol{x}^c_K)^T$ in $\mathbb{X}^c$;

6. impute each missing value in $\boldsymbol{x}_{i\nu}$ through the following: $x^{\star}_{ij} = \dfrac{1}{K}\displaystyle\sum_{k=1}^{K} x^{c}_{kj}$, by varying j in ν, thus obtaining $\boldsymbol{x}^{\star}_{i\nu}$;

7. obtain the imputed vector $\boldsymbol{x}^{\mathrm{im}}_{i}$ by merging $\boldsymbol{x}_{i\bar{\nu}}$ and $\boldsymbol{x}^{\star}_{i\nu}$;

8. compute the pseudo-observations of $\boldsymbol{x}^{\mathrm{im}}_{i}$, say $\boldsymbol{u}^{\mathrm{im}}_{i}$, which can be interpreted as an element of the lower-orthant quantile (see e.g. Embrechts and Puccetti (2006)), and its 'order' α_i given by $C^{\mathcal{B}}_{\boldsymbol{U}_{\nu}|\boldsymbol{U}_{\bar{\nu}}}(\boldsymbol{u}^{\mathrm{im}}_{i})$.

To clarify the procedure of the proposed method we provide its algorithm in the box **Algorithm 1**. It is worth noticing that: (*i*) in principle, empirical copula functions other than the beta can be used, i.e. the empirical checkerboard copula (see Segers et al. (2017) and the references therein); this is technically feasible since it has been implemented in the `CoImp` package; (*ii*) the range and the number of values in $\boldsymbol{\Psi}$ can have an impact on the imputation results (see simulation results in Sect. 3); (*iii*) setting all missing values equal to $(0.5 \pm \Psi_a)$ when selecting values for imputation may be restrictive, but it helps to reduce the computational burden of the developed method.

3 Monte Carlo study

In this section we carry out a simulation study to assess the performance of the NPCoImp method versus other imputation methods. Precisely, we compare our proposal with the following imputation methods: HD (see e.g. Kalton and Kasprzyk (1982), Fuller and Kim (2005)), also called donor imputation, kNN (Kowarik and Templ (2016)), PMM (see e.g. Rubin (1986), Little (1988)), and CoImp (Di Lascio et al. (2015)). All the competitor methods considered are (*i*) developed in a single imputation context, (*ii*) feasible for continuous variables and missing of any pattern and dimension, (*iii*) generally based on the assumption of MCAR data (i.e. data in which the missingness of a response is unrelated to both its unknown value and observed data).

As for the hot-deck imputation method, each missing value is replaced with the complete response most similar to the missing one based on the Euclidean distance and called donor. In cases where there is more than one unit with minimum distance, the donor is randomly selected in the so-called donor pool (Rubin (2004)). The kNN imputation method is again based on a donor observation but, differently from the HD method, the missing value is imputed through an aggregation of its nearest neighbors (five in our case). Here, the distance between two observations, i and i', is computed though

Algorithm 1 The NPCoImp algorithm

```
**Input:** $\mathbb{X}$ the $(n \times p)$-dimensional data matrix with missing values, $M$ the total number of single missing values, $\boldsymbol{\Psi}$ an $A$-dimensional vector of values in $]0, 0.5[$, $d_{(\cdot,\cdot)}$ a dissimilarity measure, and $K$ the number of dissimilarities to select for the imputation.
**Output:** The $(n \times p)$-dimensional imputed data matrix $\mathbb{X}^{\mathrm{im}}$ and the vector of probabilities $\boldsymbol{\alpha}$ corresponding to the lower-orthant quantiles used for imputation.
Compute $\mathbb{U}$ as the probability integral transforms (pseudo-observations) of $\mathbb{X}$
for $i = 1, \dots, n$ do
    $l = 0$
    if $\boldsymbol{x}_i$ contains missing value(s) then
        $l = l + 1$
        select the corresponding $\boldsymbol{u}_i$;
        set $\nu \subset \{1, \dots, p\}$ as the set of indices corresponding to the observed variables containing missing value(s) in $\boldsymbol{x}_i$ and $\boldsymbol{u}_i$;
        set $|\nu| = q$ the number of single missing values in $\boldsymbol{x}_i$ and $\boldsymbol{u}_i$, i.e. the dimension of the multivariate missing values;
        compute $\sum_{a=1}^{A} D_a$, where $D_a$ is given in Eq. (6) that exploits Eq. (4);
        if $\sum_{a=1}^{A} D_a = 0$ then
            $C^{\mathcal{B}}_{\boldsymbol{U}_\nu | \boldsymbol{U}_{\bar{\nu}}}(\cdot)$ is assumed to be symmetric;
            $\boldsymbol{u}^{\star}_{i\nu} = \mathbf{1}_q 0.5$;
        else
            if $\sum_{a=1}^{A} D_a > 0$ then
                $C^{\mathcal{B}}_{\boldsymbol{U}_\nu | \boldsymbol{U}_{\bar{\nu}}}(\cdot)$ is assumed to be negative symmetric;
                $\boldsymbol{u}^{\star}_{i\nu} = \mathbf{1}_q \left(0.5 - \arg\max_{\Psi_a} D_a\right)$;
```

Algorithm 1 The NPCoImp algorithm (continued)

```
            else
                if $\sum_{a=1}^{A} D_a < 0$ then
                    $C^{\mathcal{B}}_{\boldsymbol{U}_\nu|\boldsymbol{U}_{\bar{\nu}}}(\cdot)$ is assumed to be positive symmetric;
                    $\boldsymbol{u}^\star_{i\nu} = \mathbf{1}_q \left(0.5 + \arg\max_{\Psi_a} D_a\right)$;
                end if
            end if
        end if
        set $\hat{\boldsymbol{u}}_i = (\boldsymbol{u}^\star_{i\nu}, \boldsymbol{u}_{i\bar{\nu}})$;
        compute $d_{(\hat{\boldsymbol{u}}_i, \boldsymbol{u}^c_{i'})}$, by varying $i'$ in $(1, \ldots, n_c)$ (i.e. for each row in
        $\mathbb{U}^c$);
        select $(\boldsymbol{u}^c_1, \ldots, \boldsymbol{u}^c_k, \ldots, \boldsymbol{u}^c_K)^T$ minimizing the computed dissimilarities, and the corresponding $(\boldsymbol{x}^c_1, \ldots, \boldsymbol{x}^c_k, \ldots, \boldsymbol{x}^c_K)^T$ in $X^c$;
        impute $\boldsymbol{x}_{i\nu}$ through $\boldsymbol{x}^\star_{i\nu} = \left(\frac{1}{K}\sum_{k=1}^{K} x^c_{k1}, \ldots, \frac{1}{K}\sum_{k=1}^{K} x^c_{kq}\right)$;
        obtain the imputed vector $\boldsymbol{x}^{\text{im}}_i = (\boldsymbol{x}^\star_{i\nu}, \boldsymbol{x}_{i\bar{\nu}})$;
        store $\boldsymbol{x}^{\text{im}}_i$ in the $i$-th row of $\mathbb{X}^{\text{im}}$;
        compute the pseudo-observation of $\boldsymbol{x}^{\text{im}}_i$, say $\boldsymbol{u}^{\text{im}}_i$, and
        $C^{\mathcal{B}}_{\boldsymbol{U}_\nu|\boldsymbol{U}_{\bar{\nu}}}(\boldsymbol{u}^{\text{im}}_i) = \alpha_l$.
    end if
end for
Obtain $\mathbb{X}^{\text{im}}$ and $\boldsymbol{\alpha}$.
```

an extension of the Gower distance (Gower (1971)) and it is given by the following weighted mean of the contributions of each variable:

