entropy

MDPI

Article

Estimating Gaussian Copulas with Missing Data with and without Expert Knowledge

Maximilian Kertel [1,2,*] and Markus Pauly [2,3]

[1] BMW Group, Battery Cell Competence Centre, 80788 Munich, Germany
[2] Department of Statistics, TU Dortmund University, 44227 Dortmund, Germany
[3] Research Center Trustworthy Data Science and Security, UA Ruhr, 44227 Dortmund, Germany
[*] Correspondence: maximilian.kertel@bmw.de

Abstract: In this work, we present a rigorous application of the Expectation Maximization algorithm to determine the marginal distributions and the dependence structure in a Gaussian copula model with missing data. We further show how to circumvent a priori assumptions on the marginals with semiparametric modeling. Further, we outline how expert knowledge on the marginals and the dependency structure can be included. A simulation study shows that the distribution learned through this algorithm is closer to the true distribution than that obtained with existing methods and that the incorporation of domain knowledge provides benefits.

Keywords: missing at random; expert knowledge; expectation maximization; semiparametric estimation

check for updates

Citation: Kertel, M.; Pauly, M. Estimating Gaussian Copulas with Missing Data with and without Expert Knowledge. *Entropy* **2022**, *24*, 1849. https://doi.org/10.3390/e24121849

Academic Editor: Boris Ryabko

Received: 27 October 2022
Accepted: 10 December 2022
Published: 19 December 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Even though the amount of data is increasing due to new technologies, big data are by no means good data. For example, missing values are ubiquitous in various fields, from the social sciences [1] to manufacturing [2]. For explanatory analysis or decision making, one is often interested in the joint distribution of a multivariate dataset, and its estimation is a central topic in statistics [3]. At the same time, there exists background knowledge in many domains that can help to compensate for the potential shortcomings of datasets. For instance, domain experts have an understanding of the causal relationships in the data generation process [4]. It is the scope of this paper to unify expert knowledge and datasets with missing data to derive approximations of the underlying joint distribution.

To estimate the multivariate distribution, we use copulas, where the dependence structure is assumed to belong to a parametric family, while the marginals are estimated nonparametrically. Genest et al. [5] showed that for complete datasets, a two-step approach consisting of the estimation of the marginals with an empirical cumulative distribution function (ecdf) and subsequent derivation of the dependence structure is consistent. This idea is even transferable to high dimensions [6].

In the case of missing values, the situation becomes more complex. Here, nonparametric methods do not scale well with the number of dimensions [7]. On the other hand, assuming that the distribution belongs to a parametric family, it can often be derived by using the EM algorithm [8]. However, this assumption is, in general, restrictive. Due to the encouraging results for complete datasets, there have been several works that have investigated the estimation of the joint distribution under a copula model. The authors of [9,10] even discussed the estimation in a missing-not-at-random (MNAR) setting. While MNAR is less restrictive than missing at random (MAR), it demands the explicit modeling of the missing mechanism [11]. On the contrary, the authors of [12,13] provided results in cases in which data were missing completely at random (MCAR). This strong assumption is rarely fulfilled in practice. Therefore, we assume an MAR mechanism in what follows [11].

Another interesting contribution [14] assumed external covariates, such that the probability of a missing value depended exclusively on them and not on the variables under investigation. They applied inverse probability weighting (IPW) and the two-step approach of [5]. While they proved a consistent result, it is unclear how this approach can be adapted to a setting without those covariates. IPW for general missing patterns is computationally demanding, and no software exists [15,16]. Thus, IPW is mostly applied with monotone missing patterns that appear, for example, in longitudinal studies [17]. The popular work of [18] proposed an EM algorithm in order to derive the joint distribution in a Gaussian copula model with data MAR [11]. However, their approach had weaknesses:

1. The presented algorithm was inexact. Among other things, the algorithm simplified by assuming that the marginals and the copula could be estimated separately (compare Equation (6) in [18] and Equation (11) in this paper).
2. If there was no a priori knowledge of the parametric family of all marginals, Ref. [18] proposed using the ecdf of the observed data points. Afterwards, they exclusively derived the parameters of the copula. This estimator of the marginals was biased [19,20], which is often overlooked in the copula literature, e.g., [21] (Section 4.3), [22] (Section 3), [23] (Section 3), or [24] (Section 3).
3. The description of the simulation study was incomplete and the results were not reproducible.

The aim of this paper is to close these gaps, and our contributions are the following:

1. We give a rigorous derivation of the EM algorithm under a Gaussian copula model. Similarly to [5], it consists of two separate steps, which estimate the marginals and the copula, respectively. However, these two steps alternate.
2. We show how prior knowledge about the marginals and the dependency structure can be utilized in order to achieve better results.
3. We propose a flexible parametrization of the marginals when a priori knowledge is absent. This allows us to learn the underlying marginal distributions; see Figure 1.
4. We provide a Python library that implements the proposed algorithm.

The structure of this paper is as follows. In Section 2, we review some background information about the Gaussian copula. We proceed by presenting the method (Section 3). In Section 4, we investigate its performance and the effect of domain knowledge in simulation studies. We conclude in Section 5. All technical aspects and proofs in this paper are given in Appendices A and B.

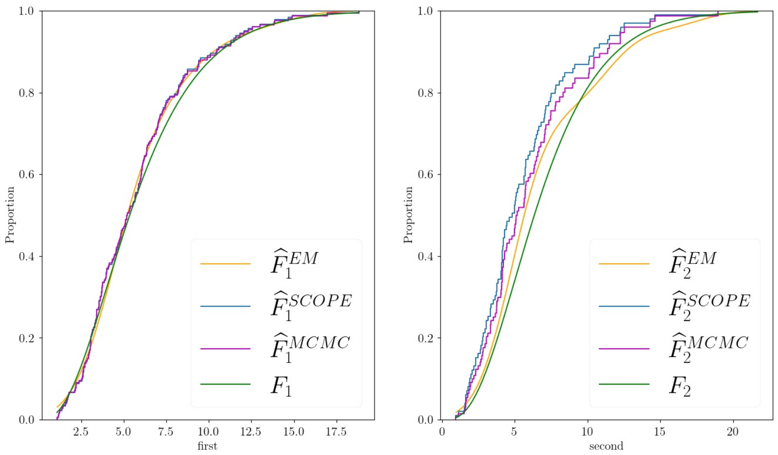

Figure 1. Estimates of the proposed EM algorithm ($\widehat{F}_i^{EM}$, orange line), the Standard Copula Estimator ($\widehat{F}_i^{SCOPE}$, blue line, corresponds to ecdf), the Markov chain–Monte Carlo approach ($\widehat{F}_i^{MCMC}$, purple line) for the marginals X_i, $i = 1, 2$, and the truth (F_i, green line) of a two-dimensional example dataset generated as described in Section 4.2 with $N = 200$, $\rho = 0.5$, and $\beta = (0, 2)$.

2. The Gaussian Copula Model

2.1. Notation and Assumptions

In the following, we consider a p-dimensional dataset $\{\mathbf{x}^1, \dots, \mathbf{x}^N\} \subset \mathbb{R}^p$ of size N, where $\mathbf{x}^1, \dots, \mathbf{x}^N$ are i.i.d. samples from a p-dimensional random vector $\mathbf{X} = (X_1, \dots, X_p)$ with a joint distribution function F and marginal distribution functions $F_1, \dots, F_p$. We denote the entries of $\mathbf{x}^\ell$ by $\mathbf{x}^\ell = \left(x_1^\ell, \dots, x_p^\ell\right) \forall \ell = 1, \dots, N$. The parameters of the marginals are represented by $\theta = (\theta_1, \dots, \theta_p)$, where θ_j is the parameter of F_j, so we write $F_j^{\theta_j}$, where θ_j can be a vector itself.

For $\ell \in \{1, \dots, p\}$, we define $\mathbf{obs}(\ell) \subset \{1, \dots, p\}$ as the index set of the observed and $\mathbf{mis}(\ell) \subset \{1, \dots, p\}$ as the index set of the missing columns of $\mathbf{x}^\ell$. Hence, $\mathbf{mis}(\ell) \cup \mathbf{obs}(\ell) = \{1, \dots, p\}$ and $\mathbf{mis}(\ell) \cap \mathbf{obs}(\ell) = \emptyset$. $\mathbf{R} = (R_1, \dots, R_p) \in \{0, 1\}^p$ is a random vector for which $R_i = 0$ if X_i is missing and $R_i = 1$ if X_i can be observed. Further, we define ϕ to be the density function and Φ to be the distribution function of the one-dimensional standard normal distribution. $\Phi_{\mu,\Sigma}$ stands for the distribution function of a p-variate normal distribution with covariance $\Sigma \in \mathbb{R}^{p \times p}$ and mean $\boldsymbol{\mu} \in \mathbb{R}^p$. To simplify the notation, we define $\Phi_\Sigma := \Phi_{0,\Sigma}$. For a matrix $A \in \mathbb{R}^{p \times p}$, the entry of the i-th row and the j-th column is denoted by A_{ij}, while for index sets $\mathbf{S}, \mathbf{T} \subset \{1, \dots, p\}$, $A_{\mathbf{S},\mathbf{T}}$ is the submatrix of A with the row number in $\mathbf{S}$ and column number in $\mathbf{T}$. For a (random) vector $\mathbf{x}$ ($\mathbf{X}$), $\mathbf{x}_\mathbf{S}$ ($\mathbf{X}_\mathbf{S}$) is the subvector containing entries with the index in $\mathbf{S}$.

Throughout, we assume F to be strictly increasing and continuous in every component. Therefore, F_j is strictly increasing and continuous for all $j \in \{1, \dots, p\}$, and so is the existing inverse function F_j^{-1}. For $\mathbf{S} = \{s_1, \dots, s_k\} \subset \{1, \dots, p\}$, we define $F_\mathbf{S} : \mathbf{R}^{|S|} \to \mathbf{R}^{|S|}$ by

$$F_\mathbf{S}(x_{s_1}, \dots, x_{s_k}) = \left(F_{s_1}(x_{s_1}), \dots, F_{s_k}(x_{s_k})\right).$$

This work assumes that data are Missing at Random (MAR), as defined by [11], i.e.,

$$\mathbb{P}_{\mathbf{X},\mathbf{R}}(\mathbf{R} = \mathbf{r} | \mathbf{X}_\mathbf{r} = \mathbf{x}_{-\mathbf{r}}, \mathbf{X}_\mathbf{r} = \mathbf{x}_\mathbf{r}) = \mathbb{P}_{\mathbf{X},\mathbf{R}}(\mathbf{R} = \mathbf{r} | \mathbf{X}_\mathbf{r} = \mathbf{x}_\mathbf{r}), \tag{1}$$

where $\mathbf{X}_\mathbf{r} := \mathbf{X}_{\{i:\, \mathbf{r}_i = 1\}}$ are the observed and $\mathbf{X}_{-\mathbf{r}} := \mathbf{X}_{\{i:\, \mathbf{r}_i = 0\}}$ are the missing entries of $\mathbf{X}$.

