D-Vine Quantile Regression Based Multiple Imputation Using a Fully Conditional Specification Approach with Applications to ESG Data

Master's Thesis

Benedikt Simon Flierl

Supervisor:	Prof. Claudia Czado, Ph. D.
Advisor:	Prof. Claudia Czado, Ph. D.
Submission Date:	January 22, 2025

Declaration

I hereby declare that this thesis is entirely the result of my own work except where otherwise indicated. I have only used the resources given in the list of references.

Munich, January 22, 2025

Abstract

Multiple imputation is a statistical framework for analyzing incomplete data sets. The basic idea is to impute an incomplete data set several times considering the observed dependence structure, analyze each of these completed data sets separately, and then combine the results. So far, only a few exceptions in the literature focus on the generation of the imputations with vine copulas, although vine copulas are highly flexible models for multidimensional dependence. In this thesis, we propose a novel multiple imputation method based on the fully conditional specification approach using D-vine quantile regression models. We conduct a simulation study to evaluate the performance of this vine based method and compare it to well-established multiple imputation methods. Our findings indicate that, under certain conditions, the D-vine quantile regression approach can yield enhanced performance. Furthermore, we present a real-data application in finance concerning the determination of ESG scores, where our approach demonstrates superiority over other methods.

Contents

List of Tables

List of Figures

Glossary

Basic Random Quantities

$\mathbf{Q} = (Q_1, \dots, Q_k)$
└── Vector of k quantities of interest

$\mathbf{R} = (R_{ij})_{i=1,\dots,n,\; j=1,\dots,d}$
└── Response matrix of $\mathbf{Y} = (Y_{ij})_{i=1,\dots,n,\; j=1,\dots,d}$:
$R_{ij} = 1$ if Y_{ij} is observed and $R_{ij} = 0$ otherwise

$\mathbf{R}_{i:} = (R_{i1}, \dots, R_{id})$
└── ith row of $\mathbf{R}$

$\mathbf{R}_{:j} = (R_{1j}, \dots, R_{nj})^\top$
└── jth column of $\mathbf{R}$

$Y = (Y_1, \dots, Y_d)$
└── Random vector which describes the distribution of the rows $\mathbf{Y}_{i:}$ of $\mathbf{Y}$

$\mathbf{Y} = (Y_{ij})_{i=1,\dots,n,\; j=1,\dots,d}$
└── (Hypothetically) Complete data set

$\mathbf{Y}_{i:} = (Y_{i1}, \dots, Y_{id})$
└── ith row of $\mathbf{Y}$, ith observation

$\mathbf{Y}_{:j} = (Y_{1j}, \dots, Y_{nj})^\top$
└── jth column of $\mathbf{Y}$

$\mathbf{Y}_{:j^c} = (\mathbf{Y}_{:1}, \dots, \mathbf{Y}_{:j-1}, \mathbf{Y}_{:j+1}, \dots, \mathbf{Y}_{:d})$
└── $\mathbf{Y}$ with removed jth column

$\mathbf{Y}_{\text{obs}} = \{Y_{ij} \colon R_{ij} = 1\}_{i=1,\dots,n,\; j=1,\dots,d}$
└── Observed part of the data set $\mathbf{Y}$

$\mathbf{Y}_{\text{mis}} = \{Y_{ij} \colon R_{ij} = 0\}_{i=1,\dots,n,\; j=1,\dots,d}$
└── Missing part of the data set $\mathbf{Y}$

$\mathbf{Y}_{:j,\text{obs}} = \{Y_{ij} \colon R_{ij} = 1\}_{i=1,\dots,n}$
└── Observed part of column $\mathbf{Y}_{:j}$ of $\mathbf{Y}$

$\mathbf{Y}_{:j,\text{mis}} = \{Y_{ij} \colon R_{ij} = 0\}_{i=1,\dots,n}$
└── Missing part of column $\mathbf{Y}_{:j}$ of $\mathbf{Y}$

$\mathbf{Y}^{(m)}_{\text{mis}} = \{Y^{(m)}_{ij} \colon R_{ij} = 0\}_{i=1,\dots,n,\; j=1,\dots,d}$
└── mth imputation

$\mathbf{Y}^{(m)} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}^{(m)}_{\text{mis}})$
└──── mth imputed data set

$\boldsymbol{\Theta}$ ──── Parameter vector of $\mathbf{Y}$

$\boldsymbol{\Phi}$ ──── Parameter vector of $\mathbf{R}$

Complete-Data Estimators

$\widehat{\mathbf{Q}} = \widehat{\mathbf{Q}}(\mathbf{Y})$
└──── Complete-data estimator of $\mathbf{Q}$

$\widehat{\mathbf{Q}}^{(m)} = \widehat{\mathbf{Q}}(\mathbf{Y} = \mathbf{Y}^{(m)})$
└──── Estimator $\widehat{\mathbf{Q}}$ applied on the mth imputed data set $\mathbf{Y}^{(m)}$

$\widehat{Q}_l = \widehat{Q}_l(\mathbf{Y})$
└──── Complete-data estimator of Q_l

$\widehat{Q}^{(m)}_l = \widehat{Q}_l(\mathbf{Y} = \mathbf{Y}^{(m)})$
└──── Estimator $\widehat{Q}_l$ applied on the mth imputed data set $\mathbf{Y}^{(m)}$

$\widehat{\mathbf{W}} = \widehat{\mathbf{W}}(\mathbf{Y})$
└──── Complete-data estimator of the variance of $\widehat{\mathbf{Q}}$

$\widehat{\mathbf{W}}^{(m)} = \widehat{\mathbf{W}}(\mathbf{Y} = \mathbf{Y}^{(m)})$
└──── Estimator $\widehat{\mathbf{W}}$ applied on the mth imputed data set $\mathbf{Y}^{(m)}$

$\widehat{W}_l = \widehat{W}_l(\mathbf{Y})$
└──── Complete-data estimator of the variance of $\widehat{Q}_l$

$\widehat{W}^{(m)}_l = \widehat{W}_l(\mathbf{Y} = \mathbf{Y}^{(m)})$
└──── Estimator $\widehat{W}_l$ applied on the mth imputed data set $\mathbf{Y}^{(m)}$

Estimators for Repeated-Imputation Inference of $\mathbf{Q}$

$\mathbf{B} = \mathbf{B}(\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(M)}) = \frac{\sum_{m=1}^{M}(\widehat{\mathbf{Q}}^{(m)} - \overline{\mathbf{Q}})^\top(\widehat{\mathbf{Q}}^{(m)} - \overline{\mathbf{Q}})}{M-1}$
└──── Between-imputation variance estimator of $\overline{\mathbf{Q}}$

$\mathbf{B}_\infty = \mathbf{B}_\infty(\mathbf{Y}_{\text{obs}}) = \lim_{M\to\infty} \frac{M-1}{M}\mathbf{B}$
└──── Estimator $\mathbf{B}$ based on infinitely many imputed data sets

$N = N(\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(M)}) = \frac{N' N_{\text{obs}}}{N' + N_{\text{obs}}}$
└──── Estimator of the degrees of freedom for the estimation of $\mathbf{Q}$ with $\overline{\mathbf{Q}}$:

$$N_{\text{obs}} = \frac{\nu_{\text{com}}+1}{\nu_{\text{com}}+3} \cdot \nu_{\text{com}}(1-\Lambda), \qquad N' = \frac{M-1}{\Lambda^2}, \qquad \Lambda = \left(1 + \frac{1}{M}\right)\frac{\operatorname{tr}(\mathbf{B}\mathbf{T}^{-1})}{k}$$

and ν_{com} are the degrees of freedom for the estimation of $\mathbf{Q}$ with $\widehat{\mathbf{Q}}$ from the hypothetically complete data set $\mathbf{Y}$

$\overline{\mathbf{Q}} = \overline{\mathbf{Q}}(\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(M)}) = \frac{1}{M}\sum_{m=1}^{M} \widehat{\mathbf{Q}}^{(m)}$
└── Repeated-imputation estimator of $\mathbf{Q}$

$\overline{\mathbf{Q}}_\infty = \overline{\mathbf{Q}}_\infty(\mathbf{Y}_{\text{obs}}) = \lim_{M\to\infty} \overline{\mathbf{Q}}$
└── Estimator $\overline{\mathbf{Q}}$ based on infinitely many imputed data sets

$\mathbf{T} = \mathbf{T}(\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(M)}) = \overline{\mathbf{W}} + \mathbf{B} + \frac{\mathbf{B}}{M}$
└── (Total) Variance estimator of $\overline{\mathbf{Q}}$

$\mathbf{T}_\infty = \mathbf{T}_\infty(\mathbf{Y}_{\text{obs}}) = \overline{\mathbf{W}}_\infty + \mathbf{B}_\infty$
└── Estimator $\mathbf{T}$ based on infinitely many imputed data sets

$\overline{\mathbf{W}} = \overline{\mathbf{W}}(\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(M)}) = \frac{1}{M}\sum_{m=1}^{M} \widehat{\mathbf{W}}^{(m)}$
└── Within-imputation variance estimator of $\overline{\mathbf{Q}}$

$\overline{\mathbf{W}}_\infty = \overline{\mathbf{W}}_\infty(\mathbf{Y}_{\text{obs}}) = \lim_{M\to\infty} \overline{\mathbf{W}}$
└── Estimator $\overline{\mathbf{W}}$ based on infinitely many imputed data sets

Estimators for Element-Wise Repeated-Imputation Inference of $\mathbf{Q} = (Q_l)_{l=1,\ldots,k}$

$B_l = B_l(\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(M)}) = \frac{\sum_{m=1}^{M}(\widehat{Q}_l^{(m)} - \overline{Q}_l)^2}{M-1}$
└── Between-imputation variance estimator of $\overline{Q}_l$

$N_l = N_l(\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(M)}) = \frac{N_l' N_{l,\text{obs}}}{N_l' + N_{l,\text{obs}}}$
└── Estimator of the degrees of freedom for the estimation of Q_l with $\overline{Q}_l$:

$$N_{l,\text{obs}} = \frac{\nu_{\text{com}}+1}{\nu_{\text{com}}+3} \cdot \nu_{\text{com}}(1 - \Lambda_l), \quad N_l' = \frac{M-1}{\Lambda_l^2}, \quad \Lambda_l = \left(1 + \frac{1}{M}\right)\frac{B_l}{T_l}$$

and ν_{com} are the degrees of freedom for the estimation of $Q_1, \ldots, Q_k$ with $\widehat{Q}_1, \ldots, \widehat{Q}_k$ from the hypothetically complete data set $\mathbf{Y}$

$\overline{Q}_l = \overline{Q}_l(\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(M)}) = \frac{1}{M}\sum_{m=1}^{M} \widehat{Q}_l^{(m)}$
└── Repeated-imputation estimator of Q_l

$T_l = T_l(\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(M)}) = \overline{W}_l + B_l + \frac{B_l}{M}$
└── (Total) Variance estimator of $\overline{Q}_l$

$\overline{W}_l = \overline{W}_l(\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(M)}) = \frac{1}{M}\sum_{m=1}^{M} \widehat{W}_l^{(m)}$
└── Within-imputation variance estimator of $\overline{Q}_l$

Important Indices and Index Sets

i ── Index for rows of $\mathbf{Y}$ and $\mathbf{R}$

$I_{\mathbf{r}}^{\text{obs}}(\mathbf{j}) = \{i \in \{1, \ldots, n\} \colon r_{ij} = 1 \text{ for all } j \in \mathbf{j}\}$
└── All observations where the values at the positions $j \in \mathbf{j}$ are observed

$I_{\mathbf{r}}^{\text{mis}}(\mathbf{j}) = \{i \in \{1, \ldots, n\} \colon r_{ij} = 0 \text{ for all } j \in \mathbf{j}\}$
└── All observations where the values at the positions $j \in \mathbf{j}$ are missing

j ── Index for columns of $\mathbf{Y}$ and $\mathbf{R}$, index for elements of $\mathbf{Y}$

l	Index for elements of $\boldsymbol{\mathsf{Q}}$

Other Important Quantities

d	Number of columns of $\boldsymbol{\mathsf{Y}}$ and $\boldsymbol{\mathsf{R}}$, dimension of $\mathbf{Y}$
k	Dimension of $\boldsymbol{\mathsf{Q}}$, number of quantities of interest
K	Number of nearest neighbors
M	Number of imputations and imputed data sets
$\mathcal{M}$	General imputation method
n	Number of rows of $\boldsymbol{\mathsf{Y}}$ and $\boldsymbol{\mathsf{R}}$
T	Number of the iterations of the quasi-Gibbs sampler

1 Introduction

In recent years, environmental, social and governance (ESG) criteria have increasingly become an important aspect in responsible investing. Thus, many data providers have established in this field publishing metrics that shall make companies comparable in the view of ESG criteria. However, for various reasons, there may be discrepancies in the availability of ESG-related data from one company to another resulting in missing data. Among others this situation might be caused by:

1. **Lack of regulatory and standardization:** Kotsantonis and Serafeim (2019) state that the lack of regulation and standardization for ESG disclosures leads to differences in the data availability across different companies.
2. **Different corporate cultures:** For Chinese companies, Bai, Shang, and Huang (2024) observe a relationship between corporate culture and and data availability.
3. **Varying company sizes:** According to the study of Drempetic, Klein, and Zwergel (2020) bigger companies have in general more data available. A reason might be that they have more resources to provide ESG data.

Missing data can be a serious problem making the comparison and statistical analysis of ESG scores more difficult and less reliable (Sahin et al. 2022). One way of dealing with missing data in general is so-called **multiple imputation**. Multiple imputation is a statistical framework for analyzing incomplete data sets first considered and developed by Rubin under Bayesian considerations (Rubin 1978; Rubin 1987). Such a specific framework is necessary since missing values introduce an additional source of uncertainty in the data. This uncertainty in turn has to be taken into account if any statistical results shall be valid. Otherwise, the uncertainty in the data is possibly underestimated.

The core idea behind the multiple imputation framework is to impute an incomplete data set multiple times, accounting for the observed dependence structure, analyze each completed data set separately, and then combine the results. Although several methods exist for leveraging the dependence structure in data when generating imputations, vine copulas have not been widely explored in the imputation context, despite their high flexibility in modeling multidimensional dependence. The goal of this thesis is to introduce a novel method for generating imputations using vine copulas and to compare its performance with that of well-established multiple imputation techniques. Furthermore, we aim to explore whether multiple imputation can improve the accuracy of ESG score calculations.

Chapter 2 discusses the modeling of incomplete data sets within a Bayesian framework. In Chapter 3, we provide an overview of the general concept of multiple imputation, discuss how to generate imputations and review widely used imputation methods. Chapter 4 introduces a novel multiple imputation method based on the fully conditional specification approach using D-vine quantile regression models. In Chapter 5, we compare this vine based imputation method with other well-established imputation methods in a simulation study. Chapter 6 presents an application of multiple imputation to ESG data in connection with the ESG score calculation. Finally, Chapter 7 provides a summary of the results and offers a brief outlook on future research.

Throughout this thesis we will use the following basic notation: Random variables are denoted by italic capital letters such as Y, random vectors by bold capital letters such as $\mathbf{Y}$ and other collections of random variables, i. e. random matrices or random sets, with bold capital letters sans serifs such as $\boldsymbol{\mathsf{Y}}$. In general, for realizations of random quantities the base symbols are changed from capital letters to the corresponding small letters, i. e. for the quantities defined above we would have y, $\mathbf{y}$ and $\boldsymbol{\mathsf{y}}$. Parameter random vectors and their realizations are denoted analogously but with Greek letters, i. e. $\boldsymbol{\Theta}$ and $\boldsymbol{\theta}$. Independent of the types of the variables, we will denote both, densities and probability mass functions, with $f(\cdot)$ and their conditional versions with $f(\cdot|\cdot)$. Even if we do not restrict ourselves to absolutely continuous random variables, we will always speak of a density for the sake of simplicity. Throughout the thesis, $\mathbb{1}\{\cdot\}$ denotes the indicator function that takes value 1 if the event in braces occurs and 0 otherwise. All computations in this thesis are performed using the R programming language (R Core Team 2023).

2 Incomplete Data Sets: Modeling and Bayesian Analysis

2.1 Bayesian Inference and Prediction in a Nutshell

As already mentioned in the introduction, the concept of multiple imputation builds on Bayesian considerations. In order to embed the situation of incomplete data sets in a Bayesian context, we first introduce a general Bayesian framework following Gelman, Carlin, et al. (2013, pp. 6–7), which is then applied to incomplete data sets. To this end, let $\mathbf{X}$ and $\widetilde{\mathbf{X}}$ be possibly dependent collections of random variables representing observed and unobserved data. In the context of Bayesian inference and prediction, the data $(\mathbf{X}, \widetilde{\mathbf{X}})$ depends on a parameter vector $\boldsymbol{\Psi}$ which is assumed to be random. This leads to a joint probability model for $(\mathbf{X}, \widetilde{\mathbf{X}}, \boldsymbol{\Psi})$ expressed by the joint density $f_{\mathbf{X},\widetilde{\mathbf{X}},\boldsymbol{\Psi}}(\mathbf{x}, \widetilde{\mathbf{x}}, \boldsymbol{\psi})$. By definition of the conditional density, we can model this joint density as the product of the marginal density $f_{\boldsymbol{\Psi}}(\boldsymbol{\psi})$ of $\boldsymbol{\Psi}$ (**prior density**) and the conditional density $f_{\mathbf{X},\widetilde{\mathbf{X}}|\boldsymbol{\Psi}}(\mathbf{x}, \widetilde{\mathbf{x}} \mid \boldsymbol{\psi})$ of $(\mathbf{X}, \widetilde{\mathbf{X}})$ given $\boldsymbol{\Psi}$ (**sampling density**):

$$f_{\mathbf{X},\widetilde{\mathbf{X}},\boldsymbol{\Psi}}(\mathbf{x}, \widetilde{\mathbf{x}}, \boldsymbol{\psi}) := f_{\boldsymbol{\Psi}}(\boldsymbol{\psi}) f_{\mathbf{X},\widetilde{\mathbf{X}}|\boldsymbol{\Psi}}(\mathbf{x}, \widetilde{\mathbf{x}} \mid \boldsymbol{\psi}).$$

The prior density can be seen as an initial guess how $\boldsymbol{\Psi}$ is distributed. The sampling density describes how the data is generated given the model parameters.

In **Bayesian inference**, the goal is to update the initial guess for the distribution of $\boldsymbol{\Psi}$ based on the observed data. One is therefore interested in the conditional density $f_{\boldsymbol{\Psi}|\mathbf{X}}(\boldsymbol{\psi} \mid \mathbf{x})$ of $\boldsymbol{\Psi}$ given $\mathbf{X}$, called **posterior density**. It calculates as

$$f_{\boldsymbol{\Psi}|\mathbf{X}}(\boldsymbol{\psi} \mid \mathbf{x}) = \frac{f_{\mathbf{X},\boldsymbol{\Psi}}(\mathbf{x}, \boldsymbol{\psi})}{f_{\mathbf{X}}(\mathbf{x})} = \frac{f_{\boldsymbol{\Psi}}(\boldsymbol{\psi}) f_{\mathbf{X}|\boldsymbol{\Psi}}(\mathbf{x} \mid \boldsymbol{\psi})}{f_{\mathbf{X}}(\mathbf{x})},$$

where both the marginal sampling density $f_{\mathbf{X}|\boldsymbol{\Psi}}(\mathbf{x} \mid \boldsymbol{\psi})$ of $\mathbf{X}$,

$$f_{\mathbf{X}|\boldsymbol{\Psi}}(\mathbf{x} \mid \boldsymbol{\psi}) = \int f_{\mathbf{X},\widetilde{\mathbf{X}}|\boldsymbol{\Psi}}(\mathbf{x}, \widetilde{\mathbf{x}} \mid \boldsymbol{\psi}) \, \mathrm{d}\widetilde{\mathbf{x}},$$

and the marginal density $f_{\mathbf{X}}(\mathbf{x})$ of $\mathbf{X}$,

$$f_{\mathbf{X}}(\mathbf{x}) = \int \int f_{\mathbf{X},\widetilde{\mathbf{X}},\boldsymbol{\Psi}}(\mathbf{x}, \widetilde{\mathbf{x}}, \boldsymbol{\psi}) \, \mathrm{d}\widetilde{\mathbf{x}} \, \mathrm{d}\boldsymbol{\psi} = \int f_{\boldsymbol{\Psi}}(\boldsymbol{\psi}) \int f_{\mathbf{X},\widetilde{\mathbf{X}}|\boldsymbol{\Psi}}(\mathbf{x}, \widetilde{\mathbf{x}} \mid \boldsymbol{\psi}) \, \mathrm{d}\widetilde{\mathbf{x}} \, \mathrm{d}\boldsymbol{\psi},$$

can be expressed solely in terms of the prior and the sampling density. The posterior density is analyzed to make inference about the parameter vector $\boldsymbol{\Psi}$.

Often, not only $\boldsymbol{\Psi}$ itself but other quantities $\widetilde{\mathbf{Q}} := \widetilde{\mathbf{Q}}(\boldsymbol{\Psi})$ related to $\boldsymbol{\Psi}$ are of interest. Thus we define the conditional distribution of $\widetilde{\mathbf{Q}}$ given X via

$$\mathbb{P}(\widetilde{\mathbf{Q}} \in B \mid \mathsf{X} = \mathsf{x}) := \int_{\{\boldsymbol{\psi}\colon\, \widetilde{\mathbf{Q}}(\boldsymbol{\psi}) \in B\}} f_{\boldsymbol{\Psi}|\mathsf{X}}(\boldsymbol{\psi} \mid \mathsf{x})\, \mathrm{d}\boldsymbol{\psi},$$

where B is a set of possible values of $\widetilde{\mathbf{Q}}$. The corresponding conditional density $f_{\widetilde{\mathbf{Q}}|\mathsf{X}}(\widetilde{\mathbf{q}}|\mathsf{x})$ we call posterior density of $\widetilde{\mathbf{Q}}$. Since the posterior density of $\boldsymbol{\Psi}$ equals the posterior density of $\widetilde{\mathbf{Q}}$ when setting $\widetilde{\mathbf{Q}} = \widetilde{\mathbf{Q}}(\boldsymbol{\Psi}) = \boldsymbol{\Psi}$ it is enough to consider the posterior density of $\widetilde{\mathbf{Q}}$ for Bayesian inference.

Standard estimates in Bayesian inference include (Rubin 1987, pp. 59–62):

(a) **Highest posterior density regions**: Let $1-\alpha \in (0,1)$ be fixed. Then a set C of possible values of $\widetilde{\mathbf{Q}}$ is a highest posterior density region of coverage $1-\alpha$ if

  (i) $\mathbb{P}(\widetilde{\mathbf{Q}} \in C \mid \mathsf{X} = \mathsf{x}) = 1-\alpha$ and

  (ii) $f_{\widetilde{\mathbf{Q}}|\mathsf{X}}(\widetilde{\mathbf{q}}_1 \mid \mathsf{x}) > f_{\widetilde{\mathbf{Q}}|\mathsf{X}}(\widetilde{\mathbf{q}}_2 \mid \mathsf{x})$ for every $\widetilde{\mathbf{q}}_1 \in C$ and $\widetilde{\mathbf{q}}_2 \notin C$,

that is the posterior probability that C contains $\widetilde{\mathbf{Q}}$ equals $1-\alpha$ and every point in C has higher posterior density than every point outside C.

(b) **Significance levels – p-values**: Let $\mathbb{P}(f_{\widetilde{\mathbf{Q}}|\mathsf{X}}(\widetilde{\mathbf{Q}} \mid \mathsf{x}) > f_{\widetilde{\mathbf{Q}}|\mathsf{X}}(\widetilde{\mathbf{q}}_0 \mid \mathsf{x})) = 1-\alpha$ be the posterior probability that $\widetilde{\mathbf{Q}}$ has higher posterior density than a fixed null value $\widetilde{\mathbf{q}}_0$. Then α is called the significance level or p-value of the null value $\widetilde{\mathbf{q}}_0$.

(c) **Point estimates** of $\widetilde{\mathbf{Q}}$ are calculated as the posterior mode, mean or median.

The highest posterior density region and p-values for a normally and t-distributed $\widetilde{\mathbf{Q}}$ are given in the following example (Rubin 1987, pp. 60–61).

Example 2.1.1.

(a) Suppose the posterior distribution of $\widetilde{\mathbf{Q}}$ given $\mathsf{X} = \mathsf{x}$ follows a $\widetilde{k}$-dimensional multivariate normal distribution with mean vector $\mathbf{m}$ and variance matrix v, i. e. $(\widetilde{\mathbf{Q}} \mid \mathsf{X} = \mathsf{x}) \sim \mathcal{N}(\mathbf{m}, \mathsf{v})$. Then the highest posterior density region with coverage $1-\alpha$ is the set of all $\widetilde{\mathbf{q}}$ such that

$$(\widetilde{\mathbf{q}} - \mathbf{m})\mathsf{v}^{-1}(\widetilde{\mathbf{q}} - \mathbf{m})^\top < \chi^2_{\widetilde{k}}(\alpha),$$

where $\chi^2_{\widetilde{k}}(\alpha)$ is the α-quantile of the chi-squared distribution on $\widetilde{k}$ degrees of freedom. The significance level or p-value of the null value $\widetilde{\mathbf{q}}_0$ is given by

$$\mathbb{P}\big(\chi^2_{\widetilde{k}} > (\widetilde{\mathbf{q}}_0 - \mathbf{m})\mathsf{v}^{-1}(\widetilde{\mathbf{q}}_0 - \mathbf{m})^\top\big),$$

where $\chi^2_{\widetilde{k}}$ is a chi-squared distributed random variable on $\widetilde{k}$ degrees of freedom.

(b) Suppose the posterior distribution of $\widetilde{\mathbf{Q}}$ given $\mathbf{X} = \mathbf{x}$ follows a $\widetilde{k}$-dimensional multivariate t-distribution with location $\mathbf{l}$, scale matrix $\mathbf{s}^{1/2}$ and ν degrees of freedom, i. e. $(\widetilde{\mathbf{Q}} \mid \mathbf{X} = \mathbf{x}) \sim t_\nu(\mathbf{l}, \mathbf{s})$. Then the highest posterior density region with coverage $1 - \alpha$ is the set of all $\widetilde{\mathbf{q}}$ such that

$$\frac{(\widetilde{\mathbf{q}} - \mathbf{l})\mathbf{s}^{-1}(\widetilde{\mathbf{q}} - \mathbf{l})^\top}{\widetilde{k}} < F_{\widetilde{k},\nu}(\alpha),$$

where $F_{\widetilde{k},\nu}(\alpha)$ is the α-quantile of the F-distribution on $\widetilde{k}$ and ν degrees of freedom. The significance level or p-value of the null value $\widetilde{\mathbf{q}}_0$ is given by

$$\mathbb{P}\left(F_{\widetilde{k},\nu} > \frac{(\widetilde{\mathbf{q}} - \mathbf{l})\mathbf{s}^{-1}(\widetilde{\mathbf{q}} - \mathbf{l})^\top}{\widetilde{k}}\right),$$

where $F_{\widetilde{k},\nu}$ is an F-distributed random variable on $\widetilde{k}$ and ν degrees of freedom. △

In **Bayesian prediction**, the goal is to make inference about the unobserved data $\widetilde{\mathbf{X}}$ based on the observed data $\mathbf{X}$. Thus, one is interested in the conditional density $f_{\widetilde{\mathbf{X}}|\mathbf{X}}(\widetilde{\mathbf{x}} \mid \mathbf{x})$ of $\widetilde{\mathbf{X}}$ given $\mathbf{X}$, called **posterior predictive density**. It calculates as

$$f_{\widetilde{\mathbf{X}}|\mathbf{X}}(\widetilde{\mathbf{x}} \mid \mathbf{x}) = \int f_{\widetilde{\mathbf{X}},\boldsymbol{\Psi}|\mathbf{X}}(\widetilde{\mathbf{x}}, \boldsymbol{\psi} \mid \mathbf{x})\,\mathrm{d}\boldsymbol{\psi} = \int f_{\widetilde{\mathbf{X}}|\mathbf{X},\boldsymbol{\Psi}}(\widetilde{\mathbf{x}} \mid \mathbf{x}, \boldsymbol{\psi}) f_{\boldsymbol{\Psi}|\mathbf{X}}(\boldsymbol{\psi} \mid \mathbf{x})\,\mathrm{d}\boldsymbol{\psi},$$

where the conditional distribution of $\widetilde{\mathbf{X}}$ given $\mathbf{X}$,

$$f_{\widetilde{\mathbf{X}}|\mathbf{X},\boldsymbol{\Psi}}(\widetilde{\mathbf{x}} \mid \mathbf{x}, \boldsymbol{\psi}) = \frac{f_{\mathbf{X},\widetilde{\mathbf{X}}|\boldsymbol{\Psi}}(\mathbf{x}, \widetilde{\mathbf{x}} \mid \boldsymbol{\psi})}{f_{\mathbf{X}|\boldsymbol{\Psi}}(\mathbf{x} \mid \boldsymbol{\psi})} = \frac{f_{\mathbf{X},\widetilde{\mathbf{X}}|\boldsymbol{\Psi}}(\mathbf{x}, \widetilde{\mathbf{x}} \mid \boldsymbol{\psi})}{\int f_{\mathbf{X},\widetilde{\mathbf{X}}|\boldsymbol{\Psi}}(\mathbf{x}, \widetilde{\mathbf{x}} \mid \boldsymbol{\psi})\,\mathrm{d}\widetilde{\mathbf{x}}},$$

can be expressed solely in terms of the sampling density.

2.2 Incomplete Data Sets: A Mathematical Description

To transfer our previous considerations to the situation of incomplete data sets, we first need to model incomplete data sets via collections of random variables. To this end, let $\mathbf{Y} := (Y_{ij})_{i=1,\dots,n,\, j=1,\dots,d}$ be a $n \times d$ random matrix representing our **(hypothetically) complete data set**. We denote the jth column of $\mathbf{Y}$ by $\mathbf{Y}_{:j} := (Y_{1j}, \dots, Y_{nj})^\top$. Similarly, the ith row of $\mathbf{Y}$ is denoted by $\mathbf{Y}_{i:} := (Y_{i1}, \dots, Y_{id})$. We assume that the data set $\mathbf{Y}$ is an i. i. d. sample of a d-dimensional random vector $\mathbf{Y} := (Y_1, \dots, Y_d)$ such that the rows $\mathbf{Y}_{1:}, \dots, \mathbf{Y}_{n:}$ of $\mathbf{Y}$ are independent and identically distributed as $\mathbf{Y}$. We restrict the random variables $Y_1, \dots, Y_d$ to be either absolutely continuous or discrete. If a variable $Y_1, \dots, Y_d$ is discrete it has to be at least ordinally scaled.

To decide which data has been observed or not, we define the $n \times d$ random matrix $\mathbf{R} := (R_{ij})_{i=1,\dots,n,\, j=1,\dots,d}$, the **response** matrix, consisting of indicator random variables

$$R_{ij} := \begin{cases} 1 & \text{if } Y_{ij} \text{ has been observed} \\ 0 & \text{otherwise} \end{cases}.$$

Analogously to $\mathbf{Y}$, for $\mathbf{R}$ its jth column is denoted by $\mathbf{R}_{:j} := (R_{1j}, \dots, R_{nj})^\top$ and its ith row by $\mathbf{R}_{i:} := (R_{i1}, \dots, R_{id})$. We assume that for each row $\mathbf{Y}_{i:}$ of $\mathbf{Y}$ at least one value has been observed, that is for all $i = 1, \dots, n$ we have $\mathbf{R}_{i:} \neq \mathbf{0}$. Also, it is reasonable to assume that for each column $\mathbf{Y}_{:j}$ of $\mathbf{Y}$ at least one value has been observed, that is for all $j = 1, \dots, d$ we have $\mathbf{R}_{:j} \neq \mathbf{0}^\top$. Together, $(\mathbf{Y}, \mathbf{R})$ forms a model for an incomplete data set of dimension $n \times d$.

Remark 2.2.1 (Response Pattern). From the definition of $\mathbf{R}$ it follows that its rows $\mathbf{R}_{1:}, \dots, \mathbf{R}_{n:}$ always map into to the set $\mathcal{R} := \{0,1\}^d \setminus \{\mathbf{0}\}$. A vector $\mathbf{r} \in \mathcal{R}$ we call **response pattern**. Thus, $\mathcal{R}$ is the set of all possible response patterns. △

With the help of the response matrix $\mathbf{R}$ we distinguish the **observed part** $\mathbf{Y}_{\text{obs}}$ of $\mathbf{Y}$ as

$$\mathbf{Y}_{\text{obs}} := \{Y_{ij} \colon R_{ij} = 1\}_{i=1,\dots,n,\, j=1,\dots,d},$$

and the **missing part** $\mathbf{Y}_{\text{mis}}$ as

$$\mathbf{Y}_{\text{mis}} := \{Y_{ij} \colon R_{ij} = 0\}_{i=1,\dots,n,\, j=1,\dots,d}.$$

From here we can decompose $\mathbf{Y}$ as $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$.

Example 2.2.2. We consider an incomplete realized data set $(\mathbf{y}, \mathbf{r})$ with $n = 15$ observations. The hypothetically complete data set $\mathbf{y}$ is simulated from the random vector $\mathbf{Y} = (Y_1, Y_2, Y_3)$ of dimension $d = 3$ distributed according to the D-vine with marginals given in Table 2.1 and vine copula shown in Figure 2.1 (for an introduction to D-vines see Section 4.1).

Variable	Distribution
Y_1	$\mathcal{N}(4,1)$
Y_2	$\mathcal{N}(6,1)$
Y_3	$\mathbb{P}(Y_3 = \mathrm{F}) = \frac{1}{2} = \mathbb{P}(Y_3 = \mathrm{T})$

Table 2.1: Marginal distributions of the data generating vine in Example 2.2.2.

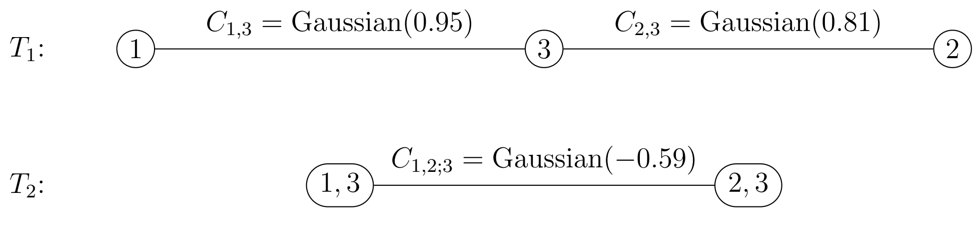

T_1:
1
$C_{1,3}$ = Gaussian(0.95)
3
$C_{2,3}$ = Gaussian(0.81)
2
T_2:
1, 3
$C_{1,2;3}$ = Gaussian(−0.59)
2, 3

Figure 2.1: Vine copula of the data generating vine in Example 2.2.2.

Especially, the third variable Y_3 is discrete and takes the two values F and T. The missing data is generated missing at random (see Definition 2.3.1).

The realizations $\mathbf{y}$ and $\mathbf{r}$ are given as

$$\mathbf{y} = \begin{bmatrix} 4.93 & 7.58 & \text{T} \\ 3.22 & 5.06 & \text{F} \\ 2.46 & 4.78 & \text{F} \\ 5.51 & 7.38 & \text{T} \\ 3.01 & 6.60 & \text{F} \\ 4.38 & 5.41 & \text{T} \\ 5.20 & 7.66 & \text{T} \\ 3.64 & 5.93 & \text{T} \\ 4.51 & 6.61 & \text{T} \\ 3.92 & 6.93 & \text{T} \\ 2.69 & 5.17 & \text{F} \\ 4.22 & 6.14 & \text{F} \\ 3.68 & 6.46 & \text{T} \\ 4.29 & 6.44 & \text{F} \\ 2.32 & 4.40 & \text{F} \end{bmatrix}, \qquad \mathbf{r} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}.$$

For $\mathbf{y}$, the observed part $\mathbf{y}_{\text{obs}}$ is marked by a gray background while the missing part $\mathbf{y}_{\text{mis}}$ is marked by a red background. For better visualization the realized response $\mathbf{r}$ is colored accordingly.

For $\mathbf{r}$, from the possible seven response patterns

$$\mathbf{r} \in \mathcal{R} = \{(1,1,1),(0,1,1),(1,0,1),(1,1,0),(0,0,1),(0,1,0),(1,0,0)\},$$

four response patterns occurred, namely the response patterns

$$\mathbf{r} \in \{(1,1,1),(1,0,1),(1,1,0),(1,0,0)\}. \qquad \triangle$$

Remark 2.2.3. Similarly to the observed part $\mathbf{Y}_{\text{obs}}$ and the missing part of $\mathbf{Y}_{\text{mis}}$ of $\mathbf{Y}$, we define the observed part $\mathbf{Y}_{:j,\text{obs}}$ of column $\mathbf{Y}_{:j}$ as

$$\mathbf{Y}_{:j,\text{obs}} := \{Y_{ij} \colon R_{ij} = 1\}_{i=1,\ldots,n},$$

and the missing part $\mathbf{Y}_{:j,\text{mis}}$ as

$$\mathbf{Y}_{:j,\text{mis}} := \{Y_{ij} \colon R_{ij} = 0\}_{i=1,\ldots,n},$$

such that we can decompose column $\mathbf{Y}_{:j}$ as $\mathbf{Y}_{:j} = (\mathbf{Y}_{:j,\text{obs}}, \mathbf{Y}_{:j,\text{mis}})$. $\triangle$

Notation 2.2.4. Let $\mathbf{r} = (r_{ij})_{i=1,\ldots,n,\ j=1,\ldots,d} \in \{0,1\}^{n \times d}$ be a realization of the response matrix $\mathbf{R}$. Let $J \subseteq \{1,\ldots,d\}$ and the vector $\mathbf{j}$ be an arbitrary permutation of J.

(a) We define the set

$$I_{\mathbf{r}}^{\text{obs}}(\mathbf{j}) := \{i \in \{1,\ldots,n\} \colon r_{ij} = 1 \text{ for all } j \in \mathbf{j}\}$$

which indexes all observations with observed values at the positions defined by $\mathbf{j}$.

(b) Analogously, the set

$$I_{\mathbf{r}}^{\mathrm{mis}}(\mathbf{j}) \coloneqq \{i \in \{1, \ldots, n\} \colon r_{ij} = 0 \text{ for all } j \in \mathbf{j}\}$$

indexes all observations with missing values at the positions defined by $\mathbf{j}$. △

A special case of Notation 2.2.4 is the situation where $\mathbf{j} = j \in \{1, \ldots, d\}$. Then, the set $I_{\mathbf{r}}^{\mathrm{obs}}(j)$ indexes exactly those observations with an observed value for variable Y_j. On the other hand, the set $I_{\mathbf{r}}^{\mathrm{mis}}(j)$ indexes all observations for which the value for variable Y_j is missing. Therefore, given a realized incomplete data set $(\mathbf{y}, \mathbf{r})$ we can also express the observed part $\mathbf{y}_{:j,\mathrm{obs}}$ of column $\mathbf{y}_{:j}$ as $\mathbf{y}_{:j,\mathrm{obs}} = \{y_{ij} \colon i \in I_{\mathbf{r}}^{\mathrm{obs}}(j)\}$ and the missing part $\mathbf{y}_{:j,\mathrm{mis}}$ as $\mathbf{y}_{:j,\mathrm{mis}} = \{y_{ij} \colon i \in I_{\mathbf{r}}^{\mathrm{mis}}(j)\}$.