$$d_{(i,i')} = \frac{\sum_{j=1}^{p} w_j \delta_{ii'j}}{\sum_{j=1}^{p} w_j} \tag{8}$$

where w_j is the weight associated to the j-th variable and $\delta_{ii'j}$ is the contribution of the j-th variable to the distance between observations i and i' that can be computed as:

$$\delta_{ii'j} = \frac{\mid x_{ij} - x_{i'j} \mid}{r_j}$$

where x_{ij} and $x_{i'j}$ are the values of i and i' on the j-th variable that has range r_j. As $\delta_{ii'j} \in [0, 1]$, $\forall i, i', j$, $d_{(i,i')} \in [0, 1]$. The PMM imputation method estimates on the complete data a multivariate linear regression model for each variable with missing values: the latter is used as dependent variable, while the others observed variables are used as covariates. The missing value is replaced with the observed value that corresponds to the closest predicted mean, randomly selected from a small set of nearby candidate donors (five in our case). Finally, as for the CoImp, we set 0.5 the values for the nearest neighbour component of the smoothing parameter for the local likelihood estimator and we let the CoImp method selects automatically the most suitable copula model among the Elliptical and Archimedean copula families and their rotated version (Nelsen (2002); Brechmann and Schepsmeier (2013)). The CoImp is the most interesting competitor of our proposal since it is based on copulas and on the conditional distribution of the missing observations given the available ones. Our main purpose is, indeed, to propose an imputation method that, while still based on conditional copulas, allows for an expanded and unconstrained set of multivariate dependence structures for the DGP, is faster than CoImp, and outperforms it. Finally, for the NPCoImp method, we set $\mathbf{\Psi} = \{0.050, 0.051, \ldots, 0.449, 0.450\}$, use the Gower distance in Eq. (8) with unit weights, and set the number of distances used for imputation to 10% of n.

We perform a large simulation study using two different DGPs and by varying several parameters. Specifically, we generate data from $i)$ a mixture copula composed by a three-variate Clayton copula and a three-variate rotated-Gumbel copula, with uniform margins X_j, for $j = 1, 2, 3$ and $ii)$ a mixture copula with two components given by a four-variate Frank and a rotated Frank copulas, with uniform margins X_j, for $j = 1, 2, 3, 4$. An example of a sample generated from the two considered DGPs is provided in Fig. 1. We generate a random sample from each of the above described DGP by varying

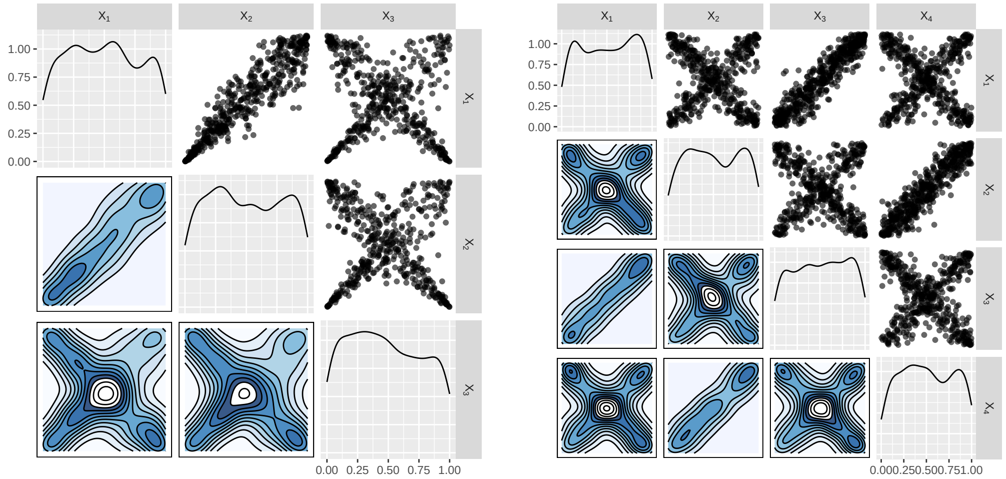

Figure 1: Scatter plots (upper triangular), contour plots (lower triangular), and density functions (the diagonal) for the three-mixture copula DGP with a Clayton and rotated-Gumbel copulas (left panel), and for the four-variate mixture copula DGP with a Frank and a rotated Frank copulas (right panel), with $\tau = 0.75$. Sample size $n = 500$.

1. the dependence parameter of each copula model in the mixture such that the Kendall's $\tau = 0.25, 0.50, 0.75$;
2. the sample size $n = 50, 250, 500$.

We then introduce artificial MCAR values into each of the generated data set by varying the number of multivariate missing data, i.e. matrix rows with missing values, in $(0.20n, 0.40n)$. It is worth noticing that the total number of single missing values M can vary in each replication since we set the number of matrix rows with multivariate missing value and not the number of single missing values. Next, we apply the five methods described above to the data set as to obtain the M imputed (single) observations, say x_m^{im}, $m = 1, \ldots, M$. The number of replications H is set to 500 in order to take into account the source of variability deriving from the randomness of the mechanism of generation of the missing data. Finally, the goodness of the imputation methods is assessed by means of the following performance measures that compare the imputed dataset with the original one in terms of microdata and dependence structure:

1. the mean absolute error (MAE hereafter)

$$\text{MAE} = \frac{1}{H}\sum_{h=1}^{H}\left[\frac{1}{M}\sum_{m=1}^{M}\left|x_m^{im} - x_m^{ob}\right|\right];$$

where x_m^{im} is the m-th value imputed through the method considered and x_m^{ob} is the observed value;

2. the unscaled mean bounded relative absolute error (UMBRAE hereafter)

$$\text{UMBRAE} = \frac{1}{H}\sum_{h=1}^{H}\left[\frac{1}{M}\sum_{m=1}^{M}\frac{\left|x_m^{im} - x_m^{ob}\right|}{\left|x_m^{im} - x_m^{ob}\right| + \left|x_m^{im_{ben}} - x_m^{ob}\right|}\right],$$

where x_m^{im} is the m-th (single) value imputed through the method under consideration, $x_m^{im_{ben}}$ is the m-th value imputed by using the benchmark method, which is the CoImp imputation method in our case, and x_m^{ob} is the corresponding observed value; hence, when UMBRAE = 1, the method considered performs roughly the same as the CoImp imputation method; when UMBRAE < 1, the method considered performs roughly (1 − UMBRAE)% better than the benchmark method; when UMBRAE > 1, the method considered is roughly (UMBRAE − 1)% worse than the benchmark method (see Chen et al. (2017) for details);

3. the relative bias (RB hereafter) and the relative root mean squared error (RRMSE hereafter) for the Blomqvist's beta introduced by Nelsen (2002):

$$\text{RB}_\beta = \frac{1}{H}\sum_{h=1}^{H}\left(\frac{\hat{\beta}_h - \beta}{\beta}\right); \qquad \text{RRMSE}_\beta = \sqrt{\frac{1}{H}\sum_{h=1}^{H}\left(\frac{\hat{\beta}_h - \beta}{\beta}\right)^2} \quad (9)$$

where β, which takes value in $[-1, 1]$, is the true value of the multivariate Blomqvist's beta, i.e the multivariate version of medial correlation coefficient, and $\hat{\beta}_h$ is its estimated value for the h-th simulated sample. More in detail, let $\boldsymbol{U}$ be a p-dimensional random vector of uniform $[0, 1]$ variables with p-copula $C(\cdot)$ and survival p-copula $\widetilde{C}(\cdot)$, then the multivariate version of Blomqvist's beta in terms of p-copulas (see Úbeda-Flores (2005)), denoted by $\beta_{p,C}$ is given by:

$$\beta_{p,C} = \frac{2^{p-1}[C(\mathbf{1/2}) + \widetilde{C}(\mathbf{1/2})] - 1}{2^{p-1} - 1}$$

and it is such that $\beta_{p,C} = 0$ when $C(\cdot)$ is the distribution of independent random variables and $\beta_{p,C} = 1$ for perfect positive dependence. We specify that, for each simulated scenario, in the calculation of Eqs. (9), β denotes the Blomqvist's coefficient estimated before introducing missing values, i.e. from the originally drawn data matrix, while $\hat{\beta}_h$ represents the Blomqvist's coefficient estimated after imputation, i.e. from the imputed data matrix.