2.2. Properties

Sklar's theorem [25] decomposes F into its marginals $F_1, \dots, F_p$ and its dependency structure C with

$$F(x_1, \dots, x_p) = C\left(F_1(x_1), \dots, F_p(x_p)\right). \tag{2}$$

Here, C is a copula, which means it is a p-dimensional distribution function with support $[0, 1]^p$ whose marginal distributions are uniform. In this paper, we focus on Gaussian copulas, where

$$C_\Sigma(u_1, \dots, u_p) = \Phi_\Sigma\left(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_p)\right) \tag{3}$$

and Σ is a covariance matrix with $\Sigma_{jj} = 1 \; \forall j \in \{1, \dots, p\}$. Beyond all multivariate normal distributions, there are distributions with non-normal marginals whose copula is Gaussian. Hence, the Gaussian copula model provides an extension of the normality assumption. Consider a random vector $\mathbf{X}$ whose copula is C_Σ. Under the transformation

$$\mathbf{Z} := \Phi^{-1} \circ F(\mathbf{X}) := \left(\Phi^{-1} \circ F_1(X_1), \dots, \Phi^{-1} \circ F_p(X_p)\right),$$

it holds that

$$\begin{aligned}F_{\mathbf{Z}}(z_1\dots,z_p) &= \mathbb{P}\big(Z_1 \le z_1,\dots,Z_p \le z_p\big)\\ &= \mathbb{P}\Big(X_1 \le F_1^{-1}(\Phi(z_1)),\dots,X_p \le F_p^{-1}(\Phi(z_p))\Big)\\ &= \Phi_\Sigma\Big(\Phi^{-1}\Big(F_1\Big(F_1^{-1}(\Phi(z_1))\Big)\Big),\dots,\Phi^{-1}\Big(F_p\Big(F_p^{-1}(\Phi(z_p))\Big)\Big)\Big)\\ &= \Phi_\Sigma(z_1,\dots,z_p)\end{aligned} \tag{4}$$

and hence, $\mathbf{Z}$ is normally distributed with mean 0 and covariance Σ. The two-step approaches given in [5,6] use this property and apply the following scheme:

1. Find consistent estimates $\hat{F}_1,\dots,\hat{F}_p$ for the marginal distributions $F_1,\dots,F_p$.
2. Find Σ by estimating the covariance of the random vector

$$\mathbf{Z} = \Big(\Phi^{-1}\Big(\widehat{F}_1(X_1)\Big),\dots,\Phi^{-1}\Big(\widehat{F}_p(X_p)\Big)\Big).$$

From now on, we assume that the marginals of $\mathbf{X}$ have existing density functions $f_1,\dots,f_p$. Then, by using Equation (4) and a change of variables, we can derive the joint density function

$$f_{F_1,\dots,F_p,\Sigma}(x_1,\dots,x_p) = f(x_1,\dots,x_p) = |\Sigma|^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}\mathbf{z}^T\Big(\Sigma^{-1}-I\Big)\mathbf{z}\Big)\prod_{j=1}^{p} f_j(x_j), \tag{5}$$

where $\mathbf{z} := \big(\Phi^{-1}(F_1(x_1)),\dots,\Phi^{-1}\big(F_p(x_p)\big)\big)$. As for the multivariate normal distribution, we can identify the conditional independencies ([6]) from the inverse of the covariance matrix $K := \Sigma^{-1}$ by using the property

$$K_{jk} = K_{kj} = 0 \iff X_j \perp X_k|\{X_i : i \in \{1,\dots,p\}\setminus\{j,k\}\}. \tag{6}$$

K is called the precision matrix. In order to slim down the notation, we define

$$\Phi^{-1}(F_{\mathbf{S}}(\mathbf{x}_{\mathbf{S}})) := \Big(\Phi^{-1}(F_{s_1}(x_{s_1})),\dots,\Phi^{-1}\big(F_{s_k}(x_{s_k})\big)\Big)$$

and similarly

$$F_{\mathbf{S}}^{-1}(\Phi(\mathbf{z}_{\mathbf{S}})) := \Big(F_{s_1}^{-1}(\Phi(z_{s_1})),\dots,F_{s_1}^{-1}\big(\Phi(z_{s_k})\big)\Big).$$

The former function transforms the data of a Gaussian copula distribution to be normally distributed. The latter mapping takes multivariate normally distributed data and returns data following a Gaussian copula distribution with marginals $F_{s_1},\dots,F_{s_k}$. The conditional density functions have a closed form.

Proposition 1 (Conditional Distribution of Gaussian Copula). *Let* $\mathbf{S} = \{s_1,\dots,s_k\}$ *and* $\mathbf{T} = \{t_1,\dots,t_{k'}\}$ *be such that* $\mathbf{T}\dot{\cup}\mathbf{S} = \{1,\dots,p\}$.

1. *The conditional density of* $\mathbf{X}_{\mathbf{T}}|\mathbf{X}_{\mathbf{S}} = \mathbf{x}_{\mathbf{S}}$ *is given by*

$$f(\mathbf{x}_{\mathbf{T}}|\mathbf{X}_{\mathbf{S}} = \mathbf{x}_{\mathbf{S}}) = |\Sigma'|^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}(\mathbf{z}_{\mathbf{T}}-\boldsymbol{\mu})^T\Sigma'^{-1}(\mathbf{z}_{\mathbf{T}}-\boldsymbol{\mu})\Big)\exp\Big(\frac{1}{2}\mathbf{z}_{\mathbf{T}}^T\mathbf{z}_{\mathbf{T}}\Big)\prod_{j\in\mathbf{T}} f_j(x_j),$$

where $\boldsymbol{\mu} = \Sigma_{\mathbf{T},\mathbf{S}}\Sigma_{\mathbf{S},\mathbf{S}}^{-1}\mathbf{z}_{\mathbf{S}}$, $\Sigma' = \Sigma_{\mathbf{T},\mathbf{T}} - \Sigma_{\mathbf{T},\mathbf{S}}\Sigma_{\mathbf{S},\mathbf{S}}^{-1}\Sigma_{\mathbf{S},\mathbf{T}}$, $\mathbf{z}_{\mathbf{T}} = \Phi^{-1}(F_{\mathbf{T}}(\mathbf{x}_{\mathbf{T}}))$ *and* $\mathbf{z}_{\mathbf{S}} = \Phi^{-1}(F_{\mathbf{S}}(\mathbf{x}_{\mathbf{S}}))$.

2. $\Phi^{-1}(F_{\mathbf{T}}(\mathbf{X}_{\mathbf{T}}))|\mathbf{X}_{\mathbf{S}} = \mathbf{x}_{\mathbf{s}}$ *is normally distributed with mean* $\boldsymbol{\mu}$ *and covariance* Σ'.
3. *The expectation of* $h(\mathbf{X}_{\mathbf{T}})$ *with respect to the density* $f(\mathbf{x}_{\mathbf{T}}|\mathbf{X}_{\mathbf{S}} = \mathbf{x}_{\mathbf{S}})$ *can be expressed by*

$$\int h(\mathbf{x}_{\mathbf{T}})f(\mathbf{x}_{\mathbf{T}}|\mathbf{X}_{\mathbf{S}} = \mathbf{x}_{\mathbf{S}})d\mathbf{x}_{\mathbf{T}} = \int h\Big(F_{\mathbf{T}}^{-1}(\Phi(\mathbf{z}_{\mathbf{T}}))\Big)\phi_{\boldsymbol{\mu},\Sigma'}(\mathbf{z}_{\mathbf{T}})d\mathbf{z}_{\mathbf{T}}.$$

Proposition 1 shows that the conditional distribution's copula is Gaussian as well. More importantly, we can derive an algorithm for sampling from the conditional distribution.

Algorithm 1: Sampling from the conditional distribution of a Gaussian copula

```
Input: $\mathbf{x_S}, \Sigma, F_1, \ldots, F_p$
Result: $m$ samples of $\mathbf{X_T} | \mathbf{X_S} = \mathbf{x_S}$
Calculate $\mathbf{z_S} := \Phi^{-1}(F_\mathbf{S}(\mathbf{x_S}))$ ;
Calculate $\boldsymbol{\mu}$ and $\Sigma'$ as in Proposition 1 using $\mathbf{z_S}$ and $\Sigma$;
Draw samples $\{\mathbf{z}^1, \ldots, \mathbf{z}^m\}$ from $\mathcal{N}(\mu, \Sigma')$;
return $\{F_\mathbf{T}^{-1}(\Phi(\mathbf{z}^1)), \ldots, F_\mathbf{T}^{-1}(\Phi(\mathbf{z}^m))\}$
```

The very last step follows with Proposition 1, as it holds for any measurable $A \subset \mathbb{R}^{k'}$:

$$\mathbb{P}(\mathbf{X_T} \in A | \mathbf{X_S} = \mathbf{x_S}) = \int 1_A(\mathbf{x_T}) f(\mathbf{x_T} | \mathbf{X_S} = \mathbf{x_S}) d\mathbf{x_T} = \int 1_A\left(F_\mathbf{T}^{-1}(\Phi(\mathbf{z_T}))\right) \phi_{\mu,\Sigma'}(\mathbf{z_T}) d\mathbf{z_T}.$$

3. The EM Algorithm in the Gaussian Copula Model

3.1. The EM Algorithm

Let $\{\mathbf{y}^1, \ldots, \mathbf{y}^N\} \subset \mathbb{R}^p$ be a dataset following a distribution with parameter ψ and corresponding density function $g_\psi(\cdot)$, where observations are MAR. The EM algorithm [8] finds a local optimum of the log-likelihood function

$$\begin{aligned}\sum_{\ell=1}^{N} \ln\left(g_\psi\left(\mathbf{y}^\ell_{\mathbf{obs}(\ell)}\right)\right) &= \sum_{\ell=1}^{N} \int \ln\left(g_\psi\left(\left(\mathbf{y}^\ell_{\mathbf{obs}(\ell)}, \mathbf{y}_{\mathbf{mis}(\ell)}\right)\right)\right) \\ &\qquad g_\psi\left(\mathbf{y}_{\mathbf{mis}(\ell)} | \mathbf{Y}^\ell_{\mathbf{obs}(\ell)} = \mathbf{y}^\ell_{\mathbf{obs}(\ell)}\right) d\mathbf{y}_{\mathbf{mis}(\ell)} \\ &= \sum_{\ell=1}^{N} \mathbb{E}_\psi\left(\ln\left(g_\psi\left(\left(\mathbf{y}_{\mathbf{obs}(\ell)}, \mathbf{y}_{\mathbf{mis}(\ell)}\right)\right)\right) | \mathbf{Y}^\ell_{\mathbf{obs}} = \mathbf{y}^\ell_{\mathbf{obs}(\ell)}\right).\end{aligned}$$

After choosing a start value ψ^0, it does so by iterating the following two steps.

1. E-Step: Calculate

$$\begin{aligned}\lambda(\psi | \mathbf{y}^1, \ldots, \mathbf{y}^N, \psi^t) &:= \sum_{\ell=1}^{N} \mathbb{E}_{\psi^t}\left(\ln\left(g_\psi\left(\left(\mathbf{y}_{\mathbf{obs}(\ell)}, \mathbf{y}_{\mathbf{mis}(\ell)}\right)\right)\right) | \mathbf{Y}^\ell_{\mathbf{obs}} = \mathbf{y}^\ell_{\mathbf{obs}(\ell)}\right) \\ &= \sum_{\ell=1}^{N} \lambda(\psi | \mathbf{y}^\ell, \psi^t).\end{aligned} \tag{7}$$

2. M-Step: Set

$$\psi^{t+1} = \operatorname*{argmax}_\psi \lambda(\psi | \mathbf{y}^1, \ldots, \mathbf{y}^N, \psi^t) \tag{8}$$

and $t = t + 1$.

For our purposes, there are two extensions of interest:

- If there is no closed formula for the right-hand side of Equation (7), one can apply Monte Carlo integration [26] as an approximation. This is called the Monte Carlo EM algorithm.
- If $\psi = (\psi_1, \ldots, \psi_v)$ and the joint maximization of (8) with respect to ψ is not feasible, Ref. [27] proposed a sequential maximization. Thus, we optimize (8) with respect to ψ_i while holding $\psi_1 = \psi_1^{t+1}, \ldots, \psi_{i-1} = \psi_{i-1}^{t+1}, \psi_{i+1} = \psi_{i+1}^t, \ldots, \psi_v = \psi_v^t$ fixed before we continue with ψ_{i+1}. This is called the Expectation Conditional Maximization (ECM) algorithm.

3.2. Applying the ECM Algorithm on the Gaussian Copula Model

As we need a full parametrization of the Gaussian copula model for the EM algorithm, we assume parametric marginal distributions $F_1^{\theta_1}, \dots, F_p^{\theta_p}$ with densities $f_1^{\theta_1}, \dots, f_p^{\theta_p}$. According to Equation (5), the joint density with respect to the parameters $\theta = (\theta_1, \dots, \theta_p)$ and Σ has the form

$$f_{\theta,\Sigma}(x_1, \dots, x_p) = |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{z}_\theta^T\left(\Sigma^{-1} - I\right)\mathbf{z}_\theta\right)\prod_{j=1}^{p} f_j^{\theta_j}(x_j), \tag{9}$$

where $\mathbf{z}_\theta := \left(\Phi^{-1}\left(F_1^{\theta_1}(x_1)\right), \dots, \Phi^{-1}\left(F_p^{\theta_p}(x_p)\right)\right)$. Section 3.3 will describe how we can keep the flexibility for the marginals despite the parametrization. However, first, we outline the EM algorithm for general parametric marginal distributions.