2.3 Bayesian Inference and Prediction for Incomplete Data Sets

To embed $\mathbf{Y} = (\mathbf{Y}_{\mathrm{obs}}, \mathbf{Y}_{\mathrm{mis}})$ and $\mathbf{R}$ in the Bayesian setting from the previous section we consider a parameter vector $(\boldsymbol{\Theta}, \boldsymbol{\Phi})$ for $(\mathbf{Y}, \mathbf{R})$, where $\boldsymbol{\Theta}$ is the parameter vector governing the data model $\mathbf{Y}$ and $\boldsymbol{\Phi}$ is the parameter vector governing the model for the response mechanism $\mathbf{R}$. The joint model of $(\mathbf{Y}, \mathbf{R}, \boldsymbol{\Theta}, \boldsymbol{\Phi})$ is defined in terms of densities as the product of the **prior density** $f_{\boldsymbol{\Theta},\boldsymbol{\Phi}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ and the **sampling density** $f_{\mathbf{Y},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}(\mathbf{y}, \mathbf{r} \mid \boldsymbol{\theta}, \boldsymbol{\phi})$:

$$f_{\mathbf{Y},\mathbf{R},\boldsymbol{\Theta},\boldsymbol{\Phi}}(\mathbf{y}, \mathbf{r}, \boldsymbol{\theta}, \boldsymbol{\phi}) \coloneqq f_{\boldsymbol{\Theta},\boldsymbol{\Phi}}(\boldsymbol{\theta}, \boldsymbol{\phi}) f_{\mathbf{Y},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}(\mathbf{y}, \mathbf{r} \mid \boldsymbol{\theta}, \boldsymbol{\phi}).$$

In the terminology of the previous section, $\mathbf{Y}_{\mathrm{mis}}$ corresponds to $\widetilde{\mathbf{X}}$, $(\mathbf{Y}_{\mathrm{obs}}, \mathbf{R})$ corresponds to $\mathbf{X}$ and $(\boldsymbol{\Theta}, \boldsymbol{\Phi})$ corresponds to $\boldsymbol{\Psi}$. Consequently, the **posterior density** is given as

$$f_{\boldsymbol{\Theta},\boldsymbol{\Phi}|\mathbf{Y}_{\mathrm{obs}},\mathbf{R}}(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{y}_{\mathrm{obs}}, \mathbf{r}) = \frac{f_{\boldsymbol{\Theta},\boldsymbol{\Phi}}(\boldsymbol{\theta}, \boldsymbol{\phi}) f_{\mathbf{Y}_{\mathrm{obs}},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}(\mathbf{y}_{\mathrm{obs}}, \mathbf{r} \mid \boldsymbol{\theta}, \boldsymbol{\phi})}{f_{\mathbf{Y}_{\mathrm{obs}},\mathbf{R}}(\mathbf{y}_{\mathrm{obs}}, \mathbf{r})},$$

with the marginal sampling density

$$f_{\mathbf{Y}_{\mathrm{obs}},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}(\mathbf{y}_{\mathrm{obs}}, \mathbf{r} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) = \int f_{\mathbf{Y},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}(\mathbf{y}, \mathbf{r} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) \, \mathrm{d}\mathbf{y}_{\mathrm{mis}}$$

of $(\mathbf{Y}_{\mathrm{obs}}, \mathbf{R})$ and the marginal density

$$f_{\mathbf{Y}_{\mathrm{obs}},\mathbf{R}}(\mathbf{y}_{\mathrm{obs}}, \mathbf{r}) = \int\int f_{\boldsymbol{\Theta},\boldsymbol{\Phi}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \int f_{\mathbf{Y},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}(\mathbf{y}, \mathbf{r} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) \, \mathrm{d}\mathbf{y}_{\mathrm{mis}} \, \mathrm{d}\boldsymbol{\phi} \, \mathrm{d}\boldsymbol{\theta}.$$

of $(\mathbf{Y}_{\mathrm{obs}}, \mathbf{R})$. Since our focus lies on the data generating process, instead of the full posterior density, we are interested in the marginal posterior density of $\boldsymbol{\Theta}$,

$$f_{\boldsymbol{\Theta}|\mathbf{Y}_{\mathrm{obs}},\mathbf{R}}(\boldsymbol{\theta} \mid \mathbf{y}_{\mathrm{obs}}, \mathbf{r}) = \int f_{\boldsymbol{\Theta},\boldsymbol{\Phi}|\mathbf{Y}_{\mathrm{obs}},\mathbf{R}}(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{y}_{\mathrm{obs}}, \mathbf{r}) \, \mathrm{d}\boldsymbol{\phi}.$$

The relevant **posterior predictive density** is

$$f_{\mathbf{Y}_{\mathrm{mis}}|\mathbf{Y}_{\mathrm{obs}},\mathbf{R}}(\mathbf{y}_{\mathrm{mis}} \mid \mathbf{y}_{\mathrm{obs}}, \mathbf{r}).$$

In the context of multiple imputation, it is usually assumed that the missing data is **missing at random**. If the missing data is missing at random, the response mechanism may depend on the observed data but not on the missing data. This assumption is formally defined in Little and Rubin (2020, Equation (6.51)) as follows.

Definition 2.3.1 (Missing at Random)**.** The missing data is missing at random at $(\mathbf{y}_{\mathrm{obs}}, \mathbf{r})$ if for all $\mathbf{y}_{\mathrm{mis}}$ and $\boldsymbol{\phi}$ the equation

$$f_{\mathbf{R}|\mathbf{Y},\boldsymbol{\Phi}}(\mathbf{r} \mid \mathbf{y}, \boldsymbol{\phi}) = f_{\mathbf{R}|\mathbf{Y}_{\mathrm{obs}},\boldsymbol{\Phi}}(\mathbf{r} \mid \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\phi})$$

holds, where $f_{\mathbf{R}|\mathbf{Y},\boldsymbol{\Phi}}(\mathbf{r} \mid \mathbf{y}, \boldsymbol{\phi})$ denotes the conditional density of $\mathbf{R}$ given $\mathbf{Y}$ and $\boldsymbol{\Phi}$ and $f_{\mathbf{R}|\mathbf{Y}_{\mathrm{obs}},\boldsymbol{\Phi}}(\mathbf{r} \mid \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\phi})$ denotes the conditional density of $\mathbf{R}$ given $\mathbf{Y}_{\mathrm{obs}}$ and $\boldsymbol{\Phi}$. △

The following theorem shows that, if the missing data is missing at random at $(\mathbf{y}_{\mathrm{obs}}, \mathbf{r})$ and $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ are a priori independent, for Bayesian inference and prediction it is sufficient to omit $\mathbf{R}$ and $\boldsymbol{\Phi}$. This means that in this specific situation one could work based on a joint model only for $(\mathbf{Y}, \boldsymbol{\Theta})$ with corresponding posterior density $f_{\boldsymbol{\Theta}|\mathbf{Y}_{\mathrm{obs}}}$ and posterior predictive density $f_{\mathbf{Y}_{\mathrm{mis}}|\mathbf{Y}_{\mathrm{obs}}}$.

Theorem 2.3.2. *Assume the missing data is missing at random at* $(\mathbf{y}_{\mathrm{obs}}, \mathbf{r})$ *as well as* $\boldsymbol{\Theta}$ *and* $\boldsymbol{\Phi}$ *are a priori independent, that is we have* $f_{\boldsymbol{\Theta},\boldsymbol{\Phi}} \equiv f_{\boldsymbol{\Theta}} f_{\boldsymbol{\Phi}}$*. Then:*

(a) $f_{\boldsymbol{\Theta}|\mathbf{Y}_{\mathrm{obs}},\mathbf{R}}(\boldsymbol{\theta} \mid \mathbf{y}_{\mathrm{obs}}, \mathbf{r}) = f_{\boldsymbol{\Theta}|\mathbf{Y}_{\mathrm{obs}}}(\boldsymbol{\theta} \mid \mathbf{y}_{\mathrm{obs}})$ *and*

(b) $f_{\mathbf{Y}_{\mathrm{mis}}|\mathbf{Y}_{\mathrm{obs}},\mathbf{R}}(\mathbf{y}_{\mathrm{mis}} \mid \mathbf{y}_{\mathrm{obs}}, \mathbf{r}) = f_{\mathbf{Y}_{\mathrm{mis}}|\mathbf{Y}_{\mathrm{obs}}}(\mathbf{y}_{\mathrm{mis}} \mid \mathbf{y}_{\mathrm{obs}})$.

Part (a) of the Theorem 2.3.2 is presented as Corollary 6.1B in Little and Rubin (2020) while part (b) is discussed in Rubin (1987, p. 53).

Proof.

(a) The sampling density $f_{\mathbf{Y},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}$ factorizes as

$$\begin{aligned} f_{\mathbf{Y},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}(\mathbf{y}, \mathbf{r} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) &= f_{\mathbf{R}|\mathbf{Y},\boldsymbol{\Theta},\boldsymbol{\Phi}}(\mathbf{r} \mid \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\phi}) f_{\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\Phi}}(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) \\ &= f_{\mathbf{R}|\mathbf{Y},\boldsymbol{\Phi}}(\mathbf{r} \mid \mathbf{y}, \boldsymbol{\phi}) f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y} \mid \boldsymbol{\theta}) \\ &= f_{\mathbf{R}|\mathbf{Y}_{\mathrm{obs}},\boldsymbol{\Phi}}(\mathbf{r} \mid \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\phi}) f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y} \mid \boldsymbol{\theta}), \end{aligned} \tag{2.1}$$

where the second equality follows since $\boldsymbol{\Theta}$ has no direct influence on $\mathbf{R}$ and $\boldsymbol{\Phi}$ has no direct influence on $\mathbf{Y}$, and the third equality follows since the missing data is missing at random at $(\mathbf{y}_{\text{obs}}, \mathbf{r})$. Consequently, the marginal sampling density $f_{\mathbf{Y}_{\text{obs}},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}$ can be written as

$$\begin{aligned}
f_{\mathbf{Y}_{\text{obs}},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}(\mathbf{y}_{\text{obs}}, \mathbf{r} \mid \boldsymbol{\phi}, \boldsymbol{\theta}) &= \int f_{\mathbf{Y},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}(\mathbf{y}, \mathbf{r} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) \,\mathrm{d}\mathbf{y}_{\text{mis}} \\
&= \int f_{\mathbf{R}|\mathbf{Y}_{\text{obs}},\boldsymbol{\Phi}}(\mathbf{r} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\phi}) f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y} \mid \boldsymbol{\theta}) \,\mathrm{d}\mathbf{y}_{\text{mis}} \\
&= f_{\mathbf{R}|\mathbf{Y}_{\text{obs}},\boldsymbol{\Phi}}(\mathbf{r} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\phi}) \int f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y} \mid \boldsymbol{\theta}) \,\mathrm{d}\mathbf{y}_{\text{mis}} \\
&= f_{\mathbf{R}|\mathbf{Y}_{\text{obs}},\boldsymbol{\Phi}}(\mathbf{r} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\phi}) f_{\mathbf{Y}_{\text{obs}}|\boldsymbol{\Theta}}(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta}).
\end{aligned}$$

Together with the a priori independence of $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ this leads to

$$\begin{aligned}
f_{\boldsymbol{\Theta}|\mathbf{Y}_{\text{obs}},\mathbf{R}}(\boldsymbol{\theta} \mid \mathbf{y}_{\text{obs}}, \mathbf{r}) &= \int f_{\boldsymbol{\Theta},\boldsymbol{\Phi}|\mathbf{Y}_{\text{obs}},\mathbf{R}}(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{y}_{\text{obs}}, \mathbf{r}) \,\mathrm{d}\boldsymbol{\phi} \\
&= \frac{\int f_{\boldsymbol{\Theta},\boldsymbol{\Phi}}(\boldsymbol{\theta}, \boldsymbol{\phi}) f_{\mathbf{Y}_{\text{obs}},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}(\mathbf{y}_{\text{obs}}, \mathbf{r} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) \,\mathrm{d}\boldsymbol{\phi}}{f_{\mathbf{Y}_{\text{obs}},\mathbf{R}}(\mathbf{y}_{\text{obs}}, \mathbf{r})} \\
&= \frac{\int f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) f_{\boldsymbol{\Phi}}(\boldsymbol{\phi}) f_{\mathbf{R}|\mathbf{Y}_{\text{obs}},\boldsymbol{\Phi}}(\mathbf{r} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\phi}) f_{\mathbf{Y}_{\text{obs}}|\boldsymbol{\Theta}}(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\phi}}{f_{\mathbf{Y}_{\text{obs}},\mathbf{R}}(\mathbf{y}_{\text{obs}}, \mathbf{r})} \\
&= f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) f_{\mathbf{Y}_{\text{obs}}|\boldsymbol{\Theta}}(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta}) \cdot \frac{\int f_{\mathbf{R},\boldsymbol{\Phi}|\mathbf{Y}_{\text{obs}}}(\mathbf{r}, \boldsymbol{\phi} \mid \mathbf{y}_{\text{obs}}) \,\mathrm{d}\boldsymbol{\phi}}{f_{\mathbf{Y}_{\text{obs}},\mathbf{R}}(\mathbf{y}_{\text{obs}}, \mathbf{r})} \\
&= f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) f_{\mathbf{Y}_{\text{obs}}|\boldsymbol{\Theta}}(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta}) \cdot \frac{f_{\mathbf{R}|\mathbf{Y}_{\text{obs}}}(\mathbf{r} \mid \mathbf{y}_{\text{obs}})}{f_{\mathbf{Y}_{\text{obs}},\mathbf{R}}(\mathbf{y}_{\text{obs}}, \mathbf{r})} \\
&= \frac{f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) f_{\mathbf{Y}_{\text{obs}}|\boldsymbol{\Theta}}(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta})}{f_{\mathbf{Y}_{\text{obs}}}(\mathbf{y}_{\text{obs}})} \\
&= f_{\boldsymbol{\Theta}|\mathbf{Y}_{\text{obs}}}(\boldsymbol{\theta} \mid \mathbf{y}_{\text{obs}}).
\end{aligned}$$

(b) From the definition of the conditional density the posterior predictive density is

$$\begin{aligned}
&f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}},\mathbf{R}}(\mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}, \mathbf{r}) \\
&= \frac{f_{\mathbf{Y},\mathbf{R}}(\mathbf{y}, \mathbf{r})}{f_{\mathbf{Y}_{\text{obs}},\mathbf{R}}(\mathbf{y}_{\text{obs}}, \mathbf{r})} \\
&= \frac{\int\int f_{\mathbf{Y},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}(\mathbf{y}, \mathbf{r} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) f_{\boldsymbol{\Theta},\boldsymbol{\Phi}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \,\mathrm{d}\boldsymbol{\theta} \,\mathrm{d}\boldsymbol{\phi}}{\int\int\int f_{\mathbf{Y},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}(\mathbf{y}, \mathbf{r} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) f_{\boldsymbol{\Theta},\boldsymbol{\Phi}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \,\mathrm{d}\boldsymbol{\theta} \,\mathrm{d}\boldsymbol{\phi} \,\mathrm{d}\mathbf{y}_{\text{mis}}} \\
&\overset{(2.1)}{=} \frac{\int\int f_{\mathbf{R}|\mathbf{Y}_{\text{obs}},\boldsymbol{\Phi}}(\mathbf{r} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\phi}) f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y} \mid \boldsymbol{\theta}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) f_{\boldsymbol{\Phi}}(\boldsymbol{\phi}) \,\mathrm{d}\boldsymbol{\theta} \,\mathrm{d}\boldsymbol{\phi}}{\int\int\int f_{\mathbf{R}|\mathbf{Y}_{\text{obs}},\boldsymbol{\Phi}}(\mathbf{r} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\phi}) f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y} \mid \boldsymbol{\theta}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) f_{\boldsymbol{\Phi}}(\boldsymbol{\phi}) \,\mathrm{d}\boldsymbol{\theta} \,\mathrm{d}\boldsymbol{\phi} \,\mathrm{d}\mathbf{y}_{\text{mis}}} \\
&= \frac{\int f_{\mathbf{R}|\mathbf{Y}_{\text{obs}},\boldsymbol{\Phi}}(\mathbf{r} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\phi}) f_{\boldsymbol{\Phi}}(\boldsymbol{\phi}) \,\mathrm{d}\boldsymbol{\phi}}{\int f_{\mathbf{R}|\mathbf{Y}_{\text{obs}},\boldsymbol{\Phi}}(\mathbf{r} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\phi}) f_{\boldsymbol{\Phi}}(\boldsymbol{\phi}) \,\mathrm{d}\boldsymbol{\phi}} \cdot \frac{\int f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y} \mid \boldsymbol{\theta}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta}}{\int\int f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y} \mid \boldsymbol{\theta}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta} \,\mathrm{d}\mathbf{y}_{\text{mis}}} \\
&= \frac{f_{\mathbf{Y}}(\mathbf{y})}{f_{\mathbf{Y}_{\text{obs}}}(\mathbf{y}_{\text{obs}})} \\
&= f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}(\mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}).
\end{aligned}$$

□

Assumption 2.3.3. The missing data is missing at random at $(\mathbf{y}_{\text{obs}}, \mathbf{r})$ and $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ are a priori independent. △

Let $\mathbf{Q} \coloneqq (Q_1, \ldots, Q_k)$ denote a k-dimensional vector of quantities of interest depending on $\boldsymbol{\Theta}$, i. e.$\mathbf{Q} = \mathbf{Q}(\boldsymbol{\Theta})$. We define the conditional distribution of $\mathbf{Q}$ given $\mathbf{Y}_{\text{obs}}$ via

$$\mathbb{P}(\mathbf{Q} \in B \mid \mathbf{Y}_{\text{obs}} = \mathbf{y}_{\text{obs}}) \coloneqq \int_{\{\boldsymbol{\theta}\colon\, \mathbf{Q}(\boldsymbol{\theta}) \in B\}} f_{\boldsymbol{\Theta}|\mathbf{Y}_{\text{obs}}}(\boldsymbol{\theta} \mid \mathbf{y}_{\text{obs}}) \, \mathrm{d}\boldsymbol{\theta},$$

where B is a set of possible values of $\mathbf{Q}$. The conditional density $f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}(\mathbf{q} \mid \mathbf{y}_{\text{obs}})$ corresponding to this conditional distribution we call posterior density of $\mathbf{Q}$ given $\mathbf{Y}_{\text{obs}}$.

3 Multiple Imputation

3.1 Basic Mechanism

Suppose for a moment that $\mathbf{Y}$ would be completely observed. From a frequentist perspective, for making inference one would typically define a point estimator $\widehat{\mathbf{Q}} := \widehat{\mathbf{Q}}(\mathbf{Y})$ for $\mathbf{Q}$ as well as an estimator $\widehat{\mathbf{W}} := \widehat{\mathbf{W}}(\mathbf{Y})$ for the variance of $\widehat{\mathbf{Q}}$ under the sampling density $f_{\mathbf{Y}|\mathbf{\Theta}}$ and then rely on the assumption that the distribution of $\mathbf{Q} - \widehat{\mathbf{Q}}$ given $\mathbf{\Theta} = \boldsymbol{\theta}$ can be approximated by a k-dimensional normal distribution, that is

$$\big(\mathbf{Q} - \widehat{\mathbf{Q}} \mid \mathbf{\Theta} = \boldsymbol{\theta}\big) \overset{a}{\sim} \mathcal{N}\big(\mathbf{0}, \widehat{\mathbf{W}}\big) \quad \text{for large sample sizes } n.$$

For the purpose of multiple imputation, the before given normal approximation is interpreted from a Bayesian perspective as

$$\big(\mathbf{Q} - \widehat{\mathbf{Q}} \mid \mathbf{Y} = \mathbf{y}\big) \overset{a}{\sim} \mathcal{N}\big(\mathbf{0}, \widehat{\mathbf{W}}\big) \quad \text{for large sample sizes } n,$$

implicitly assuming that $\widehat{\mathbf{Q}}$ and $\widehat{\mathbf{W}}$ well approximate the posterior mean $\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{y}]$ and the posterior variance $\mathrm{Var}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{y}]$ of $\mathbf{Q}$ given $\mathbf{Y} = \mathbf{y}$ (Schafer 1997, pp. 107–108).

Within this setup, multiple imputation consists of three parts (Rubin 1987, p. 67):

1. **Imputation**: Generate $M \geq 2$ imputations $\mathbf{Y}_{\text{mis}}^{(m)} := \{Y_{ij}^{(m)} \colon R_{ij} = 0\}_{i=1,\ldots,n,\, j=1,\ldots,d}$, $m = 1, \ldots, M$, from the posterior predictive density $f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}$, which gives M completed data sets $\mathbf{Y}^{(m)} := (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}^{(m)})$, $m = 1, \ldots, M$.
2. **Analysis**: For each $m = 1, \ldots, M$, apply the estimators $\widehat{\mathbf{Q}}$ and $\widehat{\mathbf{W}}$ to the imputed data set $\mathbf{Y}^{(m)}$ giving $\widehat{\mathbf{Q}}^{(m)} := \widehat{\mathbf{Q}}(\mathbf{Y} = \mathbf{Y}^{(m)})$ and $\widehat{\mathbf{W}}^{(m)} := \widehat{\mathbf{W}}(\mathbf{Y} = \mathbf{Y}^{(m)})$.
3. **Pooling**: Combine $\widehat{\mathbf{Q}}^{(1)}, \ldots, \widehat{\mathbf{Q}}^{(M)}$ and $\widehat{\mathbf{W}}^{(1)}, \ldots, \widehat{\mathbf{W}}^{(M)}$ into

   $$\overline{\mathbf{Q}} := \overline{\mathbf{Q}}(\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(M)}) := \frac{1}{M} \sum_{m=1}^{M} \widehat{\mathbf{Q}}^{(m)},$$

   an estimator of the posterior mean $\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}}[\mathbf{Q} \mid \mathbf{Y}_{\text{obs}}]$ of $\mathbf{Q}$ given $\mathbf{Y}_{\text{obs}}$, and

   $$\mathbf{T} := \mathbf{T}(\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(M)}) := \overline{\mathbf{W}} + \mathbf{B} + \frac{\mathbf{B}}{M},$$

an estimator of the posterior variance $\mathrm{Var}_{f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}}[\mathbf{Q} \mid \mathbf{Y}_{\text{obs}}]$ of $\mathbf{Q}$ given $\mathbf{Y}_{\text{obs}}$. Here, the total variance $\mathbf{T}$ is composed of the within-imputation variance $\overline{\mathbf{W}}$,

$$\overline{\mathbf{W}} := \overline{\mathbf{W}}(\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(M)}) := \frac{1}{M} \sum_{m=1}^{M} \widehat{\mathbf{W}}^{(m)},$$

and the between-imputation variance $\mathbf{B}$,

$$\mathbf{B} := \mathbf{B}(\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(M)}) := \frac{1}{M-1} \sum_{m=1}^{M} (\widehat{\mathbf{Q}}^{(m)} - \overline{\mathbf{Q}})^\top (\widehat{\mathbf{Q}}^{(m)} - \overline{\mathbf{Q}}).$$

The above pooling rules are also called **Rubin's rules**.

The definition of $\overline{\mathbf{Q}}$ and $\mathbf{T}$ is based on the following Theorem 3.1.1 (Rubin 1987, Results 3.1 and 3.2) and Corollary 3.1.2 (Rubin 1987, Equations (3.2.8) and (3.2.8)). They imply that, if we generate multiple imputations $\mathbf{Y}_{\text{mis}}^{(m)}$, $m = 1, \dots, M$, from the posterior predictive density $f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}$ and if the estimators $\widehat{\mathbf{Q}}$ and $\widehat{\mathbf{W}}$ well approximate the posterior mean of $\mathbf{Q}$ given $\mathbf{Y}$ in the complete-data case, one indeed obtains

$$\overline{\mathbf{Q}} \approx \mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}}[\mathbf{Q} \mid \mathbf{Y}_{\text{obs}}] \quad \text{and} \quad \mathbf{T} \approx \mathrm{Var}_{f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}}[\mathbf{Q} \mid \mathbf{Y}_{\text{obs}}].$$

This means that in the situation of incomplete data with $\overline{\mathbf{Q}}$ and $\mathbf{T}$ we can reliable estimate the posterior mean and variance of $\mathbf{Q}$ given the observed data $\mathbf{Y}_{\text{obs}}$.

Theorem 3.1.1. *The following identities hold:*

(a) $f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}(\mathbf{q} \mid \mathbf{y}_{\text{obs}}) = \int f_{\mathbf{Q}|\mathbf{Y}}(\mathbf{q} \mid \mathbf{y}_{\text{mis}}, \mathbf{y}_{\text{obs}}) f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}(\mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}) \, \mathrm{d}\mathbf{y}_{\text{mis}}$,

(b) $\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}}[\mathbf{Q} \mid \mathbf{Y}_{\text{obs}}] = \mathbb{E}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}\big[\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{Y}] \bigm| \mathbf{Y}_{\text{obs}}\big]$ *and*

(c) $\mathrm{Var}_{f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}}[\mathbf{Q}|\mathbf{Y}_{\text{obs}}] = \mathbb{E}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}\big[\mathrm{Var}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q}|\mathbf{Y}] \bigm| \mathbf{Y}_{\text{obs}}\big] + \mathrm{Var}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}\big[\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q}|\mathbf{Y}] \bigm| \mathbf{Y}_{\text{obs}}\big]$.

Proof.

(a) Using the law of total probability it holds that

$$\begin{aligned} f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}(\mathbf{q} \mid \mathbf{y}_{\text{obs}}) &= \frac{f_{\mathbf{Q},\mathbf{Y}_{\text{obs}}}(\mathbf{q}, \mathbf{y}_{\text{obs}})}{f_{\mathbf{Y}_{\text{obs}}}(\mathbf{y}_{\text{obs}})} \\ &= \frac{\int f_{\mathbf{Q},\mathbf{Y}}(\mathbf{q}, \mathbf{y}_{\text{mis}}, \mathbf{y}_{\text{obs}}) \, \mathrm{d}\mathbf{y}_{\text{mis}}}{f_{\mathbf{Y}_{\text{obs}}}(\mathbf{y}_{\text{obs}})} \\ &= \frac{\int f_{\mathbf{Q}|\mathbf{Y}}(\mathbf{q} \mid \mathbf{y}_{\text{mis}}, \mathbf{y}_{\text{obs}}) f_{\mathbf{Y}}(\mathbf{y}_{\text{mis}}, \mathbf{y}_{\text{obs}}) \, \mathrm{d}\mathbf{y}_{\text{mis}}}{f_{\mathbf{Y}_{\text{obs}}}(\mathbf{y}_{\text{obs}})} \\ &= \int f_{\mathbf{Q}|\mathbf{Y}}(\mathbf{q} \mid \mathbf{y}_{\text{mis}}, \mathbf{y}_{\text{obs}}) f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}(\mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}) \, \mathrm{d}\mathbf{y}_{\text{mis}}. \end{aligned}$$

(b) Let $\mathbf{Y}_{\text{obs}} = \mathbf{y}_{\text{obs}}$ be arbitrary but fixed. Using (a) we get

$$\begin{aligned}\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}}[\mathbf{Q} \mid \mathbf{y}_{\text{obs}}] &= \int \mathbf{q} f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}(\mathbf{q} \mid \mathbf{y}_{\text{obs}})\,\mathrm{d}\mathbf{q} \\ &= \int \mathbf{q}\Big(\int f_{\mathbf{Q}|\mathbf{Y}}(\mathbf{q} \mid \mathbf{y}_{\text{mis}}, \mathbf{y}_{\text{obs}}) f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}(\mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}})\Big)\,\mathrm{d}\mathbf{y}_{\text{mis}}\,\mathrm{d}\mathbf{q} \\ &= \int \Big(\int \mathbf{q} f_{\mathbf{Q}|\mathbf{Y}}(\mathbf{q} \mid \mathbf{y}_{\text{mis}}, \mathbf{y}_{\text{obs}})\,\mathrm{d}\mathbf{q}\Big) f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}(\mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}})\,\mathrm{d}\mathbf{y}_{\text{mis}} \\ &= \mathbb{E}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}\big[\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{Y}_{\text{mis}}, \mathbf{y}_{\text{obs}}] \bigm| \mathbf{y}_{\text{obs}}\big].\end{aligned}$$

Since $\mathbf{Y}_{\text{obs}} = \mathbf{y}_{\text{obs}}$ was chosen arbitrary the statement follows.

(c) First we recognize that analogously to the proof of (b) we can derive that

$$\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}}[\mathbf{Q}^\top\mathbf{Q} \mid \mathbf{Y}_{\text{obs}}] = \mathbb{E}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}\big[\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q}^\top\mathbf{Q} \mid \mathbf{Y}] \bigm| \mathbf{Y}_{\text{obs}}\big].$$

by replacing the first $\mathbf{q}$ in the integrand with $\mathbf{q}^\top\mathbf{q}$ (but not the $\mathbf{q}$ in the density). Then, together with the shortcut formula for the variance it follows that we can decompose

$$\begin{aligned}&\mathbb{E}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}\big[\mathrm{Var}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{Y}] \bigm| \mathbf{Y}_{\text{obs}}\big] \\ &= \mathbb{E}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}\big[\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q}^\top\mathbf{Q} \mid \mathbf{Y}] \bigm| \mathbf{Y}_{\text{obs}}\big] \\ &\quad - \mathbb{E}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}\big[\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{Y}]^\top\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{Y}] \bigm| \mathbf{Y}_{\text{obs}}\big] \\ &= \mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}}[\mathbf{Q}^\top\mathbf{Q} \mid \mathbf{Y}_{\text{obs}}] - \mathbb{E}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}\big[\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{Y}]^\top\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{Y}] \bigm| \mathbf{Y}_{\text{obs}}\big].\end{aligned}$$

On the other hand, again using the shortcut formula for the variance, we get the variance decomposition

$$\begin{aligned}&\mathrm{Var}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}\big[\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{Y}] \bigm| \mathbf{Y}_{\text{obs}}\big] \\ &= \mathbb{E}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}\big[\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{Y}]^\top\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{Y}] \bigm| \mathbf{Y}_{\text{obs}}\big] \\ &\quad - \mathbb{E}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}\big[\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{Y}] \bigm| \mathbf{Y}_{\text{obs}}\big]^\top\mathbb{E}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}\big[\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{Y}] \bigm| \mathbf{Y}_{\text{obs}}\big] \\ &= \mathbb{E}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}\big[\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{Y}]^\top\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{Y}] \bigm| \mathbf{Y}_{\text{obs}}\big] \\ &\qquad - \mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}}[\mathbf{Q} \mid \mathbf{Y}_{\text{obs}}]^\top\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}}[\mathbf{Q} \mid \mathbf{Y}_{\text{obs}}].\end{aligned}$$

Summing up the last two equations completes the proof:

$$\begin{aligned}&\mathbb{E}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}\big[\mathrm{Var}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{Y}] \bigm| \mathbf{Y}_{\text{obs}}\big] + \mathrm{Var}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}\big[\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{Y}] \bigm| \mathbf{Y}_{\text{obs}}\big] \\ &= \mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}}[\mathbf{Q}^\top\mathbf{Q} \mid \mathbf{Y}_{\text{obs}}] - \mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}}[\mathbf{Q} \mid \mathbf{Y}_{\text{obs}}]^\top\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}}[\mathbf{Q} \mid \mathbf{Y}_{\text{obs}}] \\ &= \mathrm{Var}_{f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}}[\mathbf{Q} \mid \mathbf{Y}_{\text{obs}}].\end{aligned}$$

□

From the strong law of large numbers it follows that

$$\overline{\mathbf{Q}}_\infty := \overline{\mathbf{Q}}_\infty(\mathbf{Y}_{\text{obs}}) := \lim_{M\to\infty} \overline{\mathbf{Q}} = \lim_{M\to\infty} \frac{1}{M}\sum_{m=1}^{M} \widehat{\mathbf{Q}}^{(m)} = \mathbb{E}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}[\widehat{\mathbf{Q}} \mid \mathbf{Y}_{\text{obs}}], \tag{3.1}$$

$$\overline{\mathbf{W}}_\infty \coloneqq \overline{\mathbf{W}}_\infty(\mathbf{Y}_{\text{obs}}) \coloneqq \lim_{M\to\infty} \overline{\mathbf{W}} = \lim_{M\to\infty} \frac{1}{M}\sum_{m=1}^{M} \widehat{\mathbf{W}}^{(m)} = \mathbb{E}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}[\widehat{\mathbf{W}} \mid \mathbf{Y}_{\text{obs}}], \tag{3.2}$$

$$\begin{aligned}\mathbf{B}_\infty \coloneqq \mathbf{B}_\infty(\mathbf{Y}_{\text{obs}}) &\coloneqq \lim_{M\to\infty} \frac{M-1}{M}\cdot \mathbf{B} \\ &= \lim_{M\to\infty} \frac{1}{M}\sum_{m=1}^{M} (\widehat{\mathbf{Q}}^{(m)} - \overline{\mathbf{Q}})^\top(\widehat{\mathbf{Q}}^{(m)} - \overline{\mathbf{Q}}) \\ &= \operatorname{Var}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}[\widehat{\mathbf{Q}} \mid \mathbf{Y}_{\text{obs}}]. \end{aligned}\tag{3.3}$$

Additionally define

$$\mathbf{T}_\infty \coloneqq \mathbf{T}_\infty(\mathbf{Y}_{\text{obs}}) \coloneqq \overline{\mathbf{W}}_\infty + \mathbf{B}_\infty. \tag{3.4}$$

Corollary 3.1.2. *Suppose* $\widehat{\mathbf{Q}} = \mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{Y}]$ *and* $\widehat{\mathbf{W}} = \operatorname{Var}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{Y}]$*, then*

(a) $\overline{\mathbf{Q}}_\infty = \mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}}[\mathbf{Q} \mid \mathbf{Y}_{\text{obs}}]$ *and*

(b) $\mathbf{T}_\infty = \operatorname{Var}_{f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}}[\mathbf{Q} \mid \mathbf{Y}_{\text{obs}}]$.

Proof.

(a) From Theorem 3.1.1 part (b) we have

$$\overline{\mathbf{Q}}_\infty = \mathbb{E}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}[\widehat{\mathbf{Q}} \mid \mathbf{Y}_{\text{obs}}] = \mathbb{E}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}\big[\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{Y}] \bigm| \mathbf{Y}_{\text{obs}}\big] = \mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}}[\mathbf{Q} \mid \mathbf{Y}_{\text{obs}}].$$

(b) From Theorem 3.1.1 part (c) we have

$$\begin{aligned}\mathbf{T}_\infty &= \overline{\mathbf{W}}_\infty + \mathbf{B}_\infty \\ &= \mathbb{E}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}[\widehat{\mathbf{W}} \mid \mathbf{Y}_{\text{obs}}] + \operatorname{Var}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}[\widehat{\mathbf{Q}} \mid \mathbf{Y}_{\text{obs}}] \\ &= \mathbb{E}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}\big[\operatorname{Var}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{Y}] \bigm| \mathbf{Y}_{\text{obs}}\big] + \operatorname{Var}_{f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}}\big[\mathbb{E}_{f_{\mathbf{Q}|\mathbf{Y}}}[\mathbf{Q} \mid \mathbf{Y}] \bigm| \mathbf{Y}_{\text{obs}}\big] \\ &= \operatorname{Var}_{f_{\mathbf{Q}|\mathbf{Y}_{\text{obs}}}}[\mathbf{Q} \mid \mathbf{Y}_{\text{obs}}]. \end{aligned}$$ □

Given a realized data set $(\mathbf{y}, \mathbf{r})$, under the large sample normal approximation and generating theoretical infinitely many imputations, it then holds that

$$\big(\mathbf{Q} - \overline{\mathbf{Q}}_\infty \bigm| \mathbf{Y}_{\text{obs}} = \mathbf{y}_{\text{obs}}\big) \overset{a}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{T}_\infty) \tag{3.5}$$

for large sample sizes n. Rubin (1987) shows that in the case of a finite number of M realized imputations, i. e. $\mathbf{y}_{\text{mis}}^{(1)}, \dots, \mathbf{y}_{\text{mis}}^{(M)}$, the normal distribution must be replaced by a t-distribution. With a refinement of the degrees of freedom of this t-distribution, Barnard and Rubin (1999) argue that $\overline{\mathbf{Q}}$ approximately fulfills

$$\big(\mathbf{Q} - \overline{\mathbf{Q}} \bigm| \mathbf{Y}_{\text{obs}} = \mathbf{y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}^{(1)} = \mathbf{y}_{\text{mis}}^{(1)}, \dots, \mathbf{Y}_{\text{mis}}^{(M)} = \mathbf{y}_{\text{mis}}^{(M)}\big) \overset{a}{\sim} t_N\big(\mathbf{0}, \mathbf{T}\big) \tag{3.6}$$

for large sample sizes n, with the estimator of the degrees of freedom N defined as

$$N \coloneqq N(\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(M)}) \coloneqq \frac{N' N_{\text{obs}}}{N' + N_{\text{obs}}},$$

where

$$N_{\text{obs}} \coloneqq \frac{\nu_{\text{com}} + 1}{\nu_{\text{com}} + 3} \cdot \nu_{\text{com}}(1 - \Lambda), \quad N' \coloneqq \frac{M-1}{\Lambda^2}, \quad \Lambda \coloneqq \left(1 + \frac{1}{M}\right) \frac{\operatorname{tr}(\mathbf{B}\mathbf{T}^{-1})}{k}$$

and ν_{com} are the degrees of freedom for the estimation of $\mathbf{Q}$ with $\widehat{\mathbf{Q}}$ from the hypothetically complete data set $\mathbf{Y}$. In the original version, Rubin (1987) estimates the degrees of freedom of the t-distribution using N'. From the posterior distributions given in Equations (3.5) and (3.6), point estimates, interval estimates and significance levels can be derived as given in Example 2.1.1.

Example 3.1.3.