For the mixture copula composed by a three-variate Clayton copula and a three-variate rotated-Gumbel copula, we also evaluate the performance of the NPCoImp method by varying $\mathbf{\Psi}$. We consider four different sets of values for $\mathbf{\Psi}$: from 0.050 to 0.450 with sequence increments of 0.01 and 0.001, and from 0.010 to 0.490 with sequence increments of 0.01 and 0.001; in such cases we set $\tau = 0.50$, $n = 250$, $H = 500$ and the number of multivariate missing data equal to $0.2n$. In addition, for a more comprehensive assessment of NPCoImp's performance, we illustrate its computational efficiency by computing the relative process time (RPT hereafter) used by the imputation methods investigated throughout this manuscript. The RPT is defined as the CPU time in seconds required for the imputation of $0.4n$ with $n = 500$ multivariate missing values (i.e. matrix rows with missing value(s)) divided by CPU time required when the number of multivariate missing data is $0.2n$. The DGP here is again the three-variate mixture of Clayton and rotated-Gumbel copulas with $\tau = 0.50$.

Tables 1, 2, and 3 show the results of the simulations for the Clayton and rotated-Gumbel mixture model with three margins. Considering low dependence level ($\tau = 0.25$), NPCoImp overcomes all the other considered methods in terms of both MAE, RB$_\beta$ and RRMSE$_\beta$, and irrespective of the number of missing data in all the scenarios. Coherently, our proposal always outperforms the benchmark method, i.e. CoImp (see the values of UMBRAE in Tab. 1). In the scenarios with mild and high dependence (i.e. $\tau = (0.5, 0.75)$), NPCoImp appears to be the best method for imputation in terms of preservation of the dependence structure; in fact, both RB$_\beta$ and RRMSE$_\beta$ show the lowest values regardless of the sample size and the number of multivariate missing values. Regarding the MAE, our proposal outperforms all other methods but kNN in all considered scenarios (with the exception of a case where $\tau = 0.50$, $n = 50$ and few missing values, in which our method performs better). Finally, the NPCoImp algorithm outperforms CoImp but seems to be slightly worse when the dependence is high ($\tau = 0.75$) and the sample size is small ($n = 50$).

Tables 4, 5, and 6 show the results of the simulations for a four-variate rotated-Frank and Frank copula mixture model. When the dependence is low or mild, we can draw the same conclusions as when Clayton and rotated-Gumbel mixture copula model was used as DGP in terms of measures of dependence and almost the same conclusions in terms of MAE. When the dependence is high, the proposed method is again the best one to preserve multivariate dependence, except in cases when the sample size is large ($n = 500$)

Performance measures	Num. of multiv. missing values	NPCoImp	HD	kNN	CoImp	PMM
		$n = 50$				
MAE	0.2n	0.408	0.496	0.422	0.445	0.481
UMBRAE		0.976	1.128	1.001	1	1.088
RB_β		-0.686	-0.847	-0.805	-0.815	-0.802
RRMSE_β		0.757	0.898	0.885	0.871	0.863
MAE	0.4n	0.411	0.496	0.416	0.429	0.482
UMBRAE		0.988	1.129	0.997	1	1.100
RB_β		-0.648	-0.874	-0.792	-0.782	-0.794
RRMSE_β		0.743	0.929	0.904	0.841	0.877
		$n = 250$				
MAE	0.2n	0.399	0.498	0.415	0.433	0.483
UMBRAE		0.968	1.115	0.967	1	1.086
RB_β		-0.593	-0.791	-0.746	-0.761	-0.753
RRMSE_β		0.606	0.799	0.763	0.771	0.764
MAE	0.4n	0.404	0.502	0.416	0.433	0.484
UMBRAE		0.970	1.124	0.976	1	1.081
RB_β		-0.605	-0.818	-0.719	-0.755	-0.755
RRMSE_β		0.625	0.828	0.754	0.769	0.770
		$n = 500$				
MAE	0.2n	0.396	0.498	0.411	0.434	0.483
UMBRAE		0.953	1.112	0.965	1	1.081
RB_β		-0.580	-0.779	-0.717	-0.744	-0.740
RRMSE_β		0.587	0.783	0.725	0.749	0.746
MAE	0.4n	0.401	0.500	0.414	0.434	0.482
UMBRAE		0.962	1.117	0.965	1	1.077
RB_β		-0.608	-0.821	-0.697	-0.746	-0.741
RRMSE_β		0.619	0.825	0.713	0.753	0.749

Table 1: Simulation study results for the three-variate Clayton and rotated-Gumbel mixture copula, as defined in the text with $\tau = 0.25$: performance measures of the NPCoImp method compared to other imputation methods.

and there are many missing values ($0.4n$). Nonetheless, NPCoImp overcomes the CoImp (see values of UMBRAE in Tabs. 4, 5, and 6), but it seems to be slightly worse in scenarios with high dependence and few missing values regardless of the sample size. In terms of MAE, under low dependence ($\tau = 0.25$), NPCoImp achieves the best imputation accuracy, except when $n = 50$ and the percentage of missing values is high, where kNN outperforms all methods. Under mild dependence, kNN outperforms NPCoImp, while under high dependence, both kNN and PMM achieve better performance than our proposed method. The latter result may be due to the fact that the components of the copula mixture are symmetric, making standard methods less sensitive to complexity. Overall, we can conclude that NPCoImp performs

Performance measures	Num. of multiv. missing values	NPCoImp	HD	kNN	CoImp	PMM
		$n = 50$				
MAE	0.2n	0.380	0.493	0.381	0.432	0.438
UMBRAE		0.974	1.173	0.947	1	1.056
RB$_\beta$		-0.493	-0.650	-0.572	-0.591	-0.585
RRMSE$_\beta$		0.539	0.689	0.626	0.637	0.629
MAE	0.4n	0.387	0.497	0.383	0.418	0.433
UMBRAE		0.985	1.187	0.964	1	1.043
RB$_\beta$		-0.493	-0.712	-0.606	-0.597	-0.585
RRMSE$_\beta$		0.565	0.754	0.693	0.653	0.643
		$n = 250$				
MAE	0.2n	0.373	0.502	0.368	0.418	0.434
UMBRAE		0.964	1.196	0.920	1	1.036
RB$_\beta$		-0.432	-0.621	-0.545	-0.561	-0.557
RRMSE$_\beta$		0.441	0.627	0.556	0.569	0.565
MAE	0.4n	0.376	0.499	0.364	0.418	0.431
UMBRAE		0.959	1.176	0.898	1	1.027
RB$_\beta$		-0.468	-0.677	-0.507	-0.551	-0.553
RRMSE$_\beta$		0.488	0.684	0.535	0.562	0.565
		$n = 500$				
MAE	0.2n	0.373	0.499	0.363	0.420	0.432
UMBRAE		0.956	1.181	0.895	1	1.027
RB$_\beta$		-0.436	-0.611	-0.526	-0.548	-0.551
RRMSE$_\beta$		0.442	0.614	0.533	0.552	0.556
MAE	0.4n	0.376	0.498	0.361	0.420	0.430
UMBRAE		0.959	1.172	0.882	1	1.024
RB$_\beta$		-0.478	-0.676	-0.497	-0.543	-0.552
RRMSE$_\beta$		0.489	0.679	0.512	0.549	0.559

Table 2: Simulation study results for the three-variate Clayton and rotated-Gumbel mixture copula, as defined in the text with $\tau = 0.50$: performance measures of the NPCoImp method compared to other imputation methods.

satisfactorily, in terms of preservation of both microdata and dependence.