3.2.1. E-Step

Set $K := \Sigma^{-1}$ and $K^t := \Sigma^{t^{-1}}$. For simplicity, we pick one summand in Equation (7). By Equation (7) and (9), it holds with $\psi = (\theta, \Sigma)$ and $\mathbf{x}^\ell$ taking the role of $\mathbf{y}^\ell$:

$$\begin{aligned}\lambda(\theta, \Sigma | \mathbf{x}^\ell, \theta^t, \Sigma^t) &= \mathbb{E}_{\theta^t,\Sigma^t}\left(\ln\left(f_{\theta,\Sigma}\left(\left(\mathbf{x}_{\mathbf{obs}(\ell)}, \mathbf{x}_{\mathbf{mis}(\ell)}\right)\right)\right)|\mathbf{X}_{\mathbf{obs}(\ell)} = \mathbf{x}^\ell_{\mathbf{obs}(\ell)}\right) \\ &= -\frac{1}{2}\ln(|\Sigma|) \\ &\quad -\frac{1}{2}\mathbb{E}_{\Sigma^t,\theta^t}\left(\mathbf{z}_\theta{}^T(K - I)\mathbf{z}_\theta|\mathbf{X}_{\mathbf{obs}(\ell)} = \mathbf{x}^\ell_{\mathbf{obs}(\ell)}\right) \\ &\quad + \sum_{j=1}^{p}\mathbb{E}_{\Sigma^t,\theta^t}\left(\ln\left(f_j^{\theta_j}(x_j)\right)|\mathbf{X}_{\mathbf{obs}(\ell)} = \mathbf{x}^\ell_{\mathbf{obs}(\ell)}\right).\end{aligned} \tag{10}$$

The first and last summand depend only on Σ and θ, respectively. Thus, of special interest is the second summand, for which we obtain the following with Proposition 1:

$$\mathbb{E}_{\Sigma^t,\theta^t}\left(\mathbf{z}_\theta^T(K - I)\mathbf{z}_\theta|\mathbf{X}_{\mathbf{obs}(\ell)} = \mathbf{x}^\ell_{\mathbf{obs}(\ell)}\right) = \int\left(\mathbf{z}_{\theta,\theta^t}^T(K - I)\mathbf{z}_{\theta,\theta^t}\right)\phi_{\boldsymbol{\mu},\Sigma^{t\prime}}\left(\mathbf{q}_{\mathbf{mis}(\ell)}\right)d\mathbf{q}_{\mathbf{mis}(\ell)}, \tag{11}$$

where

$$\mathbf{z}_{\theta,\theta^t} := \left(\Phi^{-1}\left(F_1^{\theta_1}\left(F_1^{\theta_1^t}{}^{-1}(\Phi(q_1))\right)\right), \dots, \Phi^{-1}\left(F_p^{\theta_p}\left(F_p^{\theta_p^t}{}^{-1}(\Phi(q_p))\right)\right)\right).$$

Here,

$$\boldsymbol{\mu} = \Sigma_{\mathbf{mis}(\ell),\mathbf{obs}(\ell)}\Sigma^{-1}_{\mathbf{obs}(\ell),\mathbf{obs}(\ell)}\Phi^{-1}\left(F^{\theta^t}_{\mathbf{obs}(\ell)}\left(x^\ell_{\mathbf{obs}(\ell)}\right)\right)$$

and

$$\Sigma^{t\prime} = \Sigma^t_{\mathbf{mis}(\ell),\mathbf{mis}(\ell)} - \Sigma^t_{\mathbf{mis}(\ell),\mathbf{obs}(\ell)}\left(\Sigma^t_{\mathbf{obs}(\ell),\mathbf{obs}(\ell)}\right)^{-1}\Sigma^t_{\mathbf{obs}(\ell),\mathbf{mis}(\ell)}.$$

At this point, the authors of [18] neglected that, in general,

$$F_j^{\theta_j^t} \neq F_j^{\theta_j}, j = 1, \dots, p$$

holds, and hence, (11) depends not only on Σ, but also on θ. This let us reconsider their approach, as we describe below.

3.2.2. M-Step

The joint optimization with respect to θ and Σ is difficult, as there is no closed form for Equation (10). We circumvent this problem by sequentially optimizing with respect to Σ and θ by applying the ECM algorithm. The maximization routine is the following.

1. Set $\Sigma^{t+1} = \operatorname{argmax}_\Sigma \sum_{l=1}^{N} \lambda(\theta^t, \Sigma | \mathbf{x}^\ell, \theta^t, \Sigma^t)$.

2. Set $\theta^{t+1} = \text{argmax}_\theta \sum_{l=1}^{N} \lambda(\theta, \Sigma^{t+1}|\mathbf{x}^\ell, \theta^t, \Sigma^t)$.

This is a two-step approach consisting of estimating the copula first and the marginals second. However, both steps are executed iteratively, which is typical for the EM algorithm.

Estimating Σ

As we are maximizing Equation (10) with respect to Σ with a fixed $\theta = \theta^t$, the last summand can be neglected. By a change-of-variables argument, we show the following in Theorem A1:

$$\mathbb{E}_{\Sigma^t,\theta^t}\left(\mathbf{z}_{\theta^t}{}^T(K-I)\mathbf{z}_{\theta^t}|\mathbf{X}_{\mathbf{obs}(\ell)} = \mathbf{x}^\ell_{\mathbf{obs}(\ell)}\right) = tr\left(\Sigma^{-1}V_\ell\right),$$

where V_ℓ depends on Σ^t and $\mathbf{z}_{\theta^t,\mathbf{obs}(\ell)} = \Phi^{-1}\left(F^{\theta^t}_{\mathbf{obs}(\ell)}\left(\mathbf{x}^\ell_{\mathbf{obs}(\ell)}\right)\right)$. Thus, considering all observations, we search for

$$\begin{aligned}\Sigma^{t+1} &= \underset{\Sigma,\Sigma_{\ell\ell}=1\forall\ell=1,\ldots,p}{\text{argmax}} \frac{1}{N}\sum_{l=1}^{N}\lambda(\theta^t,\Sigma|\mathbf{x}^\ell,\theta^t,\Sigma^t) \\ &= \underset{\Sigma,\Sigma_{\ell\ell}=1\forall\ell=1,\ldots,p}{\text{argmax}} \frac{1}{N}\sum_{\ell=1}^{N} -\frac{1}{2}\ln(|\Sigma|) - \frac{1}{2}tr\left(\Sigma^{-1}V_\ell\right) \\ &= \underset{\Sigma,\Sigma_{\ell\ell}=1\forall\ell=1,\ldots,p}{\text{argmax}} -\frac{1}{2}\ln(|\Sigma|) - \frac{1}{2}tr\left(\Sigma^{-1}\frac{1}{N}\sum_{\ell=1}^{N}V_\ell\right),\end{aligned} \tag{12}$$

which only depends on the statistic $S := \frac{1}{N}\sum_{\ell=1}^{N} V_\ell$. Generally, this maximization can be formalized as a convex optimization problem that can be solved by a gradient descent. However, the properties of this estimator are not understood (for example, a scaling of S by $a \in \mathbb{R}_{>0}$ leads to a different solution; see Appendix A.3). To overcome this issue, we instead approximate the solution with the correlation matrix

$$\underset{\Sigma,\Sigma_{\ell\ell}=1\forall\ell=1,\ldots,p}{\text{argmax}} -\frac{1}{2}\ln(|\Sigma|) - \frac{1}{2}tr\left(\Sigma^{-1}S\right) \approx PSP,$$

where $P \in \mathbb{R}^p$ is the diagonal matrix with entries $P_{jj} = \frac{1}{\sqrt{S_{jj}}}, \forall j = 1, \ldots, p$. This was also proposed in [28] (Section 2.2).

In cases in which there is expert knowledge on the dependency structure of the underlying distribution, one can adapt Equation (12) accordingly. We discuss this in more detail in Section 4.4.

Estimating θ

We now focus on finding θ^{t+1}, which is the maximizer of

$$\begin{aligned}\sum_{\ell=1}^{N}\lambda(\theta,\Sigma^{t+1}|\mathbf{x}^\ell,\theta^t,\Sigma^t) &= \sum_{\ell=1}^{N}\mathbb{E}_{\theta^t,\Sigma^t}\left(\ln\left(f_{\theta,\Sigma^{t+1}}\left(\mathbf{x}_{\mathbf{obs}(\ell)},\mathbf{x}_{\mathbf{mis}(\ell)}\right)\right)|\mathbf{X}_{\mathbf{obs}(\ell)} = \mathbf{x}^\ell_{\mathbf{obs}(\ell)}\right) \\ &= \sum_{\ell=1}^{N}\int \ln\left(f_{\theta,\Sigma^{t+1}}\left(\mathbf{x}^\ell_{\mathbf{obs}(\ell)},\mathbf{x}_{\mathbf{mis}(\ell)}\right)\right) \\ &\qquad f_{\theta^t,\Sigma^t}\left(\mathbf{x}_{\mathbf{mis}(\ell)}|\mathbf{X}_{\mathbf{obs}(\ell)} = \mathbf{x}^\ell_{\mathbf{obs}(\ell)}\right)d\mathbf{x}_{\mathbf{mis}(\ell)}\end{aligned}$$

with respect to θ. As there is, in general, no closed formula for the right-hand side, we use Monte Carlo integration. Again, we start by considering a single observation $\mathbf{x}^\ell$ to simplify terms. Employing Algorithm 1, we receive M samples $\mathbf{x}^\ell_{\mathbf{mis}(\ell),1}, \ldots, \mathbf{x}^\ell_{\mathbf{mis}(\ell),M}$

from the distribution of $X_{\mathbf{mis}(\ell)}|X_{\mathbf{obs}(\ell)} = x^{\ell}_{\mathbf{obs}(\ell)}$ given the parameters θ^t and Σ^t. We set $\mathbf{x}^{\ell}_{\mathbf{obs}(\ell),m} = \mathbf{x}^{\ell}_{\mathbf{obs}(\ell)}$ $\forall m = 1, \ldots, M$. Then, by Equation (9),

$$\begin{aligned}\lambda(\theta, \Sigma^{t+1}|\mathbf{x}^{\ell}, \theta^t, \Sigma^t) \approx C + \frac{1}{M}\sum_{m=1}^{M} -\frac{1}{2}&\left(\Phi^{-1}\left(F_1^{\theta_1}(x^{\ell}_{1,m})\right), \ldots, \Phi^{-1}\left(F_p^{\theta_p}(x^{\ell}_{p,m})\right)\right)^T \\ &\left(K^{t+1} - I\right) \\ &\left(\Phi^{-1}\left(F_1^{\theta_1}(x^{\ell}_{1,m})\right), \ldots, \Phi^{-1}\left(F_p^{\theta_p}(x^{\ell}_{p,m})\right)\right) \\ &+ \sum_{j=1}^{p} \ln\left(f_j^{\theta_j}(x^{\ell}_{j,m})\right).\end{aligned} \tag{13}$$

Hence, considering all observations, we set

$$\begin{aligned}\theta^{t+1} = \underset{\theta}{\operatorname{argmax}} \frac{1}{M}\sum_{\ell=1}^{N}\sum_{m=1}^{M} -\frac{1}{2}&\left(\Phi^{-1}\left(F_1^{\theta_1}(x^{\ell}_{1,m})\right), \ldots, \Phi^{-1}\left(F_p^{\theta_p}(x^{\ell}_{p,m})\right)\right)^T \\ &\left(K^{t+1} - I\right) \\ &\left(\Phi^{-1}\left(F_1^{\theta_1}(x^{\ell}_{1,m})\right), \ldots, \Phi^{-1}\left(F_p^{\theta_p}(x^{\ell}_{p,m})\right)\right) \\ &+ \sum_{j=1}^{p} \ln\left(f_j^{\theta_j}(x^{\ell}_{j,m})\right).\end{aligned} \tag{14}$$

Note that we only use the Monte Carlo samples to update the parameters of the marginal distributions θ. We would also like to point out some interesting aspects about Equations (13) and (14):

- The summand $\sum_{\ell=1}^{N}\sum_{m=1}^{M} \ln\left(f_j^{\theta_j}(x^{\ell}_{j,m})\right)$ describes how well the marginal distributions fit the (one-dimensional) data.
- The estimations of the marginals are interdependent. Hence, in order to maximize with respect to θ_j, we have to take into account all other components of θ.
- The first summand adjusts for the dependence structure in the data. If all observations at step $t+1$ are assumed to be independent, then $K^{t+1} = I$, and this term is 0.
- More generally, the derivative $\frac{\partial\lambda(\theta,\Sigma^{t+1}|\mathbf{x}^{\ell},\theta^t,\Sigma^t)}{\partial\theta_j}$ depends on θ_k if and only if $K^{t+1}_{jk} \neq 0$. This means that if Σ^{t+1} implies the conditional independence of column j and k given all other columns (Equation (6)), the optimal θ_j can be found without considering θ_k. This, e.g., is the case if we set entries of the precision matrix to 0. Thus, the incorporation of prior knowledge reduces the complexity of the identification of the marginal distributions.

The intuition behind the derived EM algorithm is simple. Given a dataset with missing values, we estimate the dependency structure. With the identified dependency structure, we can derive likely locations of the missing values. Again, these locations help us to find a better dependency structure. This leads to the proposed cyclic approach. The framework of the EM algorithm guarantees the convergence of this procedure to a local maximum for $M \to \infty$ in Equation (14).

3.3. Modelling with Semiparametric Marginals

In the case in which the missing mechanism is MAR, the estimation of the marginal distribution using only complete observations is biased. Even worse, any moment of the distribution can be distorted. Thus, one needs a priori knowledge in order to identify the parametric family of the marginals [19,20]. If their family is known, one can directly apply the algorithm of Section 3.2. If this is not the case, we propose the use of a mixture model parametrization of the form

$$F_j^{\theta_j}(x_j) = \frac{1}{g}\sum_{k=1}^{g} \Phi\left(\frac{x_j - \theta_{jk}}{\sigma_j}\right), \theta_{j1} \le \ldots \le \theta_{jg}, \forall j = 1, \ldots, p, \tag{15}$$

where σ_j is a hyperparameter and the ordering of the θ_{jk} ensures the identifiability.