(a) Under the model of Equation (3.5) the highest posterior density region with coverage $1 - \alpha \in (0, 1)$ is the set of all $\mathbf{Q}$ such that

$$(\mathbf{Q} - \overline{\mathbf{Q}}_\infty)\mathbf{T}_\infty^{-1}(\mathbf{Q} - \overline{\mathbf{Q}}_\infty)^\top < \chi_k^2(\alpha),$$

where $\chi_k^2(\alpha)$ is the α-quantile of the chi-squared distribution on k degrees of freedom. The p-value of the null value $\mathbf{q}_0$ is given by

$$\mathbb{P}\big(\chi_k^2 > (\mathbf{q}_0 - \overline{\mathbf{Q}}_\infty)\mathbf{T}_\infty^{-1}(\mathbf{q}_0 - \overline{\mathbf{Q}}_\infty)^\top\big),$$

where χ_k^2 is a chi-squared distributed random variable on k degrees of freedom. In the special case $k = 1$, a $100(1 - \alpha)\,\%$ interval estimate of $\mathbf{Q}$ is given as

$$\left[\overline{\mathbf{Q}}_\infty - z\Big(\frac{\alpha}{2}\Big)\sqrt{\mathbf{T}_\infty}, \overline{\mathbf{Q}}_\infty + z\Big(\frac{\alpha}{2}\Big)\sqrt{\mathbf{T}_\infty}\right], \tag{3.7}$$

where $z(\frac{\alpha}{2})$ is the $\frac{\alpha}{2}$-quantile of the standard normal distribution, and the significance level associated with the null value $\mathbf{q}_0$ is given by

$$\mathbb{P}\left(\chi_1^2 > \frac{(\mathbf{q}_0 - \overline{\mathbf{Q}}_\infty)^2}{\mathbf{T}_\infty}\right).$$

(b) Under the model of Equation (3.6) the highest posterior density region with coverage $1 - \alpha \in (0, 1)$ is the set of all $\mathbf{Q}$ such that

$$\frac{(\mathbf{Q} - \overline{\mathbf{Q}})\mathbf{T}^{-1}(\mathbf{Q} - \overline{\mathbf{Q}})^\top}{k} < F_{k,N}(\alpha),$$

where $F_{k,N}(\alpha)$ is the α-quantile of the F-distribution on k and N degrees of freedom. The p-value of the null value $\mathbf{q}_0$ is given by

$$\mathbb{P}\left(F_{k,N} > \frac{(\mathbf{q}_0 - \overline{\mathbf{Q}})\mathbf{T}^{-1}(\mathbf{q}_0 - \overline{\mathbf{Q}})^\top}{k}\right),$$

where $F_{k,N}$ is an F-distributed random variable on k and N degrees of freedom. In the special case $k = 1$, a $100(1-\alpha)\,\%$ interval estimate of $\mathbf{Q}$ is given as

$$\left[\overline{\mathbf{Q}} - t_N\left(\frac{\alpha}{2}\right)\sqrt{\mathbf{T}}, \overline{\mathbf{Q}} + t_N\left(\frac{\alpha}{2}\right)\sqrt{\mathbf{T}}\right], \tag{3.8}$$

where $t_N(\frac{\alpha}{2})$ is the $\frac{\alpha}{2}$-quantile of the t-distribution with N degrees of freedom, and the significance level associated with the null value $\mathbf{q}_0$ is given by

$$\mathbb{P}\left(F_{1,N} > \frac{(\mathbf{q}_0 - \overline{\mathbf{Q}})^2}{\mathbf{T}}\right). \qquad \triangle$$

Remark 3.1.4 (Element-wise multiple imputation procedure)*.* In some situations it is reasonable to determine interval estimates and p-values for the individual quantities of interest collected in $\mathbf{Q} = (Q_1, \dots, Q_k)$ (Buuren 2018, p. 49). In these cases, we need a slightly adopted procedure and notation. Especially, for each $l = 1, \dots, k$ we define a separate complete-data estimator $\widehat{Q}_l := \widehat{Q}_l(\mathbf{Y})$ for Q_l as well as a separate complete-data estimator $\widehat{W}_l := \widehat{W}_l(\mathbf{Y})$ for the variance of $\widehat{Q}_l$ under the sampling density $f_{\mathbf{Y}|\mathbf{\Theta}}$. As before, the multiple imputation procedure consists of three parts:

1. **Imputation**: Generate $M \geq 2$ imputations $\mathbf{Y}_{\text{mis}}^{(m)} := \{Y_{ij}^{(m)} : R_{ij} = 0\}_{i=1,\dots,n,\, j=1,\dots,d}$, $m = 1, \dots, M$, from the posterior predictive density $f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}$, which gives M completed data sets $\mathbf{Y}^{(m)} := (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}^{(m)})$, $m = 1, \dots, M$.

2. **Analysis**: For each $l = 1, \dots, k$ and $m = 1, \dots, M$, apply $\widehat{Q}_l$ and $\widehat{W}_l$ to the imputed data set $\mathbf{Y}^{(m)}$ giving $\widehat{Q}_l^{(m)} := \widehat{Q}_l(\mathbf{Y} = \mathbf{Y}^{(m)})$ and $\widehat{W}_l^{(m)} := \widehat{W}_l(\mathbf{Y} = \mathbf{Y}^{(m)})$.

3. **Pooling**: For each $l = 1, \dots, k$, combine $\widehat{Q}_l^{(1)}, \dots, \widehat{Q}_l^{(M)}$ and $\widehat{W}_l^{(1)}, \dots, \widehat{W}_l^{(M)}$ into

$$\overline{Q}_l := \overline{Q}_l(\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(M)}) := \frac{1}{M}\sum_{m=1}^{M} \widehat{Q}_l^{(m)}, \tag{3.9}$$

an estimator of the posterior mean $\mathbb{E}_{f_{Q_l|\mathbf{Y}_{\text{obs}}}}[Q_l \mid \mathbf{Y}_{\text{obs}}]$ of Q_l given $\mathbf{Y}_{\text{obs}}$, and

$$T_l := T_l(\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(M)}) := \overline{W}_l + B_l + \frac{B_l}{M}, \tag{3.10}$$

an estimator of the posterior variance $\operatorname{Var}_{f_{Q_l|\mathbf{Y}_{\text{obs}}}}[Q_l \mid \mathbf{Y}_{\text{obs}}]$ of Q_l given $\mathbf{Y}_{\text{obs}}$. Here, the total variance T_l is composed of the within-imputation variance $\overline{W}_l$,

$$\overline{W}_l := \overline{W}_l(\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(M)}) := \frac{1}{M}\sum_{m=1}^{M} \widehat{W}_l^{(m)},$$

and the between-imputation variance B_l,

$$B_l \coloneqq B_l(\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(M)}) \coloneqq \frac{1}{M-1} \sum_{m=1}^{M} (\widehat{Q}_l^{(m)} - \overline{Q}_l)^2.$$

Then, given an incomplete data set $(\mathbf{y}, \mathbf{r})$ and realized imputations $\mathbf{y}_{\text{mis}}^{(1)}, \dots, \mathbf{y}_{\text{mis}}^{(M)}$, for each $l = 1, \dots, k$ the quantity of interest Q_l approximately fulfills

$$\left(\sqrt{T_l}(Q_l - \overline{Q}_l) \;\middle|\; \mathbf{Y}_{\text{obs}} = \mathbf{y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}^{(1)} = \mathbf{y}_{\text{mis}}^{(1)}, \dots, \mathbf{Y}_{\text{mis}}^{(M)} = \mathbf{y}_{\text{mis}}^{(M)}\right) \overset{a}{\sim} t_{N_l}(0,1)$$

for large sample sizes n with the estimator of the degrees of freedom N_l defined as

$$N_l \coloneqq N_l(\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(M)}) \coloneqq \frac{N_l' N_{l,\text{obs}}}{N_l' + N_{l,\text{obs}}}, \tag{3.11}$$

where

$$N_{l,\text{obs}} \coloneqq \frac{\nu_{\text{com}} + 1}{\nu_{\text{com}} + 3} \cdot \nu_{\text{com}}(1 - \Lambda_l), \quad N_l' \coloneqq \frac{M-1}{\Lambda_l^2}, \quad \Lambda_l \coloneqq \left(1 + \frac{1}{M}\right) \frac{B_l}{T_l}$$

and ν_{com} are the degrees of freedom for the estimation of $Q_1, \dots, Q_k$ with $\widehat{Q}_1, \dots, \widehat{Q}_k$ from the hypothetically complete data set $\mathbf{Y}$. Interval estimates and p-values are derived analogously to the one-dimensional case in Example 3.1.3. △

3.2 Generating the Imputations

3.2.1 The Original Approach: Joint Models

In the previous section, we have left it open how to generate concrete imputations $\mathbf{Y}_{\text{mis}}^{(m)} = \mathbf{y}_{\text{mis}}^{(m)}$ given a realized incomplete data set $(\mathbf{y}, \mathbf{r})$. In the following, we will further investigate how to sample them from the posterior predictive distribution of $\mathbf{Y}_{\text{mis}}$ given $\mathbf{Y}_{\text{obs}}$. In the original multiple imputation framework, for this purpose Rubin (1987, Chapter 5.2) distinguishes three tasks.

1. **Modeling task**: A density

   $$f_{\mathbf{Y}}(\mathbf{y}) = \int f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y} \mid \boldsymbol{\theta}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$$

   for $\mathbf{Y}$ has to be specified, where $f_{\mathbf{Y}|\boldsymbol{\Theta}}$ denotes the conditional probability of $\mathbf{Y}$ given $\boldsymbol{\Theta}$ and $f_{\boldsymbol{\Theta}}$ denotes the prior density of the parameter vector $\boldsymbol{\Theta}$.

2. **Imputation task**: The posterior predictive density

$$f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}(\mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}) = \int f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}},\boldsymbol{\Theta}}(\mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\theta}) f_{\boldsymbol{\Theta}|\mathbf{Y}_{\text{obs}}}(\boldsymbol{\theta} \mid \mathbf{y}_{\text{obs}}) \, \mathrm{d}\boldsymbol{\theta}$$

   of the missing part $\mathbf{Y}_{\text{mis}}$ of the data given the observed part $\mathbf{Y}_{\text{obs}}$ of the data has to be derived from $f_{\mathbf{Y}}$, where $f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}},\boldsymbol{\Theta}}$ denotes the conditional density of the missing part $\mathbf{Y}_{\text{mis}}$ given the observed part $\mathbf{Y}_{\text{obs}}$ and the parameter vector $\boldsymbol{\Theta}$, and $f_{\boldsymbol{\Theta}|\mathbf{Y}_{\text{obs}}}$ denotes the posterior density of $\boldsymbol{\Theta}$ given $\mathbf{Y}_{\text{obs}}$.

3. **Estimation task**: The posterior density

$$f_{\boldsymbol{\Theta}|\mathbf{Y}_{\text{obs}}}(\boldsymbol{\theta} \mid \mathbf{y}_{\text{obs}}) = \frac{f_{\mathbf{Y}_{\text{obs}}|\boldsymbol{\Theta}}(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}{\int f_{\mathbf{Y}_{\text{obs}}|\boldsymbol{\Theta}}(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}}$$

   of $\boldsymbol{\Theta}$ given $\mathbf{Y}_{\text{obs}}$ has to be derived that random draws can be made from it.

Finally, for $m = 1, \ldots, M$ and given the observed realized values $\mathbf{y}_{\text{obs}}$, the imputation $\mathbf{y}_{\text{mis}}^{(m)}$ can be generated compositionally:

(a) Draw a parameter vector $\boldsymbol{\theta}^{(m)}$ from $f_{\boldsymbol{\Theta}|\mathbf{Y}_{\text{obs}}}(\cdot \mid \mathbf{y}_{\text{obs}})$.

(b) Draw $\mathbf{y}_{\text{mis}}^{(m)}$ from $f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}},\boldsymbol{\Theta}}\big(\cdot \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\theta}^{(m)}\big)$.

This procedure gives M draws from the distribution of $(\mathbf{Y}_{\text{mis}}, \boldsymbol{\Theta} \mid \mathbf{Y}_{\text{obs}} = \mathbf{y}_{\text{obs}})$. By simply ignoring the drawn values of $\boldsymbol{\Theta}$ one obtains M draws from the posterior predictive distribution (Rubin 1987, p. 162).

At this point, one might ask why the posterior distribution of $\mathbf{Q}(\boldsymbol{\Theta})$ is not estimated directly by sampling with the above procedure, i. e. generate $\mathbf{y}_{\text{mis}}^{(m)}$ compositionally and then derive $\mathbf{q}^{(m)}$ as the estimate $\mathbf{Q}(\mathbf{y}^{(m)})$ based on the imputed data set $\mathbf{y}^{(m)} = (\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}^{(m)})$. The reason is that in general it needs a large number of estimates $\mathbf{q}^{(m)}$ to estimate the posterior distribution by sampling even in the case of a normal distribution while the multiple imputation procedure just needs a small number of draws (Little and Rubin 2020, pp. 232–233).

Nevertheless, it is not an easy task to draw from the posterior density $f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}$. One possibility is the **data augmentation** (or **imputation posterior**) method of Tanner and Wong (1987), a Markov chain Monte Carlo approach. Another option is to use simpler methods that approximate draws from $f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}}$. Little and Rubin (2020, pp. 238–241) propose the following alternatives:

(A1) **Improper multiple imputation**: Use an estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\Theta}$, i. e. the maximum likelihood estimate or an easy-to-compute estimate such as that from the complete units, and draw $\mathbf{y}_{\text{mis}}^{(m)}$ from

$$f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}},\boldsymbol{\Theta}}\big(\cdot \mid \mathbf{y}_{\text{obs}}, \widehat{\boldsymbol{\theta}}\big).$$

Rubin (1987, Chapter 4) shows that this approximation does not provide valid frequentist inferences in general, since the uncertainty about $\boldsymbol{\Theta}$ is not propagated and thus calls it improper.

(A2) **Use the posterior distribution of $\mathbf{\Theta}$ given a subset of the data**: We draw $\boldsymbol{\theta}^{(m)}$ from the posterior distribution of $\mathbf{\Theta}$ not given the full observed data $\mathbf{Y}_{\text{obs}}$ but given a subset of $\mathbf{Y}_{\text{obs}}$. For instance, one might use the posterior distribution of $\mathbf{\Theta}$ given the complete cases $\mathbf{Y}^{\text{cc}}$ for which all variables have been observed:

(a) Draw a parameter vector $\widetilde{\boldsymbol{\theta}}^{(m)}$ from $f_{\mathbf{\Theta}|\mathbf{Y}^{\text{cc}}}(\cdot \mid \mathbf{y}^{\text{cc}})$.

(b) Draw $\mathbf{y}_{\text{mis}}^{(m)}$ from $f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}},\mathbf{\Theta}}\Big(\cdot \;\Big|\; \mathbf{y}_{\text{obs}}, \widetilde{\boldsymbol{\theta}}^{(m)}\Big)$.

In contrast to improper multiple imputation, this method does propagate uncertainty about $\mathbf{\Theta}$ but not all available information are used. This can be useful if the posterior distribution of $\mathbf{\Theta}$ has a simple form given only the complete cases.

(A3) **Use the asymptotic normal distribution of the maximum likelihood estimator**: Suppose the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}_{\text{ML}}$ of $\mathbf{\Theta}$ is available, together with an estimate of its large-sample covariance matrix $\text{Var}\big(\widehat{\boldsymbol{\theta}}_{\text{ML}}\big)$. Then $\boldsymbol{\theta}^{(m)}$ can be drawn from the asymptotic normal posterior distribution of $\widehat{\mathbf{\Theta}}_{\text{ML}}$:

(a) Draw a parameter vector $\widetilde{\boldsymbol{\theta}}^{(m)}$ from $\mathcal{N}\big(\widehat{\boldsymbol{\theta}}_{\text{ML}}, \text{Var}\big(\widehat{\boldsymbol{\theta}}_{\text{ML}}\big)\big)$.

(b) Draw $\mathbf{y}_{\text{mis}}^{(m)}$ from $f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}},\mathbf{\Theta}}\Big(\cdot \;\Big|\; \mathbf{y}_{\text{obs}}, \widetilde{\boldsymbol{\theta}}^{(d)}\Big)$.

(A4) **Refining approximate draws using importance sampling**: Assume we are using only an approximating prior density $g_{\mathbf{\Theta}}(\boldsymbol{\theta})$ to sample from $\mathbf{\Theta}$ but the conditional density of $\mathbf{Y}_{\text{mis}}$ given $\mathbf{Y}_{\text{obs}}$ and $\mathbf{\Theta}$ is correctly specified as it is the case in the two methods before. A refinement is obtained by generating a substantial set of draws $\Big\{\widetilde{\boldsymbol{\theta}}^{(m)} : m = 1, \ldots, M^*,\ M^* \gg M\Big\}$ and then selecting a subset of size M from this set with probability of the selection of draw $\widetilde{\boldsymbol{\theta}}^{(m)}$ proportional to

$$w_m \propto \frac{f_{\mathbf{\Theta}}\Big(\widetilde{\boldsymbol{\theta}}^{(m)}\Big) \cdot f_{\mathbf{Y}_{\text{obs}}|\mathbf{\Theta}}\Big(\mathbf{y}_{\text{obs}} \;\Big|\; \widetilde{\boldsymbol{\theta}}^{(m)}\Big)}{g_{\mathbf{\Theta}}\Big(\widetilde{\boldsymbol{\theta}}^{(m)}\Big)}.$$

This method is a version of sampling importance resampling algorithm (see for example Gelfand and Smith 1990).

(A5) **Use maximum likelihood estimates from bootstrapped samples of the incomplete data set**: Let $\big((\mathbf{y}_{i:}, \mathbf{r}_{i:})\big)_{i=1,\ldots,n}$ be the realized incomplete data set expressed as the collection of its n observations. Generate a bootstrapped sample $(\mathbf{y}^{\text{boot},(m)}, \mathbf{r}^{\text{boot},(m)})$ of $(\mathbf{y}, \mathbf{r})$ by sampling n observations from $\{(\mathbf{y}_{i:}, \mathbf{r}_{i:})\}_{i=1,\ldots,n}$ with replacement. Then:

(a) Set $\widetilde{\boldsymbol{\theta}}^{(m)} = \widehat{\boldsymbol{\theta}}_{\text{ML}}\big(\mathbf{y}_{\text{obs}}^{\text{boot},(m)}\big)$.

(b) Draw $\mathbf{y}_{\text{mis}}^{(m)}$ from $f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}},\mathbf{\Theta}}\Big(\cdot \;\Big|\; \mathbf{y}_{\text{obs}}, \widetilde{\boldsymbol{\theta}}^{(m)}\Big)$.

A special situation arises when so-called **monotone missingness** is present in the incomplete data. Monotone missingness is present if we can find a permutation $\sigma\colon \{1,\dots,d\} \to \{1,\dots,d\}$ such that for all $i = 1,\dots,n$ the following holds true (Little and Rubin 2020, p. 11): If there is a $j_1 \in \{1,\dots,d\}$ with $R_{ij_1} = 0$ then $R_{ij_2} = 0$ for all $j_2 \in \{1,\dots,d\}$ with $\sigma(j_2) > \sigma(j_1)$. In this case, the posterior density $f_{\mathbf{\Theta}|\mathbf{Y}_{\text{obs}}}$ may have a simple (or at least simpler) form (see for example Little and Rubin 2020, Chapter 7.4). This fact motivates two more approximation methods (Little and Rubin 2020, p. 239):

(A6) **Use the posterior distribution of $\mathbf{\Theta}$ given a subset of the data with monotone missingness:** We draw $\boldsymbol{\theta}^{(m)}$ from the posterior distribution of $\mathbf{\Theta}$ given the observed part $\mathbf{Y}_{\text{obs}}^{\text{mm}}$ of a subset $\mathbf{Y}^{\text{mm}} \subseteq \mathbf{Y}$ close to the full data with monotone missingness:

(a) Draw a parameter vector $\widetilde{\boldsymbol{\theta}}^{(m)}$ from $f_{\mathbf{\Theta}|\mathbf{Y}_{\text{obs}}^{\text{mm}}}(\cdot \mid \mathbf{y}_{\text{obs}}^{\text{mm}})$.

(b) Draw $\mathbf{y}_{\text{mis}}^{(m)}$ from $f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}},\mathbf{\Theta}}\Big(\cdot \,\Big|\, \mathbf{y}_{\text{obs}}, \widetilde{\boldsymbol{\theta}}^{(m)}\Big)$.

(A7) **Filling in data to create monotone missingness**: If monotone missingness is destroyed by only a small number of observations impute some values by single imputation to gain monotone missingness. Draw $\boldsymbol{\theta}^{(m)}$ from the posterior distribution of $\mathbf{\Theta}$ given the observed part $\mathbf{Y}_{\text{obs}}^{\text{aug-mm}}$ of this augmented data set $\mathbf{Y}^{\text{aug-mm}}$:

(a) Draw a parameter vector $\widetilde{\boldsymbol{\theta}}^{(m)}$ from $f_{\mathbf{\Theta}|\mathbf{Y}_{\text{obs}}^{\text{aug-mm}}}(\cdot \mid \mathbf{y}_{\text{obs}}^{\text{aug-mm}})$.

(b) Draw $\mathbf{y}_{\text{mis}}^{(m)}$ from $f_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}},\mathbf{\Theta}}\Big(\cdot \,\Big|\, \mathbf{y}_{\text{obs}}, \widetilde{\boldsymbol{\theta}}^{(m)}\Big)$.

So far, in the presented original approach and its approximations the joint probability $f_{\mathbf{Y}}$ formed the overall starting point for the generation of the imputations. Imputation methods which formulate such a complete joint probability model to generate the imputations from we will refer to as methods using the **joint model approach**.

3.2.2 Fully Conditional Specification Approach and Extensions

One point of criticism that is repeatedly leveled at the joint model approach is the lack of flexibility to account for important features of the data (see for instance Buuren 2007, p. 222). To overcome these limitations, another widely used approach has been developed that bypasses the modeling of a joint distribution, the so-called **fully conditional specification approach**. It is based on the core idea to generate the imputations variable-by-variable. More precise, if $\mathbf{Y}_{:j^c} := (\mathbf{Y}_{:1},\dots,\mathbf{Y}_{:j-1},\mathbf{Y}_{:j+1},\dots,\mathbf{Y}_{:d})$ denotes the data set $\mathbf{Y}$ with removed jth column $\mathbf{Y}_{:j}$, for each $j = 1,\dots,d$ a posterior predictive distribution for the missing part $\mathbf{Y}_{:j,\text{mis}}$ of column $\mathbf{Y}_{:j}$ given the observed part $\mathbf{Y}_{:j,\text{obs}}$ of column $\mathbf{Y}_{:j}$ and the remaining data $\mathbf{Y}_{:j^c}$ is defined, here represented by the density

$$\begin{aligned} &f_{\mathbf{Y}_{:j,\text{mis}}|\mathbf{Y}_{:j,\text{obs}},\mathbf{Y}_{:j^c}}(\mathbf{y}_{:j,\text{mis}} \mid \mathbf{y}_{:j,\text{obs}}, \mathbf{y}_{:j^c}) \\ &\quad = \int f_{\mathbf{Y}_{:j,\text{mis}}|\mathbf{Y}_{:j,\text{obs}},\mathbf{Y}_{:j^c},\mathbf{\Theta}_j}(\mathbf{y}_{:j,\text{mis}} \mid \mathbf{y}_{:j,\text{obs}}, \mathbf{y}_{:j^c}, \boldsymbol{\theta}_j) f_{\mathbf{\Theta}_j|\mathbf{Y}_{:j,\text{obs}},\mathbf{Y}_{:j^c}}(\boldsymbol{\theta}_j \mid \mathbf{y}_{:j,\text{obs}}, \mathbf{y}_{:j^c})\,\mathrm{d}\boldsymbol{\theta}_j, \end{aligned}$$

which is based on a random parameter vector $\boldsymbol{\Theta}_j$. Also, similarly to the estimation task of the joint model approach, for each $l = 1, \dots, k$ the posterior density

$$f_{\boldsymbol{\Theta}_j|\mathbf{Y}_{:j,\mathrm{obs}},\mathbf{Y}_{:j^c}}(\boldsymbol{\theta}_j \mid \mathbf{y}_{:j,\mathrm{obs}}, \mathbf{y}_{:j^c}) = \frac{f_{\mathbf{Y}_{:j,\mathrm{obs}},\mathbf{Y}_{:j^c}|\boldsymbol{\Theta}_j}(\mathbf{y}_{:j,\mathrm{obs}}, \mathbf{y}_{:j^c} \mid \boldsymbol{\theta}_j) f_{\boldsymbol{\Theta}_j}(\boldsymbol{\theta}_j)}{\int f_{\mathbf{Y}_{:j,\mathrm{obs}},\mathbf{Y}_{:j^c}|\boldsymbol{\Theta}_j}(\mathbf{y}_{:j,\mathrm{obs}}, \mathbf{y}_{:j^c} \mid \boldsymbol{\theta}_j) f_{\boldsymbol{\Theta}_j}(\boldsymbol{\theta}_j)\, \mathrm{d}\boldsymbol{\theta}_j}$$

of $\boldsymbol{\Theta}_j$ given $\mathbf{Y}_{:j,\mathrm{obs}}$ and $\mathbf{Y}_{:j^c}$ has to be derived that random draws can be made from it.

Then, an imputation $\mathbf{y}_{\mathrm{mis}}^{(m)}$ is generated by iteratively sampling from the posterior predictive densities $f_{\mathbf{Y}_{:j,\mathrm{mis}}|\mathbf{Y}_{:j,\mathrm{obs}},\mathbf{Y}_{:j^c}}$. Similarly to the joint model approach, each draw from a posterior predictive density $f_{\mathbf{Y}_{:j,\mathrm{mis}}|\mathbf{Y}_{:j,\mathrm{obs}},\mathbf{Y}_{:j^c}}$ is composed of a draw from the posterior density $f_{\boldsymbol{\Theta}_j|\mathbf{Y}_{:j,\mathrm{obs}},\mathbf{Y}_{:j^c}}$ and the conditional density $f_{\mathbf{Y}_{:j,\mathrm{mis}}|\mathbf{Y}_{:j,\mathrm{obs}},\mathbf{Y}_{:j^c},\boldsymbol{\Theta}_j}$. Taken together, we obtain the following procedure to generate an imputed data set $\mathbf{y}^{(m)} = (\mathbf{y}_{\mathrm{obs}}, \mathbf{y}_{\mathrm{mis}}^{(m)})$ (Little and Rubin 2020, p. 242; Buuren and Groothuis-Oudshoorn 2011, pp. 6–7).

1. Complete the observed data $\mathbf{y}_{\mathrm{obs}}$ with plausible values $\mathbf{y}_{\mathrm{mis}}^{(m,0)}$ to generate an initial complete data set $\mathbf{y}^{(m,0)} = (\mathbf{y}_{\mathrm{obs}}, \mathbf{y}_{\mathrm{mis}}^{(m,0)})$. One possibility is to draw $\mathbf{y}_{\mathrm{mis}}^{(m,0)}$ from the observed marginals (Buuren and Groothuis-Oudshoorn 2011, p. 6).

2. Run a **(pseudo-)Gibbs sampler** for T iterations. That is the tth iteration, draws

$$\begin{aligned}
\boldsymbol{\theta}_1^{(m,t)} \quad &\text{from} \quad f_{\boldsymbol{\Theta}_1|\mathbf{Y}_{:1,\mathrm{obs}},\mathbf{Y}_{:1^c}}\Big(\cdot \,\Big|\, \mathbf{y}_{:1,\mathrm{obs}}, \mathbf{y}_{:2}^{(m,t-1)}, \dots, \mathbf{y}_{:d}^{(m,t-1)}\Big),\\
\mathbf{y}_{:1,\mathrm{mis}}^{(m,t)} \quad &\text{from} \quad f_{\mathbf{Y}_{:1,\mathrm{mis}}|\mathbf{Y}_{:1,\mathrm{obs}},\mathbf{Y}_{:1^c},\boldsymbol{\Theta}_1}\Big(\cdot \,\Big|\, \mathbf{y}_{:1,\mathrm{obs}}, \mathbf{y}_{:2}^{(m,t-1)}, \dots, \mathbf{y}_{:d}^{(m,t-1)}, \boldsymbol{\theta}_1^{(m,t)}\Big),\\
&\vdots\\
\boldsymbol{\theta}_j^{(m,t)} \quad &\text{from} \quad f_{\boldsymbol{\Theta}_j|\mathbf{Y}_{:j,\mathrm{obs}},\mathbf{Y}_{:j^c}}\Big(\cdot \,\Big|\, \mathbf{y}_{:j,\mathrm{obs}}, \mathbf{y}_{:1}^{(m,t)}, \dots, \mathbf{y}_{:j-1}^{(m,t)}, \mathbf{y}_{:j+1}^{(m,t-1)}, \dots, \mathbf{y}_{:d}^{(m,t-1)}\Big),\\
\mathbf{y}_{:j,\mathrm{mis}}^{(m,t)} \quad &\text{from} \quad f_{\mathbf{Y}_{:j,\mathrm{mis}}|\mathbf{Y}_{:j,\mathrm{obs}},\mathbf{Y}_{:j^c},\boldsymbol{\Theta}_j}\Big(\cdot \,\Big|\, \mathbf{y}_{:j,\mathrm{obs}}, \mathbf{y}_{:1}^{(m,t)}, \dots, \mathbf{y}_{:j-1}^{(m,t)},\\
&\qquad\qquad \mathbf{y}_{:j+1}^{(m,t-1)}, \dots, \mathbf{y}_{:d}^{(m,t-1)}, \boldsymbol{\theta}_j^{(m,t)}\Big),\\
&\vdots\\
\boldsymbol{\theta}_d^{(m,t)} \quad &\text{from} \quad f_{\boldsymbol{\Theta}_d|\mathbf{Y}_{:d,\mathrm{obs}},\mathbf{Y}_{:d^c}}\Big(\cdot \,\Big|\, \mathbf{y}_{:d,\mathrm{obs}}, \mathbf{y}_{:1}^{(m,t)}, \dots, \mathbf{y}_{:d-1}^{(m,t)}\Big),\\
\mathbf{y}_{:d,\mathrm{mis}}^{(m,t)} \quad &\text{from} \quad f_{\mathbf{Y}_{:d,\mathrm{mis}}|\mathbf{Y}_{:d,\mathrm{obs}},\mathbf{Y}_{:d^c},\boldsymbol{\Theta}_d}\Big(\cdot \,\Big|\, \mathbf{y}_{:d,\mathrm{obs}}, \mathbf{y}_{:1}^{(m,t)}, \dots, \mathbf{y}_{:d-1}^{(m,t)}, \boldsymbol{\theta}_d^{(m,t)}\Big),
\end{aligned}$$

   where $\mathbf{y}_{:j}^{(m,t)}$ is the jth column of the mth imputation in iteration t.

The imputed data set $\mathbf{y}^{(m)}$ is then composed as

$$\mathbf{y}^{(m)} = \Big(\mathbf{y}_{\mathrm{obs}}, \mathbf{y}_{:1,\mathrm{mis}}^{(m,T)}, \dots, \mathbf{y}_{:d,\mathrm{mis}}^{(m,T)}\Big).$$

To obtain M imputed data sets we run M independent chains. Buuren and Groothuis-Oudshoorn (2011, p. 7) note that from their experience the number of iterations T can often be chosen small in the range of 10 to 20.

Remark 3.2.1.

(a) The approximations (A1) to (A7) which we presented for the sampling from $f_{\mathbf{Y}_{\mathrm{mis}}|\mathbf{Y}_{\mathrm{obs}}}$ can also be applied to the sampling from $f_{\mathbf{Y}_{:j,\mathrm{mis}}|\mathbf{Y}_{:j,\mathrm{obs}},\mathbf{Y}_{:j^c}}$.

(b) The fully conditional specification approach is also known under a variety of other names (Buuren 2007, p. 227): stochastic relaxation, variable-by-variable imputation, regression switching, sequential regressions, ordered pseudo-Gibbs sampler, partially incompatible Markov chain Monte Carlo, iterated univariate imputation, chained equations. △

The fully conditional specification approach is more flexible than the joint model approach, but the conditional models taken together – even if each of it is a completely specified probability model – do often not correspond to a proper joint distribution for $\mathbf{Y}$ (Murray 2018, p. 152). A compromise between those two approaches is to formulate a joint model via a sequential approach motivated by the relationship

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{Y}_{:1}}(\mathbf{y}_{:1}) \cdot f_{\mathbf{Y}_{:2}|\mathbf{Y}_{:1}}(\mathbf{y}_{:2} \mid \mathbf{y}_{:1}) \cdot \ldots \cdot f_{\mathbf{Y}_{:d}|\mathbf{Y}_{:1},\ldots,\mathbf{Y}_{:d-1}}(\mathbf{y}_{:d} \mid \mathbf{y}_{:1}, \ldots, \mathbf{y}_{:d-1}).$$

Here, if the conditional models are proper probability models, they result in a proper joint probability model. The substantial question in this approach is how to choose the order of the variables in the sequence since different orderings will in general lead to different joint distributions (Murray 2018, p. 153). Thus, the sequential approach is particularly reasonable if the missing data pattern is monotone because then the question about the order of the variables is obsolete.

In the remainder, we will refer to procedures which generate imputations for $\mathbf{Y}_{\mathrm{mis}}$ given $\mathbf{Y}_{\mathrm{obs}}$ generally as **(multivariate) imputation methods**. If missingness occurs only in one column $\mathbf{Y}_{:j}$, we refer to procedures which generate imputations for $\mathbf{Y}_{:j,\mathrm{mis}}$ given $\mathbf{Y}_{:j,\mathrm{obs}}$ and $\mathbf{Y}_{:j^c}$ as **univariate imputation methods**. It is obvious that every multivariate imputation method can also be used as a univariate imputation method.

3.3 Literature Review of Imputation Methods and their Implementation in the R Programming Language

3.3.1 Univariate Imputation Methods

As seen in the previous section, univariate imputation methods form the basis for the fully conditional specification approach by combining them during the quasi-Gibbs sampling. Therefore, in what follows, we will review typical univariate imputation methods applied in the the fully conditional specification approach. Most of them are implemented in the R-packages `mice` (Buuren and Groothuis-Oudshoorn 2011) and `mi` (Su et al. 2011) which provide multiple imputation based on the fully conditional specification approach. We therefore refrain from listing all individual R-functions of the individual methods. Essentially, two approaches are widely used in practice to form univariate imputation methods: sampling from (approximated) Bayesian generalized linear models and using nearest neighbor methods.

(Approximated) Bayesian Generalized Linear Models

For imputing **continuous variables**, Rubin (1987, pp. 166–167) proposes to draw the imputations from a linear regression model where the model parameters follow the Jeffreys prior. While this approach is fully Bayesian, Buuren (2018, p. 67) just takes the maximum likelihood estimates as regression parameters which corresponds to approximation (A1) from Section 3.2.1. Heitjan and Little (1991), however, expand this procedure by using a bootstrapped sample of the original data for each imputation to estimate the regression parameters from. This corresponds to approximation (A5) from Section 3.2.1. Gelman, Jakulin, et al. (2008) impute **dichotomous** variables by sampling from a logistic regression model where the model parameters are estimated via maximum-a-posterior placing Student-t priors on the regression coefficients. This corresponds to approximation (A1) from Section 3.2.1. Rubin (1987, pp. 169–170) proposes a logistic regression model where the posterior distribution of the model parameters equals the asymptotic normal distribution of the maximum likelihood estimates which corresponds to approximation (A3) from Section 3.2.1. This method can easily be extended to **polytomous case**, i. e. with the multinomial logit model (Brand 1999, Appendix 4.B) or the ordered logit (also proportinal odds) model (Buuren 2018, p. 88). Albert and Chib (1993) use a fully Bayesian (polytomous) probit regression model together with a normal prior for the regression coefficients. For **count data**, Raghunathan et al. (2001) propose the Poisson regression model where the posterior distribution of the parameters is approximated by the asymptotic normal distribution of the maximum likelihood estimates. This corresponds to approximation (A3) from Section 3.2.1. For **semi-continuous variables**, Rubin (1987, p. 180) presents a two stage approach combining the logistic regression model with a multivariate normal approximation of the parameters and the linear regression model with the Jeffreys prior from above. Su et al. (2011) extend the approach of Gelman, Jakulin, et al. (2008) for other generalized linear models.

Nearest Neighbor Methods / Hot Deck Methods

Nearest neighbor methods (or hot deck methods) are typically based on an implicit probability model. In general, nearest neighbor methods impute missing values by sampling from a pool of already observed values of “nearby” complete observations (Murray 2018, p. 148), called donors. To find the nearby observations, the nearest neighbor methods define a distance metric. A well known nearest neighbor method is the **predictive mean matching** (Rubin 1987; Heitjan and Little 1991; Schenker and Taylor 1996). Here, the distance between two observations is defined as the difference of the predictive means from a linear regression model. Murray (2018, p. 148) mentions that instead of the traditional linear regression other methods could be used to make this prediction.

Other Methods

The **Bayesian bootstrap** (Rubin 1981) and the **approximate Bayesian bootstrap** (Rubin and Schenker 1986) are Bayesian resampling methods. The Bayesian bootstrap samples the imputations with probabilities drawn from a Dirichlet distribution. In the approximate Bayesian bootstrap the Dirichlet distribution is replaced by a scalded multinomial distribution. The **local residual draw** proposed by (Rubin 1987; Schenker and Taylor 1996) mixes the classical prediction of a linear regression model and the predictive mean matching. Each missing value is imputed by the sum of its predicted mean and an error term where the error term gets sampled from the residuals of the predicted means and the observed values of the donors. The pool of donors is determined in the same way as for predictive mean matching. In addition, machine learning algorithms can be used. Reiter (2005) and Burgette and Reiter (2010) propose multiple imputation based on **classification and regression trees** (Breiman et al. 1984). For a detailed treatment of these and other univariate imputation methods, see Buuren (2018, Chapter 3).

The standard multiple imputation function `mice()` of the `mice` R-package (Buuren and Groothuis-Oudshoorn 2011) uses the predictive mean matching for numeric variables, logistic regression models for binary variables, the multinomial logit model for unordered categorical data and the ordered logit model for ordered categorical data. For all methods, the parameter uncertainty in the regression parameters is generated by sampling from the asymptotic normal distribution of the maximum likelihood estimates which corresponds to approximation (A3) from Section 3.2.1.

3.3.2 Multivariate Imputation Methods

Methods Based on the Multivariate Normal Distribution

There exist several methods based on the multivariate normal distribution. Schafer (1997, chapter 5) developed a fully Bayesian method putting the normal inverted-Wishart prior on the parameters of the multivariate normal distribution. This approach is implemented in the R-package `norm` (Novo and Schafer 2023). King et al. (2001) approximate the uncertainty in the model parameters applying the asymptotic normal distribution of the maximum likelihood estimates as well as importance sampling as described in the approximations (A3) and (A4) of Section 3.2.1. In contrast, Honaker and King (2010) are using maximum likelihood estimates of the parameters based on bootstrapped samples corresponding to approximation (A5) of Section 3.2.1. This method is implemented in the R-package `Amelia` (Honaker, King, and Blackwell 2011). For high dimensional data, the fully Bayesian approach of Audigier, Husson, and Josse (2016) is based on a principal component analysis in conjunction with a normal prior.

Methods Based on Other Parametric Distributions

Often methods based on the multivariate normal distribution are also used for discrete variables by just rounding the continuous imputations based on some threshold. Nevertheless, it is more preferable to use methods with appropriate probability models for the different types of variables. For purely discrete variables **log-linear models** (Schafer 1997, Chapter 8) or **latent class models** (Vermunt et al. 2008; Gebregziabher and DeSantis 2010) have been used for multiple imputation. To generate imputations for mixed continuous and discrete variables the **general location model** (Olkin and Tate 1961; Little and Schluchter 1985; Schafer 1997, Chapter 9.2) and extensions of it (Liu and Rubin 1998) can be applied. For an overview of other multivariate parametric probability models used in imputation methods, see Murray (2018, p. 150).

Bayesian Non-Parametric Methods

While the methods mentioned so far are based on parametric probability models, there also exist methods exploiting Bayesian non-parametric models: Paddock (2002) presents a non-parametric approach based on **Pólia trees**. For categorical data, Si and Reiter (2013) and Manrique-Vallier and Reiter (2014) make use of the **Dirichlet process mixture of product multinomials**. For continuous data, truncated **Dirichlet process mixture of multivariate normal distributions** can be used (Kim et al. 2014). Combining the latter two models it is even possible to impute mixed, continuous and categorical variables (Murray and Reiter 2016; DeYoreo, Reiter, and Hillygus 2017). The method of Murray and Reiter (2016) is implemented in the R-package `MixedDataImpute`. More Bayesian non-parametric methods can be found in Murray (2018, pp. 150–151).