Table 7 presents the performance of the NPCoImp method by varying $\boldsymbol{\Psi}$. Including extreme values in Ψ appears to have a slightly negative impact on the method's performance in terms of both dependence measures and MAE. On the contrary, increasing the number of values considered for $\boldsymbol{\Psi}$ enhances NPCoImp's performance in terms of microdata preservation. In addition, the choice of $\boldsymbol{\Psi}$ does not seem to affect NPCoImp's performance relative to CoImp, which is consistently outperformed by NPCoImp across all evaluated performance measures. In conclusion, $\boldsymbol{\Psi}$ has a slight impact on the imputation performance, and the selection of Ψ_a values should primarily be guided by the specific objectives of the analysis and computational resources.

Performance measures	Num. of multiv. missing values	NPCoImp	HD	kNN	CoImp	PMM
		$n = 50$				
MAE	0.2n	0.349	0.492	0.333	0.398	0.361
UMBRAE		1.033	1.354	0.927	1	1.009
RB_β		-0.353	-0.516	-0.436	-0.447	-0.437
RRMSE_β		0.390	0.545	0.479	0.487	0.471
MAE	0.4n	0.351	0.500	0.350	0.393	0.362
UMBRAE		1.017	1.359	0.973	1	0.994
RB_β		-0.371	-0.608	-0.504	-0.449	-0.436
RRMSE_β		0.438	0.638	0.570	0.506	0.487
		$n = 250$				
MAE	0.2n	0.334	0.503	0.315	0.391	0.361
UMBRAE		0.979	1.357	0.869	1	0.970
RB_β		-0.311	-0.499	-0.419	-0.413	-0.415
RRMSE_β		0.320	0.505	0.430	0.421	0.423
MAE	0.4n	0.332	0.499	0.297	0.388	0.358
UMBRAE		0.978	1.360	0.815	1	0.973
RB_β		-0.310	-0.588	-0.395	-0.405	-0.419
RRMSE_β		0.331	0.593	0.417	0.419	0.430
		$n = 500$				
MAE	0.2n	0.329	0.498	0.296	0.388	0.357
UMBRAE		0.975	1.359	0.807	1	0.972
RB_β		-0.298	-0.494	-0.403	-0.402	-0.416
RRMSE_β		0.303	0.497	0.409	0.405	0.420
MAE	0.4n	0.329	0.501	0.293	0.390	0.357
UMBRAE		0.959	1.353	0.772	1	0.965
RB_β		-0.279	-0.582	-0.380	-0.393	-0.418
RRMSE_β		0.292	0.584	0.392	0.399	0.424

Table 3: Simulation study results for the three-variate Clayton and rotated-Gumbel mixture copula, as defined in the text with $\tau = 0.75$: performance measures of the NPCoImp method compared to other imputation methods.

Finally, Table 8 shows estimates of the relative process time (RPT hereafter) used by the imputation methods investigated throughout this manuscript. Our proposal turns out to be considerably faster than CoImp and competitive with respect to the other imputation methods considered. Although slightly slower than HD and PMM, NPCoImp exhibits less standard error, confirming a favorable trade-off between speed and stability. This suggests that our method can be suitable also for datasets with a high number of missing values that nowadays are often encountered in large surveys and big data analytics.

Performance measures	Num. of multiv. missing values	NPCoImp	HD	kNN	CoImp	PMM
		$n = 50$				
MAE	0.2n	0.533	0.664	0.547	0.579	0.634
UMBRAE		0.970	1.137	0.976	1	1.094
RB_β		-0.494	-0.677	-0.584	-0.619	-0.621
RRMSE_β		0.577	0.742	0.678	0.687	0.694
MAE	0.4n	0.555	0.664	0.551	0.571	0.635
UMBRAE		0.999	1.132	0.986	1	1.086
RB_β		-0.423	-0.721	-0.610	-0.680	-0.636
RRMSE_β		0.564	0.788	0.729	0.763	0.725
		$n = 250$				
MAE	0.2n	0.526	0.665	0.543	0.571	0.634
UMBRAE		0.954	1.121	0.963	1	1.073
RB_β		-0.364	-0.597	-0.538	-0.583	-0.553
RRMSE_β		0.384	0.607	0.555	0.593	0.566
MAE	0.4n	0.543	0.666	0.550	0.574	0.637
UMBRAE		0.965	1.108	0.964	1	1.066
RB_β		-0.290	-0.665	-0.517	-0.617	-0.566
RRMSE_β		0.338	0.677	0.552	0.630	0.584
		$n = 500$				
MAE	0.2n	0.526	0.664	0.548	0.572	0.637
UMBRAE		0.949	1.107	0.961	1	1.069
RB_β		-0.338	-0.594	-0.532	-0.567	-0.543
RRMSE_β		0.349	0.598	0.541	0.572	0.548
MAE	0.4n	0.542	0.667	0.551	0.576	0.636
UMBRAE		0.962	1.111	0.961	1	1.063
RB_β		-0.265	-0.659	-0.482	-0.609	-0.549
RRMSE_β		0.296	0.665	0.507	0.615	0.557

Table 4: Simulation study results for the four-variate rotated-Frank and Frank mixture copula, as defined in the text with $\tau = 0.25$: performance measures of the NPCoImp method compared to other imputation methods.

4 Real data applications

In this section, we apply the NPCoImp method to two empirical case studies, both related to the environmental impact of agriculture. Specifically, the first application concerns the use of plant protection products to control crop diseases, while the second focuses on air pollutants originating from the agricultural sector.

4.1 The first case study: plant protection products

Here we assess the performance of the NPCoImp method to impute missing values in data concerning the plant protection products distributed by

Performance measures	Num. of multiv. missing values	NPCoImp	HD	kNN	CoImp	PMM
		$n = 50$				
MAE	0.2n	0.495	0.664	0.485	0.560	0.548
UMBRAE		0.952	1.191	0.907	1	1.008
RB_β		-0.306	-0.439	-0.333	-0.377	-0.355
RRMSE_β		0.363	0.473	0.388	0.421	0.402
MAE	0.4n	0.519	0.662	0.494	0.556	0.554
UMBRAE		0.981	1.181	0.928	1	1.008
RB_β		-0.285	-0.549	-0.397	-0.427	-0.373
RRMSE_β		0.380	0.589	0.472	0.488	0.438
		$n = 250$				
MAE	0.2n	0.503	0.663	0.488	0.564	0.553
UMBRAE		0.948	1.147	0.896	1	0.981
RB_β		-0.251	-0.411	-0.326	-0.350	-0.328
RRMSE_β		0.265	0.418	0.336	0.357	0.335
MAE	0.4n	0.514	0.663	0.482	0.566	0.550
UMBRAE		0.952	1.143	0.881	1	0.968
RB_β		-0.273	-0.509	-0.331	-0.393	-0.334
RRMSE_β		0.300	0.516	0.358	0.404	0.347
		$n = 500$				
MAE	0.2n	0.503	0.664	0.478	0.569	0.553
UMBRAE		0.938	1.136	0.871	1	0.969
RB_β		-0.237	-0.407	-0.309	-0.353	-0.319
RRMSE_β		0.245	0.410	0.317	0.357	0.323
MAE	0.4n	0.512	0.666	0.473	0.569	0.550
UMBRAE		0.943	1.142	0.861	1	0.964
RB_β		-0.257	-0.504	-0.277	-0.395	-0.328
RRMSE_β		0.278	0.507	0.296	0.400	0.334