Using mixture models for density estimation is a well-known idea (e.g., [29–31]). As the authors of [31] noted, mixture models vary between being parametric and being non-parametric, where flexibility increases with g. It is reasonable to choose Gaussian mixture models, as their density functions are dense in the set of all density functions with respect to the L^1-norm [29] (Section 3.2). This flexibility and the provided parametrization make the mixture models a natural choice.

3.4. A Blueprint of the Algorithm

The complete algorithm is summarized in Algorithm 2. For the Monte Carlo EM algorithm, Ref. [26] proposed the stabilization of the parameters with a rather small number of samples M and to increase this number substantially in the latter steps of the algorithm. This seems to be reasonable for line 8 of Algorithm 2 as well.

If there is no a priori knowledge about the marginals, we propose that we follow Section 3.3. We choose the initial θ^0 such that the cumulative distribution function of the mixture model fits the ecdf of the observed data points. For an empirical analysis of the role of g, see Section 4.3.3. For $\sigma_1, \ldots, \sigma_p$, we use a rule of thumb inspired by [3] and set

$$\sigma_j = 1.06\frac{\widehat{\sigma_j}}{g^{1/5}},$$

where $\widehat{\sigma_j}$ is the standard deviation of the observed data points in the j-th component.

Algorithm 2: Blueprint for the EM algorithm for the Gaussian copula model

```
   Input: $\{x^1, \ldots, x^N\}, \Sigma^0, \theta^0, n_{max}, \epsilon_{converged}, M$
   Result: $\Sigma, \theta$
 1 $n_{iter} \leftarrow 0$;
 2 $\epsilon \leftarrow \infty$;
 3 $\Sigma^t \leftarrow \Sigma^0$;
 4 $\theta^t \leftarrow \theta^0$;
 5 while $n_{iter} \le n_{max}$ and $\epsilon > \epsilon_{converged}$ do
 6     $\Sigma^{t+1} \leftarrow$ solution of Equation (12);
 7     for $\ell \in \{1, \ldots, N\}$ do
 8         Draw $M$ samples of $X|\mathbf{X}_{\mathbf{obs}(\ell)} = \mathbf{x}^{\ell}_{\mathbf{obs}(\ell)}$, under $(\theta^t, \Sigma^t)$;
 9     end
10     $\theta^{t+1} \leftarrow$ solution of Equation (14);
11     $\epsilon \leftarrow \|\Sigma^{t+1} - \Sigma^t\| + \|\theta^{t+1} - \theta^t\|$;
12     $\theta^t \leftarrow \theta^{t+1}$;
13     $\Sigma^t \leftarrow \Sigma^{t+1}$;
14     $n_{iter} \leftarrow n_{iter} + 1$;
15 end
16 return $\Sigma^t, \theta^t$
```

4. Simulation Study

We analyze the performance of the proposed estimator in two studies. First, we consider scenarios for two-dimensional datasets and check the potential of the algorithm.

In the second part, we explore how expert knowledge can be incorporated and how this affects the behavior and performance. The proposed procedure, which is indexed with EM in the figures below, is compared with:

1. **Standard COPula Estimator (SCOPE)**: The marginal distributions are estimated by the ecdf of the observed data points. This was proposed by [18] if the parametric family is unknown, and it is the state-of-the art approach. Thus, we apply an EM algorithm to determine the correlation structure on the mapped data points
$$z_j^\ell = \Phi^{-1}\left(\widehat{F}_j(x_j^\ell)\right), \ell = 1, \dots, N, j = 1, \dots, p,$$
where $\widehat{F}_j$ is the ecdf of the observed data points in column j. Its corresponding results are indexed with SCOPE in the figures and tables.
2. **Known marginals**: The distribution of the marginals is completely known. The idea is to eliminate the difficulty of finding them. Here, we apply the EM algorithm for the correlation structure on
$$z_j^\ell = \Phi^{-1}\left(F_j(x_j^\ell)\right), \ell = 1, \dots, N, j = 1, \dots, p,$$
where F_j is the real marginal distribution function. Its corresponding results are indexed with a 0 in the figures and tables.
3. **Markov chain–Monte Carlo (MCMC) approach [21]**: The author proposed an MCMC scheme to estimate the copula in a Bayesian fashion. Therefore, Ref. [21] derived the distribution of the multivariate ranks. The marginals are treated as nuisance parameters. We employed the R package `sbgcop`, which is available on `CRAN`, as it provides not only a posterior distribution of the correlation matrix Σ, but also imputations for missing values. In order to compare the approach with the likelihood-based methods, we set
$$\widehat{\Sigma}_{MCMC} = \frac{1}{M}\sum_{m=1}^{M}\Sigma^m,$$
where $\{\Sigma^m : m = 1, \dots, M\}$ are samples of the posterior distribution of the correlation matrix. For the marginals, we defined
$$\widehat{F}_{j,MCMC}(x) = \frac{1}{MN}\sum_{\ell=1}^{N}\sum_{m=1}^{M} 1_{\{x_{j,m}^\ell \le x\}},$$
where $x_{j,m}^\ell$ is the m-th of the total of M imputations for x_j^ℓ and $x_{j,m}^\ell = x_j^\ell \; \forall m = 1, \dots, M$ if x_j^ℓ can be observed. The samples were drawn from the posterior distribution. The corresponding results were indexed with the MCMC approach in the figures and tables.

Sklar's theorem shows that the joint distribution can be decomposed into the marginals and the copula. Thus, we analyze them separately.

4.1. Adapting the EM Algorithm

In Sections 4.3 and 4.4, we chose $g = 15$, for which we saw a sufficient flexibility. A sensitivity analysis of the procedure with respect to g can be found in Section 4.3.3. The initial θ^0 was chosen by fitting the marginals to the existing observations, and Σ^0 was the identity matrix. For the number of Monte Carlo samples M, we observed that with $M = 20$, θ stabilized after around 10 steps. Cautiously, we ran 20 steps before we increased M to 1000, for which we run another five steps. We stopped the algorithm when the condition $||\Sigma^{t+1} - \Sigma^t||_1 < 10^{-5}$ was fulfilled.

4.2. Data Generation

We considered a two-dimensional dataset (we would have liked to include the setup of the simulation study of [18]; however, neither could the missing mechanism be extracted from the paper nor did the authors provide it on request) with a priori unknown marginals F_1 and F_2, whose copula was Gaussian with the correlation parameter $\rho \in [-1,1]$. The marginals were chosen to be χ^2 with six and seven degrees of freedom. The data matrix $D \in \mathbb{R}^{N\times 2}$ kept N (complete) observations of the random vector. We enforced the following MAR mechanism:

1. Remove every entry in D with probability $0 \leq p_{MCAR} < 1$. We denote the resulting data matrix (with missing entries) as $D^{MCAR} = \left(D^{MCAR}_{\ell j}\right)_{\ell=1,\dots,N, j=1,2}$.
2. If $D^{MCAR}_{\ell 1}$ and $D^{MCAR}_{\ell 2}$ are observed, remove $D^{MCAR}_{\ell 2}$ with probability

$$\begin{aligned}\mathbb{P}(R_2 = 0|X_1 = D_{\ell 1}, X_2 = D_{\ell 2}) &= \mathbb{P}(R_2 = 0|X_1 = D_{\ell 1}) \\ &= \frac{1}{1+\exp(-\beta_0 - \beta_1 \Phi^{-1}(F_1(D_{\ell 1})))}\end{aligned}$$

We call the resulting data matrix D^{MAR}.

The missing patterns were non-monotone. Aside from p_{MCAR}, the parameters β_0 and β_1 controlled how many entries were absent in the final dataset. Assuming that $\rho > 0$, $\beta_1 > 0$, and $|\beta_0|$ was not too large, the ecdf of the observed values of X_2 was shifted to the left compared to the true distribution function (changing the signs of β_1 and/or ρ may change the direction of the shift, but the situation is analogous). This can be seen in Figure 1, where we chose $N = 200$, $\rho = 0.5$, $\beta = (\beta_0, \beta_1) = (0, 2)$. The marginal distribution of X_1 could be estimated well by the ecdf of the observed data.

4.3. Results

This subsection explores how different specifications of the data-generating process presented in Section 4.2 influenced the estimation of the joint distribution. First, we investigate the influence of the share of missing values (controlled via β) and the dependency (controlled via ρ) by fixing the number of observations (denoted by N) to 100. Then, we vary N to study the behavior of the algorithms for larger sample sizes. Afterwards, we carry out a sensitivity analysis of the EM algorithm with respect to g, the number of mixtures. Finally, we study the computational demands of the algorithms.

4.3.1. The Effects of Dependency and Share of Missing Values

We investigate two different choices for the setup in Section 4.2 by setting the parameters to $\rho = 0.1$, $\beta = (-1, 1)$ and $\rho = 0.5$, $\beta = (0, 2)$. For both, we draw 1000 datasets with $N = 100$ each and apply the estimators. To evaluate the methods, we look at two different aspects.

First, we compare the estimators for ρ with respect to bias and standard deviations. The results are depicted in the corresponding third columns of Table 1 and are summarized as boxplots in Figure A1 in Appendix B.3. We see that no method is clearly superior. While the EM algorithm has a stronger bias for $\rho = 0.5$ than that of SCOPE, it also has a smaller standard deviation. The MCMC approach shows the largest bias. As even known marginals (ρ_0) do not lead to substantially better estimators compared to SCOPE (ρ_{SCOPE}) or the proposed (ρ_{EM}) approach, we deduce that (at least in this setting) the estimators for the marginals are almost negligible. MCMC performs notably worse.

Second, we investigate the Cramer–von Mises statistics ω between the estimated and the true marginal distribution (ω^1 statistic for the first marginal, ω^2 statistic for the second marginal). The results are shown in Table 1 (corresponding first two columns) and are summarized as boxplots in Figure A2 in Appendix B.3. While for $\rho = 0.1$, the proposed estimator behaves only slightly better than SCOPE, we see that the benefit becomes larger in the case of high correlation and more missing values, especially when estimating the second marginal. This is in line with the intuition that if the correlation is vanishing, the two random variables X_1 and X_2 become independent. Thus, R_2, the missing value indicator, and X_2 become independent. (Note that there is a difference from the case in which $\rho \neq 0$, and hence, the missingness probability R_2 isconditionally independent from X_2 given X_1.) In that case, we can estimate the marginal of X_2 using the ecdf of the observed data points. Hence, SCOPE's estimates of the marginals should be good for small values of ρ. An illustration can be found in Figure 2. Again, the MCMC approach performs the worst.

Table 1. Comparison of the algorithms with respect to the Cramer–von Mises distance between the estimated and the true first (ω^1) and true second marginal distributions (ω^2), as well as the correlation (ρ). Shown are the mean and standard deviation of the proposed EM algorithm (EM), the method based on known marginals (0), the Standard Copula Estimator (SCOPE), and the Markov chain–Monte Carlo approach (MCMC) for 1000 datasets generated as described in Section 4.3.1.

		Mean			**Standard Deviation**		
Setting	**Method**	ω^1	ω^2	ρ	ω^1	ω^2	ρ
$\rho = 0.1, \beta = (-1, 1)$	EM	8.55	10.41	0.107	9.30	11.67	0.139
	0	-	-	0.109	-	-	0.144
	SCOPE	9.13	12.25	0.105	8.47	11.00	0.144
	MCMC	18.21	24.99	0.094	16.62	21.89	0.127
$\rho = 0.5, \beta = (0, 2)$	EM	8.03	16.48	0.455	8.68	19.47	0.139
	0	-	-	0.498	-	-	0.138
	SCOPE	9.06	45.25	0.486	8.25	36.11	0.143
	MCMC	17.90	59.34	0.393	16.13	57.15	0.131

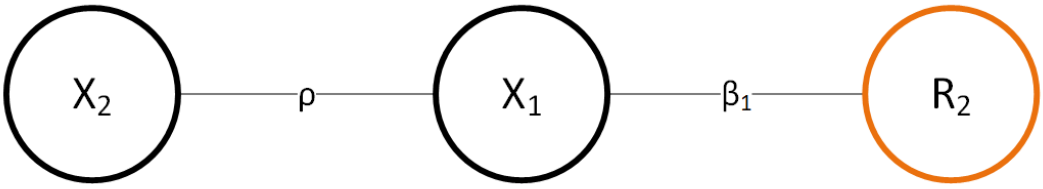

Figure 2. Dependency graph for X_1, X_2, and R_2. X_2 is independent of R_2 if either X_1 and X_2 are independent ($\rho = 0$) or if X_1 and R_2 are independent ($\beta_1 = 0$).