Methods Based on Copulas

Recently, methods for the multiple imputation of continuous variables have been developed where the joint distributions are modeled via copulas. Di Lascio, Giannerini, and Reale (2015) consider **multivariate Gaussian, t-, Gumbel, Clayton and Frank copulas**. The estimation of the margins and the copula is done by a inference for margins approach using all the observed data. The imputations are sampled from the respective conditional distributions which are derived for each of the appearing response patterns via Bayes rule. This method is implemented in the `CoImp` R-package (Di Lascio, Gatto, and Giannerini 2024) with an extension to the **rotated Gumbel copula and a non-parametric copula**. The method also works for exclusively discrete variables. A fully Bayesian approach is presented by Hollenbach et al. (2021) who consider joint models based on **Gaussian copulas** and the inverted-Wishart prior. The sampling of the imputations is done by a Gibbs sampler. This method can even handle continuous and discrete variables at the same time making use of the rank likelihood. Hasler, Craiu, and Rivest (2018) use a **D-vine copula** construction for the multiple imputation of missing data with monotone missing patterns where all variables are continuous. They use a semi-parametric approach for the estimation of the margins and the copula parameters. Chapon, Ouarda, and Hamdi (2023) also use **D-vine copulas** to model the underlying joint distributions. The selection of the involved pair-copulas and the adjustment of their parameters is done jointly via an reversible jump Markov chain Monte Carlo algorithm based on the work of Min and Czado (2011) which samples of the posterior distribution. The imputations are sampled from the conditional distributions available from the D-vine but get additionally disturbed by a normally distributed error.

3.4 Randomization Valid Inference

Even if the theory in Section 3.1 presents a Bayesian approach for deriving point estimates, interval estimates and p-values for $\mathbf{Q}$, it is common to evaluate the performance of imputation methods from a frequentist perspective. This change of perspective is reasonable because, under weak conditions, Bayesian inference corresponds to frequentist inference and vice verse (Rubin 1987, Section 2.10). To this end, let the parameter vectors $\boldsymbol{\Theta} = \boldsymbol{\theta}$ and $\boldsymbol{\Phi} = \boldsymbol{\phi}$ be fixed at their true values. In this situation we also know the true value $\mathbf{q} := \mathbf{Q}(\boldsymbol{\theta})$ of $\mathbf{Q}$ since we chose the quantity of interest to be a function of $\boldsymbol{\Theta}$. Following Rubin (1987), an imputation method is suitable if the resulting repeated-imputation inference is **randomization valid** for $\mathbf{Q} = \mathbf{q}$. Rubin (1987) formulates randomization validity in the following way.

Definition 3.4.1 (Randomization validity). Let $\boldsymbol{\Theta} = \boldsymbol{\theta}$ and $\boldsymbol{\Phi} = \boldsymbol{\phi}$ be fixed at their true values and let the estimators $\overline{\mathbf{Q}}_\infty$ and $\mathbf{T}_\infty$ be defined as in Equations (3.1) and (3.4). The repeated-imputation inference is randomization valid for $\mathbf{Q} = \mathbf{q}$ if, under the probability model defined by $f_{\mathbf{Y},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}(\cdot,\cdot \mid \boldsymbol{\theta},\boldsymbol{\phi})$, the estimator $\overline{\mathbf{Q}}_\infty$ is normally distributed with mean

$$\mathbb{E}_{f_{\mathbf{Y},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}}[\overline{\mathbf{Q}}_\infty \mid \boldsymbol{\theta},\boldsymbol{\phi}] = \mathbf{q},$$

and the variance of $\overline{\mathbf{Q}}_\infty$ is well approximated by the estimator $\mathbf{T}_\infty$ in the sense that

$$\mathbb{E}_{f_{\mathbf{Y},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}}[\mathbf{T}_\infty \mid \boldsymbol{\theta},\boldsymbol{\phi}] = \mathrm{Var}_{f_{\mathbf{Y},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}}[\overline{\mathbf{Q}}_\infty \mid \boldsymbol{\theta},\boldsymbol{\phi}]. \qquad \triangle$$

Remark 3.4.2. Rubin (1987, Chapter 4.2) and Rubin (1996) further divide randomization validity of $\overline{\mathbf{Q}}_\infty$ into two conditions: (a) Randomization validity of $\widehat{\mathbf{Q}}$ and (b) Properness of the imputation model. The first condition holds – under the assumption of asymptotic normality – if

$$\begin{aligned}\mathbb{E}_{f_{\mathbf{Y}|\boldsymbol{\Theta}}}[\widehat{\mathbf{Q}} \mid \boldsymbol{\theta}] &= \mathbf{q},\\ \mathbb{E}_{f_{\mathbf{Y}|\boldsymbol{\Theta}}}[\widehat{\mathbf{W}} \mid \boldsymbol{\theta}] &= \mathrm{Var}_{f_{\mathbf{Y}|\boldsymbol{\Theta}}}[\widehat{\mathbf{Q}} \mid \boldsymbol{\theta}].\end{aligned}$$

On the other hand, omitting some other but technical conditions that are generally not relevant in practice (Rubin 1996), the imputation model is **proper** if

$$\mathbb{E}_{f_{\mathbf{R}|\mathbf{Y},\boldsymbol{\Phi}}}[\overline{\mathbf{Q}}_\infty \mid \mathbf{Y},\boldsymbol{\phi}] = \widehat{\mathbf{Q}},$$

$$\begin{aligned}\mathbb{E}_{f_{\mathbf{R}|\mathbf{Y},\boldsymbol{\Phi}}}[\overline{\mathbf{W}}_\infty \mid \mathbf{Y}, \boldsymbol{\phi}] &= \widehat{\mathbf{W}},\\ \mathbb{E}_{f_{\mathbf{R}|\mathbf{Y},\boldsymbol{\Phi}}}[\mathbf{B}_\infty \mid \mathbf{Y}, \boldsymbol{\phi}] &= \mathrm{Var}_{f_{\mathbf{R}|\mathbf{Y},\boldsymbol{\Phi}}}[\overline{\mathbf{Q}}_\infty \mid \mathbf{Y}, \boldsymbol{\phi}],\end{aligned}$$

where $\overline{\mathbf{Q}}_\infty$, $\overline{\mathbf{W}}_\infty$ and $\mathbf{B}_\infty$ are defined in Equations (3.1), (3.2) and (3.3). If both, randomization validity of $\widehat{\mathbf{Q}}$ and properness of the imputation model are fulfilled, $\overline{\mathbf{Q}}_\infty$ is randomization valid (Rubin 1987, Result 4.1; Rubin 1996). Essentially, it holds that (Rubin 1996, p. 478):

$$\mathbb{E}_{f_{\mathbf{Y},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}}[\overline{\mathbf{Q}}_\infty \mid \boldsymbol{\theta}, \boldsymbol{\phi}] = \mathbb{E}_{f_{\mathbf{Y}|\boldsymbol{\Theta}}}\big[\mathbb{E}_{f_{\mathbf{R}|\mathbf{Y},\boldsymbol{\Phi}}}[\overline{\mathbf{Q}}_\infty \mid \mathbf{Y}, \boldsymbol{\phi}] \bigm| \boldsymbol{\theta}\big] = \mathbb{E}_{f_{\mathbf{Y}|\boldsymbol{\Theta}}}[\widehat{\mathbf{Q}} \mid \boldsymbol{\theta}] = \mathbf{q}$$

and

$$\begin{aligned}&\mathbb{E}_{f_{\mathbf{Y},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}}[\mathbf{T}_\infty \mid \boldsymbol{\theta}, \boldsymbol{\phi}]\\ &\overset{(3.4)}{=} \mathbb{E}_{f_{\mathbf{Y},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}}[\overline{\mathbf{W}}_\infty \mid \boldsymbol{\theta}, \boldsymbol{\phi}] + \mathbb{E}_{f_{\mathbf{Y},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}}[\mathbf{B}_\infty \mid \boldsymbol{\theta}, \boldsymbol{\phi}]\\ &= \mathbb{E}_{f_{\mathbf{Y}|\boldsymbol{\Theta}}}\big[\mathbb{E}_{f_{\mathbf{R}|\mathbf{Y},\boldsymbol{\Phi}}}[\overline{\mathbf{W}}_\infty \mid \mathbf{Y}, \boldsymbol{\phi}] \bigm| \boldsymbol{\theta}\big] + \mathbb{E}_{f_{\mathbf{Y}|\boldsymbol{\Theta}}}\big[\mathbb{E}_{f_{\mathbf{R}|\mathbf{Y},\boldsymbol{\Phi}}}[\mathbf{B}_\infty \mid \mathbf{Y}, \boldsymbol{\phi}] \bigm| \boldsymbol{\theta}\big]\\ &= \mathbb{E}_{f_{\mathbf{Y}|\boldsymbol{\Theta}}}[\widehat{\mathbf{W}} \mid \boldsymbol{\theta}] + \mathbb{E}_{f_{\mathbf{Y}|\boldsymbol{\Theta}}}\big[\mathrm{Var}_{f_{\mathbf{R}|\mathbf{Y},\boldsymbol{\Phi}}}[\overline{\mathbf{Q}}_\infty \mid \mathbf{Y}, \boldsymbol{\phi}] \bigm| \boldsymbol{\theta}\big]\\ &= \mathrm{Var}_{f_{\mathbf{Y}|\boldsymbol{\Theta}}}[\widehat{\mathbf{Q}} \mid \boldsymbol{\theta}] + \mathbb{E}_{f_{\mathbf{Y}|\boldsymbol{\Theta}}}\big[\mathrm{Var}_{f_{\mathbf{R}|\mathbf{Y},\boldsymbol{\Phi}}}[\overline{\mathbf{Q}}_\infty \mid \mathbf{Y}, \boldsymbol{\phi}] \bigm| \boldsymbol{\theta}\big]\\ &= \mathrm{Var}_{f_{\mathbf{Y}|\boldsymbol{\Theta}}}\big[\mathbb{E}_{f_{\mathbf{R}|\mathbf{Y},\boldsymbol{\Phi}}}[\overline{\mathbf{Q}}_\infty \mid \mathbf{Y}, \boldsymbol{\phi}] \bigm| \boldsymbol{\theta}\big] + \mathbb{E}_{f_{\mathbf{Y}|\boldsymbol{\Theta}}}\big[\mathrm{Var}_{f_{\mathbf{R}|\mathbf{Y},\boldsymbol{\Phi}}}[\overline{\mathbf{Q}}_\infty \mid \mathbf{Y}, \boldsymbol{\phi}] \bigm| \boldsymbol{\theta}\big]\\ &= \mathrm{Var}_{f_{\mathbf{Y},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}}[\overline{\mathbf{Q}}_\infty \mid \boldsymbol{\theta}, \boldsymbol{\phi}]. \qquad \triangle\end{aligned}$$

If randomization validity holds, then from the frequentist perspective the $100(1-\alpha)\,\%$ interval estimate given in Equation (3.7) is indeed a $100(1-\alpha)\,\%$ confidence interval (Rubin 1987, Section 4.2). Thus, in the case of $k = 1$, a procedure to investigate the randomization validity of a given method for a specific quantity of interest $\mathbf{Q} = \mathbf{q}$ is given as follows (Buuren 2018, Chapter 2.5): Sample repeatedly from $f_{\mathbf{Y},\mathbf{R}|\boldsymbol{\Theta},\boldsymbol{\Phi}}(\cdot, \cdot \mid \boldsymbol{\theta}, \boldsymbol{\phi})$ to generate incomplete data sets. Complete each incomplete data set M times and apply Rubin's rules to get point estimates and interval estimates of a certain level $1-\alpha$ of $\mathbf{q}$ for each originally incomplete data set. If the average point estimates are close to $\mathbf{q}$ and the interval estimates contain $\mathbf{q}$ in roughly $100(1-\alpha)\,\%$ of the cases this would support randomization validity. For $k > 1$, Buuren (2018, Chapter 2.5) suggests to repeat the above procedure element by element applying Remark 3.1.4.

By definition, randomization validity is not a general property of an imputation method but, in theory, potentially has to be verified individually for each data model, each quantity of interest and each concrete parameter realization. Since this is almost impossible, it has become established in practice to test randomization validity only in a few and often simple situations. If in these situations randomization validity does not hold for a specific imputation method it is generally assumed that this imputation is not a suitable imputation method. Conversely, if randomization validity is fulfilled in these situations, this serves as an indication of randomization validity in other, more general situations.

4 Multiple Imputation via Iterated D-Vine Based Quantile Regression

4.1 A Refresher on D-Vines and D-Vine Copulas

A d-dimensional copula C is a multivariate distribution function on the d-dimensional unit hypercube $[0,1]^d$ for which the marginals are uniformly distributed. The importance of copulas is based on the following fundamental representation theorem for multivariate distributions proved by Sklar (1959).

Theorem 4.1.1. *For every random vector $\mathbf{Y} = (Y_1, \dots, Y_d)$ with joint distribution function F and marginal distribution functions $F_1, \dots, F_d$, there exists a d-dimensional copula C, such that*

$$F(\mathbf{y}) = C\big(F_1(y_1), \dots, F_d(y_d)\big). \tag{4.1}$$

In the case of an absolutely continuous $\mathbf{Y}$, the copula in the above decomposition is unique. In addition, its joint density f can then be decomposed similarly into

$$f(\mathbf{y}) = \prod_{j=1}^{d} f_j(y_j) \cdot c\big(F_1(y_1), \dots, F_d(y_d)\big), \tag{4.2}$$

where $c(v_1, \dots, v_d) \coloneqq \frac{\partial^d}{\partial v_1 \cdots \partial v_d} C(v_1, \dots, v_d)$ is the copula density and $f_1, \dots, f_d$ are the marginal densities.

Let us provisionally assume, until stated otherwise, that $\mathbf{Y}$ is absolutely continuous. In Equation (4.1) the **probability integral transform** is applied to the marginals: $U_j \coloneqq F_j(Y_j)$, $j = 1, \dots, d$. Since $\mathbf{Y}$ is absolutely continuous all U_j are uniformly distributed and consequently the joint distribution function of the random vector $\mathbf{U} \coloneqq (U_1, \dots, U_d)$ is the copula C associated with $\mathbf{Y}$. In the following, we denote the index set of $\mathbf{Y}$ and $\mathbf{U}$ with $D \coloneqq \{1, \dots, d\}$. Additionally for an arbitrary $j \in D$ we define $D_{-j} \coloneqq D \setminus \{j\}$.

Notation 4.1.2. Let $j_1, j_2, j \in D$ and $J \subsetneq D$ with $j, j_1, j_2 \notin J$. Then:

(a) The copula associated with the conditional distribution of (Y_{j_1}, Y_{j_2}) given $\mathbf{Y}_J = \mathbf{y}_J$ is denoted by $C_{Y_{j_1}, Y_{j_2}; \mathbf{Y}_J}(\cdot, \cdot; \mathbf{y}_J)$. The corresponding copula density is denoted by $c_{Y_{j_1}, Y_{j_2}; \mathbf{Y}_J}(\cdot, \cdot; \mathbf{y}_J)$. We abbreviate both by $C_{j_1, j_2; J}(\cdot, \cdot; \mathbf{y}_J)$ and $c_{j_1, j_2; J}(\cdot, \cdot; \mathbf{y}_J)$.

(b) The conditional distribution of the random variable Y_j given $\mathbf{Y}_J = \mathbf{y}_J$ is denoted by $F_{Y_j|\mathbf{Y}_J}(\cdot \mid \mathbf{y}_J)$. We abbreviate this by $F_{j|J}(\cdot \mid \mathbf{y}_J)$.

(c) The conditional distribution of the random variable U_j given $\mathbf{U}_J = \mathbf{u}_J$ is denoted by $C_{U_j|\mathbf{U}_J}(\cdot \mid \mathbf{u}_J)$. We abbreviate this by $C_{j|J}(\cdot \mid \mathbf{u}_J)$.

(d) We define the h-functions associated with a pair-copula $C_{j_1,j_2;J}$ as

$$\begin{aligned} h_{j_1|j_2;J}(u_{j_1} \mid u_{j_2}; \mathbf{u}_J) &:= \frac{\partial C_{j_1,j_2;J}(u_{j_1}, u_{j_2}; \mathbf{u}_J)}{\partial u_{j_2}}, \\ h_{j_2|j_1;J}(u_{j_2} \mid u_{j_1}; \mathbf{u}_J) &:= \frac{\partial C_{j_1,j_2;J}(u_{j_1}, u_{j_2}; \mathbf{u}_J)}{\partial u_{j_1}}. \end{aligned} \qquad \triangle$$

Notation 4.1.3. Let $J \subseteq D$ with $|J| = \widetilde{s}$ and let $\mathbf{j} = (j_1, \ldots, j_{\widetilde{s}})$ be an arbitrary permutation of J. For two indices $s_1 \leq s_2 \leq \widetilde{s}$, we define $\mathbf{j}_{s_1:s_2} := \{j_{s_1}, \ldots, j_{s_2}\}$. $\triangle$

The following theorem shows that the copula density in Equation (4.2) can be replaced by a product of exclusively (conditional) bivariate copula densities (Kraus and Czado 2017, Equation (2.1)).

Theorem 4.1.4 (Drawable vine (D-vine) density). *Let $\mathbf{Y} = (Y_1, \ldots, Y_d)$ be an absolutely continuous random vector with joint density f and let $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_d)$ be an arbitrary permutation of the index set D. Then f can be decomposed as*

$$\begin{aligned} f(\mathbf{y}) = \prod_{k=1}^{d} f_{\ell_k}(y_{\ell_k}) \cdot \prod_{i=1}^{d-1} \prod_{j=i+1}^{d} c_{\ell_i,\ell_j;\boldsymbol{\ell}_{i+1:j-1}} \Big(& F_{\ell_i|\boldsymbol{\ell}_{i+1:j-1}}(y_{\ell_i} \mid \mathbf{y}_{\boldsymbol{\ell}_{i+1:j-1}}), \\ & F_{\ell_j|\boldsymbol{\ell}_{i+1:j-1}}(y_{\ell_j} \mid \mathbf{y}_{\boldsymbol{\ell}_{i+1:j-1}}); \mathbf{y}_{\boldsymbol{\ell}_{i+1:j-1}} \Big). \end{aligned}$$

The distribution associated with this decomposition is called a drawable vine (D-vine) with order $\boldsymbol{\ell}$. If all margins are uniform one speaks of a D-vine copula with order $\boldsymbol{\ell}$.

D-vines build a subclass of the more general pair-copula construction called **regular vines (R-vines)** for which Bedford and Cooke (2002) introduced a graphic theoretic representation. A general D-vine in this representation is shown in Figure 4.1. It is worth mentioning that all the conditional distributions $F_{\ell_i|\boldsymbol{\ell}_{i+1:j-1}}$ and $F_{\ell_j|\boldsymbol{\ell}_{i+1:j-1}}$ appearing in the D-vine density can be expressed using only the pair-copulas already specified in the decomposition. This can be reached by applying the following recursion, which was first stated by Joe (1996) and follows directly from the chain rule of differentiation.

Theorem 4.1.5 (Recursion for conditional distribution functions). *Let $\mathbf{Y}$ be absolutely continuous. Additionally, let $J \subsetneq D$, $i \in D \setminus J$, $l \in J$ and $J_{-l} := J \setminus \{l\}$. Then,*

$$\begin{aligned} F_{i|J}(y_i \mid \mathbf{y}_J) &= \frac{\partial C_{i,l;J_{-l}}\big(F_{i|J_{-l}}(y_i \mid \mathbf{y}_{J_{-l}}), F_{l|J_{-l}}(y_l \mid \mathbf{y}_{J_{-l}}); \mathbf{y}_{J_{-l}}\big)}{\partial F_{l|J_{-l}}(y_l \mid \mathbf{y}_{J_{-l}})} \\ &= h_{i|l;J_{-l}}\big(F_{i|J_{-l}}(y_i \mid \mathbf{y}_{J_{-l}}), F_{l|J_{-l}}(y_l \mid \mathbf{y}_{J_{-l}}); \mathbf{y}_{J_{-l}}\big). \end{aligned} \tag{4.3}$$

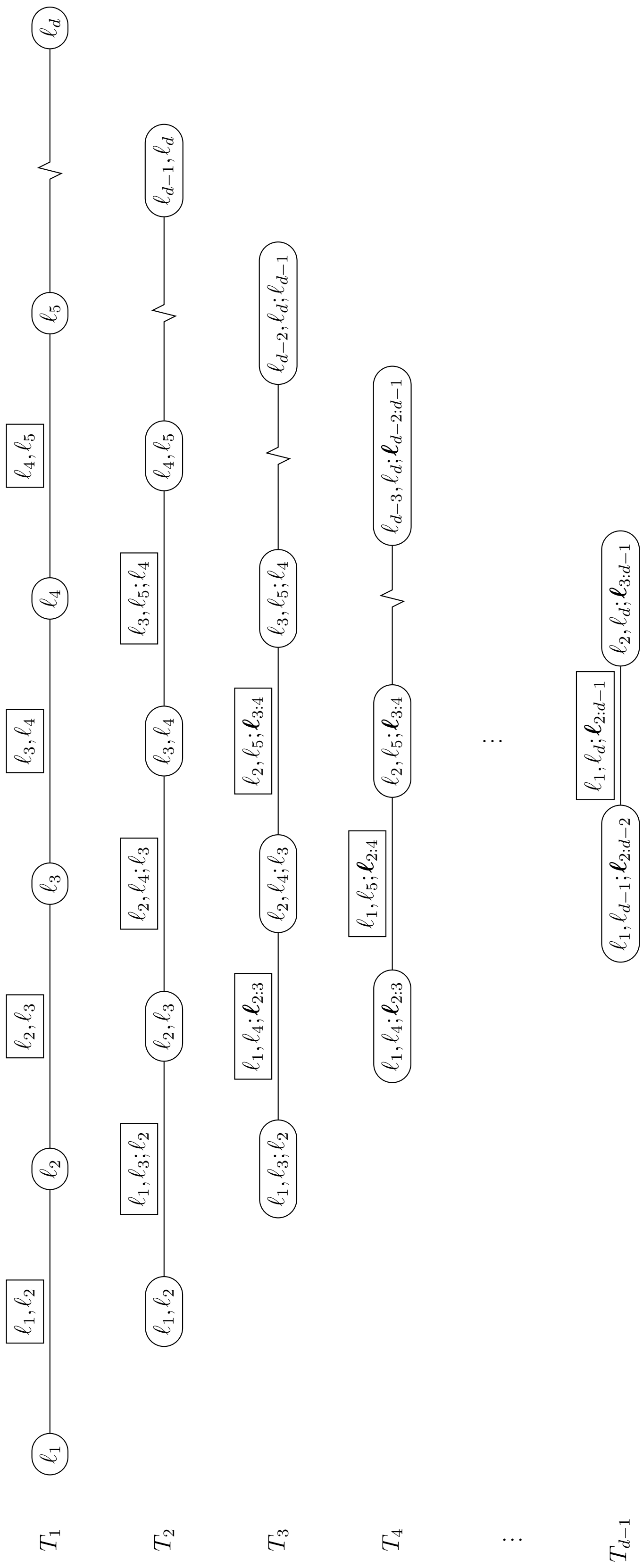

T_1
T_2
T_3
T_4
$\cdots$
T_{d-1}
ℓ_1
ℓ_2
ℓ_3
ℓ_4
ℓ_5
ℓ_d
ℓ_1,ℓ_2
ℓ_2,ℓ_3
ℓ_3,ℓ_4
ℓ_4,ℓ_5
ℓ_{d-1},ℓ_d
$\ell_1,\ell_3;\ell_2$
$\ell_2,\ell_4;\ell_3$
$\ell_3,\ell_5;\ell_4$
$\ell_{d-2},\ell_d;\ell_{d-1}$
$\ell_1,\ell_4;\boldsymbol{\ell}_{2:3}$
$\ell_2,\ell_5;\boldsymbol{\ell}_{3:4}$
$\ell_{d-3},\ell_d;\boldsymbol{\ell}_{d-2:d-1}$
$\ell_1,\ell_5;\boldsymbol{\ell}_{2:4}$
$\cdots$
$\ell_1,\ell_{d-1};\boldsymbol{\ell}_{2:d-2}$
$\ell_2,\ell_d;\boldsymbol{\ell}_{3:d-1}$
$\ell_1,\ell_d;\boldsymbol{\ell}_{2:d-1}$

Figure 4.1: Graph theoretic D-vine representation of a D-vine with order $\boldsymbol{\ell}=(\ell_1,\ldots,\ell_d)$.

When working with D-vines it is common to assume that the copulas $C_{j_1,j_2;J}$ associated with conditional distributions do not depend on the conditioning value $\mathbf{y}_J$, i.e. $C_{j_1,j_2;J}(\cdot,\cdot;\mathbf{y}_J) \equiv C_{j_1,j_2;J}(\cdot,\cdot)$. This assumption is called the **simplifying assumption** (for a further discussion of the simplifying assumption see Czado and Nagler 2022). It also implies that the h-functions are independent of $\mathbf{u}_J$. Especially, it holds that

$$\begin{aligned} h_{j_1|j_2;J}(u_{j_1} \mid u_{j_2}; \mathbf{u}_J) &= \frac{\partial C_{j_1,j_2;J}(u_{j_1}, u_{j_2})}{\partial u_{j_2}} = C_{j_1|j_2;J}(u_{j_1} \mid u_{j_2}), \\ h_{j_2|j_1;J}(u_{j_2} \mid u_{j_1}; \mathbf{u}_J) &= \frac{\partial C_{j_1,j_2;J}(u_{j_1}, u_{j_2})}{\partial u_{j_1}} = C_{j_2|j_1;J}(u_{j_2} \mid u_{j_1}). \end{aligned}$$

Assumption 4.1.6. The simplifying assumption holds. △

4.2 Sampling From Conditional Distributions Using D-Vines Copulas

In the previous section, we have seen how to express the joint distribution F of $\mathbf{Y} = (Y_1,\dots,Y_d)$ in terms of its marginal distribution functions $F_1,\dots,F_d$ and a copula C. Now, for the purpose of multiple imputation based on the fully conditional specification approach, given a fixed but arbitrary $j \in D$, we are primary not interested in F but in the conditional distribution function $F_{j|D_{-j}}$ of Y_j given the remaining variables $\mathbf{Y}_{D_{-j}}$ and how to sample from it during the quasi-Gibbs sampling. Sampling from $F_{j|D_{-j}}$ can be done via the conditional quantile function of Y_j given $\mathbf{Y}_{D_{-j}}$:

$$q_j(\alpha \mid \mathbf{y}_{D_{-j}}) := F^{-1}_{j|D_{-j}}(\alpha \mid \mathbf{y}_{D_{-j}}), \quad \alpha \in (0,1).$$

As shown in Kraus and Czado (2017, p. 3), using the probability integral transforms U_j and $\mathbf{U}_{D_{-j}}$, the conditional distribution function of Y_j given $\mathbf{Y}_{D_{-j}}$ can be expressed in terms of the conditional distribution function of U_j given $\mathbf{U}_{D_{-j}}$:

$$\begin{aligned} F_{j|D_{-j}}(y_j \mid \mathbf{y}_{D_{-j}}) &= \mathbb{P}(Y_j \leq y_j \mid \mathbf{Y}_{D_{-j}} = \mathbf{y}_{D_{-j}}) \\ &= \mathbb{P}(U_j \leq u_j \mid \mathbf{U}_{D_{-j}} = \mathbf{u}_{D_{-j}}) \\ &= C_{j|D_{-j}}(u_j \mid \mathbf{u}_{D_{-j}}). \end{aligned}$$

Inverting the last equation yields an expression for $q_j(\alpha \mid \mathbf{y}_{D_{-j}})$ in terms of the inverse marginal distribution function F_j^{-1} of Y_j and the conditional quantile function $C^{-1}_{j|D_{-j}}$ of U_j given $\mathbf{U}_{D_{-j}}$:

$$q_j(\alpha \mid \mathbf{y}_{D_{-j}}) = F_j^{-1}\Big(C^{-1}_{j|D_{-j}}(\alpha \mid \mathbf{u}_{D_{-j}})\Big).$$

If we construct C as a D-vine copula with order $\boldsymbol{\ell} = (\ell_1,\dots,\ell_d)$ where $\boldsymbol{\ell}$ is an arbitrary permutation of D with $\ell_1 = j$ the conditional distribution function $C_{j|D_{-j}}(u_j \mid \mathbf{u}_{D_{-j}})$

can be derived from the conditional pair-copula $C_{j,\ell_d;\boldsymbol{\ell}_{2:d-1}}(u_j, u_{\ell_d})$ involved in the copula construction. Especially, the recursion given in Equation (4.3) allows to express the conditional distribution function $C_{j|D_{-j}}(u_j \mid \mathbf{u}_{D_{-j}})$ in a closed form as a composition of h-functions. Consequently, the conditional quantile function $C^{-1}_{j|D_{-j}}(\alpha \mid \mathbf{u}_{-j})$ can be expressed in a closed form as a composition of inverse h-functions.

Of course there is no need to predict Y_j by conditioning on all of the remaining variables $\mathbf{Y}_{D_{-j}}$. One can take any subset $\widetilde{D} \subseteq D$ fulfilling $j \in \widetilde{D}$ resulting in the conditional quantile function of Y_j given $\mathbf{Y}_{\widetilde{D}_{-j}}$ as

$$q_j(\alpha \mid \mathbf{y}_{\widetilde{D}_{-j}}) \coloneqq F^{-1}_{j|\widetilde{D}_{-j}}(\alpha \mid \mathbf{y}_{\widetilde{D}_{-j}}) = F_j^{-1}\Big(C^{-1}_{j|\widetilde{D}_{-j}}(\alpha \mid \mathbf{u}_{\widetilde{D}_{-j}})\Big), \tag{4.4}$$

where $\widetilde{D}_{-j} \coloneqq \widetilde{D} \setminus \{j\}$. As before, now constructing a D-vine copula for $\mathbf{U}_{\widetilde{D}}$ with order $\widetilde{\boldsymbol{\ell}} \coloneqq (\widetilde{\ell}_1, \ldots, \widetilde{\ell}_{\widetilde{d}})$ where $\widetilde{\boldsymbol{\ell}}$ is an arbitrary permutation of $\widetilde{D}$ with $\widetilde{\ell}_1 = j$ and $\widetilde{d} \coloneqq |\widetilde{D}|$ is the cardinality of $\widetilde{D}$, we can derive the conditional quantile function $C^{-1}_{j|\widetilde{D}_{-j}}(\alpha \mid \mathbf{u}_{\widetilde{D}_{-j}})$ in a closed form. Since we derive $q_j(\alpha \mid \mathbf{y}_{\widetilde{D}_{-j}})$ from a D-vine copula we also refer to it as a **D-vine quantile regression model** with response Y_j and covariates $\mathbf{Y}_{\widetilde{D}_{-j}}$.

As outlined in Kraus and Czado (2017, p. 3), estimating the marginal distribution functions F_j, $j \in \widetilde{D}$, and the conditional distribution function $C_{j|\widetilde{D}_{-j}}$ gives an estimate of $q_j(\alpha \mid \mathbf{y}_{\widetilde{D}_{-j}})$ by plugging them into Equation (4.4):

$$\widehat{q}_j(\alpha \mid \mathbf{y}_{\widetilde{D}_{-j}}) \coloneqq \widehat{F}_j^{-1}\Big(\widehat{C}^{-1}_{j|\widetilde{D}_{-j}}(\alpha \mid \widehat{\mathbf{u}}_{\widetilde{D}_{-j}})\Big), \tag{4.5}$$

where $\widehat{\mathbf{u}}_{\widetilde{D}_{-j}} \coloneqq (\widehat{u}_j)_{j\in\widetilde{D}_{-j}}$ and $\widehat{u}_j \coloneqq \widehat{F}_j(y_j)$ is the estimated probability integral transform of y_j for all $j \in \widetilde{D}_{-j}$. Further, Kraus and Czado (2017, p. 4) state, that this representation of $\widehat{q}_j(\alpha \mid \mathbf{y}_{\widetilde{D}_{-j}})$ allows to divide the estimation process into two steps: the estimation of the marginal distribution functions and the estimation of the D-vine copula that specifies the pair-copulas needed to evaluate $\widehat{C}^{-1}_{j|\widetilde{D}_{-j}}(\alpha \mid \widehat{\mathbf{u}}_{\widetilde{D}_{-j}})$.

In the following, we present the estimation of a D-vine quantile regression model including a selection algorithm for the appropriate choice of the set $\widetilde{D}$ together with an order $\boldsymbol{\ell}$ to construct the model D-vine copula. We follow the approach of Kraus and Czado (2017) adapted to the situation of missing data given a realized incomplete data set $(\mathbf{y}, \mathbf{r})$ of dimension $n \times d$ where $\mathbf{y} = (y_{ij})_{i=1,\ldots,n,\, j=1,\ldots,d}$.

First Step: Estimation of the Marginals

Kraus and Czado (2017, p. 5) fit the marginal distribution functions using the below defined estimator which is derived from the kernel density estimator (Parzen 1962).

Definition 4.2.1 (Kernel estimator of a univariate cumulative distribution function)**.** Let $(X_i)_{i=1,\ldots,n}$ be an i. i. d. sample of a random variable X. The kernel estimator $\widehat{F}_X(x)$ of the cumulative distribution function $F_X(x)$ of X is defined as

$$\widehat{F}_X(x) \coloneqq \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \quad x \in \mathbb{R},$$

where $K(x) \coloneqq \int_{-\infty}^{x} k(t)\,\mathrm{d}t$ with $k(\cdot)$ being a symmetric probability density function and $h > 0$ a bandwidth parameter. △

In the present situation, for each $j = 1, \dots, d$ we must take into account that the sample $\mathbf{y}_{:j} = (y_{1j}, \dots, y_{nj})^\top$ of Y_j may have missing values. To estimate the marginal distribution function F_j we can only use the observed values $\mathbf{y}_{:j,\mathrm{obs}} = \{y_{ij}\}_{i \in I_{\mathbf{r}}^{\mathrm{obs}}(j)}$. Appropriately adjusting the formula for the kernel estimator we thus obtain the estimate

$$\widehat{F}_j(y_j) \coloneqq \frac{1}{|I_{\mathbf{r}}^{\mathrm{obs}}(j)|} \sum_{i \in I_{\mathbf{r}}^{\mathrm{obs}}(j)} K\left(\frac{y_j - y_{ij}}{h}\right).$$

With the help of the estimates $\widehat{F}_j,\ j = 1, \dots, d$, the data $\mathbf{y}$ is then transformed to pseudo copula data $\widehat{\mathbf{u}} = (\widehat{u}_{ij})_{i=1,\dots,n,\, j=1,\dots,d}$ where $\widehat{u}_{ij} \coloneqq \widehat{F}_j(y_{ij})$. The incomplete pseudo copula data $(\widehat{\mathbf{u}}, \mathbf{r})$ is used as an approximately i. i. d. sample from the random vector $\mathbf{U}$.

Second Step: Estimation of the D-Vine Copula

Analogous to the procedure of Kraus and Czado (2017, p. 5), we iteratively select the most influential covariates. Starting without any covariate, at each step that covariate is added to the model which results in the greatest improvement to the model's fit. The model's fit is assessed using a weighted conditional log-likelihood, which is based on the estimated D-vine copula with order $\widetilde{\boldsymbol{\ell}}$, pair-copula families $\widehat{\boldsymbol{\mathcal{F}}}$ and the corresponding copula parameters $\widehat{\boldsymbol{\theta}}$, all given the pseudo copula data $\widehat{\mathbf{u}}$. It is defined as

$$\operatorname{cll}\left(\widetilde{\boldsymbol{\ell}}, \widehat{\boldsymbol{\mathcal{F}}}, \widehat{\boldsymbol{\theta}}; \widehat{\mathbf{u}}, \mathbf{r}\right) \coloneqq \frac{n}{|I_{\mathbf{r}}^{\mathrm{obs}}(\widetilde{\boldsymbol{\ell}})|} \sum_{i \in I_{\mathbf{r}}^{\mathrm{obs}}(\widetilde{\boldsymbol{\ell}})} \log c_{j|\widetilde{D}_{-j}}\left(\widehat{u}_{ij} \;\middle|\; \widehat{\mathbf{u}}_{i\widetilde{D}_{-j}}; \widetilde{\boldsymbol{\ell}}, \widehat{\boldsymbol{\mathcal{F}}}, \widehat{\boldsymbol{\theta}}\right), \tag{4.6}$$

where $\widehat{\mathbf{u}}_{iJ} \coloneqq (\widehat{u}_{ij})_{j \in J}$ for any subset $J \subseteq D$ and $I_{\mathbf{r}}^{\mathrm{obs}}(\widetilde{\boldsymbol{\ell}})$ indexes the observations for which all variables Y_j and thus U_j, $j \in \widetilde{\boldsymbol{\ell}}$, have been observed. For complete data sets, Equation (4.6) equals the conditional log-likelihood of Kraus and Czado (2017, Equation (3.7)) since then $I_{\mathbf{r}}^{\mathrm{obs}}(\widetilde{\boldsymbol{\ell}}) = \{1, \dots, n\}$ for all possible orders $\widetilde{\boldsymbol{\ell}}$. As given in Kraus and Czado (2017, p. 5), $c_{j|\widetilde{D}_{-j}}\left(\widehat{u}_{ij} \;\middle|\; \widehat{\mathbf{u}}_{i\widetilde{D}_{-j}}; \widetilde{\boldsymbol{\ell}}, \widehat{\boldsymbol{\mathcal{F}}}, \widehat{\boldsymbol{\theta}}\right)$ can be expressed as the product of densities over all pair-copulas of the estimated D-vine copula which contain U_j:

$$\begin{aligned}
c_{j|\widetilde{D}_{-j}}\left(\widehat{u}_{ij} \middle| \widehat{\mathbf{u}}_{i\widetilde{D}_{-j}}; \widetilde{\boldsymbol{\ell}}, \widehat{\boldsymbol{\mathcal{F}}}, \widehat{\boldsymbol{\theta}}\right) ={}& c_{j,\widetilde{\ell}_2}\left(\widehat{u}_{ij}, \widehat{u}_{i\widetilde{\ell}_2}; \widehat{\mathcal{F}}_{j,\widetilde{\ell}_2}, \widehat{\boldsymbol{\theta}}_{j,\widetilde{\ell}_2}\right) \\
&\cdot \prod_{m=3}^{\widetilde{d}} c_{j,\widetilde{\ell}_m;\widetilde{\boldsymbol{\ell}}_{2:m-1}}\left(\widehat{C}_{j|\widetilde{\boldsymbol{\ell}}_{2:m-1}}(\widehat{u}_{ij}|\widehat{\mathbf{u}}_{i\widetilde{\boldsymbol{\ell}}_{2:m-1}}), \widehat{C}_{\widetilde{\ell}_m|\widetilde{\boldsymbol{\ell}}_{2:m-1}}(\widehat{u}_{i\widetilde{\ell}_m}|\widehat{\mathbf{u}}_{i\widetilde{\boldsymbol{\ell}}_{2:m-1}}); \widehat{\mathcal{F}}_{j,\widetilde{\ell}_m;\widetilde{\boldsymbol{\ell}}_{2:m-1}}, \widehat{\boldsymbol{\theta}}_{j,\widetilde{\ell}_m;\widetilde{\boldsymbol{\ell}}_{2:m-1}}\right)
\end{aligned}$$

where $\widehat{\mathcal{F}}_\bullet$ and $\widehat{\boldsymbol{\theta}}_\bullet$ denote the family and the family parameter(s) of a pair-copula $C_\bullet$ of the estimated D-vine copula.