Table 5: Simulation study results for the four-variate rotated-Frank and Frank mixture copula, as defined in the text with $\tau = 0.50$: performance measures of the NPCoImp method compared to other imputation methods.

companies for agricultural use. In the last years, there is a lot of discussion on plant protection products, which are mixtures of active substances widely used for the protection of plants, plant products or crops from harmful agents, and their use in agriculture to control pests and diseases and, in turn, reduce crop losses. However, the use of these compounds, commonly known as "pesticides", has a strong impact not only on humans for which represent a potential risk to his health, but also on the environment and its biota, as they are highly toxic and persistent. Despite being economically accessible, fast acting and easy to handle, their negative influence on humans and the environment has led to the search for effective alternatives that are more sustainable and respectful of nature. On 22 June 2022 the Commission published its Commu-

Performance measures	Num. of multiv. missing values	NPCoImp	HD	kNN	CoImp	PMM
		$n = 50$				
MAE	0.2n	0.451	0.662	0.405	0.512	0.434
UMBRAE		1.016	1.391	0.863	1	0.933
RB_β		-0.212	-0.343	-0.228	-0.251	-0.231
RRMSE_β		0.254	0.370	0.264	0.288	0.266
MAE	0.4n	0.466	0.661	0.428	0.520	0.439
UMBRAE		0.998	1.329	0.898	1	0.914
RB_β		-0.224	-0.467	-0.315	-0.301	-0.234
RRMSE_β		0.292	0.500	0.371	0.361	0.284
		$n = 250$				
MAE	0.2n	0.454	0.662	0.408	0.504	0.427
UMBRAE		1.013	1.348	0.870	1	0.896
RB_β		-0.180	-0.316	-0.234	-0.220	-0.204
RRMSE_β		0.191	0.321	0.242	0.227	0.210
MAE	0.4n	0.456	0.660	0.392	0.512	0.431
UMBRAE		0.990	1.324	0.830	1	0.894
RB_β		-0.225	-0.435	-0.255	-0.245	-0.216
RRMSE_β		0.249	0.440	0.276	0.256	0.224
		$n = 500$				
MAE	0.2n	0.460	0.667	0.393	0.509	0.428
UMBRAE		1.011	1.343	0.833	1	0.891
RB_β		-0.175	-0.311	-0.221	-0.216	-0.203
RRMSE_β		0.182	0.314	0.226	0.220	0.205
MAE	0.4n	0.459	0.665	0.372	0.511	0.430
UMBRAE		0.992	1.330	0.781	1	0.890
RB_β		0.230	-0.428	-0.208	-0.233	-0.210
RRMSE_β		0.242	0.431	0.226	0.239	0.214

Table 6: Simulation study results for the four-variate rotated-Frank and Frank mixture copula, as defined in the text with $\tau = 0.75$: performance measures of the NPCoImp method compared to other imputation methods.

$\boldsymbol{\Psi}$	MAE	UMBRAE	RB_β	RRMSE_β
$(0.050, 0.060, \ldots, 0.440, 0.450)$	0.3739	0.9665	-0.4303	0.4395
$(0.050, 0.051, \ldots, 0.449, 0.450)$	0.3735	0.9645	-0.4319	0.4412
$(0.010, 0.020, \ldots, 0.480, 0.490)$	0.3740	0.9565	-0.4669	0.4759
$(0.010, 0.011, \ldots, 0.489, 0.490)$	0.3737	0.9694	-0.4684	0.4773

Table 7: Simulation study results for a three-variate Clayton and rotated-Gumbel mixture copula, as defined in the text with $\tau = 0.50$, $n = 250$, and the number of multivariate missing data equals to $0.2n$: performance measures of NPCoImp by varying $\boldsymbol{\Psi}$ (see the text for details).

	NPCoImp	HD	kNN	CoImp	PMM
RPT	1.712	1.179	1.771	3.757	1.073
SE	0.027	0.057	0.046	1.368	0.057

Table 8: Monte Carlo estimate of the relative process time (RPT) for $H = 50$ samples of size $n = 500$ from a three-variate Clayton and rotated-Gumbel mixture copula with $\tau = 0.50$ and $0.4n$ multivariate missing values and its standard error (SE). The RPT is the CPU time required for the imputation of $0.4n$ multivariate missing values divided by CPU time when the multivariate missing values are $0.2n$.

nication 305/2022 proposing a new regulation under the Farm to Fork Strategy (Communication 381/2020) to take action to cut by 50% the overall use of plant protection products by 2030. In Italy, about 150,000 tons/year of plant protection products are used (for more details, see the Permanent Censuses Agriculture of the Italian National Statistical Institute, ISTAT hereafter; `https://www.istat.it/it/censimenti-permanenti/agricoltura`). Imputing with reliable values is particularly relevant for the assessment of the pesticides use trends and potential cases of excessive or extreme usage, and the evaluation of the impact of new regulations aimed at reducing the use of environmentally and human-harmful products.

Here we consider data provided by the ISTAT (`http://dati.istat.it/`) and concerning the 106 Italian provinces observed in 2021. Specifically, we analyze three variables that are the substances or active ingredients contained in plant protection products (in Kg): insecticides, i.e. pesticides against harmful insects and mites, herbicides, i.e. pesticides against weeds, and other plant protection products, i.e. pesticides with various active ingredients (including organic substances) versus other harmful organisms. The Kendall's correlation coefficient of each pair of the variables considered, which ranges from 0.578 to 0.675, and the scatter plots showed in Fig. (2) support the use of an imputation method that accounts for the complex dependence structure of the DGP. As in the simulation study, we set $\mathbf{\Psi} = (0.050, 0.051, \ldots, 0.449, 0.450)$, artificially introduce MCAR missing values, and vary the number of multivariate missing data in $(20\%, 40\%)$ of the sample size n. We then compare the performance of NPCoImp with the imputation methods considered in Section 3, using the performance measures defined therein. Note that in the case of 40% multivariate missing values, we excluded replications $h = 150, 395$ from the results because CoImp imputes using exactly the same observed value, leading to a 0/0 value for the UMBRAE.

Table 9 shows the imputation results. Regardless of the percentage of

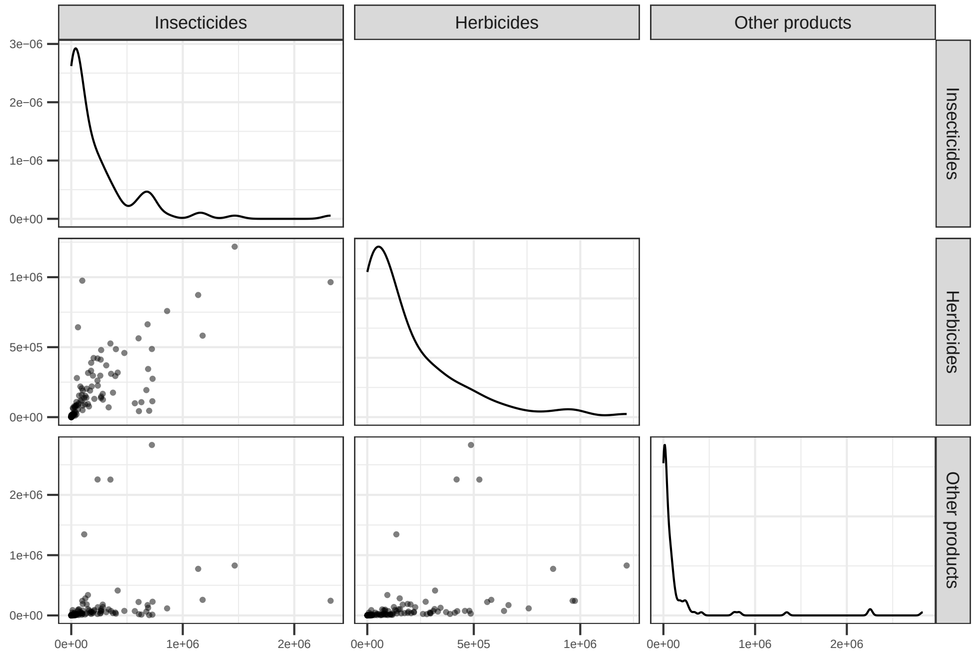

Figure 2: Scatter plots (lower triangular) and density functions (the diagonal) of variables related to plant protection products. See the text for details.