4.3.2. Varying the Sample Size N

To investigate the behavior of the methods for larger sample sizes, we repeat the experiment from Section 4.2 with $\rho = 0.5, \beta = (0, 2)$ for the sample sizes $N = 100, 200, 500, 1000$. The results are depicted in Table 2 and Figures A3–A5 in Appendix B.3. The bias of SCOPE and EM algorithm for ρ seem to vanish for large N, while the MCMC approach remains biased. Studying the estimation of the true marginals, the approximation of the second marginal via MCMC and SCOPE improves only slowly and is still poor for the largest sample sizes $N = 1000$. In contrast, the EM algorithm performs best in small sample sizes, and the mean (of ω^1 and ω^2) and standard deviations (of all three values) move towards 0 for increasing N.

Table 2. Comparison of the algorithms with respect to the Cramer–von Mises distance between the estimated and the true first (ω^1) and true second marginal distributions (ω^2), as well as the correlation (ρ). Shown are the mean and standard deviation of the proposed EM algorithm (EM), the method based on known marginals (0), the Standard Copula Estimator (SCOPE), and the Markov chain–Monte Carlo approach (MCMC) for 1000 datasets generated as described in Section 4.2 with $\rho = 0.5$ and $\beta = (0, 2)$ and varying sample sizes $N = 100, 200, 500, 1000$.

		Mean			**Standard Deviation**		
N	**Method**	ω^1	ω^2	ρ	ω^1	ω^2	ρ
$N = 100$	EM	8.03	16.48	0.455	8.68	19.47	0.139
	0	-	-	0.498	-	-	0.138
	SCOPE	9.06	45.25	0.486	8.25	36.11	0.143
	MCMC	17.90	59.34	0.393	16.13	57.15	0.131
$N = 200$	EM	4.91	8.53	0.469	5.46	8.88	0.098
	0	-	-	0.500	-	-	0.094
	SCOPE	4.76	37.38	0.493	4.18	25.35	0.096
	MCMC	9.27	42.91	0.370	8.01	36.23	0.089
$N = 500$	EM	3.01	3.83	0.480	2.92	3.59	0.063
	0	-	-	0.499	-	-	0.060
	SCOPE	2.05	31.92	0.497	1.85	14.95	0.060
	MCMC	4.01	31.41	0.0360	3.49	20.51	0.051
$N = 1000$	EM	2.25	2.74	0.486	1.92	2.40	0.047
	0	-	-	0.500	-	-	0.042
	SCOPE	1.08	30.60	0.499	0.93	11.13	0.043
	MCMC	1.99	28.13	0.365	1.84	14.49	0.037

4.3.3. The Impacts of Varying the Number of Mixtures g

The proposed EM algorithm relies on the hyperparameter g, the number of mixtures in Equation (15). To analyze the behavior of the EM algorithm with respect to g, we additionally run the EM algorithm with $g = 5$ and $g = 30$ on the 1000 datasets of Section 4.2 for $\rho = 0.5$, $\beta = (0, 2)$, and $N = 100$. We did not adjust the number of steps in the EM algorithm to keep the results comparable. The results can be found in Table 3. We see that the choice of g does not have a large effect on the estimation of ρ. However, an increased g leads to better estimates for X_1. This is in line with the intuition that the ecdf of the first components is an unbiased estimate for the distribution function of X_1, and setting g to the number of samples corresponds to the kernel density estimator. On the other hand, the estimator for X_2 benefits slightly from $g = 5$, as ω^2_{EM} has a lower mean and standard deviation compared to the choice $g = 30$. However, this effect is small and almost non-existent when we compare $g = 5$ with $g = 15$. As the choice $g = 15$ leads to better estimates of the first marginal compared to $g = 5$, we see this choice as a good compromise for our setting. For applications without prior knowledge, we recommend considering g as additional tuning parameter (via cross-validation).

Table 3. Comparison of the proposed EM algorithm with respect to the Cramer–von Mises distance between the estimated and the true first (ω^1) and true second marginal distributions (ω^2), as well as the correlation (ρ), for different numbers of mixtures g in Equation (15). Shown are the mean and standard deviation for $g = 5, 15, 30$ and for 1000 datasets generated as described in Section 4.2 with $\rho = 0.5$ and $\beta = (0, 2)$.

	Mean			**Standard Deviation**		
# Mixtures	ω^1	ω^2	ρ	ω^1	ω^2	ρ
$g = 5$	13.82	16.38	0.469	14.17	19.69	0.145
$g = 15$	8.03	16.48	0.455	8.68	19.47	0.139
$g = 30$	7.17	18.73	0.454	7.48	20.98	0.140

4.3.4. Run Time

We analyze the computational demands of the different algorithms by comparing their run times in the study of Section 4.3.1 with $\rho = 0.5$ and $\beta = (0, 2)$ (the settings $\rho = 0.1$ and $\beta = (-1, 1)$ lead to similar results and are omitted). The run times of all presented algorithms depend not only on the dataset, but also on the parameters (e.g., convergence criterion and Σ^0 for SCOPE). Thus, we do not aim for an extensive study, but focus on the magnitudes. We compare the proposed EM algorithm with a varying number of mixtures ($g = 5, 15, 30$) with MCMC and SCOPE. The results are shown in Table 4. We see that the EM algorithm has the longest run time, which depends on the number of mixtures g. The MCMC approach and the proposed EM algorithm have a higher computational demand than SCOPE, as they are trying to model the interaction between the copula and the marginals. As mentioned in the onset, we could reduce the run time of the EM algorithm by going down to only 10 steps instead of 20.

Table 4. Comparison of the algorithms with respect to the run time in seconds. Shown are the mean and standard deviation of the proposed EM algorithm (EM) with the number of mixtures g set to $5, 15, 30$, the Standard Copula Estimator (SCOPE), and the Markov chain–Monte Carlo approach (MCMC) for 1000 datasets generated as described in Section 4.2 with $\rho = 0.5$ and $\beta = (0, 2)$.

	Run Time in Seconds	
Method	**Mean**	**Standard Deviation**
EM ($g = 5$)	21.78	3.27
EM ($g = 15$)	55.94	11.39
EM ($g = 30$)	161.57	38.00
SCOPE	0.45	0.11
MCMC	12.98	0.87

4.4. Inclusion of Expert Knowledge

In the presence of prior knowledge on the dependency structure, the presented EM algorithm is highly flexible. While information on the marginals can be used to parametrize the copula model, expert knowledge on the dependency structure can be incorporated by adapting Equation (12). In the case of soft constraints on the covariance or precision matrix, one can replace Equation (12) with a penalized covariance estimation, where the penalty reflects the expert assessment [32,33]. Similarly, one can define a prior distribution on the covariance matrices and set Σ^{t+1} as the mode of the posterior distribution (the MAP estimate) of Σ given the statistic S of Equation (12).

Another possibility could be that we are aware of conditional independencies in the data-generating process. This is, for example, the case when causal relationships are known [4]. To exemplify the latter, we consider a three-dimensional dataset $\mathbf{X}$ with the

Gaussian copula C_Σ and marginals X_1, X_2, X_3, which are χ^2 distributed with six, seven, and five degrees of freedom. The precision is set to

$$K = \Sigma^{-1} = \Delta^{1/2} \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{pmatrix} \Delta^{1/2},$$

where $\Delta^{1/2}$ is a diagonal matrix, which ensures that the diagonal elements of Σ are 1. We see that X_2 and X_3 are conditionally independent given X_1. The missing mechanism is similar to the one in Section 4.2. The missingness of X_3 depends on X_1 and X_2, while the probability of a missing X_1 or X_2 is independent of the others. The mechanism is, again, MAR. Details can be found in Appendix B.2. We compare the proposed method with prior knowledge on the zeros in the precision matrix (indexed by KP, EM in the figures) with the EM, SCOPE, and MCMC algorithms without background knowledge. We again sample 1000 datasets with 50 observations each from the real distribution. The background knowledge on the precision is used by restricting the non-zero elements in Equation (12). Therefore, we apply the procedure presented in [34] (Chapter 17.3.1) to find Σ^{t+1}. The means and standard deviations of the estimates are presented in Table 5.

First, we evaluate the estimated dependency structures by calculating the Frobenius norm of the estimation error $\Sigma - \widehat{\Sigma}$. The EM algorithm with background knowledge (KP, EM) performs best and is more stable than its competitors. Apart from MCMC, the other procedures behave similarly, which indicates again that the exact knowledge of the marginal distributions is not too relevant for identifying the dependency structure. MCMC performs the worst.

Table 5. Comparison of the algorithms with respect to the Cramer–von Mises distance between the estimated and the true first marginal distribution (ω^1), true second marginal distribution (ω^2), and true third marginal distribution (ω^3), as well as the correlation (ρ). Shown are the mean and standard deviation of the proposed EM algorithm (EM), the proposed EM algorithm with prior knowledge on the conditional independencies (KP, EM), the method based on known marginals (0), the Standard Copula Estimator (SCOPE), and the Markov chain–Monte Carlo approach (MCMC) for 1000 datasets generated as described in Section 4.4.

	Mean				**Standard Deviation**			
Method	$\boldsymbol{\omega^1}$	$\boldsymbol{\omega^2}$	$\boldsymbol{\omega^3}$	$\boldsymbol{\|\|\widehat{\Sigma} - \Sigma\|\|_2}$	$\boldsymbol{\omega^1}$	$\boldsymbol{\omega^2}$	$\boldsymbol{\omega^3}$	$\boldsymbol{\|\|\widehat{\Sigma} - \Sigma\|\|_2}$
EM	12.12	13.38	21.15	0.229	13.89	14.25	22.44	0.113
KP, EM	12.04	13.28	19.66	0.182	13.93	14.37	20.88	0.111
0	-	-	-	0.227	-	-	-	0.108
SCOPE	17.57	17.55	26.69	0.232	16.75	15.55	24.84	0.113
MCMC	36.85	35.70	80.22	0.263	32.82	33.24	78.57	0.140

Second, we see that the proposed EM estimators return marginal distributions that are closer to the truth, while the estimate with background knowledge (KP, EM) performs the best. Thus, the background knowledge on the copula also transfers into better estimates for the marginal distribution—in particular, for X_3. This is due to Equation (14) and the comments thereafter. The zeros in the precision structure indicate which other marginals are relevant in order to identify the parameter of a marginal. In our case, X_2 provides no additional information for X_3. This information is provided to the EM algorithm through the restriction of the precision matrix.

Finally, we compare the EM estimates of the joint distribution. The relative entropy or Kullback–Leibler divergence is a popular tool for estimating the difference between two distributions [35,36], where one of them is absolutely continuous with respect to the other. A lower number indicates a higher similarity. Due to the discrete structure of the marginals of SCOPE and MCMC, we cannot calculate their relative entropy with respect to the truth.

However, we would like to analyze how the estimate of the proposed procedure improves if we include expert knowledge. The results are depicted in Table 6. Again, we observe that the incorporation of extra knowledge improves the estimates. This is in line with Table 5, as the estimation of all components in the joint distribution of Equation (3) is improved by the domain knowledge.

Table 6. Comparison of the algorithms with respect to the Kullback–Leibler divergence (D_{KL}) between the true joint distribution (F) and the estimates. Shown are the mean and standard deviation of the proposed EM algorithm (EM) and the proposed EM algorithm with prior knowledge on the conditional independencies (KP, EM) for 1000 datasets generated as described in Section 4.4.

	Mean($D_{KL}(F, \cdot)$)	**Standard Deviation($D_{KL}(F, \cdot)$)**
EM	1.37	0.53
KP, EM	1.26	0.32

5. Discussion

In this paper, we investigated the estimation of the Gaussian copula and the marginals with an incomplete dataset, for which we derived a rigorous EM algorithm. The procedure iteratively searches for the marginal distributions and the copula. It is, hence, similar to known methods for complete datasets. We saw that if the data are missing at random, a consistent estimate of a marginal distribution depends on the copula and other marginals.

The EM algorithm relies on a complete parametrization of the marginals. The parametric family of the marginals is, in general, a priori unknown and cannot be identified through the observed data points. For this case, we presented a novel idea of employing mixture models. Although this is practically always a misspecification, our simulation study revealed that the combination of our EM algorithm and marginal mixture models delivers better estimates for the joint distribution than currently used procedures do. In principle, uncertainty quantification of the parameters derived by the proposed EM algorithm can be achieved by bootstrapping [37].

There are different possibilities for incorporating expert knowledge. Information on the parametric family of the marginals can be used for their parametrization. However, causal and structural understandings of the data-generating process can also be utilized [4,38,39]. For example, this can be achieved by restricting the correlation matrix or its inverse, the precision matrix. We presented how one can restrict the non-zero elements of the precision, which enforces conditional independencies. Our simulation study showed that this leads not only to an improved estimate for the dependency structure, but also to better estimates for the marginals. This translates into a lower relative entropy between the real distribution and the estimate. We also discussed how soft constraints on the dependency structure can be included.