Remark 4.2.2. Note that all occurring pair-copulas are estimated on the maximal available information. For bivariate copulas C_{j_1,j_2} this means that they are estimated on the observations indexed by $i \in I_{\mathbf{r}}^{\text{obs}}\big((j_1, j_2)\big)$, that is Y_{j_1} and Y_{j_2} have been observed. For conditional bivariate copulas $C_{j_1,j_2;\mathbf{j}}$ this means that they are estimated on the observations indexed by $i \in I_{\mathbf{r}}^{\text{obs}}\big((j_1, j_2, \mathbf{j})\big)$ where all variables $\{Y_j\}_{j\in\{j_1,j_2,\mathbf{j}\}}$ have been observed. △

The selection algorithm now works as follows (Kraus and Czado 2017, p. 5): Assume that at the beginning of the tth iteration of the algorithm the current optimal D-vine copula has order $\widetilde{\boldsymbol{\ell}}^{(t)} = (\widetilde{\ell}_1^{(t)}, \dots, \widetilde{\ell}_t^{(t)})$ where $\widetilde{\ell}_1^{(t)} = 1$. For each of the remaining variables U_ℓ that have not been selected yet, so $\ell \in D \setminus \{\widetilde{\ell}_m^{(t)}\}_{m=1,\dots,t}$, the necessary pair-copulas to extend the current copula to a D-vine copula with order $(\widetilde{\ell}_1^{(t)}, \dots, \widetilde{\ell}_t^{(t)}, \ell)$ are fitted and the conditional log-likelihood of the resulting D-vine quantile regression model is computed. The current D-vine copula is updated by adding the variable that yields the highest conditional log-likelihood, which completes iteration t. If during iteration t none of the remaining variables is able to improve the model's fit, the algorithm terminates and returns the model based on the non-extended D-vine copula of order $\widetilde{\boldsymbol{\ell}}^{(t)}$.

If one wants to account for the number of parameters $|\widehat{\boldsymbol{\theta}}|$ used in the construction of the D-vine copula, the AIC-corrected conditional log-likelihood cll^{AIC} or the BIC-corrected conditional log-likelihood cll^{BIC} can be considered instead of the conditional log-likelihood (Kraus and Czado 2017, p. 5):

$$\begin{aligned}
\text{cll}^{\text{AIC}}(\widetilde{\boldsymbol{\ell}}, \widehat{\boldsymbol{\mathcal{F}}}, \widehat{\boldsymbol{\theta}}; \widehat{\mathbf{u}}, \mathbf{r}) &\coloneqq -2\text{cll}(\widetilde{\boldsymbol{\ell}}, \widehat{\boldsymbol{\mathcal{F}}}, \widehat{\boldsymbol{\theta}}; \widehat{\mathbf{u}}, \mathbf{r}) + 2|\widehat{\boldsymbol{\theta}}|,\\
\text{cll}^{\text{BIC}}(\widetilde{\boldsymbol{\ell}}, \widehat{\boldsymbol{\mathcal{F}}}, \widehat{\boldsymbol{\theta}}; \widehat{\mathbf{u}}, \mathbf{r}) &\coloneqq -2\text{cll}(\widetilde{\boldsymbol{\ell}}, \widehat{\boldsymbol{\mathcal{F}}}, \widehat{\boldsymbol{\theta}}; \widehat{\mathbf{u}}, \mathbf{r}) + \log(n)|\widehat{\boldsymbol{\theta}}|.
\end{aligned}$$

In this thesis we will only consider the AIC-corrected conditional log-likelihood cll^{AIC}.

Remark 4.2.3. In general, the estimation of the D-vine copulas within the algorithm is better the more observations can be used for it. Thus, it is reasonable to correct the conditional log-likelihood of Equation (4.6) taking into account the proportion of all n observations the conditional log-likelihood is based on. We make the proposal to correct the conditional log-likelihood as follows: For a fixed $\alpha \geq 0$ set

$$\text{cll}^{\text{corrected}}(\widetilde{\boldsymbol{\ell}}, \widehat{\boldsymbol{\mathcal{F}}}, \widehat{\boldsymbol{\theta}}; \widehat{\mathbf{u}}, \mathbf{r}) \coloneqq \text{cll}(\widetilde{\boldsymbol{\ell}}, \widehat{\boldsymbol{\mathcal{F}}}, \widehat{\boldsymbol{\theta}}; \widehat{\mathbf{u}}, \mathbf{r}) - \left(\frac{n}{|I_{\mathbf{r}}^{\text{obs}}(\widetilde{\boldsymbol{\ell}})|}\right)^{\alpha} + 1.$$

This means that the greater α is chosen the more influence non-usable observations have. For $\alpha = 0$ there is no correction. Analogously, the AIC-corrected and BIC-corrected versions can be additionally corrected with respect to the proportion of the observations the conditional log-likelihood is based on:

$$\begin{aligned}
\text{cll}^{\text{AIC}}(\widetilde{\boldsymbol{\ell}}, \widehat{\boldsymbol{\mathcal{F}}}, \widehat{\boldsymbol{\theta}}; \widehat{\mathbf{u}}, \mathbf{r}) &\coloneqq -2\text{cll}^{\text{corrected}}(\widetilde{\boldsymbol{\ell}}, \widehat{\boldsymbol{\mathcal{F}}}, \widehat{\boldsymbol{\theta}}; \widehat{\mathbf{u}}, \mathbf{r}) + 2|\widehat{\boldsymbol{\theta}}|,\\
\text{cll}^{\text{BIC}}(\widetilde{\boldsymbol{\ell}}, \widehat{\boldsymbol{\mathcal{F}}}, \widehat{\boldsymbol{\theta}}; \widehat{\mathbf{u}}, \mathbf{r}) &\coloneqq -2\text{cll}^{\text{corrected}}(\widetilde{\boldsymbol{\ell}}, \widehat{\boldsymbol{\mathcal{F}}}, \widehat{\boldsymbol{\theta}}; \widehat{\mathbf{u}}, \mathbf{r}) + \log(n)|\widehat{\boldsymbol{\theta}}|.
\end{aligned}$$

Since in the later quasi-Gibbs sampling for the estimation of the D-vine quantile regression models missing values will only occur in the response variable and thus the conditional log-likelihoods are calculated and compared on the same set of observations for all possibly added covariates, this correction is not necessary in the following. △

Schallhorn et al. (2017) extended the approach of Kraus and Czado (2017) to mixed continuous and discrete variables by using suitable differences instead of derivatives in the h-functions. Thus, the provisional assumption of the absolute continuity of $\mathbf{Y}$ can be removed from this point onwards.

Our **implementation** of the above presented procedure including discrete variables uses the function `kde1d()` of the R-package `kde1d` (Nagler and Vatter 2022) to fit the marginal distributions. The estimation of the pair-copulas during the construction of the optimal D-vine copula is based on the `bicop()` function of the `rvinecopulib` R-package (Nagler and Vatter 2023). We are also using the functions `dbicop()` and `hbicop()` of the same package to evaluate pair-copula densities and h-functions. Our implementation is an extension of the `vinereg()` function of the R-package `vinereg` (Nagler and Kraus 2024) for which only the complete observations enter the estimation of a D-vine quantile regression model. The resulting D-vine quantile regression models can be evaluated by the `predict.vinereg()` function of the `vinereg` R-package (Nagler and Kraus 2024).

Example 4.2.4. Let $(\mathbf{y}, \mathbf{r})$ be given as in Example 2.2.2 and assume that we want to fit a D-vine quantile regression model with response Y_1 and possible covariates Y_2 and Y_3. Below, we demonstrate the algorithm for choosing the AIC-optimal D-vine copula.

$(\mathbf{y}_{\text{obs}}, \mathbf{r})$

4.93	7.58	T
3.22	5.06	F
2.46	4.78	F
5.51		T
3.01	6.60	F
4.38	5.41	T
5.20	7.66	T
3.64		T
4.51		
3.92	6.93	T
2.69	5.17	F
4.22	6.14	F
3.68	6.46	T
4.29	6.44	
2.32	4.40	F

Iteration 1: For the two possible D-vine copulas with order $(1, 2)$ and $(1, 3)$ the AIC-optimal copulas $\widehat{C}_{1,2}$ and $\widehat{C}_{1,3}$ are estimated. For the estimation of $\widehat{C}_{1,2}$ we can use all observations that are simultaneously complete for Y_1 and Y_2 given by the index set

$$I_{\mathbf{r}}^{\text{obs}}\big((1,2)\big) = \{1, 2, 3, 5, 6, 7, 10, \dots, 15\}.$$

Analogously, $\widehat{C}_{1,3}$ is fitted based on all observations with observed variables Y_1 and Y_3:

$$I_{\mathbf{r}}^{\text{obs}}\big((1,3)\big) = \{1,\ldots,8,10,\ldots,13,15\}.$$

The resulting pair-copulas are the BB7 copula with parameters 2.62 and 2.51 and a log-likelihood of 8.11 ($\widehat{C}_{1,2}$) and the Clayton copula with parameter 3.25 and a log-likelihood of 4.88 ($\widehat{C}_{1,3}$). Since we used 12 observations for the estimation of $\widehat{C}_{1,2}$ and 13 for $\widehat{C}_{1,3}$, the AIC-corrected conditional log-likelihoods calculate as

$$-2\cdot\frac{15}{12}\cdot 8.11 + 2\cdot 2 \approx -16.28 \quad \text{and} \quad -2\cdot\frac{15}{13}\cdot 4.88 + 2\cdot 1 \approx -9.26.$$

Consequently, the algorithm decides for Y_2 as the first variable to construct the D-vine copula since it minimizes the AIC-corrected conditional log-likelihood.

Iteration 2: There is only one variable left to possibly extend the current AIC-optimal D-vine copula with order $(1,2)$ to a D-vine copula with order $(1,2,3)$. To construct this D-vine copula the pair-copulas $\widehat{C}_{2,3}$ and $\widehat{C}_{1,3;2}$ have to be fitted. The copula $\widehat{C}_{2,3}$ is estimated based on the observations given by the index set

$$I_{\mathbf{r}}^{\text{obs}}\big((2,3)\big) = \{1,2,3,5,6,7,10,\ldots,13,15\}$$

resulting in a Gaussian copula with parameter 0.77. For the estimation of the copula $\widehat{C}_{1,3;2}$ we can only use the observations where all the variables involved in the copula, Y_1, Y_2 and Y_3, have been observed. Thus, $\widehat{C}_{1,3;2}$ is fitted using the observations from

$$I_{\mathbf{r}}^{\text{obs}}\big((1,3,2)\big) = \{1,2,3,5,6,7,10,\ldots,13,15\}.$$

The resulting copula $\widehat{C}_{1,3;2}$ equals the Gumbel copula with parameter 1.47. Now, the AIC-corrected conditional log-likelihood is calculated also using only the 11 observations of $I_{\mathbf{r}}^{\text{obs}}\big((1,3,2)\big)$. For $\widehat{C}_{1,2}$ we obtain a log-likelihood of 7.40 while for $\widehat{C}_{1,3;2}$ the log-likelihood equals 1.36. Thus, the AIC-corrected conditional log-likelihood of the D-vine quantile regression model with order $(1,2,3)$ calculates as

$$-2\cdot\frac{15}{11}\cdot(7.40+1.36) + 2\cdot 4 \approx -15.89.$$

Since this value is greater than the AIC-corrected conditional log-likelihood of the model with order $(1,2)$ the algorithm stops not extending the D-vine copula. Thus, the final D-vine copula consists of the BB7 copula $\widehat{C}_{1,2}$ with parameters 2.62 and 2.51. △

4.3 An Imputation Method Based on the Fully Conditional Specification Approach Using D-Vine Quantile Regression Models

In the previous section we constructed the centerpiece of our imputation method: A D-vine quantile regression model can be used as the posterior predictive distribution in a univariate imputation method. Therefore, we can apply the fully conditional specification approach to develop a D-vine quantile regression based multivariate imputation procedure. The last missing component to run the quasi-Gibbs sampler of the fully specification approach is the selection of appropriate initial values. While we could simply sample the missing values for each variable from their observed values, as mentioned in Section 3.2.2, we discard this approach in order not to loose the information given by the dependencies between the variables. Instead, for each missing value we will determine $K > 1$ nearest neighbors, based on **Gower's dissimilarity coefficient** as a distance measure, from which we sample the initial values. Gower's dissimilarity coefficient has been originally introduced by Gower (1971) and extended particularly for ordinal variables by Kaufman and Rousseeuw (1990, pp. 35–36). The following definition transfers the extension of Kaufman and Rousseeuw (1990) to the case of incomplete data sets and is realized in the `gower.dist()` function (`StatMatch` R-package, D'Orazio 2024) which we use for the implementation of our imputation method.

Definition 4.3.1 (Gower's dissimilarity coefficient)**.** Let $(\mathbf{y}_{i_1:}, \mathbf{r}_{i_1:})$ and $(\mathbf{y}_{i_2:}, \mathbf{r}_{i_2:})$ be two observations of the incomplete data set $(\mathbf{y}, \mathbf{r})$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$ be a d-dimensional vector of non-negative weights for each variable $Y_1, \ldots, Y_d$. Then, Gower's dissimilarity coefficient for the observations $(\mathbf{y}_{i_1:}, \mathbf{r}_{i_1:})$ and $(\mathbf{y}_{i_2:}, \mathbf{r}_{i_2:})$ calculates as

$$d(i_1, i_2, \boldsymbol{\lambda}) := \frac{\sum_{j=1}^{d} d(y_{i_1j}, y_{i_2j}) \cdot \delta(r_{i_1j}, r_{i_2j}) \cdot \lambda_j}{\sum_{j=1}^{d} \delta(r_{i_1j}, r_{i_2j}) \cdot \lambda_j},$$

where $\delta(r_{i_1j}, r_{i_2j})$ is defined as

$$\delta(r_{i_1j}, r_{i_2j}) := \begin{cases} 0, & \text{if } r_{i_1j} = 0 \text{ or } r_{i_2j} = 0 \\ 1, & \text{else} \end{cases}.$$

For numeric variables the distance $d(y_{i_1j}, y_{i_2j})$ is defined as

$$d(y_{i_1j}, y_{i_2j}) := \frac{|y_{i_1j} - y_{i_2j}|}{R_j},$$

where $R_j := \max_{i \in I_{\mathbf{r}}^{\text{obs}}(j)} y_{ij} - \min_{i \in I_{\mathbf{r}}^{\text{obs}}(j)} y_{ij}$ is the range of the observed values in the jth column $\mathbf{y}_{:j}$. For ordinal scaled variables, the ranks corresponding to the ordered categories of the jth variable are used instead of y_{ij}. In the situation of $d(i_1, i_2, \boldsymbol{\lambda}) = \frac{0}{0}$ we set $d(i_1, i_2, \boldsymbol{\lambda}) := 1$. △

Determining the K nearest neighbors of a missing value y_{ij}

Suppose for an index pair $(i, j) \in \{1, \ldots, n\} \times \{1, \ldots, d\}$ we have a missing value y_{ij}, that is $r_{ij} = 0$. Let $|\rho_S(\mathbf{y}_{:j}, \mathbf{y}_{:j^*})|$ denote the absolute value of the sample Spearman's rank correlation coefficient between the columns $\mathbf{y}_{:j}$ and $\mathbf{y}_{:j^*}$ of $\mathbf{y}$ which is calculated using all pairwise complete observations and averaging possible ties. Note that we have assumed that the variables are at least ordinal scaled such that Spearman's rank correlation coefficient can always be calculated. Then we determine the K nearest neighbors N_{ij} of y_{ij} in the following way:

1. For each $i^* \in I_{\mathbf{r}}^{\text{obs}}(j)$, that is for every observation where the variable Y_j has been observed, calculate Gower's dissimilarity $d(i, i^*, \boldsymbol{\lambda})$ as defined in Definition 4.3.1 with $\boldsymbol{\lambda} := \boldsymbol{\lambda}(j) := \big(|\rho_S(\mathbf{y}_{:j}, \mathbf{y}_{:j^*})|\big)_{j^*=1,\dots,d}$ and collect these dissimilarities in the vector $\mathbf{d} := \mathbf{d}(i, j) := \big(d(i, i^*, \boldsymbol{\lambda})\big)_{i^* \in I_{\mathbf{r}}^{\text{obs}}(j)}$.
2. Let $d_{(K)}$ be the Kth element of an ascending permutation of the vector $\mathbf{d}$ for a $K > 1$. Then, we define the K nearest neighbors N_{ij} of y_{ij} as the vector

$$N_{ij} := (y_{i^*j})_{i^* \in I_{\mathbf{r}}^{\text{obs}}(j):\, d(i,i^*,\boldsymbol{\lambda}) \le d_{(K)}}.$$

It can happen that the vector N_{ij} has more than K elements. Nevertheless, for the sake of simplicity we refer to N_{ij} as the K nearest neighbors of y_{ij}.

Example 4.3.2. We determine the $K = 5$ nearest neighbors for each of the missing values $y_{4,2}$, $y_{8,2}$, $y_{9,2}$, $y_{9,3}$ and $y_{14,3}$ of the incomplete data set $(\mathbf{y}, \mathbf{r})$ from Example 2.2.2.

$(\mathbf{y}_{\text{obs}}, \mathbf{r})$

4.93	7.58	T
3.22	5.06	F
2.46	4.78	F
5.51		T
3.01	6.60	F
4.38	5.41	T
5.20	7.66	T
3.64		T
4.51		
3.92	6.93	T
2.69	5.17	F
4.22	6.14	F
3.68	6.46	T
4.29	6.44	
2.32	4.40	F

We take a closer look at the missing value $y_{4,2}$ with index pair $(i, j) = (4, 2)$. The observations with observed value for variable Y_2 are given by the index set

$$I_{\mathbf{r}}^{\text{obs}}(2) = \{1, 2, 3, 5, 6, 7, 10, \dots, 15\}.$$

To calculate Spearman's rank correlation coefficients we transform the discrete variable Y_3 setting F $= 1$ and T $= 2$. Then, the needed absolute values of the sample Spearman's rank correlation coefficients equal

- $|\rho_S(\mathbf{y}_{:2}, \mathbf{y}_{:1})| = 0.73$ between column $\mathbf{y}_{:2}$ and column $\mathbf{y}_{:1}$,
- $|\rho_S(\mathbf{y}_{:2}, \mathbf{y}_{:2})| = 1$ between column $\mathbf{y}_{:2}$ and column $\mathbf{y}_{:2}$ and
- $|\rho_S(\mathbf{y}_{:2}, \mathbf{y}_{:3})| = 0.69$ between column $\mathbf{y}_{:2}$ and $\mathbf{y}_{:3}$.

Thus, the vector of weights is given as $\boldsymbol{\lambda} = (0.73, 1, 0.69)$. Now, we can proceed calculating Gower's dissimilarity coefficients according to Definition 4.3.1. We have

$$d(4,1,\boldsymbol{\lambda}) = \frac{\frac{|5.51-4.93|}{5.51-2.32} \cdot 1 \cdot 0.73 + 0 + \frac{|2-2|}{2-1} \cdot 1 \cdot 0.69}{1 \cdot 0.73 + 0 + 1 \cdot 0.69} \approx 0.0935.$$

Similarly, we obtain

$$\begin{aligned} d(4,2,\boldsymbol{\lambda}) &\approx 0.8550, & d(4,3,\boldsymbol{\lambda}) &\approx 0.9774, & d(4,5,\boldsymbol{\lambda}) &\approx 0.8888, \\ d(4,6,\boldsymbol{\lambda}) &\approx 0.1821, & d(4,7,\boldsymbol{\lambda}) &\approx 0.0500, & d(4,10,\boldsymbol{\lambda}) &\approx 0.2562, \\ d(4,11,\boldsymbol{\lambda}) &\approx 0.9404, & d(4,12,\boldsymbol{\lambda}) &\approx 0.6938, & d(4,13,\boldsymbol{\lambda}) &\approx 0.2949, \\ d(4,14,\boldsymbol{\lambda}) &\approx 0.3824, & d(4,15,\boldsymbol{\lambda}) &= 1.0000. \end{aligned}$$

Sorting those values ascending we get

$$0.0500 \le 0.0935 \le 0.1821 \le 0.2562 \le 0.2949 \le \ldots \le 1.0000.$$

Consequently, for $K = 5$, the nearest neighbors $N_{4,2}$ of $y_{4,2}$ are

$$\begin{aligned} N_{4,2} &= (y_{i,2})_{i \in \{1,2,3,5,6,7,10,\ldots,15\}:\, d(4,i,\boldsymbol{\lambda}) \le 0.2949} \\ &= (y_{1,2}, y_{6,2}, y_{7,2}, y_{10,2}, y_{13,2}) \\ &= (7.58, 5.41, 7.66, 6.93, 6.46). \end{aligned}$$

Analogously, for the other missing values we get the nearest neighbors

$$\begin{aligned} N_{8,2} &= (7.58, 5.41, 6.93, 6.46, 6.44), & N_{9,2} &= (7.58, 5.41, 6.93, 6.14, 6.44), \\ N_{9,3} &= (\mathrm{T}, \mathrm{T}, \mathrm{T}, \mathrm{T}, \mathrm{F}), & N_{14,3} &= (\mathrm{T}, \mathrm{T}, \mathrm{T}, \mathrm{F}, \mathrm{T}). \end{aligned}$$ △

Imputation Procedure

We now describe a multivariate imputation method based on the fully conditional specification approach using D-vine quantile regression models (for a detailed code, see Algorithm 1). The parameter uncertainty is generated according to approximation (A5) from Section 3.2.1 via bootstrapping.

Let be given a realized incomplete data set $(\mathbf{y}, \mathbf{r})$ where $\mathbf{y}$ is a sample from the random vector $\mathbf{Y} = (Y_1, \ldots, Y_d)$ with observed part $\mathbf{y}_{\text{obs}}$ and missing part $\mathbf{y}_{\text{mis}}$. Assume that we aim to generate M imputed data sets from $(\mathbf{y}, \mathbf{r})$. Initially, for each missing data point $y_{ij} \in \mathbf{y}_{\text{mis}}$, the K nearest neighbors N_{ij} are identified. Then, for each of the M imputed data sets which we want to generate we run a separate quasi-Gibbs sampling chain. By running M independent chains of the quasi-Gibbs sampler, we generate M imputed data sets in total. In the following, we focus on a single chain of this process:

- **Initialization of the quasi-Gibbs sampler:** To initialize the quasi-Gibbs sampler, for each missing data point $y_{ij} \in \mathbf{y}_{\text{mis}}$, we uniformly sample a value from its K nearest neighbors N_{ij}. The observed values $\mathbf{y}_{\text{obs}}$ are carried over. Together, we obtain an initial completed data set.

Algorithm 1: Multivariate imputation method based on the fully conditional specification approach using D-vine quantile regression models

```
Input: Incomplete data set $(\mathbf{y}, \mathbf{r})$ of dimension $n \times d$, number of imputed data
       sets $M$, number of iterations $T$, number of nearest neighbors $K$.
/* Determining the $K$ nearest neighbors */
foreach $y_{ij} \in \mathbf{y}_{\text{mis}}$ do
    Determine the $K$ nearest neighbors $N_{ij}$ of $y_{ij}$;
for $m = 1, \ldots, M$ do
    /* Setting the initial values */
    $(\mathbf{y}_{\text{obs}}^{(m)}, \mathbf{r}^{(m)}) \leftarrow (\mathbf{y}_{\text{obs}}, \mathbf{r})$;
    foreach $y_{ij} \in \mathbf{y}_{\text{mis}}$ do
        $y \leftarrow$ Sample from discrete Unif$(N_{ij})$;
        $(y_{ij}^{(m)}, r_{ij}^{(m)}) \leftarrow (y, 1)$;
    /* Quasi-Gibbs sampling */
    for $t = 1, \ldots, T$ do
        for $j = 1, \ldots, d$ do
            if $I_{\mathbf{r}}^{\text{mis}}(j) \neq \emptyset$ then
                foreach $i \in I_{\mathbf{r}}^{\text{mis}}(j)$ do
                    $r_{ij}^{(m)} \leftarrow 0$;
                $\widetilde{I} \leftarrow n$-dimensional sample from discrete Unif$(\{1, \ldots, n\})$;
                $(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}) \leftarrow \big(\mathbf{y}_{i:}^{(m)}, \mathbf{r}_{i:}^{(m)}\big)_{i \in \widetilde{I}}$;
                Estimate an optimal D-vine quantile regression model $\widehat{q}_j(\cdot \mid \mathbf{y}_{\widetilde{D}_{-j}})$
                  from $(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})$ as presented in Section 4.2 where $\widetilde{D} \subseteq \{1, \ldots, d\}$ and
                  $\widetilde{D}_{-j} = \widetilde{D} \setminus \{j\}$;
                foreach $i \in I_{\mathbf{r}}^{\text{mis}}(j)$ do
                    $\alpha \leftarrow$ Sample from Unif$(0, 1)$;
                    $(y_{ij}^{(m)}, r_{ij}^{(m)}) \leftarrow \Big(\widehat{q}_j\big(\alpha \mid \mathbf{y}_{\widetilde{D}_{-j}} = (y_{ij^*}^{(m)})_{j^* \in \widetilde{D}_{-j}}\big), 1\Big)$ ;
return $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}$;
```

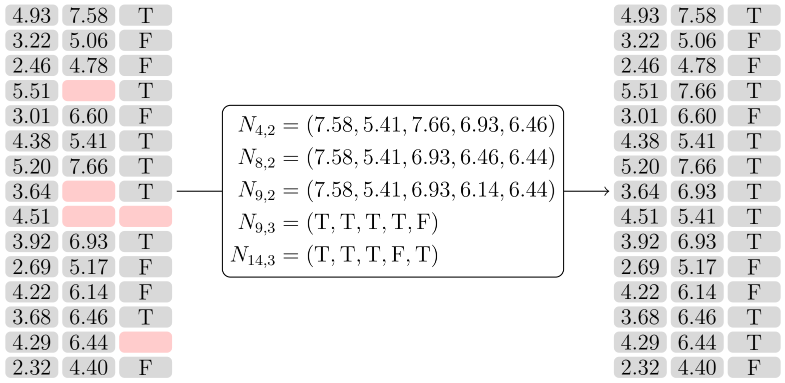
- **Iterative process of the quasi-Gibbs sampler:** At the beginning of each iteration t, a completed data set is provided. Starting with Y_1, we sequentially update each variable Y_j as follows. If the original data set $(\mathbf{y}, \mathbf{r})$ contains missing values for variable Y_j we remove the corresponding data points y_{ij} of this variable from the last completed data set, yielding an incomplete data set. On a bootstrapped sample of this incomplete data set with the same size as the original incomplete data set $(\mathbf{y}, \mathbf{r})$, we fit an AIC-optimal D-vine quantile regression model with Y_j as the response variable. The incomplete data set is then completed by sampling from the corresponding quantile function, where the quantiles $\alpha \sim \text{Unif}(0,1)$ are randomly selected for each missing value, resulting in a completed data set again. After updating Y_j, the process moves to the next variable Y_{j+1} and repeats the same steps. The process continues until the data set has been updated for all variables $Y_1, \ldots, Y_d$, completing the tth iteration.

- **Completion of the imputed data set:** The quasi-Gibbs sampler runs for T iterations. The final completed data set after the update of variable Y_d in iteration T forms the imputed data set.

In our R-implementation, the imputations are saved in a `mids` object from the `mice` package (Buuren and Groothuis-Oudshoorn 2011). This allows to apply all processing and diagnostics functions of the package to the resulting imputations. Even though our implementation allows for D-vine quantile regression models with non-parametric pair-copulas, we will restrict ourselves exclusively to parametric pair-copulas in this thesis.

Example 4.3.3. We illustrate Algorithm 1 for the incomplete data set $(\mathbf{y}, \mathbf{r})$ from Example 2.2.2 generating $M = 1$ imputed data set. The quasi-Gibbs sampler runs $T = 2$ iterations. We consider the $K = 5$ nearest neighbors.

Initial Values (Figure 4.2): The algorithm starts with uniformly sampling the initial values for the missing data points $y_{4,2}$, $y_{8,2}$, $y_{9,2}$, $y_{9,3}$ and $y_{14,3}$ from their $K = 5$ nearest neighbors $N_{4,2}$, $N_{8,2}$, $N_{9,2}$, $N_{9,3}$ and $N_{14,3}$ which were determined in Example 4.3.2. We obtain a completed data set with $y_{4,2} = 7.66$, $y_{8,2} = 6.93$, $y_{9,2} = 5.41$ and $y_{9,3} = y_{14,3} = \text{T}$.



4.93	7.58	T
3.22	5.06	F
2.46	4.78	F
5.51		T
3.01	6.60	F
4.38	5.41	T
5.20	7.66	T
3.64		T
4.51		
3.92	6.93	T
2.69	5.17	F
4.22	6.14	F
3.68	6.46	T
4.29	6.44	
2.32	4.40	F

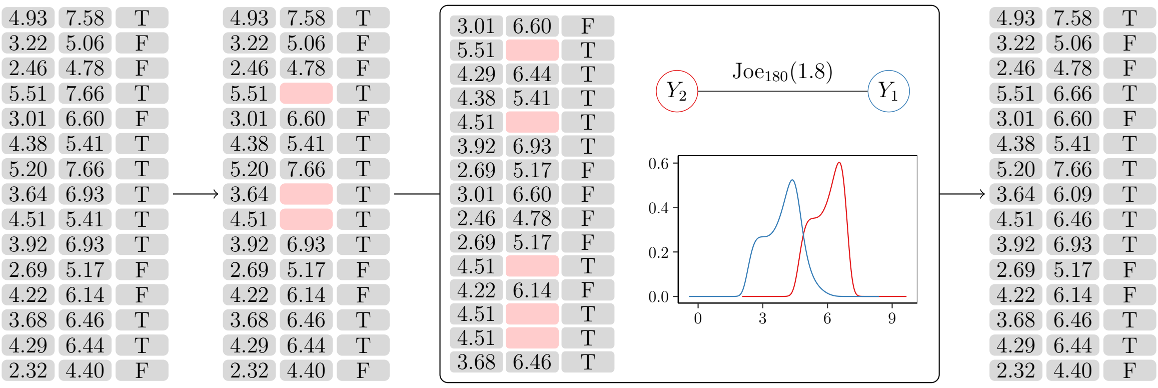
$N_{4,2} = (7.58, 5.41, 7.66, 6.93, 6.46)$
$N_{8,2} = (7.58, 5.41, 6.93, 6.46, 6.44)$
$N_{9,2} = (7.58, 5.41, 6.93, 6.14, 6.44)$
$N_{9,3} = (\text{T}, \text{T}, \text{T}, \text{T}, \text{F})$
$N_{14,3} = (\text{T}, \text{T}, \text{T}, \text{F}, \text{T})$

4.93	7.58	T
3.22	5.06	F
2.46	4.78	F
5.51	7.66	T
3.01	6.60	F
4.38	5.41	T
5.20	7.66	T
3.64	6.93	T
4.51	5.41	T
3.92	6.93	T
2.69	5.17	F
4.22	6.14	F
3.68	6.46	T
4.29	6.44	T
2.32	4.40	F

Figure 4.2: Sampling the initial values in Example 4.3.3.

Quasi-Gibbs Sampler:

- **Iteration 1, imputing variable Y_2 (Figure 4.3):** From the last completed data set the original missing values $y_{4,2}$, $y_{8,2}$ and $y_{9,2}$ of variable Y_2 are removed giving an incomplete data set. On a bootstrapped sample of this incomplete data set an AIC-optimal D-vine quantile regression model with response Y_2 is fitted: While Y_2 is coupled with Y_1 via a 180 degrees rotated Joe copula with parameter 1.8, variable Y_3 has not been selected as a covariate. The incomplete data set is then completed by sampling from the quantile function with quantiles $\alpha \sim \text{Unif}(0, 1)$ randomly selected for each missing value resulting in $y_{4,2} = 6.66$, $y_{8,2} = 6.09$ and $y_{9,2} = 6.46$.



4.93	7.58	T
3.22	5.06	F
2.46	4.78	F
5.51	7.66	T
3.01	6.60	F
4.38	5.41	T
5.20	7.66	T
3.64	6.93	T
4.51	5.41	T
3.92	6.93	T
2.69	5.17	F
4.22	6.14	F
3.68	6.46	T
4.29	6.44	T
2.32	4.40	F

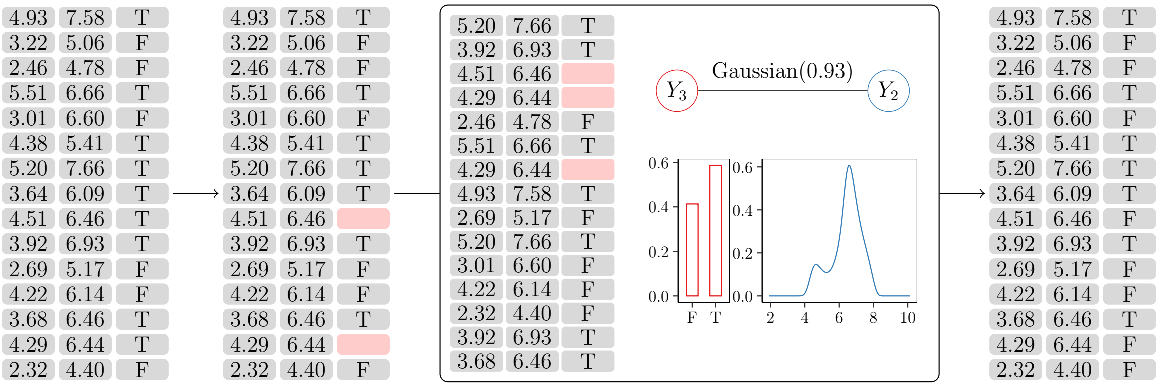
4.93	7.58	T
3.22	5.06	F
2.46	4.78	F
5.51		T
3.01	6.60	F
4.38	5.41	T
5.20	7.66	T
3.64		T
4.51		T
3.92	6.93	T
2.69	5.17	F
4.22	6.14	F
3.68	6.46	T
4.29	6.44	T
2.32	4.40	F

3.01	6.60	F
5.51		T
4.29	6.44	T
4.38	5.41	T
4.51		T
3.92	6.93	T
2.69	5.17	F
3.01	6.60	F
2.46	4.78	F
2.69	5.17	F
4.51		T
4.22	6.14	F
4.51		T
4.51		T
3.68	6.46	T

4.93	7.58	T
3.22	5.06	F
2.46	4.78	F
5.51	6.66	T
3.01	6.60	F
4.38	5.41	T
5.20	7.66	T
3.64	6.09	T
4.51	6.46	T
3.92	6.93	T
2.69	5.17	F
4.22	6.14	F
3.68	6.46	T
4.29	6.44	T
2.32	4.40	F

Figure 4.3: Imputation of variable Y_2 in iteration 1 of Example 4.3.3.

- **Iteration 1, imputing variable Y_3 (Figure 4.4):** From the last completed data set the original missing values $y_{9,3}$ and $y_{14,3}$ of variable Y_3 are removed giving an incomplete data set. From a bootstrapped sample of this incomplete data set an AIC-optimal D-vine quantile regression model with discrete response Y_3 is fitted: While Y_3 is coupled with Y_2 via a Gaussian copula with parameter 0.93, variable Y_1 has not been selected as a covariate. The incomplete data set is then completed by sampling from the quantile function with quantiles $\alpha \sim \text{Unif}(0, 1)$ randomly selected for each missing value resulting in $y_{9,3} = \text{F}$, $y_{14,3} = \text{F}$.



4.93	7.58	T
3.22	5.06	F
2.46	4.78	F
5.51	6.66	T
3.01	6.60	F
4.38	5.41	T
5.20	7.66	T
3.64	6.09	T
4.51	6.46	T
3.92	6.93	T
2.69	5.17	F
4.22	6.14	F
3.68	6.46	T
4.29	6.44	T
2.32	4.40	F

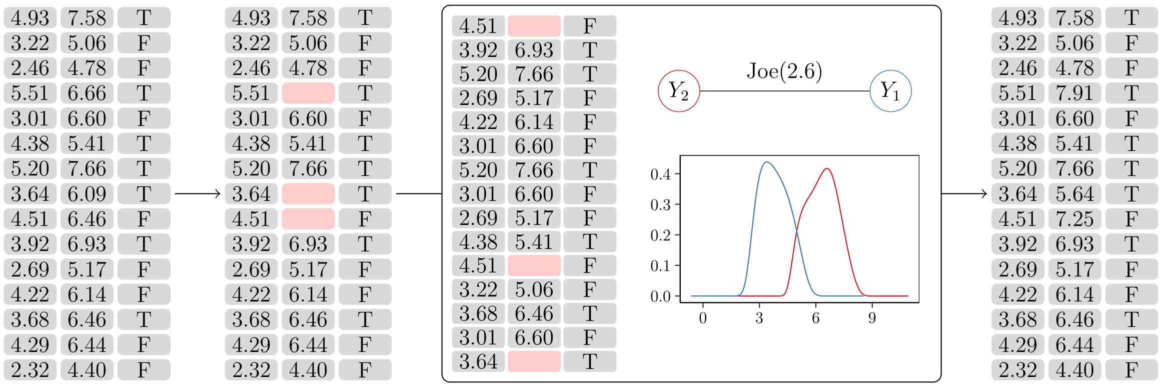
4.93	7.58	T
3.22	5.06	F
2.46	4.78	F
5.51	6.66	T
3.01	6.60	F
4.38	5.41	T
5.20	7.66	T
3.64	6.09	T
4.51	6.46	
3.92	6.93	T
2.69	5.17	F
4.22	6.14	F
3.68	6.46	T
4.29	6.44	
2.32	4.40	F

5.20	7.66	T
3.92	6.93	T
4.51	6.46	
4.29	6.44	
2.46	4.78	F
5.51	6.66	T
4.29	6.44	
4.93	7.58	T
2.69	5.17	F
5.20	7.66	T
3.01	6.60	F
4.22	6.14	F
2.32	4.40	F
3.92	6.93	T
3.68	6.46	T

4.93	7.58	T
3.22	5.06	F
2.46	4.78	F
5.51	6.66	T
3.01	6.60	F
4.38	5.41	T
5.20	7.66	T
3.64	6.09	T
4.51	6.46	F
3.92	6.93	T
2.69	5.17	F
4.22	6.14	F
3.68	6.46	T
4.29	6.44	F
2.32	4.40	F

Figure 4.4: Imputation of variable Y_3 in iteration 1 of Example 4.3.3.

- **Iteration 2, imputing variable Y_2 (Figure 4.5):** From the last completed data set the original missing values $y_{4,2}$, $y_{8,2}$ and $y_{9,2}$ of variable Y_2 are removed giving an incomplete data set. On a bootstrapped sample of this incomplete data set an AIC-optimal D-vine quantile regression model with response Y_2 is fitted: While Y_2 is coupled with Y_1 via a Joe copula with parameter 2.6, variable Y_3 has not been selected as a covariate. The incomplete data set is then completed by sampling from the quantile function with quantiles $\alpha \sim \text{Unif}(0,1)$ randomly selected for each missing value resulting in $y_{4,2} = 7.91$, $y_{8,2} = 5.64$ and $y_{9,2} = 7.25$.