Num. of multiv. missing values	Performance Measure	NPCoImp	HD	kNN	CoImp	PMM
0.2×106	MAE	221829.9	391218.6	256815.9	328128.1	310289.4
	UMBRAE	0.727	1.192	0.832	1	0.920
	RB_β	-0.082	-0.223	-0.138	-0.212	-0.145
	$RRMSE_\beta$	0.092	0.229	0.147	0.219	0.155
0.4×106	MAE	228216.3	390333.7	231303.3	363665.2	315172.8
	UMBRAE	0.793	1.123	0.697	1	0.891
	RB_β	-0.199	-0.352	-0.103	-0.340	-0.199
	$RRMSE_\beta$	0.215	0.359	0.141	0.348	0.221

Table 9: Imputation performance on data concerning the plant protection products (see text for details): NPCoImp versus the imputation methods described in Section 3, evaluated based on the performance measures defined therein.

missing values introduced, the NPCoImp algorithm performs better than all the other methods considered in preserving microdata. In addition, it outperforms the CoImp in both the considered scenarios. Moreover, the NPCoImp appears to be the best method while preserving the dependence structure when the percentage of missing values is low. However, in the case of 40% multivariate missing values, the kNN method outperforms NPCoImp in terms of RB_β and RRMSE_β, coeherently with the simulation results.

4.2 The second case study: air quality

Here we present a case study related to the role of the agricultural sector on air quality. The Italian region Lombardia is one of the most polluted in Europe due to poor air circulation and high emission levels. Air pollutants may be categorised as primary or secondary. Primary pollutants are directly emitted to the atmosphere, whereas secondary pollutants are formed in the atmosphere from precursor gases through chemical reactions and microphysical processes. Ammonia in one of the key precursor gases for secondary particulate matter (PM hereafter). This is true for both large PM with aerodynamic diameter less than $10\mu m$ and fine PM with aerodynamic diameter less than $2.5\mu m$. In Europe, around 90% of ammonia emissions (source: European Environmental Agency) and in Lombardia up to 97% of ammonia emissions originate from the agricultural sector (source: Regional Environmental Protection Agency, ARPA hereafter). The ammonia reacts with nitric acid and the product of that reaction can contribute up to 60% of the PM with aerodynamic diameter less than $10\mu m$ mass concentration (source: ARPA Lombardia).

We here analyze five daily time series concerning the concentrations of particles, such as the dioxides of nitrogen, and the emissions of ammonia for agricultural soils and agricultural waste burning in the Italian region Lombardia and its neighbourhood. The data have been provided by the AgrImOnIA – Agriculture Impact On Italian Air project (see, `https://agrimonia.net/`) and the open-access spatio-temporal dataset is available at `https://zenodo.org/records/7593803`. In detail, ARPA Lombardia (`https://www.arpalombardia.it/`) provided hourly measurements of the following variables for Lombardia sensors network:

PM_{10}: the concentration of particles with an aerodynamic diameter of less than 10 micrometers (μm) expressed in $\mu g/m^3$;

$\mathrm{PM}_{2.5}$: the concentration of particles with an aerodynamic diameter of less than 2.5 micrometers (μm) expressed in $\mu g/m^3$;

NO_2: the concentration of dioxide of nitrogen expressed in $\mu g/m^3$.

Copernicus Atmosphere Monitoring Service (CAMS hereafter) global emission inventories (see `https://www.copernicus.eu`) provided the data for the emissions of ammonia data originating from agriculture sector, i.e. the following variables:

NH_3-Soils: the emissions of ammonia originating from the agriculture soils expressed in mg/m^2;

NH_3-Waste: the emissions of ammonia originating from the burning of agriculture waste expressed in mg/m^2.

The data have been detected by $S = 141$ ground-level monitoring stations, irregularly located over the considered land in 2020. The whole land includes 93 stations within the Lombardia region and 48 stations in the neighbouring area, obtained by applying a 0.3° buffer over the regional borders. This buffer encompasses from the following Italian regions: Piemonte, Veneto, Liguria, Emilia Romagna, Trentino-Alto Adige, and Ticino. Before applying the imputation methods to the five variables/series considered, we removed the serial dependence within each of the 5 hourly time series considered that concern the period from January 1st, 2020 to December 31st, 2020. Next, a suitable SARIMA$(p, d, q)(P, D, Q)_s$ model has been identified for each time series according to the Akaike Information Criterion. In addition, for the variable NO_2 a seasonal difference operator of order 1 has been applied to remove seasonality, as suggested by the inspection of the autocorrelation and the partial autocorrelation functions of the series. The identified models are an $ARIMA(1, 1, 3)$ for PM_{10}, an $ARIMA(4, 1, 1)$ for $PM_{2.5}$, a $SARIMA(4, 0, 4)(0, 1, 0)_7$ for NO_2, an $ARIMA(2, 2, 2)$ for NH_3-soils, and an $ARIMA(2, 2, 0)$ for NH_3-waste. The models' residuals are not autocorrelated (we do not reject the null hypothesis of the Student-t and the Ljung-Box tests), but from the scatter plots of the residual time series (Fig. 3) it can be seen that when the residuals are taken in pairs they exhibit a residual correlation in some cases. We introduce artificial MCAR multivariate missing values, vary the number of multivariate missing values in $(0.20n, 0.40n)$ where $n = 359$, and set $\mathbf{\Psi} = (0.050, 0.051, \ldots, 0.449, 0.450)$. Also in this case study some replications lead to a 0/0 value for the UMBRAE when considering CoImp and they have been excluded from the simulation results; specifically, the replication $h = 151$ in the case of 20% multivariate missing values, and replications $h = 133, 286, 340$ in the case of 40% multivariate missing values. Table 10 shows the performance of the imputation methods considered. Regardless of the percentage of missing values, the NPCoImp algorithm appears to be the best method to impute data while preserving the dependence structure of the DGP: RB_β, RRMSE_β, and the MAE reach the lowest values when the NPCoImp is used. Notably, NPCoImp outperforms even kNN, which, by contrast, seemed to perform better in simulations for microdata preservation. Moreover, the NPCoImp outperforms its main competitor, the CoImp, in all the investigated scenarios.