We note that the focus of this paper is on estimating the joint distribution without precise specification of its subsequent use. Therefore, we did not discuss imputation methods (see, e.g., [40–43]). However, Gaussian copula models were employed as a device for multiple imputation (MI) with some success [22,24,44]. The resulting complete datasets can be used for inference. All approaches that we are aware of estimate the marginals by using the ecdf of the observed data points. The findings in Section 4 translate into better draws for the missing values.

Additionally, the joint distribution can be utilized for regressing a potentially multivariate $\mathbf{Y}$ on $\mathbf{Z}$ even if data are missing. By applying the EM algorithm on $\mathbf{X} := (\mathbf{Y}, \mathbf{Z})$ and by Proposition 1, one even obtains the whole conditional distribution of $\mathbf{Y}$ given $\mathbf{Z} = \mathbf{z}$.

We have shown how to incorporate a causal understanding of the data-generating process. However, in the potential outcome framework of [45], the derivation of a causal relationship can also be interpreted as a missing data problem in which the missing patterns are "misaligned" [46]. Our algorithm is applicable for this.

Author Contributions: Conceptualization, M.K.; methodology, M.K.; software, M.K.; validation, M.P. and M.K.; formal analysis, M.K.; investigation, M.K.; resources, M.P. and M.K.; data curation, M.K.; writing—original draft preparation, M.K.; writing—review and editing, M.P. and M.K; visualization, M.K.; supervision, M.P.; project administration, M.P. All authors have read and agreed to the published version of the manuscript.

Funding: Maximilian Kertel's work is funded by the BMW Group.

Data Availability Statement: The data generation procedures of the simulation studies and the proposed algorithm are availabe at https://github.com/mkrtl/misscop, accessed on 26 October 2022.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Technical Results

Appendix A.1. Proof of Conditional Distribution

Proof of Proposition 1. We prove in the order of the proposition, which is a multivariate generalization of [47].

1. We inspect the conditional density function:

$$
\begin{aligned}
f(\mathbf{x_T}|\mathbf{X_S}=\mathbf{x_S}) &= \frac{|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\mathbf{z}^T\left(\Sigma^{-1}-I\right)\mathbf{z}\right)\prod_{j=1}^{p}f_j(x_j)}{|\Sigma_{\mathbf{S},\mathbf{S}}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\mathbf{z}_\mathbf{S}^T\left(\Sigma_{\mathbf{S},\mathbf{S}}^{-1}-I\right)\mathbf{z_S}\right)\prod_{j\in\mathbf{S}}f_j(x_j)} \\
&= \frac{|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\mathbf{z}^T\Sigma^{-1}\mathbf{z}\right)\exp(\frac{1}{2}\mathbf{z}^T\mathbf{z})\prod_{j=1}^{p}f_j(x_j)}{|\Sigma_{\mathbf{S},\mathbf{S}}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\mathbf{z}_\mathbf{S}^T\Sigma_{\mathbf{S},\mathbf{S}}^{-1}\mathbf{z_S}\right)\exp(\frac{1}{2}\mathbf{z}_\mathbf{S}^T\mathbf{z_S})\prod_{j\in\mathbf{S}}f_j(x_j)} \\
&= \frac{|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\mathbf{z}^T\Sigma^{-1}\mathbf{z}\right)\exp(\frac{1}{2}\mathbf{z}_\mathbf{T}^T\mathbf{z_T})\prod_{j\in\mathbf{T}}f_j(x_j)}{|\Sigma_{\mathbf{S},\mathbf{S}}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\mathbf{z}_\mathbf{S}^T\Sigma_{\mathbf{S},\mathbf{S}}^{-1}\mathbf{z_S}\right)}
\end{aligned}
$$

   Using well-known factorization lemmas and using the Schur complement (see, for example, [48] (Section 4.3.4)) applied on Σ^{-1}, we encounter

$$
f(\mathbf{x_T}|\mathbf{X_S}=\mathbf{x_S}) = |\Sigma'|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{z_T}-\boldsymbol{\mu})^T\Sigma'^{-1}(\mathbf{z_T}-\boldsymbol{\mu})\right)\exp\left(\frac{1}{2}\mathbf{z}_\mathbf{T}^T\mathbf{z_T}\right)\prod_{j\in\mathbf{T}}f_j(x_j). \quad \text{(A1)}
$$

2. The distribution of

$$
\Phi^{-1}(F_\mathbf{T}(\mathbf{X_T}))|\mathbf{X_S}=\mathbf{x_S}
$$

   follows with a change-of-variable argument. Using Equation (A1), we observe for any measurable set A that

$$
\begin{aligned}
&\mathbb{P}\left(\left(\Phi^{-1}(F_\mathbf{T}(\mathbf{X_T}))|\mathbf{X_S}=\mathbf{x_s}\right)\in A\right) \\
&= \int_{F^{-1}(\Phi(A))}|\Sigma'|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{z_T}-\boldsymbol{\mu})^T\Sigma'^{-1}(\mathbf{z_T}-\boldsymbol{\mu})\right)\exp\left(\frac{1}{2}\mathbf{z}_\mathbf{T}^T\mathbf{z_T}\right)\prod_{j\in\mathbf{T}}f_j(x_j)d\mathbf{x_T} \\
&= \int_A \phi_{\boldsymbol{\mu},\Sigma'}(\mathbf{q_T})d\mathbf{q_T},
\end{aligned}
$$

   where, in the second equation, we used the transformation $\mathbf{q_T}=\Phi^{-1}(F_\mathbf{T}(\mathbf{x_T}))$ and the fact that

$$
\left|D\left(\phi^{-1}(F_\mathbf{T}(\mathbf{x_T}))\right)\right| = 2\pi^{\frac{|\mathbf{T}|}{2}}\exp\left(\frac{1}{2}\left(\Phi^{-1}(F_\mathbf{T}(\mathbf{x_T}))\right)^T\left(\Phi^{-1}(F_\mathbf{T}(\mathbf{x_T}))\right)\right)\prod_{j\in\mathbf{T}}f_j(x_j).
$$

3. This proof is analogous to the one above, and we finally obtain

$$\int h(\mathbf{x_T}) f(\mathbf{x_T}|\mathbf{X_S} = \mathbf{x_S}) d\mathbf{x_T} = \int h\Big(F^{-1}(\Phi(\mathbf{z_T}))\Big)\phi_{\mu,\Sigma'}(\mathbf{z_T}) d\mathbf{z_T}.$$

The result can be generalized to the case in which $\mathbf{S} \cup \mathbf{T} \neq \{1, \dots, p\}$. □

Appendix A.2. Closed-Form Solution of the E-Step for $\theta = \theta^t$

Theorem A1. *We assume w.l.o.g. that* $\mathbf{x}^\ell = (\mathbf{x}^\ell_{\mathbf{obs}(\ell)}, \mathbf{x}^\ell_{\mathbf{mis}(\ell)})$ *and set*

$$\mathbf{z}_{\theta^t} := \Big(\mathbf{z}_{\mathbf{obs}(\ell),\theta^t}, \mathbf{z}_{\mathbf{mis}(\ell)}\Big) := \Big(\Phi^{-1}\Big(F^{\theta^t}_{\mathbf{obs}(\ell)}(\mathbf{x}^\ell_{\mathbf{obs}(\ell)})\Big), \mathbf{z}_{\mathbf{mis}(\ell)}\Big).$$

Then, it holds that

$$\mathbb{E}_{\Sigma^t,\theta^t}\Big({\mathbf{z}_{\theta^t}}^T \Sigma^{-1} \mathbf{z}_{\theta^t} | \mathbf{X}_{\mathbf{obs}(\ell)} = \mathbf{x}^\ell_{\mathbf{obs}(\ell)}\Big) = tr\Big(\Sigma^{-1} V\Big),$$

where $V = \begin{pmatrix} \mathbf{z}_{\mathbf{obs}(\ell),\theta^t} {\mathbf{z}_{\mathbf{obs}(\ell),\theta^t}}^T & \mathbf{z}_{\mathbf{obs}(\ell),\theta^t} \boldsymbol{\mu}^T \\ \boldsymbol{\mu} {\mathbf{z}_{\mathbf{obs}(\ell),\theta^t}}^T & \Sigma' + \boldsymbol{\mu}\boldsymbol{\mu}^T \end{pmatrix}$, $\boldsymbol{\mu} = \Sigma^t_{\mathbf{mis}(\ell),\mathbf{obs}(\ell)} {\Sigma^t_{\mathbf{obs}(\ell),\mathbf{obs}(\ell)}}^{-1} \mathbf{z}_{\mathbf{obs}(\ell),\theta^t}$ *and* $\Sigma' = \Sigma^t_{\mathbf{mis}(\ell),\mathbf{mis}(\ell)} - \Sigma^t_{\mathbf{mis}(\ell),\mathbf{obs}(\ell)} {\Sigma^t_{\mathbf{obs}(\ell),\mathbf{obs}(\ell)}}^{-1} \Sigma^t_{\mathbf{obs}(\ell),\mathbf{mis}(\ell)}$.

Proof. We define $F_{\theta^t}(\mathbf{x}_{\mathbf{mis}(\ell)}) := F_{\theta^t}(\mathbf{x}^\ell_{\mathbf{obs}(\ell)}, \mathbf{x}_{\mathbf{mis}(\ell)})$. Then,

$$\begin{aligned}
&\mathbb{E}_{\Sigma^t,\theta^t}\Big(\mathbf{z}^T_{\theta^t} \Sigma^{-1} \mathbf{z}_{\theta^t} | \mathbf{X}_{\mathbf{obs}(\ell)} = \mathbf{x}^\ell_{\mathbf{obs}(\ell)}\Big) \\
&= \mathbb{E}_{\Sigma^t,\theta^t}\Big(\Big(\Phi^{-1}\Big(F_{\theta^t}(\mathbf{x}_{\mathbf{mis}(\ell)})\Big)\Big)^T \Sigma^{-1} \Big(\Phi^{-1}\Big(F_{\theta^t}(\mathbf{x}_{\mathbf{mis}(\ell)})\Big)\Big) | \mathbf{X}_{\mathbf{obs}(\ell)} = \mathbf{x}^\ell_{\mathbf{obs}(\ell)}\Big) \\
&= \int \Big(\Phi^{-1}\Big(F_{\theta^t}(\mathbf{x}_{\mathbf{mis}(\ell)})\Big)\Big)^T \Sigma^{-1} \Big(\Phi^{-1}\Big(F_{\theta^t}(\mathbf{x}_{\mathbf{mis}(\ell)})\Big)\Big) \\
&\qquad f_{\theta^t,\Sigma^t}\Big(\mathbf{x}_{\mathbf{mis}(\ell)} | \mathbf{X}_{\mathbf{obs}(\ell)} = \mathbf{x}^\ell_{\mathbf{obs}(\ell)}\Big) d\mathbf{x}_{\mathbf{mis}(\ell)}.
\end{aligned}$$

We now apply Proposition 1 and encounter

$$\begin{aligned}
&\int \Big(\Phi^{-1}\Big(F_{\theta^t}(\mathbf{x}_{\mathbf{mis}(\ell)})\Big)\Big)^T \Sigma^{-1} \Phi^{-1}\Big(F_{\theta^t}(\mathbf{x}_{\mathbf{mis}(\ell)})\Big) \\
&\qquad f_{\theta^t,\Sigma^t}\Big(\mathbf{x}_{\mathbf{mis}(\ell)} | \mathbf{X}_{\mathbf{obs}(\ell)} = \mathbf{x}^\ell_{\mathbf{obs}(\ell)}\Big) d\mathbf{x}_{\mathbf{mis}(\ell)} \\
&= \int \mathbf{z}^T_{\theta^t} \Sigma^{-1} \mathbf{z}_{\theta^t} \phi_{\Sigma',\boldsymbol{\mu}}(\mathbf{z}_{\mathbf{mis}(\ell)}) d\mathbf{z}_{\mathbf{mis}(\ell)} \\
&= \int tr(\mathbf{z}_{\theta^t} \mathbf{z}^T_{\theta^t} \Sigma^{-1}) \phi_{\Sigma',\boldsymbol{\mu}}(\mathbf{z}_{\mathbf{mis}(\ell)}) d\mathbf{z}_{\mathbf{mis}(\ell)} \\
&= tr\Big(\Sigma^{-1} \int \mathbf{z}_{\theta^t} \mathbf{z}^T_{\theta^t} \phi_{\Sigma',\boldsymbol{\mu}}(\mathbf{z}_{\mathbf{mis}(\ell)}) d\mathbf{z}_{\mathbf{mis}(\ell)}\Big).
\end{aligned}$$