4.93	7.58	T
3.22	5.06	F
2.46	4.78	F
5.51	6.66	T
3.01	6.60	F
4.38	5.41	T
5.20	7.66	T
3.64	6.09	T
4.51	6.46	F
3.92	6.93	T
2.69	5.17	F
4.22	6.14	F
3.68	6.46	T
4.29	6.44	F
2.32	4.40	F

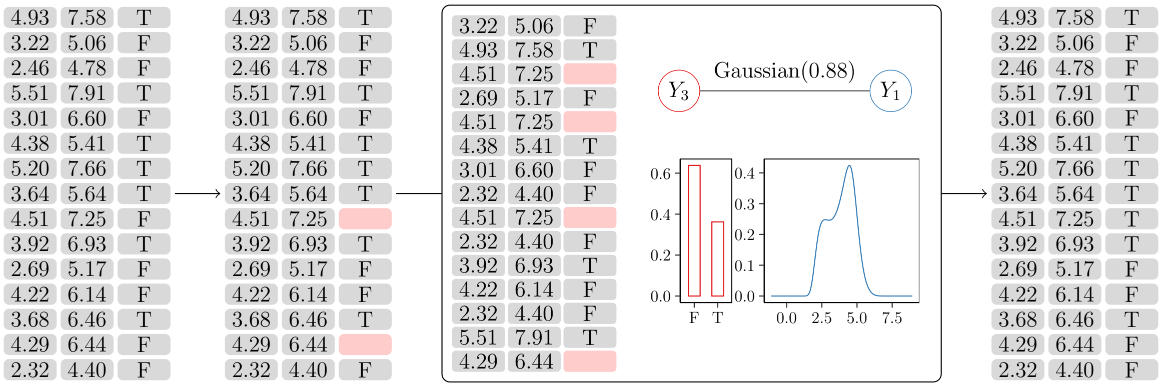
4.93	7.58	T
3.22	5.06	F
2.46	4.78	F
5.51		T
3.01	6.60	F
4.38	5.41	T
5.20	7.66	T
3.64		T
4.51		F
3.92	6.93	T
2.69	5.17	F
4.22	6.14	F
3.68	6.46	T
4.29	6.44	F
2.32	4.40	F

4.51		F
3.92	6.93	T
5.20	7.66	T
2.69	5.17	F
4.22	6.14	F
3.01	6.60	F
5.20	7.66	T
3.01	6.60	F
2.69	5.17	F
4.38	5.41	T
4.51		F
3.22	5.06	F
3.68	6.46	T
3.01	6.60	F
3.64		T

4.93	7.58	T
3.22	5.06	F
2.46	4.78	F
5.51	7.91	T
3.01	6.60	F
4.38	5.41	T
5.20	7.66	T
3.64	5.64	T
4.51	7.25	F
3.92	6.93	T
2.69	5.17	F
4.22	6.14	F
3.68	6.46	T
4.29	6.44	F
2.32	4.40	F

Figure 4.5: Imputation of variable Y_2 in iteration 2 of Example 4.3.3.

- **Iteration 2, imputing variable Y_3 (Figure 4.6):** From the last completed data set the original missing values $y_{9,3}$ and $y_{14,3}$ of variable Y_3 are removed giving an incomplete data set. From a bootstrapped sample of this incomplete data set an AIC-optimal D-vine quantile regression model with discrete response Y_3 is fitted: While Y_3 is coupled with Y_1 via a Gaussian copula with parameter 0.88, variable Y_2 has not been selected as a covariate. The incomplete data set is then completed by sampling from the quantile function with quantiles $\alpha \sim \text{Unif}(0,1)$ randomly selected for each missing value resulting in $y_{9,3} = \text{T}$, $y_{14,3} = \text{F}$.



4.93	7.58	T
3.22	5.06	F
2.46	4.78	F
5.51	7.91	T
3.01	6.60	F
4.38	5.41	T
5.20	7.66	T
3.64	5.64	T
4.51	7.25	F
3.92	6.93	T
2.69	5.17	F
4.22	6.14	F
3.68	6.46	T
4.29	6.44	F
2.32	4.40	F

4.93	7.58	T
3.22	5.06	F
2.46	4.78	F
5.51	7.91	T
3.01	6.60	F
4.38	5.41	T
5.20	7.66	T
3.64	5.64	T
4.51	7.25	
3.92	6.93	T
2.69	5.17	F
4.22	6.14	F
3.68	6.46	T
4.29	6.44	
2.32	4.40	F

3.22	5.06	F
4.93	7.58	T
4.51	7.25	
2.69	5.17	F
4.51	7.25	
4.38	5.41	T
3.01	6.60	F
2.32	4.40	F
4.51	7.25	
2.32	4.40	F
3.92	6.93	T
4.22	6.14	F
2.32	4.40	F
5.51	7.91	T
4.29	6.44	

4.93	7.58	T
3.22	5.06	F
2.46	4.78	F
5.51	7.91	T
3.01	6.60	F
4.38	5.41	T
5.20	7.66	T
3.64	5.64	T
4.51	7.25	T
3.92	6.93	T
2.69	5.17	F
4.22	6.14	F
3.68	6.46	T
4.29	6.44	F
2.32	4.40	F

Figure 4.6: Imputation of variable Y_3 in iteration 2 of Example 4.3.3..

The last completed data set forms the final imputed data set $\mathbf{y}^{(1)}$.

The imputation process with all interim completed data sets is summarized as a whole in Figure 4.7. Here, if a data set has cells half colored grey and half colored red this means, that for incoming arrows the data set is complete and for outgoing arrows the data set is incomplete with missing values for those cells. The original incomplete data set $(\mathbf{y}, \mathbf{r})$ with the missing values $y_{4,2}$, $y_{8,2}$, $y_{9,2}$, $y_{9,3}$ and $y_{14,3}$ is given in the upper left panel. The final imputed data set $\mathbf{y}^{(1)}$ is given in the upper right panel. △

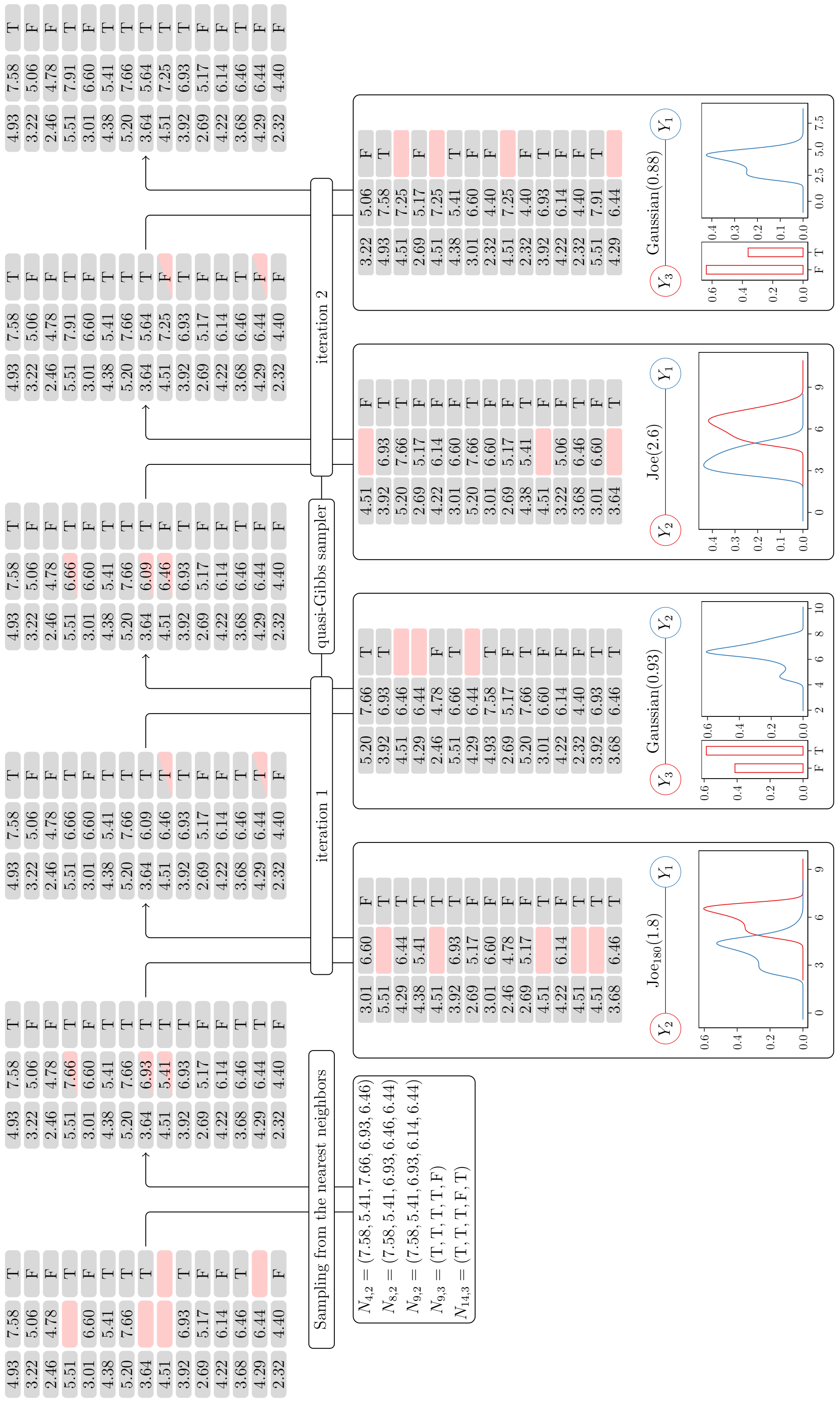

Sampling from the nearest neighbors
$N_{4,2} = (7.58, 5.41, 7.66, 6.93, 6.46)$
$N_{8,2} = (7.58, 5.41, 6.93, 6.46, 6.44)$
$N_{9,2} = (7.58, 5.41, 6.93, 6.14, 6.44)$
$N_{9,3} = (T, T, T, T, F)$
$N_{14,3} = (T, T, T, F, T)$
iteration 1
quasi-Gibbs sampler
iteration 2
Y_2 Joe$_{180}$(1.8) Y_1
Y_3 Gaussian(0.93) Y_2
Y_2 Joe(2.6) Y_1
Y_3 Gaussian(0.88) Y_1

Figure 4.7: Graphical illustration of the multiple imputation of Example 4.3.3.

5 Comparison of Different Imputation Methods: A Simulation Study

5.1 Design

Imputation Methods

Withing the simulation study we compare the following four imputation methods:

1. The **method** $\mathcal{M}_{\mathrm{v}}$ is equivalent to the vine based method given in Algorithm 1. We set the number of nearest neighbors to $K = 5$. The quasi-Gibbs sampler is running for $T = 20$ iterations.
2. The **method** $\mathcal{M}_{\mathrm{m}}$ stands for the imputation with the function `mice()` of the `mice` R-package (Buuren and Groothuis-Oudshoorn 2011). As for the method $\mathcal{M}_{\mathrm{v}}$, the number of iterations for the quasi-Gibbs sampling is set to $T = 20$. Apart from that the standard settings are applied.
3. The **method** $\mathcal{M}_{\mathrm{a}}$ denotes the imputation with the function `amelia()` of the `Amelia` R-package (Honaker, King, and Blackwell 2011) with standard settings but discrete imputation for discrete variables.
4. The **method** $\mathcal{M}_{\mathrm{c}}$ is a slightly adopted version of the imputation with the function `CoImp()` of the `CoImp` R-package (Di Lascio, Gatto, and Giannerini 2024). Since the standard version of `CoImp()` cannot handle a mixture of continuous and discrete variables, we are treating the discrete variables in the same way `amelia()` does: An appropriately scaled version of the initially continuously valued imputation is used as the probability of success in a binomial distribution. The draw from this binomial distribution is then translated back into one of the original discrete values.

With all methods we are generating $M = 10$ imputations.

Data Generating Distribution

The data on which the simulation study is based is generated from an R-vine of dimension $d = 5$ with mixed continuous and discrete marginal distributions. The marginal distributions together with their means and standard deviations are summarized in Table 5.1. The vine copula consisting of the vine structure and the pair-copula parametrizations

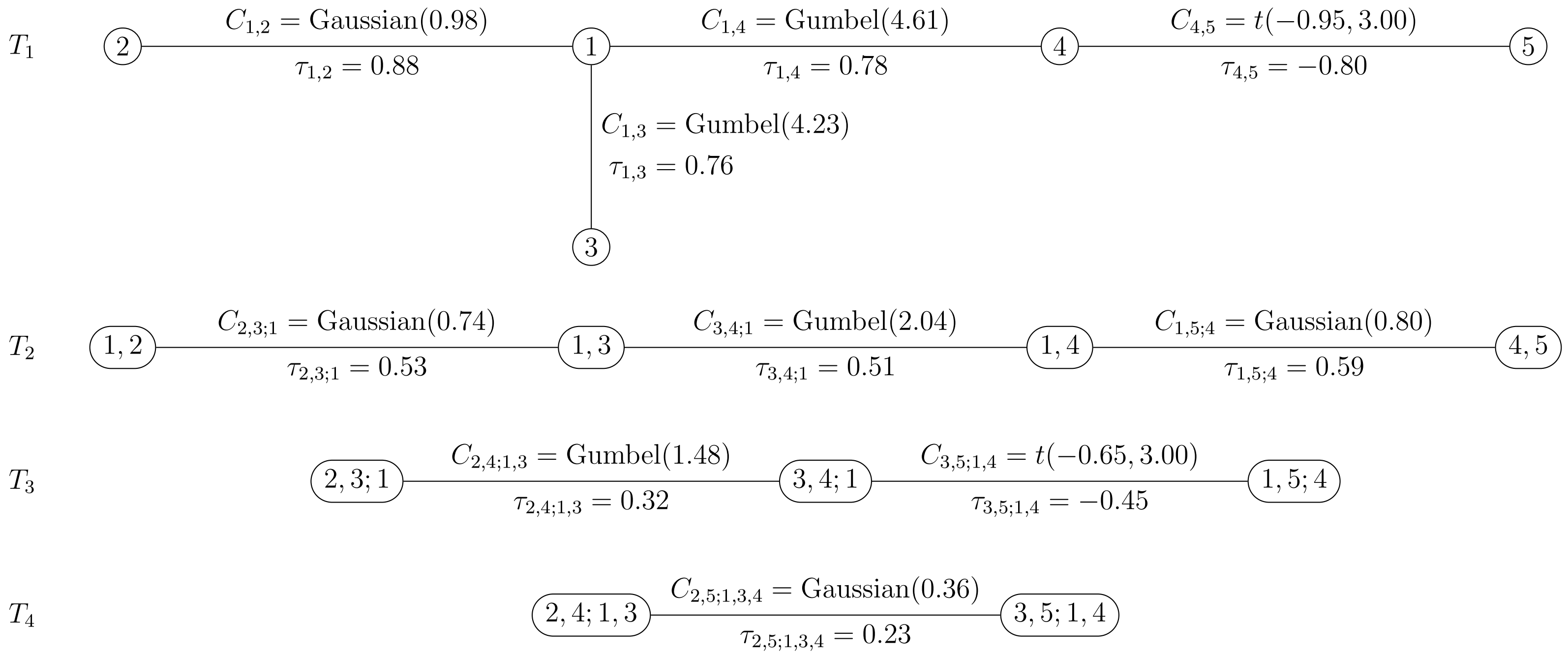

T_1
T_2
T_3
T_4
2
1
4
5
3
$C_{1,2} = \text{Gaussian}(0.98)$
$\tau_{1,2} = 0.88$
$C_{1,4} = \text{Gumbel}(4.61)$
$\tau_{1,4} = 0.78$
$C_{4,5} = t(-0.95, 3.00)$
$\tau_{4,5} = -0.80$
$C_{1,3} = \text{Gumbel}(4.23)$
$\tau_{1,3} = 0.76$
$1,2$
$1,3$
$1,4$
$4,5$
$C_{2,3;1} = \text{Gaussian}(0.74)$
$\tau_{2,3;1} = 0.53$
$C_{3,4;1} = \text{Gumbel}(2.04)$
$\tau_{3,4;1} = 0.51$
$C_{1,5;4} = \text{Gaussian}(0.80)$
$\tau_{1,5;4} = 0.59$
$2,3;1$
$3,4;1$
$1,5;4$
$C_{2,4;1,3} = \text{Gumbel}(1.48)$
$\tau_{2,4;1,3} = 0.32$
$C_{3,5;1,4} = t(-0.65, 3.00)$
$\tau_{3,5;1,4} = -0.45$
$2,4;1,3$
$3,5;1,4$
$C_{2,5;1,3,4} = \text{Gaussian}(0.36)$
$\tau_{2,5;1,3,4} = 0.23$

Figure 5.1: Vine copula of the data generating vine in the simulation study.

	Y_1	Y_2	Y_3	Y_4	Y_5
Distribution	$\mathcal{N}(0,1)$	$\text{Binom}(4,0.5)$	$\mathcal{LN}(0,0.69)$	$\mathcal{N}(0,1)$	$\mathcal{LN}(0,0.69)$
Mean	$\mu_1 = 0$	$\mu_2 = 2$	$\mu_3 \approx 1.27$	$\mu_4 = 0$	$\mu_1 \approx 1.27$
Standard deviation	$\sigma_1 = 1$	$\sigma_2 = 1$	$\sigma_3 = 1$	$\sigma_4 = 1$	$\sigma_5 = 1$

Table 5.1: Marginal distributions of the data generating vine in the simulation study.

with corresponding (un-)conditional Kendall's τ's is shown in Figure 5.1. Overall, there is a strong dependency between the variables.

We randomly selected the vine copula and the marginals from the following procedure:

- The marginals were drawn uniformly from the set of the univariate distributions $\{\text{Binom}(4, 0.5), \mathcal{N}(0, 1), \mathcal{LN}(0, 0.69)\}$. In the case that either all marginals were chosen from a continuous distribution or all marginals were chosen from a discrete distribution, we repeated the drawing process of the marginal distribution family. This ensured that at least one marginal distribution was chosen continuous as well as that at least one marginal distribution was chosen discrete.

- The vine copula structure was selected uniformly from all possible regular vine structures of dimension $d = 5$ using the function `rvine_matrix_sim()` of the `rvinecopulib` R-package (Nagler and Vatter 2023).

- The pair-copula families were sampled uniformly from the Gaussian copula, the t-copula with 3 degrees of freedom, the Gumbel copula and the 90 degrees rotated Gumbel copula. The (remaining) pair-copula parameters were set via inversion of Kendall's τ where the Kendall's τ of each pair-copula itself was randomly selected: The absolute value of τ in the first tree was drawn from $\text{Unif}(0.7, 0.9)$, in the second tree from $\text{Unif}(0.5, 0.7)$, in the third tree from $\text{Unif}(0.3, 0.5)$ and in the forth tree from $\text{Unif}(0.1, 0.3)$. Kendall's τ was assigned a positive value with probability 0.7 and a negative value with probability 0.3.

Quantities of Interest

We choose the $k = 20$ quantities of interest $Q_1, \ldots, Q_{20}$ to be the (conditional) Kendall's τ's defined by the pair-copulas of the data generating vine as well as the theoretical means and standard deviations of its marginal distributions. The true values $q_1, \ldots, q_{20}$ of the quantities of interest $Q_1, \ldots, Q_{20}$ are given in Table 5.2.

For the later application of Rubin's rules of Remark 3.1.4 we have to define a complete-data estimator $\widehat{Q}_l(\mathbf{Y})$ for each quantity of interest Q_l:

- For the Kendall's τ's $Q_1, \ldots, Q_{10}$ the estimators are defined as follows: For a complete data set an AIC-optimal R-vine is fitted, applying `vine()` of `rvinecopulib` (Nagler and Vatter 2023), which has the same same vine structure as the data generating vine. Then the theoretical Kendall's τ's calculated by inverting the pair-copula parameters of the fitted vine serve as estimators $\widehat{Q}_1(\mathbf{Y}), \ldots, \widehat{Q}_{10}(\mathbf{Y})$.

- For the means $Q_{11}, \ldots, Q_{15}$ and the standard deviations $Q_{16}, \ldots, Q_{20}$ we simply take the sample means and sample standard deviations as estimators $\widehat{Q}_{11}(\mathbf{Y}), \ldots, \widehat{Q}_{15}(\mathbf{Y})$ and $\widehat{Q}_{16}(\mathbf{Y}), \ldots, \widehat{Q}_{20}(\mathbf{Y})$.

Additionally, we need variance estimators $\widehat{W}_l(\mathbf{Y})$ for each estimator $\widehat{Q}_l(\mathbf{Y})$. Here, for all variance estimators $\widehat{W}_1(\mathbf{Y}), \ldots, \widehat{W}_{20}(\mathbf{Y})$ we take the bootstrap variance estimator based on 200 bootstrapped samples.

	q_1	q_2	q_3	q_4	q_5
Reference parameter	$\tau_{1,2}$	$\tau_{4,5}$	$\tau_{1,3}$	$\tau_{1,4}$	$\tau_{2,3;1}$
Value	0.88	-0.80	0.76	0.78	0.53
	q_6	q_7	q_8	q_9	q_{10}
Reference parameter	$\tau_{1,5;4}$	$\tau_{3,4;1}$	$\tau_{2,4;1,3}$	$\tau_{3,5;1,4}$	$\tau_{2,5;1,3,4}$
Value	0.59	0.51	0.32	-0.45	0.23
	q_{11}	q_{12}	q_{13}	q_{14}	q_{15}
Reference parameter	μ_1	μ_2	μ_3	μ_4	μ_5
Value	0	2	1.27	0	1.27
	q_{16}	q_{17}	q_{18}	q_{19}	q_{20}
Reference parameter	σ_1	σ_2	σ_3	σ_4	σ_5
Value	1	1	1	1	1

Table 5.2: True values $q_1, \dots, q_{20}$ of the quantities of interest in the simulation study.

Response Mechanism

The missing values are generated under the MAR assumption by applying the `ampute()` function of the `mice` R-package (Buuren and Groothuis-Oudshoorn 2011). For each observation, `ampute()` uses a two step sampling approach: In a first step it is randomly decided with probability p whether the observation contains missing values or not. In the case of missing values, in a second step the concrete response pattern (see Remark 2.2.1) is randomly assigned to the observation from a given set of admissible response patterns together with their probabilities of occurrence. We restrict ourselves to the response patterns with at most three missing values. All admissible response patterns for the simulation study together with their probabilities of occurrence are given in Table 5.3.

Type	Possible response patterns	Occurrence probability	
		Per pattern	Per type
One missing value	$(0,1,1,1,1)$, $(1,0,1,1,1)$, $(1,1,0,1,1)$, $(1,1,1,0,1)$, $(1,1,1,1,0)$	$\frac{1}{15}$	$\frac{1}{3}$
Two missing values	$(0,0,1,1,1)$, $(0,1,0,1,1)$, $(0,1,1,0,1)$, $(0,1,1,1,0)$, $(1,0,0,1,1)$, $(1,0,1,0,1)$, $(1,0,1,1,0)$, $(1,1,0,0,1)$, $(1,1,0,1,0)$, $(1,1,1,0,0)$	$\frac{1}{30}$	$\frac{1}{3}$
Three missing values	$(0,0,0,1,1)$, $(0,0,1,0,1)$, $(0,0,1,1,0)$, $(0,1,0,0,1)$, $(0,1,0,1,0)$, $(0,1,1,0,0)$, $(1,0,0,0,1)$, $(1,0,0,1,0)$, $(1,0,1,0,0)$, $(1,1,0,0,0)$	$\frac{1}{30}$	$\frac{1}{3}$

Table 5.3: Admissible response patterns in the simulation study.

The above setup also determines the proportion of missing values. Depending on the missingness probability p we have an expected proportion of missing values of

$$\mathbb{E}[\text{Proportion of missing values}] = \frac{1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{3} + 3 \cdot \frac{1}{3}}{5} \cdot p \approx 0.4p.$$

In the simulation study we investigate the missingness probabilities $p \in \{0.1, 0.3, 0.5\}$. These result in the following expected proportions of missing values:

Missingness probability p	$\mathbb{E}$[Proportion of missing values]
0.1	0.04
0.3	0.12
0.5	0.20

Table 5.4: Investigated missingness probabilities in the simulation study.

Procedure

From the data generating R-vine of dimension $d = 5$, for all response mechanisms which are characterized by the missingess probabilities $p \in \{0.1, 0.3, 0.5\}$ we are generating $S = 200$ incomplete data sets with $n = 500$ observations. These are then imputed $M = 10$ times with each method $\mathcal{M} \in \{\mathcal{M}_\text{a}, \mathcal{M}_\text{c}, \mathcal{M}_\text{m}, \mathcal{M}_\text{v}\}$. Finally, Rubin's rules of of Remark 3.1.4 are applied to derive posterior distributions for the quantities of interest $\{q_l\}_{l=1,\ldots,20}$. The exact procedure is given in Algorithm 2.

Algorithm 2: Procedure of the simulation study

```
for $p \in \{0.1, 0.3, 0.5\}$ do
    for $s = 1, \ldots, S$ do
        Generate an incomplete $n \times d$ data set $(\mathbf{y}^{(p,s)}, \mathbf{r}^{(p,s)})$ from the data
          generating distribution and the response mechanism;
        foreach $\mathcal{M} \in \{\mathcal{M}_\text{a}, \mathcal{M}_\text{c}, \mathcal{M}_\text{m}, \mathcal{M}_\text{v}\}$ do
            Impute $M$ times to obtain the imputed data sets
              $\mathbf{y}^{(p,s,\mathcal{M},m)} = (\mathbf{y}_\text{obs}^{(p,s)}, \mathbf{y}_\text{mis}^{(p,s,\mathcal{M},m)})$, $m = 1, \ldots, M$;
            /* Applying Rubin's rules of Remark 3.1.4                    */
            for $l = 1, \ldots, k$ do
                for $m = 1, \ldots, M$ do
                    $\widehat{q}_l^{(p,s,\mathcal{M},m)} := \widehat{Q}_l^{(m)}(\mathbf{y}^{(p,s,\mathcal{M},m)})$, $\widehat{w}_l^{(p,s,\mathcal{M},m)} := \widehat{W}_l^{(m)}(\mathbf{y}^{(p,s,\mathcal{M},m)})$;
                $\overline{q}_l^{(p,s,\mathcal{M})} := \overline{Q}_l(\mathbf{y}^{(p,s,\mathcal{M},1)}, \ldots, \mathbf{y}^{(p,s,\mathcal{M},M)})$ (see Equation (3.9));
                $t_l^{(p,s,\mathcal{M})} := T_l(\mathbf{y}^{(p,s,\mathcal{M},1)}, \ldots, \mathbf{y}^{(p,s,\mathcal{M},M)})$ (see Equation (3.10));
                $\nu_l^{(p,s,\mathcal{M})} := N_l(\mathbf{y}^{(p,s,\mathcal{M},1)}, \ldots, \mathbf{y}^{(p,s,\mathcal{M},M)})$ (see Equation (3.11));
```

In the above algorithm, for the calculation of each $\nu_l^{(p,s,\mathcal{M})}$ we have to slightly deviate from Equation (3.11). In Equation (3.11) we actually need the number of degrees of freedom ν_{com} which would result if we were to estimate the quantities of interest for the hypothetically complete data sets $\mathbf{y}^{(p,s)}$ from $\widehat{Q}_1(\mathbf{Y}), \ldots, \widehat{Q}_{20}(\mathbf{Y})$. Typically, ν_{com} equals the number of observations n reduced by the number of parameters that have to be estimated for the estimation of the quantities of interest which is often known in advance. In our situation, for the estimation of Kendall's τ's via fitting an R-vine, the number of parameters of the marginal kernel densities as well as the number of parameters of the pair-copulas is not known beforehand since we have incomplete data. Instead, we know the number of parameters we needed for the estimation of $\{\widehat{q}_l^{(p,s,\mathcal{M},m)}\}_{l=1,\ldots,20}$ from every completed data set $\mathbf{y}^{(p,s,\mathcal{M},m)}$. Thus, we also know the corresponding number of degrees of freedom which we denote as $\nu_{\text{com}}^{(p,s,\mathcal{M},m)}$. For every p, s and $\mathcal{M}$ we define the estimator

$$\widehat{\nu}_{\text{com}}^{(p,s,\mathcal{M})} := \frac{1}{M} \sum_{m=1}^{M} \nu_{\text{com}}^{(p,s,\mathcal{M},m)}$$

which we use instead ν_{com} for the calculation of $\nu_l^{(p,s,\mathcal{M})}$.

All parameters choices for the simulation study are summarized once again in Table 5.5.

Parameter	Description
$p \in \{0.1, 0.3, 0.5\}$	Missingness probability
$M = 10$	Number of imputations
$\mathcal{M} \in \{\mathcal{M}_{\text{v}}, \mathcal{M}_{\text{m}}, \mathcal{M}_{\text{a}}, \mathcal{M}_{\text{c}}\}$	Imputation method
$S = 200$	Number of samples per missingness probability p
$T = 20$	Number of iterations for the quasi-Gibbs sampler
$K = 5$	Number of nearest neighbors
$d = 5$	Number of variables
$n = 500$	Number of observations
$k = 20$	Number of quantities of interest

Table 5.5: Summary of the parameters used in the simulation study.

Evaluation Criteria

In Section 3.4 we introduced the concept of randomization valid inference as a desirable property of imputation methods and gave a procedure to investigate it. There, without explicitly naming, we already described the **raw bias** and the **coverage rate** verbally which serve as evaluation criteria for randomization validity (Buuren 2018, pp. 52–53). For each $p \in \{0.1, 0.3, 0.5\}$, $\mathcal{M} \in \{\mathcal{M}_{\text{v}}, \mathcal{M}_{\text{m}}, \mathcal{M}_{\text{a}}, \mathcal{M}_{\text{c}}\}$ and $l = 1, \ldots, k$ we define:

(a) **Raw bias:**

$$\text{RB}_l^{(p,\mathcal{M})} := \frac{1}{S} \sum_{s=1}^{S} \overline{q}_l^{(p,s,\mathcal{M})} - q_l, \tag{5.1}$$

(b) **Coverage rate for the confidence interval at level** $1-\alpha$**:**

$$\mathrm{CR}_{\alpha,l}^{(p,\mathcal{M})} := \frac{1}{S}\sum_{s=1}^{S} \mathbb{1}\{q_l \in \mathrm{CI}_{\alpha,l}^{(p,s,\mathcal{M})}\}, \tag{5.2}$$

where the confidence interval $\mathrm{CI}_{\alpha,l}^{(p,s,\mathcal{M})}$ is calculated based on Equation (3.8) as

$$\mathrm{CI}_{\alpha,l}^{(p,s,\mathcal{M})} := \left[\overline{q}_l^{(p,s,\mathcal{M})} - t_{\nu_l^{(p,s,\mathcal{M})}}\left(\frac{\alpha}{2}\right)\sqrt{t_l^{(p,s,\mathcal{M})}}, \overline{q}_l^{(p,s,\mathcal{M})} + t_{\nu_l^{(p,s,\mathcal{M})}}\left(\frac{\alpha}{2}\right)\sqrt{t_l^{(p,s,\mathcal{M})}}\right]$$

and $t_{\nu_l^{(p,s,\mathcal{M})}}\left(\frac{\alpha}{2}\right)$ is the $\frac{\alpha}{2}$-quantile of the t-distribution on $\nu_l^{(p,s,\mathcal{M})}$ degrees of freedom.

If the raw biases are close to zero and the coverage rates are close to the chosen confidence level $1-\alpha$ it is likely that the imputation methods are randomization valid.

Rubin's rules are designed to make statistical inference about the quantities of interest generating only a handful of imputations. They rely on the assumption that the posteriors of the quantities of interest follow normal distributions. If this normality assumption is severely violated the application of Rubin's rules is not reasonable. In that case, a more adequate alternative is to generate a sufficiently large amount of imputations to make inference about the quantities of interest from the empirical posterior. Within the simulation study, for each $p \in \{0.1, 0.3, 0.5\}$ and $\mathcal{M} \in \{\mathcal{M}_\mathrm{v}, \mathcal{M}_\mathrm{m}, \mathcal{M}_\mathrm{a}, \mathcal{M}_\mathrm{c}\}$ an appropriate empirical posterior cumulative distribution function $\widehat{F}_l^{(p,\mathcal{M})}$ for the quantity of interest Q_l can be defined using the independent sample $\{\widehat{q}_l^{(p,s,\mathcal{M},m)}\}_{s=1,\dots,S,\ m=1,\dots,M}$ as

$$\widehat{F}_l^{(p,\mathcal{M})}(q) := \frac{1}{S\cdot M}\sum_{s=1}^{S}\sum_{m=1}^{M} \mathbb{1}\{\widehat{q}_l^{(p,s,\mathcal{M},m)} \leq q\}.$$

To compare these posterior cumulative distribution functions for different $p \in \{0.1, 0.3, 0.5\}$ and $\mathcal{M} \in \{\mathcal{M}_\mathrm{v}, \mathcal{M}_\mathrm{m}, \mathcal{M}_\mathrm{a}, \mathcal{M}_\mathrm{c}\}$ we will use two so-called **scoring rules**: the **interval score** and the **continuous ranked probability score** (Gneiting and Raftery 2007). Scoring rules typically asses the quality of forecast distributions. Here, we consider the posterior distribution as a forecast distribution.

Definition 5.1.1 (Interval Score, Continuous Ranked Probability Score)**.** Let Q denote a real-valued quantity of interest distributed according to the posterior cumulative distribution function F_Q. Let q be the true value of Q. Then:

(a) The **interval score** $\mathrm{IS}_\alpha(F_Q, q)$ at level $1-\alpha$ is defined as

$$\mathrm{IS}_\alpha(F_Q, q) = (u-l) + \frac{2}{\alpha}(l-q)\mathbb{1}\{q < l\} + \frac{2}{\alpha}(q-u)\mathbb{1}\{q > u\},$$

where $l := F_Q^{-1}(\frac{\alpha}{2})$ and $u := F_Q^{-1}(1-\frac{\alpha}{2})$ denote the $\frac{\alpha}{2}$- and $(1-\frac{\alpha}{2})$-quantile of Q.

(b) The **continuous ranked probability score** is defined as

$$\mathrm{CRPS}(F_Q, q) := \int_{\mathbb{R}} (F_Q(\widetilde{q}) - \mathbb{1}\{q \leq \widetilde{q}\})^2 \, \mathrm{d}\widetilde{q}. \qquad \triangle$$

Using the empirical posterior cumulative distribution functions $\widehat{F}_l^{(p,\mathcal{M})}$ in the above definition, for every $p \in \{0.1, 0.3, 0.5\}$, $\mathcal{M} \in \{\mathcal{M}_\mathrm{v}, \mathcal{M}_\mathrm{m}, \mathcal{M}_\mathrm{a}, \mathcal{M}_\mathrm{c}\}$ and $l = 1, \ldots, k$ we define the evaluation criteria:

- **sample interval score at level** $1 - \alpha$**:**

$$\widehat{\mathrm{IS}}_{\alpha,l}^{(p,\mathcal{M})} := \mathrm{IS}_\alpha(\widehat{F}_l^{(p,\mathcal{M})}, q_l), \tag{5.3}$$

- **sample continuous ranked probability score:**

$$\widehat{\mathrm{CRPS}}_l^{(p,\mathcal{M})} := \mathrm{CRPS}(\widehat{F}_l^{(p,\mathcal{M})}, q_l). \tag{5.4}$$

The lower $\widehat{\mathrm{IS}}_{\alpha,l}^{(p,\mathcal{M})}$ and $\widehat{\mathrm{CRPS}}_l^{(p,\mathcal{M})}$ the better the empirical posterior distribution $\widehat{F}_l^{(p,\mathcal{M})}$ predicts the true value q_l. The sample interval score and the sample continuous ranked probability score are calculated via the `wis()` and `crps_sample()` functions from the R-packages `scoringutils` (Bosse et al. 2024) and `scoringRules` (Jordan, Krüger, and Lerch 2019).

5.2 Analysis of the Results

Computation Time

Figure 5.2 shows box plots comparing the different imputation methods $\mathcal{M} \in \{\mathcal{M}_\mathrm{v}, \mathcal{M}_\mathrm{m}, \mathcal{M}_\mathrm{a}, \mathcal{M}_\mathrm{c}\}$ with regard to the computation time needed to impute a data set $M = 10$ times on a single core (Intel® Xeon® Processor E5-2690 v3, CPUSs@2.6 GHz). To understand the figure it is important to know that we limited the computation time for all methods together to 24 hours. We imputed the data sets with the different methods in the order $\mathcal{M}_\mathrm{v}$, $\mathcal{M}_\mathrm{m}$, $\mathcal{M}_\mathrm{a}$ and $\mathcal{M}_\mathrm{c}$. The methods $\mathcal{M}_\mathrm{v}$, $\mathcal{M}_\mathrm{m}$ and $\mathcal{M}_\mathrm{a}$ always finished in time. The method $\mathcal{M}_\mathrm{c}$ only finished 154 times for $p = 0.1$, 178 times for $p = 0.3$ and 186 times for $p = 0.5$. Since the other methods together took a maximum time of approximately 177 minutes in those cases the method $\mathcal{M}_\mathrm{c}$ would have needed at least 1263 minutes to complete the imputation.

The method $\mathcal{M}_\mathrm{a}$ was the fastest of all methods. The imputation time mostly ranges from approximately 0.15 seconds to approximately 0.4 seconds with some outliers up to 0.85 seconds. With increasing missingness probability p also the computation times increased. The second fastest methods was $\mathcal{M}_\mathrm{m}$ with computation times from 10.5 to 12

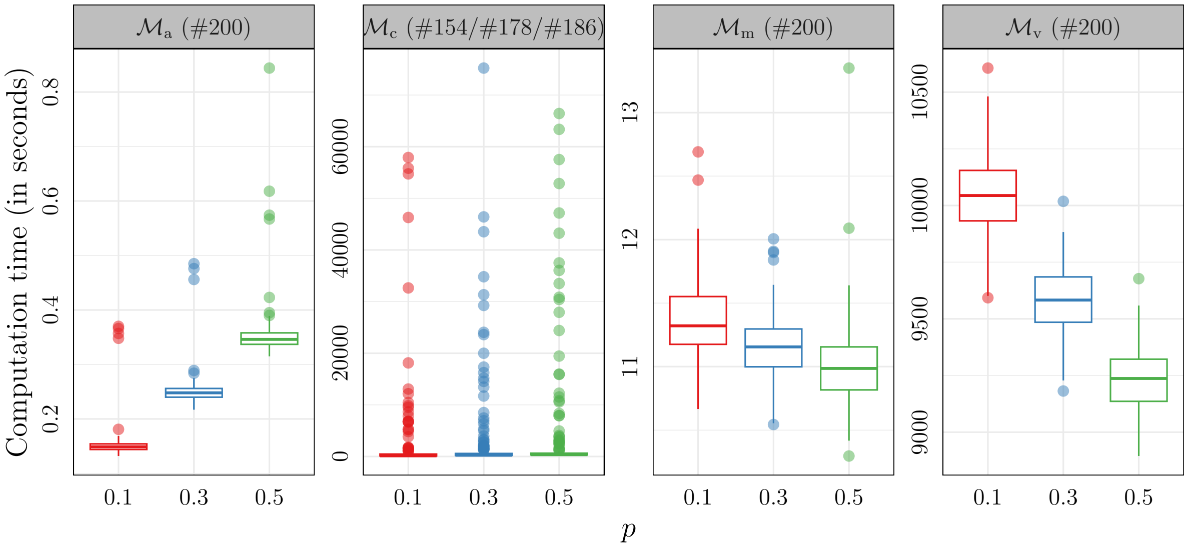


Figure 5.2: Box plots of the computation time (in seconds) needed to impute a data set $M = 10$ times using a single core applying the imputation methods $\mathcal{M}_a$, $\mathcal{M}_c$, $\mathcal{M}_m$ and $\mathcal{M}_v$ depending on the missingness probability $p \in \{0.1, 0.3, 0.5\}$ with the number of completely imputed data sets in parentheses.

seconds. Here the computation time decreased with increasing missingness probability p. We observed the longest computation times for the methods $\mathcal{M}_v$ and $\mathcal{M}_c$. While the median times for $\mathcal{M}_c$ are relatively low with approximately 132 seconds for $p = 0.1$, 243 seconds for $p = 0.3$ and 352 seconds for $p = 0.5$, there are several extreme outliers up to 21 hours. Overall, the method $\mathcal{M}_c$ is very unstable and not very reliable. For $\mathcal{M}_v$, the computation time is high but stable and ranges from approximately 141 minutes to 177 minutes. For higher p the computation times decreased. This is reasonable since the estimation of the copulas fastens the less observations have to be taken into account.