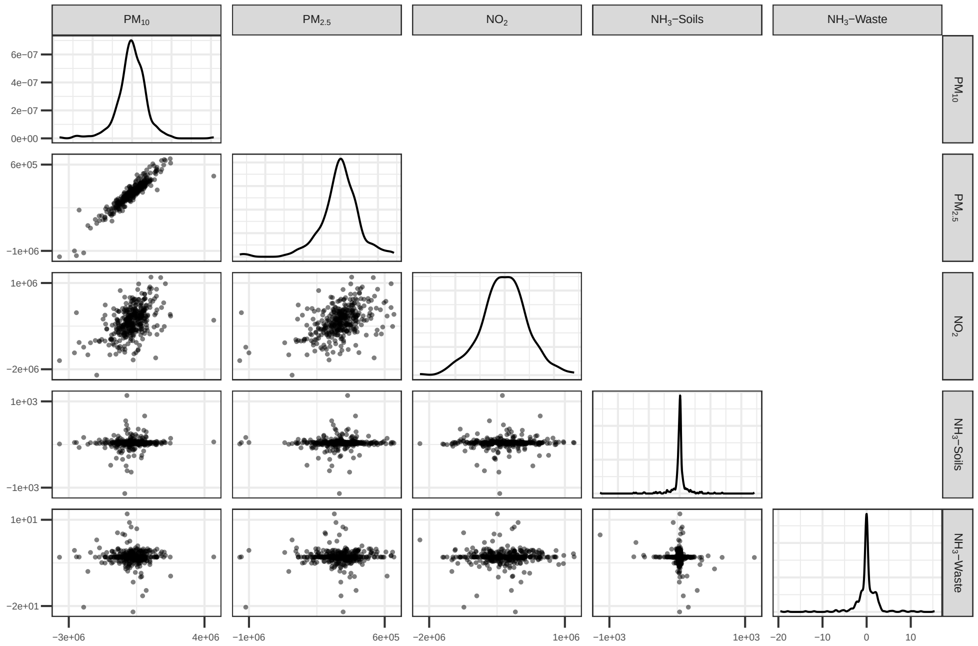

Figure 3: Scatter plots (lower triangular) and density functions (the diagonal) of the residual time series for pollutant concentration variables. See the text for details.

Num. of multiv. missing values	Performance Measure	NPCoImp	HD	kNN	CoImp	PMM
0.2×359	MAE	488307.7	788662.2	562284.3	721678.4	569571.9
	UMBRAE	0.777	1.068	0.815	1	0.876
	RB_β	-0.124	-0.429	-0.410	-0.471	-0.365
	RRMSE_β	0.158	0.436	0.419	0.476	0.372
0.4×359	MAE	504770.3	797846.5	557668.4	730865.1	573382.0
	UMBRAE	0.795	1.064	0.805	1	0.875
	RB_β	-0.056	-0.502	-0.434	-0.549	-0.367
	RRMSE_β	0.201	0.513	0.454	0.556	0.382

Table 10: Imputation performance on data concerning pollutant concentration variables (see text for details): NPCoImp versus the imputation methods described in Section 3, evaluated based on the performance measures defined therein.

5 Remarks and conclusions

In this paper, we propose a nonparametric imputation method based on the copula function, called NPCoImp. By leveraging the empirical copula, NPCoImp not only preserves the joint structure of the DGP but also maximizes flexibility in modelling the multivariate dependence structure. Therefore, NPCoImp is data-driven, has no risk of model misspecification, and it is able to impute missing data with any pattern and dimension. Additionally, the proposed imputation algorithm replaces each missing value with a single guess, identified based on a statistical criterion related to the radial symmetry of the conditional copula, in such a way as to mitigate the potential inaccuracy caused by the randomness of the imputation mechanism.

Using the conditional empirical copula to impute we overcome the two limits of the CoImp that have been noted in the literature: the slowness due to the computationally costly Hit-or-Miss Monte Carlo method and the limited families of copulas that can be used. Moreover, the Monte Carlo study carried out shows that the NPCoImp can outperform well-established methods, such as HD, kNN, and PMM, particularly in terms of preserving the dependence structure, which is our primary goal. The application of the proposed algorithm to two different case studies further confirms the effectiveness of our method compared to others, in terms of both preserving microdata and dependence structure.

Regarding future research, an aspect worth investigating is the performance of the proposed method for imputing missing in presence of outliers. This could, indeed, improve the realibility and robusteness of the NPCoImp method making more unbiased the subsequent analyses. In addition, since in some case studies the assumption of MCAR data can be unrealistic, an assessment of the NPCoImp's performance under missing at random (MAR) conditions would be useful. Furtheremore, it might be interesting to compare our approach with machine learning methods, such as random forests, that are capable of capturing complex structures in the data and predicting missing values, albeit at the cost of increased computational demands. Finally, it is worth commenting on the possible extension of the NPCoImp method within the widely used multiple imputation approach. The main advantage of multiple imputation is that it accounts for the randomness of imputed values, thereby improving inferential accuracy. However, our method would not make sense in the multiple imputation framework, as it does not introduce randomness in the imputed values. Assuming the use of the empirical beta copula and the proposed criterion for evaluating the asymmetry of its conditional distribution, the only factor affecting the imputed values is the choice of $\boldsymbol{\Psi}$, which can be chosen to ensure a certain level of imputation quality.

Acknowledgments

The authors have carried out this study within the Agritech National Research Center and received funding from the European Union - Next Generation EU (Piano Nazionale di Ripresa e Resilienza (PNRR) – Missione 4 Componente 2, Investimento 1.4 – D.D. 1032 17/06/2022, CN00000022). In particular our study represents an original paper related to the Spoke 4 “Multifunctional and resilient agriculture and forestry systems for the mitigation of climate change risks” and a baseline for the fulfilment of the milestones within the Task 4.2.3. “Big data analysis and decision support systems for the climate adaptation of agricultural and forestry systems”. This manuscript reflects only the authors’ views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Conflict of interest

The authors have no conflicts of interest to declare. All co-authors have seen and agree with the contents of the manuscript and there is no financial interest to report. We certify that the submission is original work and is not under review at any other publication.

References

Aissia, M.A.B., Chebana, F., Ouarda, T.B., 2017. Multivariate missing data in hydrology–review and applications. Advances in Water Resources 110, 299–309. doi:10.1016/j.advwatres.2017.10.002.

Arnold, B.C., Castillo, E., Sarabia, J.M., 1999. Conditional specification of statistical models. Springer.

Balusamy, B., Kadry, S., Gandomi, A.H., 2021. Big data: concepts, technology, and architecture. John Wiley & Sons. doi:10.1002/9781119701859.

Bedford, T., Cooke, R.M., 2002. Vines–a new graphical model for dependent random variables. The Annals of Statistics 30, 1031–1068. doi:10.1214/aos/1031689016.

Brechmann, E., Schepsmeier, U., 2013. Modeling dependence with c- and d-vine copulas: The r package cdvine. Journal of Statistical Software 52(3), 543–552. doi:10.18637/jss.v052.i03.

Chapon, A., Ouarda, T.B., Hamdi, Y., 2023. Imputation of missing values in environmental time series by d-vine copulas. Weather and Climate Extremes 41, 100591. doi:https://doi.org/10.1016/j.wace.2023.100591.

Chen, C., Twycross, J., Garibaldi, J.M., 2017. A new accuracy measure based on bounded relative error for time series forecasting. PloS one 12, e0174202. doi:10.1371/journal.pone.0174202.

Chen, J., Shao, J., 2000. Nearest neighbor imputation for survey data. Journal of Official Statistics 16, 113.

Chen, Z., Li, H., Bao, Y., 2019. Analyzing and modeling inter-sensor relationships for strain monitoring data and missing data imputation: a copula and functional data-analytic approach. Structural Health Monitoring 18, 1168–1188. doi:10.1177/1475921718788703.

Cont, R., Kan, Y.H.G., 2011. Statistical modeling of credit default swap portfolios. Available at SSRN 1771862 doi:10.2139/ssrn.1771862.