The last integral is understood element-wise. By the first and second moment of $\Phi_{\Sigma',\boldsymbol{\mu}}$, it follows that

$$\begin{aligned}
\int \mathbf{z}_{\theta^t} \mathbf{z}^T_{\theta^t} \phi_{\Sigma',\boldsymbol{\mu}}(\mathbf{z}_{\mathbf{mis}(\ell),\theta^t}) d\mathbf{z}_{\mathbf{mis}(\ell),\theta^t} &= \int \Big(\mathbf{z}_{\mathbf{obs}(\ell),\theta^t}, \mathbf{z}_{\mathbf{mis}(\ell),\theta^t}\Big)\Big(\mathbf{z}_{\mathbf{obs}(\ell),\theta^t}, \mathbf{z}_{\mathbf{mis}(\ell),\theta^t}\Big)^T \\
&\qquad \phi_{\Sigma',\boldsymbol{\mu}}(\mathbf{z}_{\mathbf{mis}(\ell),\theta^t}) d\mathbf{z}_{\mathbf{mis}(\ell),\theta^t} \\
&= \begin{pmatrix} \mathbf{z}_{\mathbf{obs}(\ell),\theta^t} \mathbf{z}^T_{\mathbf{obs}(\ell),\theta^t} & \mathbf{z}_{\mathbf{obs}(\ell),\theta^t} \boldsymbol{\mu}^T \\ \boldsymbol{\mu} \mathbf{z}^T_{\mathbf{obs}(\ell),\theta^t} & \Sigma' + \boldsymbol{\mu}\boldsymbol{\mu}^T \end{pmatrix}.
\end{aligned}$$

□

Appendix A.3. Maximizer of $\operatorname{argmax}_{\Sigma,\Sigma_{jj}=1\forall j=1,\dots,p}\lambda(\theta^t,\Sigma|\theta^t,\Sigma^t)$

We are interested in

$$\underset{\Sigma_{jj}=1\forall j=1,\dots,p}{\operatorname{argmax}} \; l(\Sigma) := \underset{\Sigma_{jj}=1\forall j=1,\dots,p}{\operatorname{argmax}} \; -\log(|\Sigma|) - tr\left(\Sigma^{-1}S\right),$$

where $\Sigma, S \in \mathbb{R}^{p\times p}$ are positive definite matrices. Clearly,

$$\Sigma_{jj} = 1 \iff 1 = e_j^T\Sigma e_j = tr\left(e_j^T\Sigma e_j\right) = tr\left(e_j e_j^T\Sigma\right).$$

Using the Lagrangian, we obtain the following function to optimize

$$L(\Sigma,\lambda) = -\log(|\Sigma|) - tr\left(\Sigma^{-1}S\right) + \sum_{j=1}^{p}\lambda_j\left(tr\left(e_j e_j^T\Sigma\right) - 1\right).$$

Applying the identities $\frac{\partial tr(AX)}{\partial X} = A$, $\frac{\partial tr(AX^{-1})}{\partial X} = -X^{-1}AX^{-1}$, and $\frac{\partial \log(|X|)}{\partial X} = X^{-1}$, we obtain the derivative with respect to Σ:

$$\frac{\partial L}{\partial \Sigma} = -\Sigma^{-1} + \Sigma^{-1}S\Sigma^{-1} - \left(\sum_{j=1}^{p}\lambda_j\left(e_j e_j^T\right)\right) \overset{!}{=} 0.$$

This is equivalent to

$$-K + KSK = D_\lambda,$$

where D_λ is the diagonal matrix with entries $\lambda = (\lambda_1,\dots,\lambda_p)$ and $K := \Sigma^{-1}$. We see that the scaling of S by $a \in \mathbb{R}_{>0}$ leads, in general, to a different solution K, and hence, the estimator is not invariant under strictly monotone linear transformations of S.
We can also formulate the task as a convex optimization problem:

$$\underset{\left(K^{-1}\right)_{ii}=1\ \forall i=1,\dots,p}{\operatorname{argmin}} \; -\log(|K|) + tr(KS).$$

Appendix B. Details of the Simulation Studies

Appendix B.1. Drawing Samples from the Joint Distributions

Appendix B.1.1. Estimators of the Percentile Function

- In the case of SCOPE, consider the marginal observed data points, which we assume to be ordered $y_1 \le \dots \le y_N$. We use the following linearly interpolated estimator for the percentile function:

$$\widehat{F^{-1}}(u) = \left\{ \begin{array}{ll} y_1 & \text{for } u \le \frac{1}{N+1} \\ y_N, & \text{for } u > \frac{N}{N+1} \\ \frac{u-\frac{i}{N+1}}{\frac{i+1}{N+1}-\frac{i}{N+1}}(y_{i+1}-y_i) + y_i, & \text{for } u \in \left(\frac{i}{N+1}, \frac{i+1}{N+1}\right] \end{array} \right\}$$

- To estimate the percentile function for the mixture models, we choose with equal probability (all Gaussians have equal weight) one component of the mixture and then draw a random number with its mean θ_{jk} and standard deviation σ_j, $j = 1,\dots,p$, $k = 1,\dots,g$. In this manner, we generate N' samples $y'_1,\dots,y'_{N'}$. The estimator for the percentile function is then chosen to be analogous to the one above. A higher N' leads to a more exact result. We choose N' to be 10,000.

Appendix B.1.2. Sampling

Given an estimator $\widehat{\rho}$ and estimators for the percentile functions $\widehat{F_1^{-1}}, \widehat{F_2^{-1}}$, we obtain M samples from the learned joint distribution with

$$y_\ell = (y_{\ell 1}, y_{\ell 2}) = \left(\widehat{F_1^{-1}}(u_{\ell 1}), \widehat{F_2^{-1}}(u_{\ell 2})\right) = \left(\widehat{F_1^{-1}}(\Phi(z_{\ell 1})), \widehat{F_2^{-1}}(\Phi(z_{\ell 2}))\right), \ell = 1, \ldots, M,$$

where $z_\ell = (z_{\ell 1}, z_{\ell 2}), \ell = 1, \ldots, M$ are draws from a bivariate normal distribution with mean 0 and covariance $\begin{pmatrix} 1 & \widehat{\rho} \\ \widehat{\rho} & 1 \end{pmatrix}$. In the case of the gold standard, we set $\widehat{F_j^{-1}} = F_j^{-1}, j = 1, 2$. We obtain samples of the real underlying distribution by using the correct percentile functions, as in the gold standard, and, additionally, $\widehat{\rho} = \rho$. The procedure for three dimensions is analogous.

Appendix B.2. Missing Mechanism for Section 4.4

The missing mechanism is similar to the two-dimensional case. The marginals are chosen to be χ^2 with six, seven, and five degrees of freedom. The data matrix $D \in \mathbb{R}^{N \times 3}$ keeps N (complete) observations of the random vector. We enforce the following missing data mechanism:

1. Again, we remove every entry in the data matrix D with probability $0 \leq p_{MCAR} < 1$. The resulting data matrix (with missing entries) is denoted as

$$D^{MCAR} = \left(D_{\ell j}^{MCAR}\right)_{\ell=1,\ldots,N, j=1,2,3}.$$

2. If $D_{\ell 1}^{MCAR}$, $D_{\ell 2}^{MCAR}$, and $D_{\ell 3}^{MCAR}$ are observed, we remove $D_{\ell 3}^{MCAR}$ with probability

$$\mathbb{P}(R_3 = 0 | X_1 = D_{\ell 1}, X_2 = D_{\ell 2}) = h(D_{\ell 1}, D_{\ell 2}; \beta),$$

where

$$h(D_{\ell 1}, D_{\ell 2}; \beta) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 \Phi^{-1}(F_1(D_{\ell 1})) + \beta_2 \Phi^{-1}(F_2(D_{\ell 2}))))}$$

and $\beta = (\beta_0, \beta_1, \beta_2)$.

We call the resulting data matrix D^{MAR}. Its missing patterns are, again, non-monotone, and the data are MAR, but not MCAR. In Section 4.4, we set $\beta = (0, 2, 2)$.

Appendix B.3. Complementary Figures

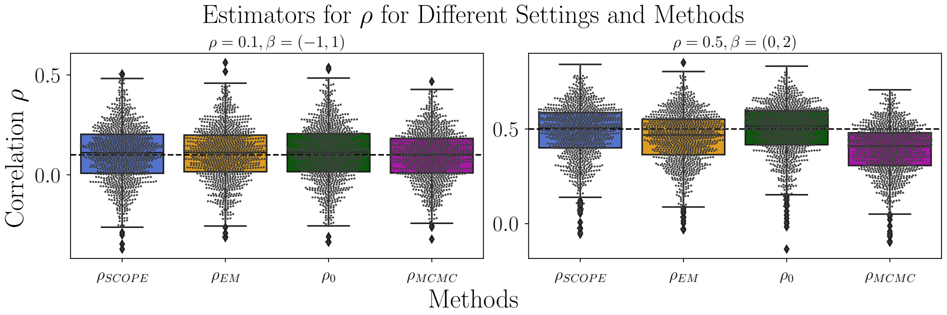

Figure A1. Comparison of the algorithms with respect to the correlation ρ. Shown are the boxplots for the Standard Copula Estimator (SCOPE), the proposed EM algorithm (EM), the method based on known marginals (0), and the Markov chain–Monte Carlo approach (MCMC) for 1000 datasets generated as described in Section 4.2, where $\rho = 0.1, \beta = (-1, 1)$ are depicted in the left canvas and $\rho = 0.5, \beta = (0, 2)$ are depicted in the right canvas. The true correlations are indicated by the dashed line.

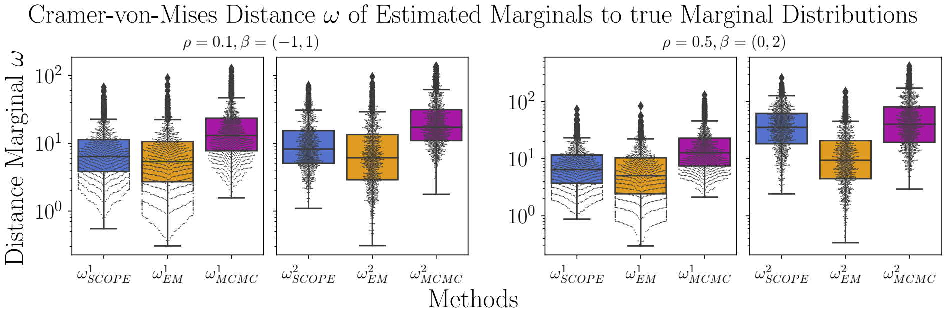

Figure A2. Comparison of the algorithms with respect to the Cramer–von Mises distance between the estimated and the true first (ω^1) and true second marginal distributions (ω^2). Shown are the boxplots on a logarithmic scale for the proposed EM algorithm (EM), the Standard Copula Estimator (SCOPE), and the Markov chain–Monte Carlo approach (MCMC) for 1000 datasets generated as described in Section 4.2, where $\rho = 0.1, \beta = (-1, 1)$ are depicted in the left canvas and $\rho = 0.5, \beta = (0, 2)$ are depicted in the right canvas.

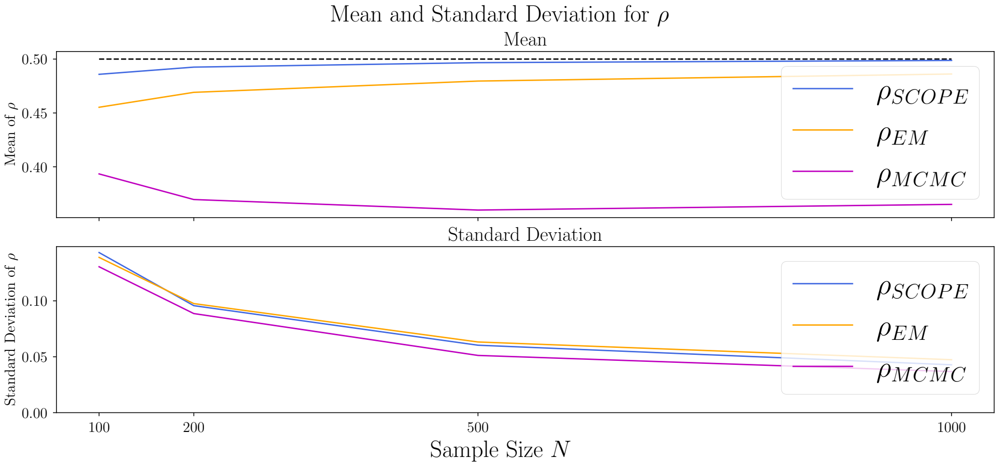

Figure A3. Comparison of the algorithms with respect to the correlation (ρ). Shown are the mean (upper canvas) and standard deviation (lower canvas) of the Standard Copula Estimator (SCOPE), the proposed EM algorithm (EM), and the Markov chain–Monte Carlo approach (MCMC) for 1000 datasets generated as described in Section 4.2 with $\rho = 0.5$ and $\beta = (0, 2)$ and for varying sample sizes $N = 100, 200, 500, 1000$, where the true ρ is 0.5 (dashed line in the upper canvas).