Convergence of the Methods Applying a Quasi-Gibbs Sampler

For the two methods $\mathcal{M}_v$ and $\mathcal{M}_m$ that make use of a quasi-Gibbs sampler we investigate the convergence by comparing the so-called **Gelman-Rubin statistics** proposed by Gelman and Rubin (1992), which allows a statement about the convergence in Markov chain Monte Carlo simulations. A Gelman-Rubin statistic close to 1 indicates convergence. Gelman, Carlin, et al. (2013, p. 287) suggest accepting convergence when the value of the Gelman-Rubin statistic is below 1.1. We used the `convergence()` function of the `mice` R-package (Buuren and Groothuis-Oudshoorn 2011) to calculate the Gelman-Rubin statistics. Figure 5.3 gives box plots of the Gelman-Rubin statistics per variable for both methods distinguished by the missingness probability p. For $\mathcal{M}_v$ the convergence is much better compared to $\mathcal{M}_m$. Especially for the variables Y_2 and Y_3 the Gelman-Rubin statistics are almost always below the threshold of 1.1. For method $\mathcal{M}_v$ as well as for method $\mathcal{M}_m$, the most problematic variables are Y_4 and Y_5.

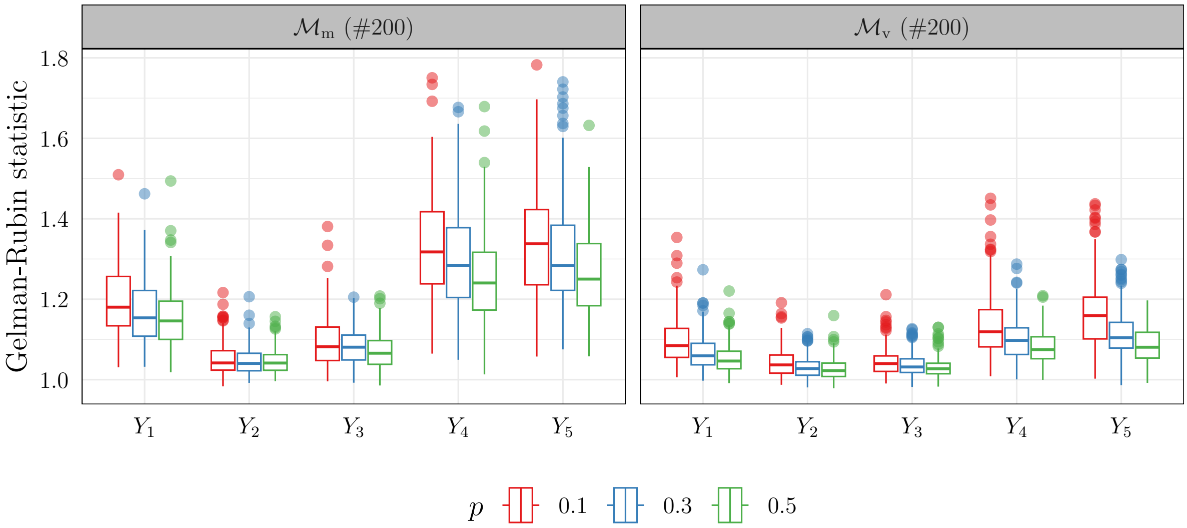


Figure 5.3: Box plots of the Gelman-Rubin statistics of the individual variables separated according to the missingness probability $p \in \{0.1, 0.3, 0.5\}$ for the imputation methods $\mathcal{M}_\mathrm{m}$ and $\mathcal{M}_\mathrm{v}$.

Randomization Validity

The raw biases are given in Figure 5.4. The methods $\mathcal{M}_\mathrm{v}$, $\mathcal{M}_\mathrm{m}$ and $\mathcal{M}_\mathrm{a}$ perform very similar. For the means and standard deviations the raw biases are almost zero. Also for the unconditional Kendall's τ's the raw biases are close to zero. For the conditional Kendall's τ's, the raw biases tend to deviate more from zero the more conditioning variables they include. This is realistic, as the Kendell's τ's of the higher trees are based on the lower trees and estimation errors therefore accumulate. It is worth mentioning that the method $\mathcal{M}_\mathrm{a}$ noticeable differs from the methods $\mathcal{M}_\mathrm{v}$ and $\mathcal{M}_\mathrm{m}$ precisely for the quantities of interest that relate to the discrete variable Y_2. This effect increases the larger p is. One possible explanation for this behavior might be that the method $\mathcal{M}_\mathrm{a}$ is originally intended for continuous variables since it is based on the multivariate normal distribution, while the other two methods have been developed also for discrete variables. For method $\mathcal{M}_\mathrm{c}$ it is clear that the raw biases do not equal zero for most of the quantities of interest. The higher p the higher the absolute deviation from zero.

The coverage rates for a nominal coverage rate $1-\alpha = 0.95$ are summarized in Figure 5.5 on page 58. For all methods the coverage rates of the means and standard deviations are clearly closer to the nominal level of 0.95 than the coverage rates of Kendall's τ's, whereby here a higher coverage rate tends to be achieved for the unconditional Kendall's τ's. As before, this seems reasonable since the Kendell's τ's of the higher trees are based on the lower trees and estimation errors therefore sum up. Striking but not surprising, all methods react negatively on increasing missingness probability p with worse coverage rates.

Nevertheless, there are also noticeable differences between the individual methods.

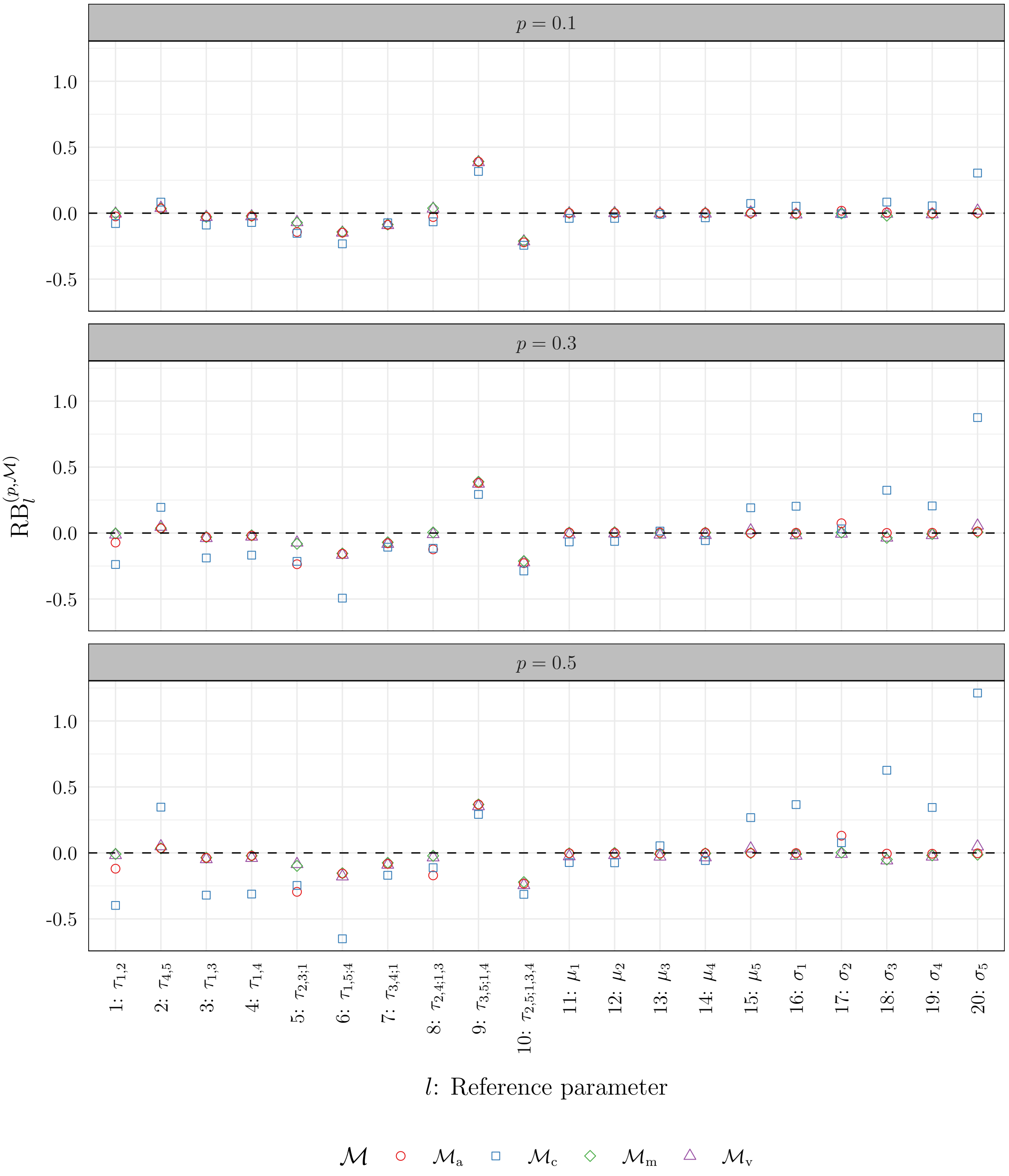


Figure 5.4: Raw biases $\mathrm{RB}_l^{(p,\mathcal{M})}$ as defined in Equation (5.1) of the quantities of interest Q_l, $l = 1, \dots, 20$, depending on the missingness probability p ($p = 0.1$: top panel, $p = 0.3$: middle panel, $p = 0.5$: bottom panel) and the imputation method $\mathcal{M}$ ($\mathcal{M}_\mathrm{a}$: red circle, $\mathcal{M}_\mathrm{c}$: blue square, $\mathcal{M}_\mathrm{m}$: green diamond, $\mathcal{M}_\mathrm{v}$: purple triangle).

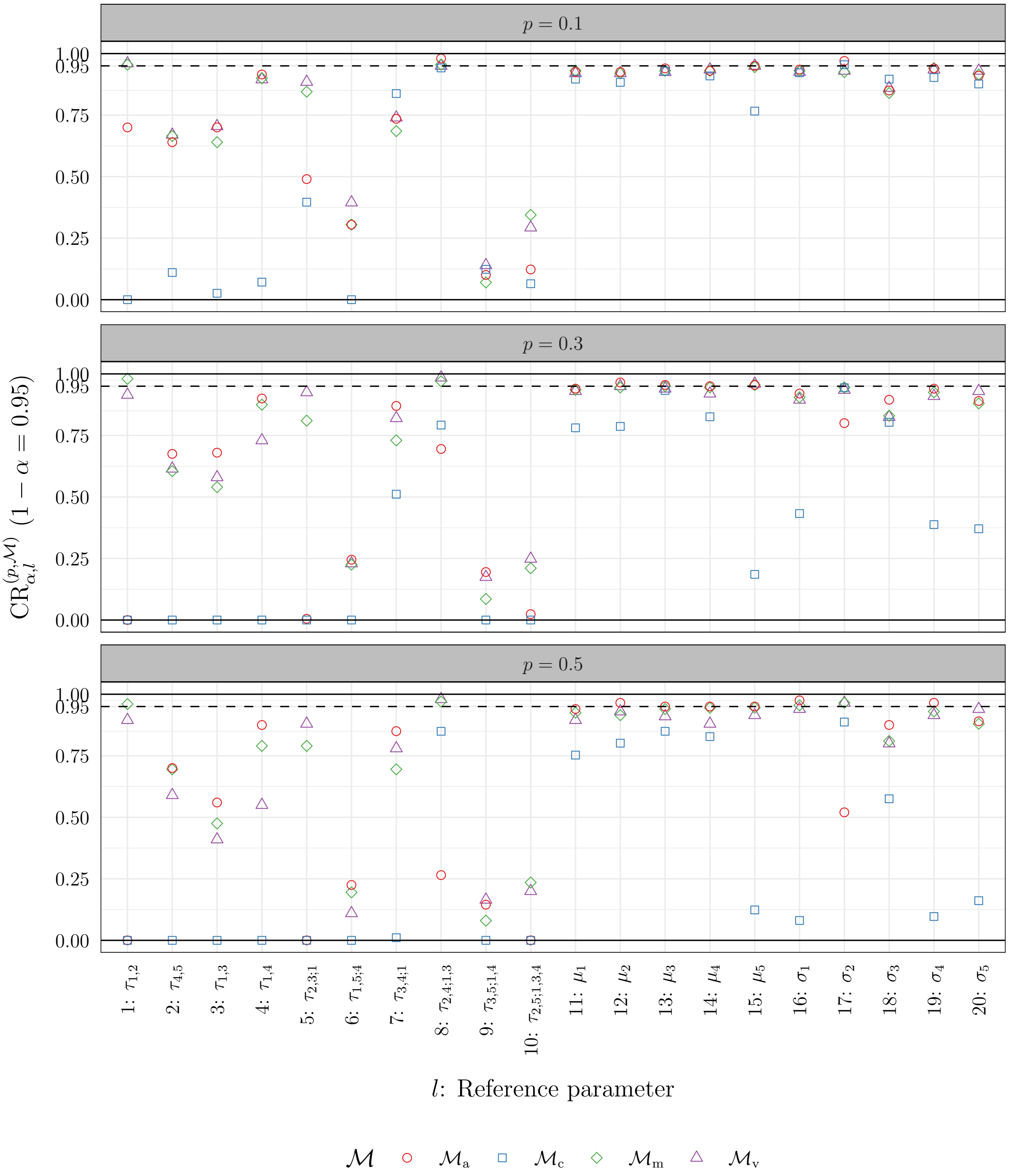


Figure 5.5: Coverage rates $\mathrm{CR}_{\alpha,l}^{(p,\mathcal{M})}$ as defined in Equation (5.2) with a nominal coverage of $1-\alpha = 0.95$ of the quantities of interest Q_l, $l = 1, \dots, 20$, depending on the missingness probability p ($p = 0.1$: top panel, $p = 0.3$: middle panel, $p = 0.5$: bottom panel) and the imputation method $\mathcal{M}$ ($\mathcal{M}_\mathrm{a}$: red circle, $\mathcal{M}_\mathrm{c}$: blue square, $\mathcal{M}_\mathrm{m}$: green diamond, $\mathcal{M}_\mathrm{v}$: purple triangle).

Again, $\mathcal{M}_\mathrm{c}$ performs clearly worse than the other methods. For many cases the coverage rate equals even zero. For the other methods the situation is somewhat more differentiated. For $p = 0.1$ and $p = 0.3$ the method $\mathcal{M}_\mathrm{v}$ performs similarly good or, for Kendall's $\tau's$ even better than method $\mathcal{M}_\mathrm{m}$. For $p = 0.5$ and the coverage rates of Kendall's τ's this situation changes since the method $\mathcal{M}_\mathrm{v}$ reacts more to the increasing missingness probability p than $\mathcal{M}_\mathrm{m}$. Except for the quantities of interest that relate to the discrete variable Y_2 the method $\mathcal{M}_\mathrm{a}$ almost always performs equal or better than methods $\mathcal{M}_\mathrm{v}$ and $\mathcal{M}_\mathrm{m}$. Also, except for quantities of interest that relate to the discrete variable Y_2 the method $\mathcal{M}_\mathrm{a}$ reacts less strongly on an increasing p.

Taking both criteria into account, we see that the method $\mathcal{M}_\mathrm{c}$ is definitively not randomization valid. Ignoring the discrete variable Y_2 the method $\mathcal{M}_\mathrm{a}$ seems the closest to randomization validity. Taking Y_2 into account, for $p = 0.1$ and $p = 0.3$ the method $\mathcal{M}_\mathrm{v}$ seems the closest to randomization validity. For $p = 0.5$ and under consideration of Y_2 the method $\mathcal{M}_\mathrm{m}$ performs best.

Sample Interval Scores and Continuous Ranked Probability Scores

Comparing the sample interval scores and continuous ranked probability scores shows a similar picture as the analysis of the raw biases and coverage rates. The sample interval scores are given in Table 5.6. The sample continuous ranked probability scores can be found in Table 5.7 on page 61. For both scores, independent of the missingness probability p, the method $\mathcal{M}_\mathrm{c}$ is almost always inferior by a wide margin to the other methods. In addition, both scores increase sharply with increasing p. For the remaining methods $\mathcal{M}_\mathrm{v}$, $\mathcal{M}_\mathrm{m}$ and $\mathcal{M}_\mathrm{a}$ one has to differentiate between (a) the parameters which only concern the continuous variables Y_1, Y_3, Y_4 and Y_5 and (b) the parameters which also concern the discrete variable Y_2.

(a) Looking at the parameters which only concern the **continuous variables**, the three methods $\mathcal{M}_\mathrm{v}$, $\mathcal{M}_\mathrm{m}$ and $\mathcal{M}_\mathrm{a}$ perform on a same level. Nevertheless, across all parameters and for all missingness probabilities p, in the most cases the sample interval scores of $\mathcal{M}_\mathrm{m}$ are slightly lower. Concerning the sample continuous ranked probability scores, for the means and standard deviations the method $\mathcal{M}_\mathrm{a}$ performs marginally better than the methods $\mathcal{M}_\mathrm{v}$ and $\mathcal{M}_\mathrm{m}$ while for Kendall's $\tau's$ the method $\mathcal{M}_\mathrm{m}$ shows minor advantages over the methods $\mathcal{M}_\mathrm{v}$ and $\mathcal{M}_\mathrm{a}$.

(b) Looking at the parameters which also concern the **discrete variable** Y_2, it becomes apparent that the method $\mathcal{M}_\mathrm{a}$ is noticeable worse than the methods $\mathcal{M}_\mathrm{v}$ and $\mathcal{M}_\mathrm{m}$. This effect increases the larger p is.

Overall, the empirical cumulative distribution function derived from the imputations with method $\mathcal{M}_\mathrm{c}$ is the least suitable to describe the true values of the quantities of interest. Compared to the other methods, $\mathcal{M}_\mathrm{a}$ only gives suitable empirical cumulative distribution functions for the parameters exclusively concerning the continuous variables. The methods $\mathcal{M}_\mathrm{v}$ and $\mathcal{M}_\mathrm{m}$ result in the most suitable empirical cumulative distribution functions over all quantities of interest.

l	Reference parameter	$p = 0.1$				$p = 0.3$				$p = 0.5$			
		$\mathcal{M}_{\text{v}}$	$\mathcal{M}_{\text{m}}$	$\mathcal{M}_{\text{a}}$	$\mathcal{M}_{\text{c}}$	$\mathcal{M}_{\text{v}}$	$\mathcal{M}_{\text{m}}$	$\mathcal{M}_{\text{a}}$	$\mathcal{M}_{\text{c}}$	$\mathcal{M}_{\text{v}}$	$\mathcal{M}_{\text{m}}$	$\mathcal{M}_{\text{a}}$	$\mathcal{M}_{\text{c}}$
1	$\tau_{1,2}$	0.0419	**0.0382**	0.0508	1.7899	0.0565	**0.0434**	1.7149	7.2984	0.0652	**0.0469**	3.4727	13.5000
2	$\tau_{4,5}$	0.0897	**0.0872**	0.0881	1.9850	0.3077	**0.0842**	0.0847	5.5676	0.4495	**0.0850**	0.0863	10.9205
3	$\tau_{1,3}$	0.0727	0.0708	**0.0704**	2.0401	0.2390	**0.0707**	0.0819	5.7261	0.4914	**0.1463**	0.1777	10.5609
4	$\tau_{1,4}$	0.0831	**0.0830**	0.0857	1.6101	**0.0730**	0.0737	0.0806	4.8067	0.2977	**0.0722**	0.0770	9.9488
5	$\tau_{2,3;1}$	0.2518	**0.2424**	1.0143	2.1251	0.2604	**0.2521**	5.4965	5.0102	0.2747	**0.2541**	7.9942	6.6106
6	$\tau_{1,5;4}$	2.4838	**2.4234**	2.4942	6.3800	2.8120	**2.6381**	2.6385	15.1214	3.4896	**2.7696**	2.7763	23.1190
7	$\tau_{3,4;1}$	0.2514	**0.2465**	0.2496	0.3383	0.2348	**0.2158**	0.2289	1.3197	0.2371	0.2264	**0.2242**	3.7811
8	$\tau_{2,4;1,3}$	0.3324	0.3354	0.3177	**0.2709**	0.3344	0.3282	0.2574	**0.2171**	0.3568	**0.3305**	2.7406	0.8189
9	$\tau_{3,5;1,4}$	8.1143	**7.9864**	8.0619	8.2853	**7.6485**	7.7299	8.1490	7.9420	**7.1796**	7.8054	7.7582	8.0386
10	$\tau_{2,5;1,3,4}$	2.7088	**2.6618**	3.7554	4.6436	2.8298	**2.3143**	5.5877	8.1872	3.1253	**2.7017**	6.1808	9.4923
11	μ_1	0.2012	0.2017	0.2009	**0.1989**	**0.1817**	0.1857	0.1846	0.2393	0.1947	0.1873	**0.1844**	0.3157
12	μ_2	0.2000	**0.1980**	0.2000	**0.1980**	0.1820	0.1781	0.1821	**0.1540**	0.1860	0.1980	0.2000	**0.1630**
13	μ_3	0.1929	0.1960	**0.1890**	0.2232	0.1806	**0.1768**	0.1789	0.3182	**0.1786**	0.1846	0.1818	0.4684
14	μ_4	0.1885	0.1883	**0.1871**	0.1938	0.1765	**0.1727**	0.1729	0.2308	0.1838	**0.1778**	0.1800	0.3103
15	μ_5	**0.1799**	0.1800	0.1816	0.2393	0.1812	**0.1669**	0.1670	2.0775	0.1841	0.1723	**0.1688**	2.2006
16	σ_1	0.2545	**0.2532**	0.2550	0.3138	0.2851	0.2669	**0.2659**	0.4613	0.2562	0.2441	**0.2369**	4.4971
17	σ_2	0.2167	0.2176	0.2240	**0.2134**	**0.2208**	0.2225	0.2328	0.2364	0.2140	**0.2090**	1.0543	0.2112
18	σ_3	0.8421	**0.7960**	0.7974	1.2953	0.8586	0.7718	**0.7270**	1.9814	0.7683	0.7807	**0.7413**	3.0884
19	σ_4	**0.2717**	0.2796	0.2781	0.3124	**0.2619**	0.2635	0.2702	0.8775	0.2554	**0.2325**	0.2335	4.1751
20	σ_5	0.7897	0.7525	**0.7500**	1.6663	0.8139	**0.7170**	0.7184	3.0014	0.8329	**0.6669**	0.6693	5.0729

Table 5.6: Sample interval scores $\widehat{\text{IS}}_{\alpha,l}^{(p,\mathcal{M})}$ at level $1-\alpha = 0.95$ as defined in Equation (5.3) for the quantities of interest Q_l, $l = 1, \ldots, 20$, per missingess probability $p \in \{0.1, 0.3, 0.5\}$ and imputation method $\mathcal{M} \in \{\mathcal{M}_{\text{v}}, \mathcal{M}_{\text{m}}, \mathcal{M}_{\text{a}}, \mathcal{M}_{\text{c}}\}$ with the lowest sample interval score over the methods marked bold for each quantity of interest and missingness probability.

l	Reference parameter	$p = 0.1$				$p = 0.3$				$p = 0.5$			
		$\mathcal{M}_\text{v}$	$\mathcal{M}_\text{m}$	$\mathcal{M}_\text{a}$	$\mathcal{M}_\text{c}$	$\mathcal{M}_\text{v}$	$\mathcal{M}_\text{m}$	$\mathcal{M}_\text{a}$	$\mathcal{M}_\text{c}$	$\mathcal{M}_\text{v}$	$\mathcal{M}_\text{m}$	$\mathcal{M}_\text{a}$	$\mathcal{M}_\text{c}$
1	$\tau_{1,2}$	0.0029	**0.0023**	0.0174	0.0678	0.0068	**0.0032**	0.0631	0.2211	0.0098	**0.0047**	0.1088	0.3792
2	$\tau_{4,5}$	0.0232	**0.0194**	0.0200	0.0708	0.0313	0.0222	**0.0212**	0.1757	0.0347	**0.0209**	**0.0209**	0.3232
3	$\tau_{1,3}$	0.0177	0.0168	**0.0167**	0.0780	0.0272	**0.0203**	**0.0203**	0.1744	0.0366	**0.0251**	0.0263	0.3029
4	$\tau_{1,4}$	0.0130	**0.0110**	0.0122	0.0604	0.0178	**0.0094**	0.0102	0.1526	0.0285	**0.0128**	0.0131	0.2918
5	$\tau_{2,3;1}$	**0.0382**	0.0413	0.1050	0.1220	**0.0413**	0.0479	0.2040	0.1873	**0.0501**	0.0615	0.2655	0.2217
6	$\tau_{1,5;4}$	0.1082	**0.1044**	0.1090	0.2060	0.1281	**0.1171**	0.1190	0.4604	0.1432	**0.1168**	0.1195	0.6281
7	$\tau_{3,4;1}$	0.0511	**0.0473**	0.0487	0.0493	0.0476	**0.0367**	0.0421	0.0835	0.0544	**0.0426**	0.0487	0.1450
8	$\tau_{2,4;1,3}$	0.0253	0.0270	**0.0233**	0.0408	0.0225	**0.0198**	0.0886	0.0835	0.0278	**0.0226**	0.1376	0.0818
9	$\tau_{3,5;1,4}$	0.3022	0.3069	0.3061	**0.2669**	0.2923	0.3024	0.3006	**0.2635**	0.2823	0.2889	0.2910	**0.2634**
10	$\tau_{2,5;1,3,4}$	**0.1766**	0.1777	0.1988	0.2091	0.1854	**0.1817**	0.2131	0.2619	0.2030	**0.1864**	0.2177	0.2897
11	μ_1	0.0119	**0.0117**	**0.0117**	0.0253	0.0120	0.0106	**0.0104**	0.0418	0.0164	0.0114	**0.0112**	0.0419
12	μ_2	**0.0113**	**0.0113**	0.0114	0.0245	**0.0104**	0.0106	0.0108	0.0425	0.0129	**0.0108**	0.0113	0.0527
13	μ_3	0.0117	**0.0114**	**0.0114**	0.0143	0.0131	0.0116	**0.0114**	0.0167	0.0182	0.0113	**0.0105**	0.0281
14	μ_4	0.0118	**0.0117**	0.0118	0.0217	0.0124	**0.0103**	0.0104	0.0352	0.0203	**0.0103**	**0.0103**	0.0350
15	μ_5	**0.0115**	0.0116	**0.0115**	0.0449	0.0132	0.0099	**0.0098**	0.1382	0.0191	0.0100	**0.0099**	0.1923
16	σ_1	0.0157	0.0157	**0.0151**	0.0299	0.0187	0.0174	**0.0170**	0.1391	0.0195	0.0160	**0.0148**	0.2765
17	σ_2	0.0129	0.0127	0.0156	**0.0126**	0.0133	**0.0130**	0.0466	0.0200	0.0131	**0.0126**	0.0987	0.0502
18	σ_3	0.0517	0.0505	**0.0476**	0.0596	0.0554	0.0509	**0.0422**	0.1320	0.0581	0.0513	**0.0438**	0.2667
19	σ_4	0.0157	0.0155	**0.0154**	0.0325	0.0181	0.0170	**0.0168**	0.1438	0.0225	0.0184	**0.0157**	0.2550
20	σ_5	0.0409	0.0412	**0.0407**	0.1346	0.0481	0.0444	**0.0443**	0.4670	0.0465	0.0444	**0.0429**	0.6725

Table 5.7: Sample continuous ranked probability scores $\widehat{\mathrm{CRPS}}_l^{(p,\mathcal{M})}$ as defined in Equation (5.4) for the quantities of interest Q_l, $l = 1, \dots, 20$, per missingess probability $p \in \{0.1, 0.3, 0.5\}$ and imputation method $\mathcal{M} \in \{\mathcal{M}_\text{v}, \mathcal{M}_\text{m}, \mathcal{M}_\text{a}, \mathcal{M}_\text{c}\}$ with the lowest sample continuous ranked probability score over the methods marked bold for each quantity of interest and missingness probability.

Summary

Within the simulation study, the method $\mathcal{M}_{\mathrm{c}}$ is not a suitable imputation method. This is mainly due to the poor performance with raw biases, coverage rates, interval scores and continuous ranked probability scores. However, the method is also computationally unreliable with a large variation in the calculation time. In contrast, the method $\mathcal{M}_{\mathrm{a}}$ is computationally very fast and generates appropriate imputations to reliable estimate the quantities of interest as long as they are concerned to continuous variables. The methods $\mathcal{M}_{\mathrm{v}}$ and $\mathcal{M}_{\mathrm{m}}$ are the best imputation methods for estimating the quantities of interest, independently of a continuous or discrete distribution of the individual variables, with slightly better results for $\mathcal{M}_{\mathrm{m}}$. While $\mathcal{M}_{\mathrm{m}}$ is much faster, the method $\mathcal{M}_{\mathrm{v}}$ shows better convergence for the quasi-Gibbs sampler.

6 Application: Imputation of ESG Data to Obtain ESG Subpillar Scores

6.1 Calculation Procedure of the ESG Score

The ESG data we will use has been provided by the London Stock Exchange Group (LSEG 2023). At the heart of this data is the ESG score. On an annual basis, it summarizes the ESG-related data published by a company in one single number and is intended to make different companies comparable. The ESG score is based on a multi-level summary of the ESG-related data: On the top level, it is calculated as a weighted sum of scores of three pillars: the Environmental (E), Social (S) and Governance (G) pillar. Each of those pillar scores is again the weighted sum of subpillar scores. To this end, the Social pillar is divided into four subpillars whereas the Environmental and Governance pillar are divided into three subpillars each – giving a total of ten subpillars. The connection between ESG score, pillar scores and subpillar scores is summarized in Figure 6.1 based on Benuzzi et al. (2024, Figure 2).

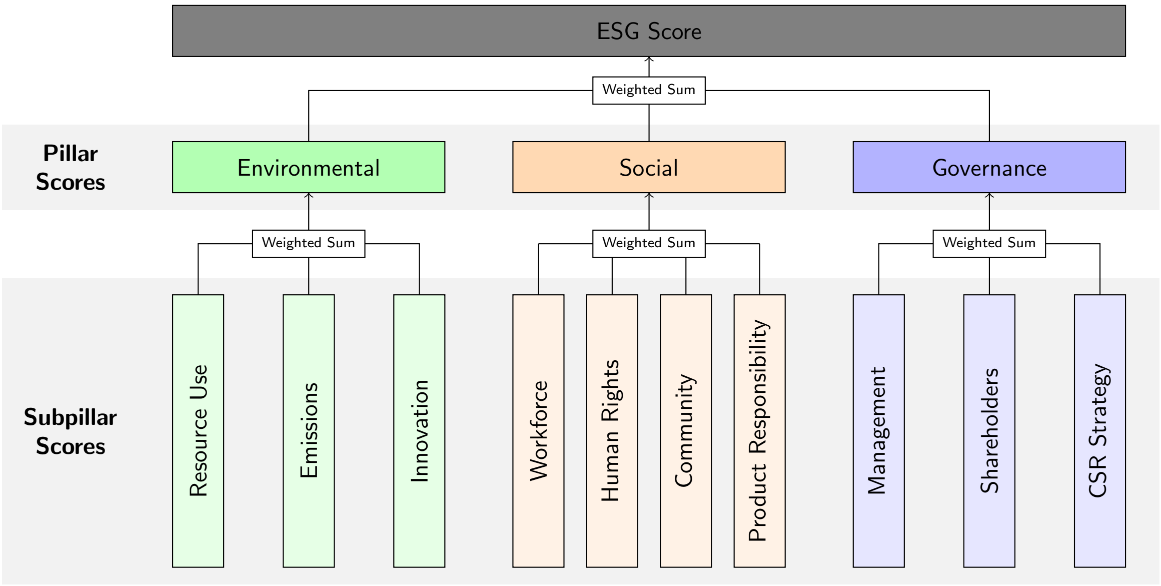


Figure 6.1: Flowchart of LSEG (2023) methodology for the ESG score.

The subpillar scores are in turn made up on more than 630 raw ESG-related indicators. For each company specific indicators are selected depending on the industry and the country of the company. From this subset of so-called scoring variables, for each of the subpillars the relevant scoring variables are selected. After this selection, the subpillar score is derived via a two level percentile ranking which compares companies of the same industry or the same country. For subpillars of the Environmental and Social pillar the companies operating withing the same industry are compared. For subpillars of the Governance pillar the companies of the same country serve as the benchmark. At first, the scoring variables are transformed into so-called indicator variables via percentile ranking. The final subpillar score is then derived via another percentile ranking applied to the sum of the indicator variables leading to a score between 0 and 100 percent. The derivation of the subpillar scores is summarized in Figure 6.2 based on Benuzzi et al. (2024, Figure 3).

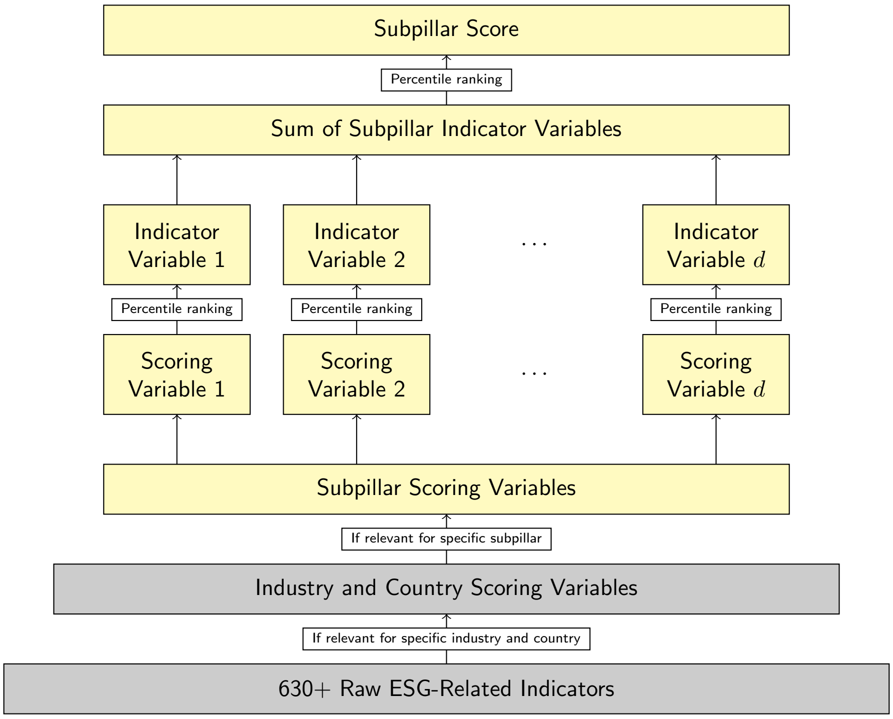

Subpillar Score
Percentile ranking
Sum of Subpillar Indicator Variables
Indicator Variable 1
Indicator Variable 2
...
Indicator Variable d
Percentile ranking
Percentile ranking
Percentile ranking
Scoring Variable 1
Scoring Variable 2
...
Scoring Variable d
Subpillar Scoring Variables
If relevant for specific subpillar
Industry and Country Scoring Variables
If relevant for specific industry and country
630+ Raw ESG-Related Indicators

Figure 6.2: Flowchart of the LSEG (2023) methodology for the subpillar scores.

The problem of missing data occurs at the level of the scoring variables. As a consequence of the calculation procedure, the missing values then affect the subpillar scores, the pillar scores and the ESG scores of potentially all companies even if companies report data for all scoring variables. From the perspective of the data provider, it would be reasonable to soften this problem by imputing the missing values in the scoring variables and to recalculate the subpillar, pillar and ESG scores afterwards on the basis of the completed data. A direct imputation of the scores for companies for which missing values occurred in the scoring variables is not possible since the other scores are affected by the missing values as well, such that potentially all scores have to be recalculated.

In practice, not only the overall ESG score is of interest but also the pillar scores and subpillar scores. For example, a fund manager could be restricted in his investments to companies for which the Emissions subpillar score lies above a defined threshold. For computational reasons, as seen in the simulation study the vine based method is computationally expensive, we will concentrate on the recalculation of a single subpillar score. To this end, let the random variables $Y_1, \dots, Y_d$ denote the d relevant scoring variables of the subpillar. The reported values of the scoring variable for the n companies of the considered industry or country are collected in the $n \times d$ incomplete data set $(\mathbf{Y}, \mathbf{R})$.

LSEG (2023) distinguishes between two types of scoring variables Y_j: numeric ones and boolean ones taking the values T(rue) and F(alse). Additionally, for each scoring variable Y_j a polarity is defined which compares each two possible values of the scoring variable. Mathematically spoken the polarity is nothing else than a total order $\preccurlyeq_j$ on the value set of Y_j. For numeric variables there are two possible polarities respective total orders either preferring larger values ($\preccurlyeq_j = \leq$) or preferring smaller values ($\preccurlyeq_j = \geq$). For boolean scoring variables either the value F is preferred over T ($\preccurlyeq_j$ is induced by T $\prec_j$ F) or the value T is preferred over F ($\preccurlyeq_j$ is induced by F $\prec_j$ T). Before we continue, we have to translate the above vague description of the subpillar score into a mathematical one. For this reason, we shortly recap how LSEG (2023) defines percentile ranks.

Definition 6.1.1 (Percentile Rank).