Costantini, E., Lang, K.M., Reeskens, T., Sijtsma, K., 2023. High-dimensional imputation for the social sciences: A comparison of state-of-the-art methods. Sociological Methods & Research , 00491241231200194doi:10.1177/00491241231200194.

Czado, C., 2019. Analyzing dependent data with vine copulas. Lecture Notes in Statistics, Springer 222. doi:https://doi.org/10.1007/978-3-030-13785-4.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society: series B (methodological) 39, 1–22. doi:10.1111/j.2517-6161.1977.tb01600.x.

Di Lascio, F.M.L., Giannerini, S., Reale, A., 2014. Imputation of complex dependent data by conditional copulas: analytic versus semiparametric approach, in: Book of proceedings of the 21st International Conference on Computational Statistics (COMPSTAT 2014), pp. 491–497.

Di Lascio, F.M.L., Giannerini, S., Reale, A., 2015. Exploring copulas for the imputation of complex dependent data. Statistical Methods & Applications 24, 159–175. doi:10.1007/s10260-014-0287-2.

Ding, W., Song, P.X.K., 2016. Em algorithm in gaussian copula with missing data. Computational Statistics & Data Analysis 101, 1–11. doi:10.1016/j.csda.2016.01.008.

Durante, F., Sempi, C., 2016. Principles of Copula Theory. CRC Press, Boca Raton, FL. doi:https://doi.org/10.1201/b18674.

Embrechts, P., Puccetti, G., 2006. Bounds for functions of multivariate risks. Journal of Multivariate Analysis 97, 526–547. doi:10.1016/j.jmva.2005.04.001.

Enders, C.K., 2022. Applied missing data analysis. Guilford Publications.

Fuller, W.A., Kim, J.K., 2005. Hot deck imputation for the response model. Survey Methodology 31, 139–149.

Gower, J.C., 1971. A general coefficient of similarity and some of its properties. Biometrics , 857–871doi:https://doi.org/10.2307/2528823.

Hammon, A., 2023. Multiple imputation of ordinal missing not at random data. AStA Advances in Statistical Analysis 107, 671–692. doi:10.1007/s10182-022-00461-9.

Hasler, C., Craiu, R.V., Rivest, L.P., 2018. Vine copulas for imputation of monotone non-response. International Statistical Review 86, 488–511. doi:10.1111/insr.12263.

Hasler, C., Tillé, Y., 2016. Balanced k-nearest neighbour imputation. Statistics 50, 1310–1331. doi:10.1080/02331888.2016.1230615.

Hüttner, A., Scherer, M., Gräler, B., 2020. Geostatistical modeling of dependent credit spreads: Estimation of large covariance matrices and imputation of missing data. Journal of Banking & Finance 118, 105897. doi:https://doi.org/10.1016/j.jbankfin.2020.105897.

Käärik, E., Käärik, M., 2009. Modeling dropouts by conditional distribution, a copula-based approach. Journal of Statistical Planning and Inference 139, 3830–3835. doi:10.1016/j.jspi.2009.05.020.

Kalton, G., Kasprzyk, D., 1982. Imputing for missing survey responses, in: Proceedings of the section on survey research methods, American Statistical Association, p. 31.

Kertel, M., Pauly, M., 2022. Estimating gaussian copulas with missing data with and without expert knowledge. Entropy 24, 1849. doi:10.3390/e24121849.

Kim, J.M., Lee, K.J., Kim, W., 2017. Variance estimation by multivariate imputation methods in complex survey designs. Model Assisted Statistics and Applications 12, 195–207. doi:10.3233/MAS-170394.

Kowarik, A., Templ, M., 2016. Imputation with the r package vim. Journal of statistical software 74, 1–16. doi:10.18637/jss.v074.i07.

Liebscher, E., 2024. Fitting copulas in the case of missing data. Statistical Papers 65, 3681–3711. doi:10.1007/s00362-024-01535-3.

Little, R.J.A., 1988. Missing-data adjustments in large surveys. Journal of Business & Economic Statistics 6, 287–296. doi:10.1080/07350015.1988.10509663.

Little, R.J.A., Rubin, D.B., 2019. Statistical analysis with missing data. volume 793. John Wiley & Sons.

Lun, Z., Khattree, R., 2022. A general approach for imputation of non-normal continuous data based on copula transformation. Communications in Statistics - Simulation and Computation 53(1), 567–594. doi:10.1080/03610918.2022.2025839.

Molenberghs, G., Kenward, M., 2007. Missing data in clinical studies. John Wiley & Sons. doi:10.1002/9780470510445.

Nelsen, R.B., 1993. Some concepts of bivariate symmetry. Journal of Nonparametric Statistics 3, 95–101.

Nelsen, R.B., 2002. Concordance and copulas: A survey. Distributions with given marginals and statistical modelling , 169–177doi:10.1007/978-94-017-0329-8_15.

Nelsen, R.B., 2006. An Introduction to Copulas. Springer Series in Statistics. second ed., Springer, New York. doi:10.1007/0-387-28678-0.

Rivero, C., Castillo, Á., Zufiria, P.J., Valdés, T., 2004. Global dynamics of a system governing an algorithm for regression with censored and non-censored data under general errors. Journal of Computational and Applied Mathematics 166, 535–551. doi:10.1016/j.cam.2003.09.048.

Robbins, M.W., Ghosh, S.K., Habiger, J.D., 2013. Imputation in high-dimensional economic data as applied to the agricultural resource management survey. Journal of the American Statistical Association 108, 81–95. doi:10.1080/01621459.2012.734158.

Rubin, D.B., 1986. Statistical matching using file concatenation with adjusted weights and multiple imputations. Journal of Business & Economic Statistics 4, 87–94. doi:10.1080/07350015.1986.10509497.

Rubin, D.B., 2004. Multiple imputation for nonresponse in surveys. volume 81. John Wiley & Sons. doi:10.1007/BF02924688.

Schafer, J.L., 1997. Analysis of incomplete multivariate data. Chapman & Hall, London. doi:10.1080/00401706.2000.10486013.

Segers, J., Sibuya, M., Tsukahara, H., 2017. The empirical beta copula. Journal of Multivariate Analysis 155, 35–51. doi:https://doi.org/10.1016/j.jmva.2016.11.010.

Shiau, J.T., Lien, Y.C., 2021. Copula-based infilling methods for daily suspended sediment loads. Water 13, 1701. doi:10.3390/w13121701.

Sklar, A., 1959. Fonctions de réparation à n dimensions et leurs marges. Publications de l'Institut de Statistique de l'Université de Paris 8, 229–231.

Tutz, G., Ramzan, S., 2015. Improved methods for the imputation of missing data by nearest neighbor methods. Computational Statistics & Data Analysis 90, 84–99. doi:https://doi.org/10.1016/j.csda.2015.04.009.

Úbeda-Flores, M., 2005. Multivariate versions of blomqvist's beta and spearman's footrule. Annals of the Institute of Statistical Mathematics 57, 781–788. doi:10.1007/BF02915438.

Van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M., Rubin, D.B., 2006. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation 76, 1049–1064. doi:10.1080/10629360600810434.

Van Buuren, S., Groothuis-Oudshoorn, K., 2011. mice: Multivariate imputation by chained equations in r. Journal of statistical software 45, 1–67. doi:10.18637/jss.v045.i03.

Wang, Y., Wan, W., Wang, R.S., Feng, E., 2009. Model, properties and imputation method of missing snp genotype data utilizing mutual information. Journal of Computational and Applied Mathematics 229, 168–174. doi:10.1016/j.cam.2008.10.020.