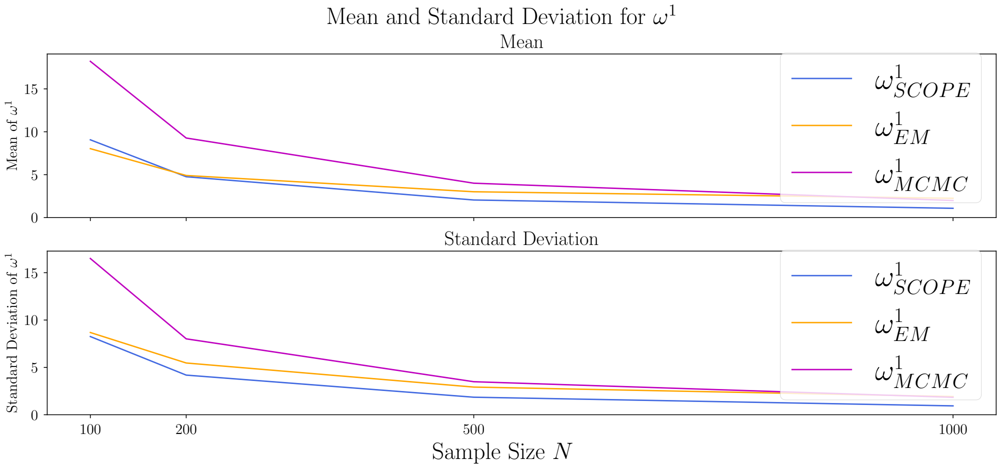

Figure A4. Comparison of the algorithms with respect to the Cramer–von Mises statistic ω^1 between the estimated and the true first marginal distribution. Shown are the mean (upper canvas) and standard deviation (lower canvas) of the Standard Copula Estimator (SCOPE), the proposed EM algorithm (EM), and the Markov chain–Monte Carlo approach (MCMC) for 1000 datasets generated as described in Section 4.2 with $\rho = 0.5$ and $\beta = (0, 2)$ and for varying sample sizes of $N = 100, 200, 500, 1000$.

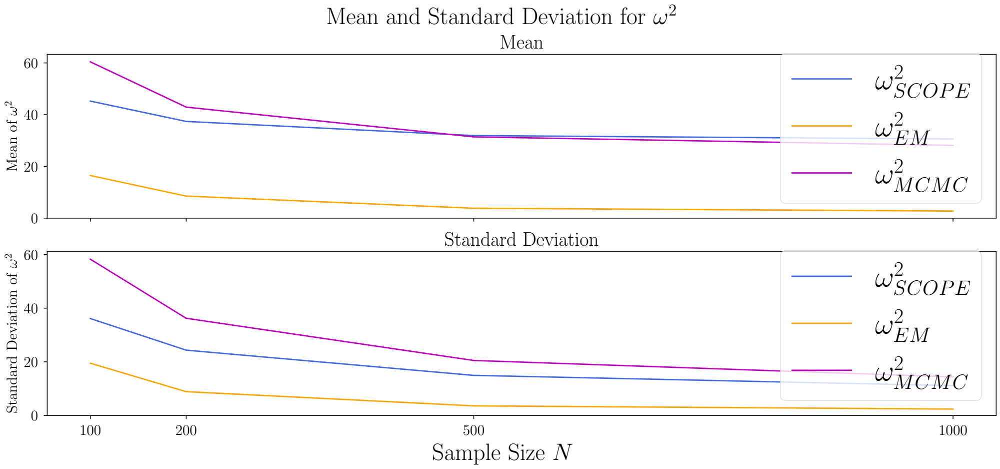

Figure A5. Comparison of the algorithms with respect to the Cramer–von Mises statistic ω^2 between the estimated and the true second marginal distribution. Shown are the mean (upper canvas) and standard deviation (lower canvas) of the Standard Copula Estimator (SCOPE), the proposed EM algorithm (EM), and the Markov chain–Monte Carlo approach (MCMC) for 1000 datasets generated as described in Section 4.2 with $\rho = 0.5$ and $\beta = (0, 2)$ and for varying sample sizes of N = 100, 200, 500, 1000.

References

1. Thurow, M.; Dumpert, F.; Ramosaj, B.; Pauly, M. Imputing missings in official statistics for general tasks–our vote for distributional accuracy. *Stat. J. IAOS* **2021**, *37*, 1379–1390. [CrossRef]
2. Liu, Y.; Dillon, T.; Yu, W.; Rahayu, W.; Mostafa, F. Missing value imputation for industrial IoT sensor data with large gaps. *IEEE Internet Things J.* **2020**, *7*, 6855–6867. [CrossRef]
3. Silverman, B. *Density Estimation for Statistics and Data Analysis*; Routledge: London, UK, 2018.
4. Kertel, M.; Harmeling, S.; Pauly, M. Learning causal graphs in manufacturing domains using structural equation models. *arXiv* **2022**, arXiv:2210.14573. https://doi.org/10.48550/ARXIV.2210.14573.
5. Genest, C.; Ghoudi, K.; Rivest, L.P. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* **1995**, *82*, 543–552. [CrossRef]
6. Liu, H.; Han, F.; Yuan, M.; Lafferty, J.; Wasserman, L. High-dimensional semiparametric gaussian copula graphical models. *Ann. Stat.* **2012**, *40*, 2293–2326. [CrossRef]

7. Titterington, D.; Mill, G. Kernel-based density estimates from incomplete data. *J. R. Stat. Soc. Ser. B Methodol.* **1983**, *45*, 258–266. [CrossRef]
8. Dempster, A.; Laird, N.; Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Methodol.* **1977**, *39*, 1–22.
9. Shen, C.; Weissfeld, L. A copula model for repeated measurements with non-ignorable non-monotone missing outcome. *Stat. Med.* **2006**, *25*, 2427–2440. [CrossRef]
10. Gomes, M.; Radice, R.; Camarena Brenes, J.; Marra, G. Copula selection models for non-Gaussian outcomes that are missing not at random. *Stat. Med.* **2019**, *38*, 480–496. [CrossRef]
11. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, *63*, 581–592. [CrossRef]
12. Cui, R.; Groot, P.; Heskes, T. Learning causal structure from mixed data with missing values using Gaussian copula models. *Stat. Comput.* **2019**, *29*, 311–333. [CrossRef]
13. Wang, H.; Fazayeli, F.; Chatterjee, S.; Banerjee, A. Gaussian copula precision estimation with missing values. In Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, Reykjavik, Iceland, 22–25 April 2014; PMLR: Reykjavik, Iceland, 2014; Volume 33, pp. 978–986.
14. Hamori, S.; Motegi, K.; Zhang, Z. Calibration estimation of semiparametric copula models with data missing at random. *J. Multivar. Anal.* **2019**, *173*, 85–109. [CrossRef]
15. Robins, J.M.; Gill, R.D. Non-response models for the analysis of non-monotone ignorable missing data. *Stat. Med.* **1997**, *16*, 39–56. [CrossRef]
16. Sun, B.; Tchetgen, E.J.T. On inverse probability weighting for nonmonotone missing at random data. *J. Am. Stat. Assoc.* **2018**, *113*, 369–379. [CrossRef] [PubMed]
17. Seaman, S.R.; White, I.R. Review of inverse probability weighting for dealing with missing data. *Stat. Methods Med. Res.* **2013**, *22*, 278–295. [CrossRef]
18. Ding, W.; Song, P. EM algorithm in gaussian copula with missing data. *Comput. Stat. Data Anal.* **2016**, *101*, 1–11. [CrossRef]
19. Efromovich, S. Adaptive nonparametric density estimation with missing observations. *J. Stat. Plan. Inference* **2013**, *143*, 637–650. [CrossRef]
20. Dubnicka, S.R. Kernel density estimation with missing data and auxiliary variables. *Aust. N. Z. J. Stat.* **2009**, *51*, 247–270. [CrossRef]
21. Hoff, P. Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Stat.* **2007**, *1*, 265–283. [CrossRef]
22. Hollenbach, F.; Bojinov, I.; Minhas, S.; Metternich, N.; Ward, M.; Volfovsky, A. Multiple imputation using gaussian copulas. *Sociol. Methods Res.* **2021**, *50*, 1259–1283. [CrossRef]
23. Di Lascio, F.; Giannerini, S.; Reale, A. Exploring copulas for the imputation of complex dependent data. *Stat. Methods Appl.* **2015**, *24*, 159–175. [CrossRef]
24. Houari, R.; Bounceur, A.; Kechadi, T.; Tari, A.; Euler, R. A new method for estimation of missing data based on sampling methods for data mining. *Adv. Intell. Syst. Comput.* **2013**, *225*, 89–100. [CrossRef]
25. Sklar, A. Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* **1959**, *8*, 229–231.
26. Wei, G.; Tanner, M. A monte carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Am. Stat. Assoc.* **1990**, *85*, 699–704. [CrossRef]
27. Meng, X.L.; Rubin, D. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **1993**, *80*, 267–278. [CrossRef]
28. Guo, J.; Levina, E.; Michailidis, G.; Zhu, J. Graphical models for ordinal data. *J. Comput. Graph. Stat.* **2015**, *24*, 183–204. [CrossRef]
29. McLachlan, G.; Lee, S.; Rathnayake, S. Finite mixture models. *Annu. Rev. Stat. Its Appl.* **2019**, *6*, 355–378. [CrossRef]
30. Hwang, J.; Lay, S.; Lippman, A. Nonparametric multivariate density estimation: A comparative study. *IEEE Trans. Signal Process.* **1994**, *42*, 2795–2810. [CrossRef]
31. Scott, D.; Sain, S. Multidimensional density estimation. *Handb. Stat.* **2005**, *24*, 229–261.
32. Zuo, Y.; Cui, Y.; Yu, G.; Li, R.; Ressom, H. Incorporating prior biological knowledge for network-based differential gene expression analysis using differentially weighted graphical LASSO. *BMC Bioinform.* **2017**, *18*, 99. [CrossRef]
33. Li, Y.; Jackson, S.A. Gene network reconstruction by integration of prior biological knowledge. *G3 Genes Genomes Genet.* **2015**, *5*, 1075–1079. [CrossRef] [PubMed]
34. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer Series in Statistics; Springer: Berlin/Heidelberg, Germany, 2009.
35. Joyce, J.M. Kullback-Leibler divergence. In *International Encyclopedia of Statistical Science*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 720–722.
36. Contreras-Reyes, J.E.; Arellano-Valle, R.B. Kullback–Leibler divergence measure for multivariate skew-normal distributions. *Entropy* **2012**, *14*, 1606–1626. [CrossRef]
37. Honaker, J.; King, G.; Blackwell, M. Amelia II: A program for missing data. *J. Stat. Softw.* **2011**, *45*, 1–47. [CrossRef]
38. Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; Müller, H. Causability and explainability of artificial intelligence in medicine. *WIREs Data Min. Knowl. Discov.* **2019**, *9*, e1312. [CrossRef] [PubMed]
39. Dinu, V.; Zhao, H.; Miller, P.L. Integrating domain knowledge with statistical and data mining methods for high-density genomic SNP disease association analysis. *J. Biomed. Inform.* **2007**, *40*, 750–760. [CrossRef]

40. Rubin, D. Multiple imputation after 18+ years. *J. Am. Stat. Assoc.* **1996**, *91*, 473–489. [CrossRef]
41. Van Buuren, S. *Flexible Imputation of Missing Data*; CRC Press: Boca Raton, FL, USA, 2018.
42. Ramosaj, B.; Pauly, M. Predicting missing values: A comparative study on non-parametric approaches for imputation. *Comput. Stat.* **2019**, *34*, 1741–1764. [CrossRef]
43. Ramosaj, B.; Amro, L.; Pauly, M. A cautionary tale on using imputation methods for inference in matched-pairs design. *Bioinformatics* **2020**, *36*, 3099–3106. [CrossRef]
44. Zhao, Y.; Udell, M. Missing value imputation for mixed data via gaussian copula. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, CA, USA, 6–10 July 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 636–646. [CrossRef]
45. Rubin, D.B. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *J. Am. Stat. Assoc.* **2005**, *100*, 322–331. [CrossRef]
46. Ding, P.; Li, F. Causal inference: A missing data perspective. *Stat. Sci.* **2017**, *33*. [CrossRef]
47. Käärik, E.; Käärik, M. Modeling dropouts by conditional distribution, a copula-based approach. *J. Stat. Plan. Inference* **2009**, *139*, 3830–3835. [CrossRef]
48. Murphy, K. *Machine Learning: A Probabilistic Perspective*; The MIT Press: Cambridge, MA, USA, 2012.