(a) Let Y_j be a **numeric** scoring variable with order $\preccurlyeq_j$. Then the percentile rank $\mathrm{PR}_{\mathbf{Y}_{:j}}^{\preccurlyeq_j}(Y_{ij})$ of $Y_{ij} \in \mathbf{Y}_{:j} = (\mathbf{Y}_{:j,\mathrm{obs}}, \mathbf{Y}_{:j,\mathrm{mis}})$ with respect to $\mathbf{Y}_{:j}$ and $\preccurlyeq_j$ is defined as

$$\mathrm{PR}_{\mathbf{Y}_{:j}}^{\preccurlyeq_j}(Y_{ij}) := \begin{cases} \frac{\sum_{Y_{lj} \in \mathbf{Y}_{:j,\mathrm{obs}}} \mathbb{1}\{Y_{lj} \prec_j Y_{ij}\} + \frac{1}{2} \sum_{Y_{lj} \in \mathbf{Y}_{:j,\mathrm{obs}}} \mathbb{1}\{Y_{lj} = Y_{ij}\}}{|\mathbf{Y}_{:j,\mathrm{obs}}|} & Y_{ij} \in \mathbf{Y}_{:j,\mathrm{obs}} \\ 0 & Y_{ij} \in \mathbf{Y}_{:j,\mathrm{mis}} \end{cases}.$$

(b) Let Y_j be an ordered **boolean** scoring variable with levels F and T where the order $\preccurlyeq_j$ is induced by F $\prec_j$ T [T $\prec_j$ F]. Then the percentile rank $\mathrm{PR}_{\mathbf{Y}_{:j}}^{\preccurlyeq_j}(Y_{ij})$ of $Y_{ij} \in \mathbf{Y}_{:j} = (\mathbf{Y}_{:j,\mathrm{obs}}, \mathbf{Y}_{:j,\mathrm{mis}})$ with respect to $\mathbf{Y}_{:j}$ and $\preccurlyeq_j$ is defined as

$$\mathrm{PR}_{\mathbf{Y}_{:j}}^{\preccurlyeq_j}(Y_{ij}) := \begin{cases} \frac{\sum_{Y_{lj} \in \mathbf{Y}_{:j,\mathrm{obs}}} \mathbb{1}\{Y_{lj} = \mathrm{F}\} + \frac{1}{2} \sum_{Y_{lj} \in \mathbf{Y}_{:j,\mathrm{obs}}} \mathbb{1}\{Y_{lj} = \mathrm{T}\} + |\mathbf{Y}_{:j,\mathrm{mis}}|}{n} & \mathbf{Y}_{:j,\mathrm{obs}} \ni Y_{ij} = \mathrm{T} \\ 0 & \text{else} \end{cases}$$

$$\left[\mathrm{PR}_{\mathbf{Y}_{:j}}^{\preccurlyeq_j}(Y_{ij}) := \begin{cases} \frac{\sum_{Y_{lj} \in \mathbf{Y}_{:j,\mathrm{obs}}} \mathbb{1}\{Y_{lj} = \mathrm{T}\} + \frac{1}{2} \sum_{Y_{lj} \in \mathbf{Y}_{:j,\mathrm{obs}}} \mathbb{1}\{Y_{lj} = \mathrm{F}\} + |\mathbf{Y}_{:j,\mathrm{mis}}|}{n} & \mathbf{Y}_{:j,\mathrm{obs}} \ni Y_{ij} = \mathrm{F} \\ 0 & \text{else} \end{cases} \right].$$

△

Let $\mathcal{O} := \{\preccurlyeq_j\}_{j=1,\dots,d}$ denote the set of the orders of the scoring variables. Then the **calculation of the subpillar scores** can be described mathematically as follows:

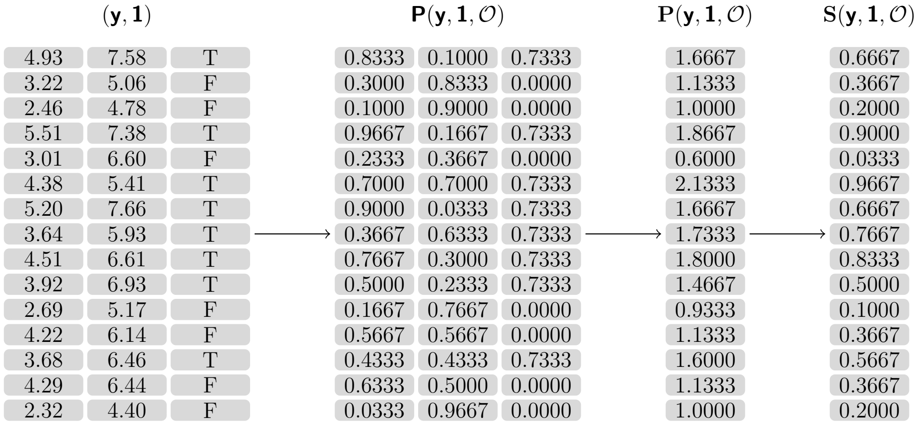
1. For each company i, the calculation of the subpillar score starts with determining the percentile ranks $P_{ij} := \mathrm{PR}^{\preccurlyeq_j}_{\mathbf{Y}_{:j}}(Y_{ij})$ of the company's observation Y_{ij} of the jth scoring variable Y_j with respect to the observations $\mathbf{Y}_{:j}$ from Y_j of all n companies over all scoring variables $Y_1, \ldots, Y_d$. We collect the percentile ranks P_{ij} for the n companies and d scoring variables in the random matrix
$$\mathbf{P} := \mathbf{P}(\mathbf{Y}, \mathbf{R}, \mathcal{O}) := (P_{ij})_{i=1,\ldots,n,\, j=1,\ldots,d} \in [0,1)^{n \times d}.$$
2. Then, for each company i, we sum up its single percentile ranks P_{ij} over all scoring variables to $P_i := \sum_{j=1}^{d} P_{ij}$ and collect these sums in the random vector
$$\mathbf{P} := \mathbf{P}(\mathbf{Y}, \mathbf{R}, \mathcal{O}) := (P_i)_{i=1,\ldots,n} \in \mathbb{R}^n_{\geq 0}.$$
3. Finally, for each company i, we calculate the percentile rank of P_i with respect to $\mathbf{P}$ resulting in the subpillar score $S_i := \mathrm{PR}^{\leq}_{\mathbf{P}}(P_i)$. The subpillar scores for all n companies are collected in the random vector
$$\mathbf{S} := \mathbf{S}(\mathbf{Y}, \mathbf{R}, \mathcal{O}) := (S_i)_{i=1,\ldots,n} \in (0,1)^{n \times d}.$$

Example 6.1.2. We consider the realized incomplete data set $(\mathbf{y}, \mathbf{r})$ from Example 2.2.2. Additionally, we assume that the set of orders $\mathcal{O} = \{\preccurlyeq_1, \preccurlyeq_2, \preccurlyeq_3\}$ is defined as follows: For Y_1 the order $\preccurlyeq_1 := \leq$ applies while for Y_2 the reverse order $\preccurlyeq_2 := \geq$ applies. For the boolean variable Y_3 the order $\preccurlyeq_3$ is induced by F $\prec_3$ T. Under these conditions we get the percentile ranks and the subpillar scores as given in Figure 6.3.

$(\mathbf{y}_{\text{obs}}, \mathbf{r})$			$\mathbf{P}(\mathbf{y}, \mathbf{r}, \mathcal{O})$			$\mathbf{P}(\mathbf{y}, \mathbf{r}, \mathcal{O})$	$\mathbf{S}(\mathbf{y}, \mathbf{r}, \mathcal{O})$
4.93	7.58	T	0.8333	0.1250	0.7667	1.7250	0.8333
3.22	5.06	F	0.3000	0.7917	0.0000	1.0917	0.4000
2.46	4.78	F	0.1000	0.8750	0.0000	0.9750	0.2333
5.51		T	0.9667	0.0000	0.7667	1.7333	0.9000
3.01	6.60	F	0.2333	0.2917	0.0000	0.5250	0.0333
4.38	5.41	T	0.7000	0.6250	0.7667	2.0917	0.9667
5.20	7.66	T	0.9000	0.0417	0.7667	1.7083	0.7667
3.64		T	0.3667	0.0000	0.7667	1.1333	0.5667
4.51			0.7667	0.0000	0.0000	0.7667	0.1000
3.92	6.93	T	0.5000	0.2083	0.7667	1.4750	0.6333
2.69	5.17	F	0.1667	0.7083	0.0000	0.8750	0.1667
4.22	6.14	F	0.5667	0.5417	0.0000	1.1083	0.5000
3.68	6.46	T	0.4333	0.3750	0.7667	1.5750	0.7000
4.29	6.44		0.6333	0.4583	0.0000	1.0917	0.4000
2.32	4.40	F	0.0333	0.9583	0.0000	0.9917	0.3000

Figure 6.3: Calculation of the percentile ranks and subpillar scores for the incomplete data set of Example 2.2.2.

Imaging now the data would have been completely observed for all companies. The resulting percentile ranks and subpillar scores are given in Figure 6.4. For 12 of the 15 companies the subpillar scores differ. As previously indicated, this also affects companies for which no missing values occurred. On the other hand, of the three companies for which the subpillar score remains unchanged, for one company not all scoring variables have been reported. △



$(\mathbf{y},\mathbf{1})$				$\mathbf{P}(\mathbf{y},\mathbf{1},\mathcal{O})$				$\mathbf{P}(\mathbf{y},\mathbf{1},\mathcal{O})$		$\mathbf{S}(\mathbf{y},\mathbf{1},\mathcal{O})$
4.93	7.58	T		0.8333	0.1000	0.7333		1.6667		0.6667
3.22	5.06	F		0.3000	0.8333	0.0000		1.1333		0.3667
2.46	4.78	F		0.1000	0.9000	0.0000		1.0000		0.2000
5.51	7.38	T		0.9667	0.1667	0.7333		1.8667		0.9000
3.01	6.60	F		0.2333	0.3667	0.0000		0.6000		0.0333
4.38	5.41	T		0.7000	0.7000	0.7333		2.1333		0.9667
5.20	7.66	T		0.9000	0.0333	0.7333		1.6667		0.6667
3.64	5.93	T	⟶	0.3667	0.6333	0.7333	⟶	1.7333	⟶	0.7667
4.51	6.61	T		0.7667	0.3000	0.7333		1.8000		0.8333
3.92	6.93	T		0.5000	0.2333	0.7333		1.4667		0.5000
2.69	5.17	F		0.1667	0.7667	0.0000		0.9333		0.1000
4.22	6.14	F		0.5667	0.5667	0.0000		1.1333		0.3667
3.68	6.46	T		0.4333	0.4333	0.7333		1.6000		0.5667
4.29	6.44	F		0.6333	0.5000	0.0000		1.1333		0.3667
2.32	4.40	F		0.0333	0.9667	0.0000		1.0000		0.2000

Figure 6.4: Calculation of the percentile ranks and subpillar scores for the hypothetically complete data set of Example 2.2.2.

6.2 Recalculation of the Emission Score

Data exploration

In the following we want to impute incomplete data for a specific subpillar and recalculate the subpillar score. For reasons of data availability we do not know exactly which scoring variables are needed for the single subpillars in the different industries. Therefore, we work only with a selection of scoring variables we could uniquely identify from an example of LSEG (2023) for the Emissions subpillar. We consider $n = 520$ companies of the machinery sector with data from the year 2021. The selected $d = 12$ scoring variables $Y_1, \dots, Y_{12}$ are given in Table 6.1. Eight out of the twelve scoring variables are boolean.

	Description	Values	Missing	Order
Y_1	Policy Emissions	$\{\text{F, T}\}$	0%	F $\prec$ T
Y_2	Targets Emissions	$\{\text{F, T}\}$	0%	F $\prec$ T
Y_3	Total CO2 Equivalent Emissions To Revenues USD in million	$[0,\infty)$	$34,62\%$	$\geq$
Y_4	Climate Change Commercial Risks Opportunities	$\{\text{F, T}\}$	0%	F $\prec$ T
Y_5	VOC or Particulate Matter Emissions Reduction	$\{\text{F, T}\}$	0%	F $\prec$ T
Y_6	Total Waste To Revenues USD in million	$[0,\infty)$	$52,50\%$	$\geq$
Y_7	Waste Recycled To Total Waste	$[0,1]$	$61,92\%$	$\leq$
Y_8	Total Hazardous Waste To Revenues USD in million	$[0,\infty)$	$61,35\%$	$\geq$
Y_9	Environmental Restoration Initiatives	$\{\text{F, T}\}$	0%	F $\prec$ T
Y_{10}	Staff Transportation Impact Reduction	$\{\text{F, T}\}$	0%	F $\prec$ T
Y_{11}	Environmental Expenditures Investments	$\{\text{F, T}\}$	0%	F $\prec$ T
Y_{12}	Environmental Partnerships	$\{\text{F, T}\}$	0%	F $\prec$ T

Table 6.1: Scoring variables $Y_1, \dots, Y_{12}$ for the recalculation of the Emissions score.

From the remaining four numeric variables three variables take positive values while one variable takes values between 0 and 1. Only four out of the twelve variables, namely the numeric ones, contain missing values. It is striking that if a variable has missing values the percentage of missing values is quite high with even more than 50% in three out of the four cases. Figure 6.5 shows the response patterns which are present in the ESG data (see Remark 2.2.1). Only 133 observations are complete. For 154 observations we have the maximal number of four missing values. Figure 6.6 shows bar plots for the boolean variables and histograms for the non-boolean variables. We transformed the numeric scoring variables via log(·) and logit(·) to bring them on unbounded support. Most of the boolean scoring variables are one-sided.

Response pattern **r**												Count
1	1	0	1	1	0	0	0	1	1	1	1	154
1	1	0	1	1	0	0	1	1	1	1	1	9
1	1	0	1	1	0	1	0	1	1	1	1	2
1	1	0	1	1	1	0	0	1	1	1	1	1
1	1	1	1	1	0	0	0	1	1	1	1	88
1	1	0	1	1	1	0	1	1	1	1	1	7
1	1	0	1	1	1	1	0	1	1	1	1	1
1	1	1	1	1	0	0	1	1	1	1	1	13
1	1	1	1	1	0	1	0	1	1	1	1	7
1	1	1	1	1	1	0	0	1	1	1	1	17
1	1	0	1	1	1	1	1	1	1	1	1	6
1	1	1	1	1	1	0	1	1	1	1	1	33
1	1	1	1	1	1	1	0	1	1	1	1	49
1	1	1	1	1	1	1	1	1	1	1	1	133

Figure 6.5: Present response patterns in the ESG data and their counts.

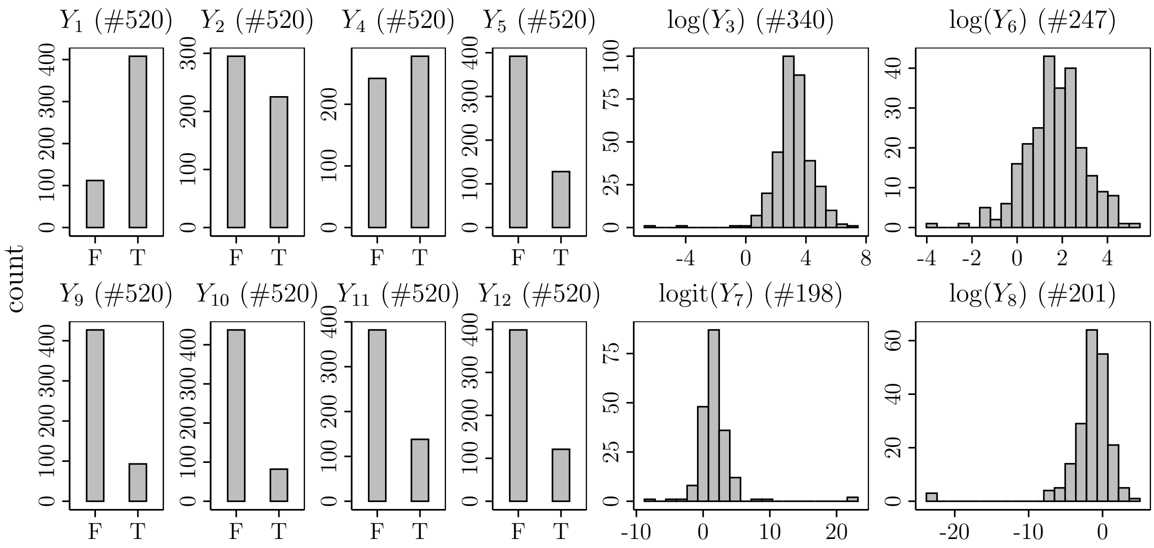


Figure 6.6: Bar plots and histograms of the scoring variables with their number of observed observations in parentheses.

Imputation Methods

We compare the following imputation methods:

- We apply the **methods $\mathcal{M}_{\mathrm{v}}$ and $\mathcal{M}_{\mathrm{m}}$** from the simulation study with their previous parameter choices, that is for $\mathcal{M}_{\mathrm{v}}$ the number of nearest neighbors equals $K = 5$ and for both methods the quasi-Gibbs sampler is running for $T = 20$ iterations. We decided for those two methods since the majority of the scoring variables are discrete and $\mathcal{M}_{\mathrm{v}}$ and $\mathcal{M}_{\mathrm{m}}$ performed best in the simulation study regarding mixed continuous and discrete variables.
- By default the methods $\mathcal{M}_{\mathrm{v}}$ and $\mathcal{M}_{\mathrm{m}}$ impute the scoring variables Y_j with missing values in the order of increasing j which here is Y_3, Y_6, Y_7, Y_8. The **methods $\widetilde{\mathcal{M}}_{\mathrm{v}}$ and $\widetilde{\mathcal{M}}_{\mathrm{m}}$** describe the modifications of $\mathcal{M}_{\mathrm{v}}$ and $\mathcal{M}_{\mathrm{m}}$ where the imputation order equals Y_7, Y_6, Y_3, Y_8. Nonetheless, the initial values for the quasi-Gibbs sampler are the same for both methods $\widetilde{\mathcal{M}}_{\mathrm{v}}$ and $\mathcal{M}_{\mathrm{v}}$ as well as for both methods $\widetilde{\mathcal{M}}_{\mathrm{m}}$ and $\mathcal{M}_{\mathrm{m}}$.
- Additionally we consider an independent imputation **method $\mathcal{M}_{\mathrm{i}}$** which samples the imputed values from the observed values of the marginals. The output of this method equals the initial values for the methods $\mathcal{M}_{\mathrm{m}}$ and $\widetilde{\mathcal{M}}_{\mathrm{m}}$.

With each method we generate $M = 100$ imputed data sets.

Procedure

Since the original ESG data is incomplete we do not know the true Emissions scores which would have realized in the complete-data case. In order to be able to make a realistic comparison, we need to generate a complete data set that is still comparable to the original ESG data. To this end, we fitted an R-vine to the original ESG data with appropriately log- and logit-transformed numeric scoring variables (the transformations have no influence on the Emissions score since strictly increasing transformations do not affect the percentile rank) using the `vine()` function of the `rvinecopulib` R-package (Nagler and Vatter 2023). This function handles missing values in the same way as we described in Remark 4.2.2 (Nagler and Vatter 2023, p. 32). This means that all observed values are used in the estimation process and not only the values of the complete observations. The resulting vine structure with the corresponding Kendall's τ's is given in Figure 6.7. From this vine we generated a complete data set $\mathbf{y}$ with observations for $n = 520$ companies. Afterwards, we combined $\mathbf{y}$ with the original realized response $\mathbf{r}$ to obtain the incomplete data set $(\mathbf{y}, \mathbf{r})$ which we will work with in the following steps. Let $\mathbf{s} := (s_i)_{i=1,\dots,n} := \mathbf{S}(\mathbf{y}, \mathbf{1}, \mathcal{O})$ be the vector of the true Emission scores derived from the hypothetical complete data set $\mathbf{y}$. The emission scores which LSEG (2023) would calculate are denoted by $\mathbf{s}^{(\mathrm{LSEG})} := (s_i^{(\mathrm{LSEG})})_{i=1,\dots,n} := \mathbf{S}(\mathbf{y}, \mathbf{r}, \mathcal{O})$.

With each imputation method $\mathcal{M} \in \{\mathcal{M}_{\mathrm{v}}, \widetilde{\mathcal{M}}_{\mathrm{v}}, \mathcal{M}_{\mathrm{m}}, \widetilde{\mathcal{M}}_{\mathrm{m}}, \mathcal{M}_{\mathrm{i}}\}$ we impute M times to obtain the completed data sets

$$\mathbf{y}^{(\mathcal{M},m)} = (\mathbf{y}_{\mathrm{obs}}, \mathbf{y}_{\mathrm{mis}}^{(\mathcal{M},m)}), \quad m = 1, \dots, M.$$

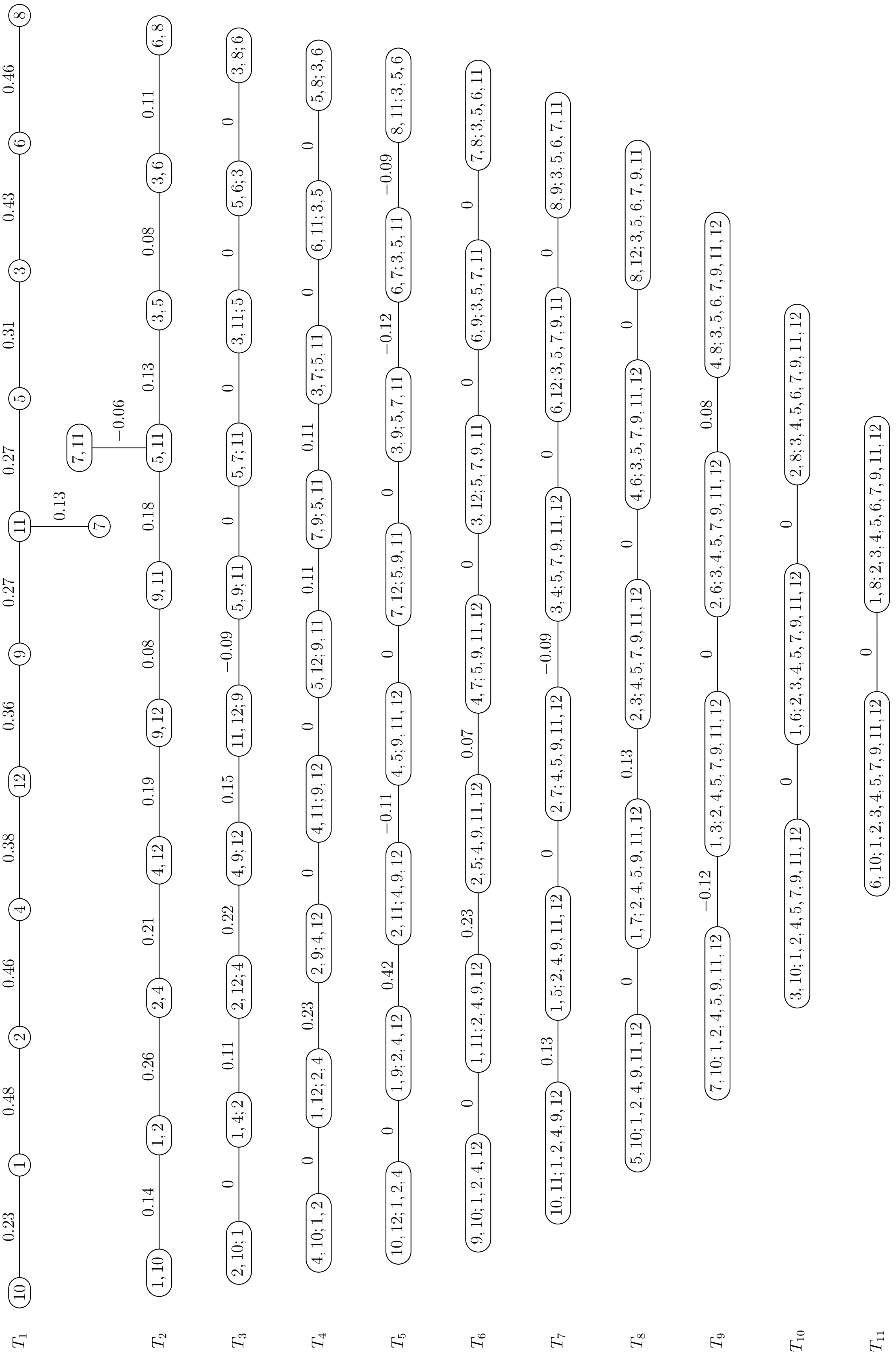


Figure 6.7: Vine structure of the fitted R-vine to the ESG data with the corresponding Kendall's τ's at the edges.

Then we calculate the corresponding Emissions score vectors

$$\mathbf{s}^{(\mathcal{M},m)} := (s_i^{(\mathcal{M},m)})_{i=1,\ldots,n} := \mathbf{S}(\mathbf{y}^{(\mathcal{M},m)}, \mathbf{1}, \mathcal{O}), \quad m = 1, \ldots, M.$$

An element $s_i^{(\mathcal{M},m)}$ is the Emissions score of company i derived via imputation method $\mathcal{M}$ from the mth imputed data set.

Analysis of the Results

For each imputation method $\mathcal{M} \in \{\mathcal{M}_{\mathrm{v}}, \widetilde{\mathcal{M}}_{\mathrm{v}}, \mathcal{M}_{\mathrm{m}}, \widetilde{\mathcal{M}}_{\mathrm{m}}, \mathcal{M}_{\mathrm{i}}\}$ and each company $i = 1, \ldots, n$, based on the Emissions scores $s_i^{(\mathcal{M},1)}, \ldots, s_i^{(\mathcal{M},M)}$ derived from the imputations, we can define an empirical posterior cumulative distribution function

$$\widehat{F}_i^{(\mathcal{M})}(s) := \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\{s_i^{(\mathcal{M},m)} < s\}$$

for the true Emissions score. As in the simulation study, we can therefore compare the imputation methods on the basis of sample continuous ranked probability scores here defined as

$$\widehat{\mathrm{CRPS}}_i^{(\mathcal{M})} := \mathrm{CRPS}(\widehat{F}_i^{(\mathcal{M})}, s_i).$$

In order to also make the LSEG (2023) methodology comparable via the sample continuous ranked probability score, for each company i we define the degenerate empirical posterior cumulative distribution function

$$\widehat{F}_i^{(\mathrm{LSEG})}(s) := \mathbb{1}\{s_i^{(\mathrm{LSEG})} < s\}.$$

The corresponding sample continuous ranked probability score we denote with

$$\widehat{\mathrm{CRPS}}_i^{(\mathrm{LSEG})} := \mathrm{CRPS}(\widehat{F}_i^{(\mathrm{LSEG})}, s_i).$$

Box plots of the realized sample continuous ranked probability scores are given in Figure 6.8. In addition to the sample continuous ranked probability score across all companies, we also consider the sample continuous ranked probability score separately for companies with the same number of non-reported scoring variables. Regardless of how many values are missing per company, there is a clear difference in the sample continuous ranked probability scores between the LSEG (2023) methodology and the imputation methods: Proceeding as LSEG (2023) is worse than applying any of the imputation methods. Especially in the group without missing values, the LSEG (2023) methodology does not seem suitable for predicting the true Emissions scores. For the imputation methods visually no big differences are recognizable. Nevertheless, it seems that the methods $\mathcal{M}_{\mathrm{v}}$, $\widetilde{\mathcal{M}}_{\mathrm{v}}$, $\mathcal{M}_{\mathrm{m}}$ and $\widetilde{\mathcal{M}}_{\mathrm{m}}$ outperform the method $\mathcal{M}_{\mathrm{i}}$ over all companies as well as for the companies with three missing values. Also $\mathcal{M}_{\mathrm{v}}$ and $\widetilde{\mathcal{M}}_{\mathrm{v}}$ seem to outperform $\mathcal{M}_{\mathrm{m}}$ and $\widetilde{\mathcal{M}}_{\mathrm{m}}$ for companies for which one scoring variable has not been reported.

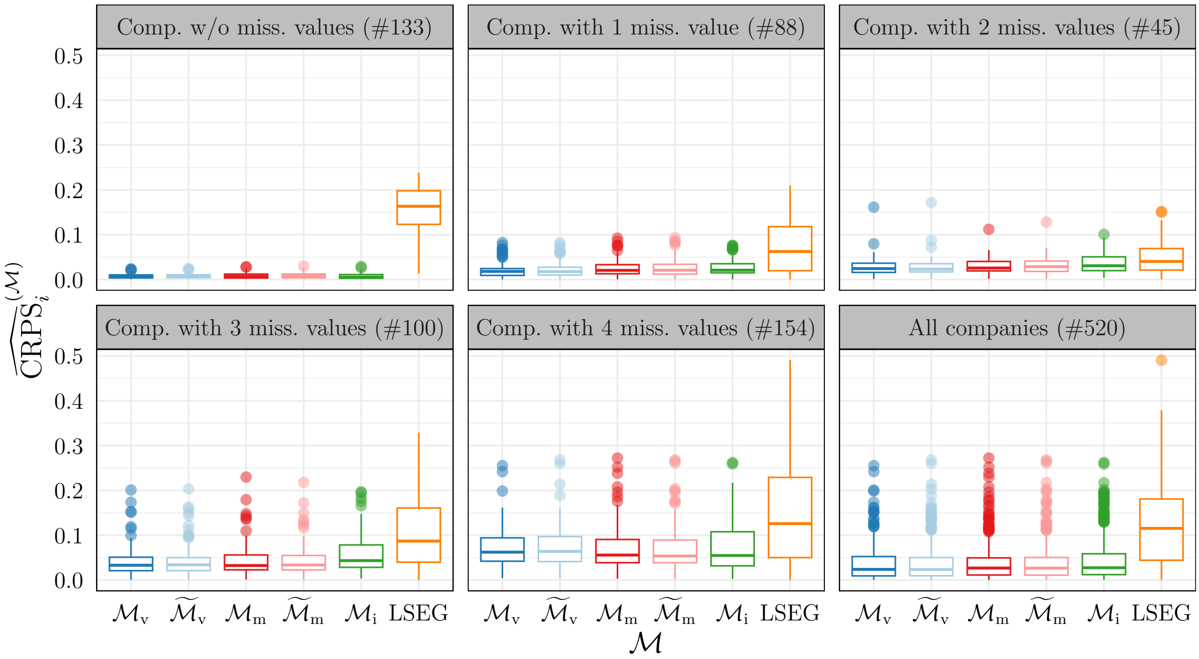


Figure 6.8: Box plots of the sample continuous ranked probability scores $\widehat{\mathrm{CRPS}}_i^{(\mathcal{M})}$ per imputation method $\mathcal{M} \in \{\mathcal{M}_\mathrm{v}, \widetilde{\mathcal{M}}_\mathrm{v}, \mathcal{M}_\mathrm{m}, \widetilde{\mathcal{M}}_\mathrm{m}, \mathcal{M}_\mathrm{i}, \mathrm{LSEG}\}$ across different groups of companies with the same number of missing values and across all companies.

Remark 6.2.1. In Figure 6.8, for notational simplicity, we refer to the LSEG (2023) methodology as an imputation method $\mathcal{M} = \mathrm{LSEG}$ even if no imputations are generated. We will proceed in the same way below. △

To statistically underpin our visual analysis, we perform paired t-tests. For any two methods $\mathcal{M}_1, \mathcal{M}_2 \in \{\mathcal{M}_\mathrm{v}, \widetilde{\mathcal{M}}_\mathrm{v}, \mathcal{M}_\mathrm{m}, \widetilde{\mathcal{M}}_\mathrm{m}, \mathcal{M}_\mathrm{i}, \mathrm{LSEG}\}$ we consider the differences in the sample continuous probability score per company i:

$$\Delta\widehat{\mathrm{CRPS}}_i(\mathcal{M}_1, \mathcal{M}_2) := \widehat{\mathrm{CRPS}}_i^{(\mathcal{M}_1)} - \widehat{\mathrm{CRPS}}_i^{(\mathcal{M}_2)}.$$

We are testing the hypotheses

$$H_0\colon \mu_{\Delta\widehat{\mathrm{CRPS}}(\mathcal{M}_1,\mathcal{M}_2)} \geq 0 \quad \text{against} \quad H_1\colon \mu_{\Delta\widehat{\mathrm{CRPS}}(\mathcal{M}_1,\mathcal{M}_2)} < 0,$$

or, formulated less formally,

$$H_0\colon \mathcal{M}_1 \text{ is worse than } \mathcal{M}_2 \quad \text{against} \quad H_1\colon \mathcal{M}_1 \text{ is better than } \mathcal{M}_2$$

with regard to the continuous ranked probability score.

The results are given as p-values in Table 6.2. We are interpreting the p-values for a significance level $\alpha = 0.1$. In this case a p-value less than 0.1 means that method $\mathcal{M}_1$ is better than $\mathcal{M}_2$. Indeed, there is statistical evidence that some imputation methods outperform others. The most important findings can be summarized as:

$\mathcal{M}_1$ / $\mathcal{M}_2$	Companies without missing values (#133)						Companies with one missing value (#88)					
	$\mathcal{M}_\mathrm{v}$	$\widetilde{\mathcal{M}}_\mathrm{v}$	$\mathcal{M}_\mathrm{m}$	$\widetilde{\mathcal{M}}_\mathrm{m}$	$\mathcal{M}_\mathrm{i}$	LSEG	$\mathcal{M}_\mathrm{v}$	$\widetilde{\mathcal{M}}_\mathrm{v}$	$\mathcal{M}_\mathrm{m}$	$\widetilde{\mathcal{M}}_\mathrm{m}$	$\mathcal{M}_\mathrm{i}$	LSEG
$\mathcal{M}_\mathrm{v}$		1.000	1.000	1.000	0.310	1.000		0.348	0.998	0.993	1.000	1.000
$\widetilde{\mathcal{M}}_\mathrm{v}$	**0.000**		0.993	0.999	**0.051**	1.000	0.652		0.999	0.997	1.000	1.000
$\mathcal{M}_\mathrm{m}$	**0.000**	**0.007**		0.981	**0.003**	1.000	**0.002**	**0.001**		0.135	0.962	1.000
$\widetilde{\mathcal{M}}_\mathrm{m}$	**0.000**	**0.001**	**0.019**		**0.000**	1.000	**0.007**	**0.003**	0.865		0.978	1.000
$\mathcal{M}_\mathrm{i}$	0.690	0.949	0.997	1.000		1.000	**0.000**	**0.000**	**0.038**	**0.022**		1.000
LSEG	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**		**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	

$\mathcal{M}_1$ / $\mathcal{M}_2$	Companies with two missing values (#45)						Companies with three missing values (#100)					
	$\mathcal{M}_\mathrm{v}$	$\widetilde{\mathcal{M}}_\mathrm{v}$	$\mathcal{M}_\mathrm{m}$	$\widetilde{\mathcal{M}}_\mathrm{m}$	$\mathcal{M}_\mathrm{i}$	LSEG	$\mathcal{M}_\mathrm{v}$	$\widetilde{\mathcal{M}}_\mathrm{v}$	$\mathcal{M}_\mathrm{m}$	$\widetilde{\mathcal{M}}_\mathrm{m}$	$\mathcal{M}_\mathrm{i}$	LSEG
$\mathcal{M}_\mathrm{v}$		0.775	0.764	0.825	0.935	0.999		0.737	0.926	0.918	1.000	1.000
$\widetilde{\mathcal{M}}_\mathrm{v}$	0.225		0.659	0.728	0.914	0.999	0.263		0.832	0.813	1.000	1.000
$\mathcal{M}_\mathrm{m}$	0.236	0.341		0.673	0.951	1.000	**0.074**	0.168		0.401	1.000	1.000
$\widetilde{\mathcal{M}}_\mathrm{m}$	0.175	0.272	0.327		0.929	1.000	**0.082**	0.187	0.599		1.000	1.000
$\mathcal{M}_\mathrm{i}$	**0.065**	**0.086**	**0.049**	**0.071**		1.000	**0.000**	**0.000**	**0.000**	**0.000**		1.000
LSEG	**0.001**	**0.001**	**0.000**	**0.000**	**0.000**		**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	

$\mathcal{M}_1$ / $\mathcal{M}_2$	Companies with four missing values (#154)						All companies (#520)					
	$\mathcal{M}_\mathrm{v}$	$\widetilde{\mathcal{M}}_\mathrm{v}$	$\mathcal{M}_\mathrm{m}$	$\widetilde{\mathcal{M}}_\mathrm{m}$	$\mathcal{M}_\mathrm{i}$	LSEG	$\mathcal{M}_\mathrm{v}$	$\widetilde{\mathcal{M}}_\mathrm{v}$	$\mathcal{M}_\mathrm{m}$	$\widetilde{\mathcal{M}}_\mathrm{m}$	$\mathcal{M}_\mathrm{i}$	LSEG
$\mathcal{M}_\mathrm{v}$		0.946	0.486	0.532	0.766	1.000		0.979	0.984	0.982	1.000	1.000
$\widetilde{\mathcal{M}}_\mathrm{v}$	**0.054**		0.167	0.200	0.588	1.000	**0.021**		0.844	0.857	1.000	1.000
$\mathcal{M}_\mathrm{m}$	0.514	0.833		0.584	0.761	1.000	**0.016**	0.156		0.527	0.999	1.000
$\widetilde{\mathcal{M}}_\mathrm{m}$	0.468	0.800	0.416		0.747	1.000	**0.018**	0.143	0.473		1.000	1.000
$\mathcal{M}_\mathrm{i}$	0.234	0.412	0.239	0.253		1.000	**0.000**	**0.000**	**0.001**	**0.000**		1.000
LSEG	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**		**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	

Table 6.2: p-values for testing $H_0\colon \mu_{\Delta\widehat{\mathrm{CRPS}}(\mathcal{M}_1,\mathcal{M}_2)} \geq 0$ against $H_1\colon \mu_{\Delta\widehat{\mathrm{CRPS}}(\mathcal{M}_1,\mathcal{M}_2)} < 0$ for any two methods $\mathcal{M}_1, \mathcal{M}_2 \in \{\mathcal{M}_\mathrm{v}, \widetilde{\mathcal{M}}_\mathrm{v}, \mathcal{M}_\mathrm{m}, \widetilde{\mathcal{M}}_\mathrm{m}, \mathcal{M}_\mathrm{i}, \mathrm{LSEG}\}$ in groups of companies with the same number of missing values and for all companies where p-values lower than 0.1 are marked bold.

- The methodology of LSEG (2023) is outperformed by all real imputation methods across all groups of companies. Any imputation always significantly improves the continuous ranked probability score.
- The methods $\mathcal{M}_{\mathrm{v}}$ and $\widetilde{\mathcal{M}}_{\mathrm{v}}$ as well as the methods $\mathcal{M}_{\mathrm{m}}$ and $\widetilde{\mathcal{M}}_{\mathrm{m}}$ outperform the method $\mathcal{M}_{\mathrm{i}}$ across most of the groups of companies. Especially considering all companies together the differences in the sample continuous ranked probability scores are significant. The main difference of the four first mentioned methods and the method $\mathcal{M}_{\mathrm{i}}$ is that they consider the the dependencies between the single variables while $\mathcal{M}_{\mathrm{i}}$ ignores it. This suggests that an imputation method that considers the dependencies between the variables should be preferred.
- The vine based methods $\mathcal{M}_{\mathrm{v}}$ and $\widetilde{\mathcal{M}}_{\mathrm{v}}$ outperform the methods $\mathcal{M}_{\mathrm{m}}$ and $\widetilde{\mathcal{M}}_{\mathrm{m}}$ in some cases. Especially considering all companies, the method $\mathcal{M}_{\mathrm{v}}$ is superior to the methods $\mathcal{M}_{\mathrm{m}}$ and $\widetilde{\mathcal{M}}_{\mathrm{m}}$. This supports the statement that the vine based methods form the best imputation methods in this application.
- Considering only the both vine based methods, $\mathcal{M}_{\mathrm{v}}$ outperforms the method $\widetilde{\mathcal{M}}_{\mathrm{v}}$. This suggest that the order in which the variables are imputed matters. In our case, $\mathcal{M}_{\mathrm{v}}$ imputes the variables in a very natural order: the variables are imputed in increasing order of the missingness fraction.

Summarizing the p-values concerning the method $\mathcal{M}_{\mathrm{v}}$, we see that the vine based method with natural imputation order outperforms all other methods even on higher significance levels. The results are given in Table 6.3.

	$\widetilde{\mathcal{M}}_{\mathrm{v}}$	$\mathcal{M}_{\mathrm{m}}$	$\widetilde{\mathcal{M}}_{\mathrm{m}}$	$\mathcal{M}_{\mathrm{i}}$	LSEG
Companies without missing values (#133)	0.000^{***}	0.000^{***}	0.000^{***}	0.690	0.000^{***}
Companies with one missing value (#88)	0.652	0.002^{***}	0.007^{***}	0.000^{***}	0.000^{***}
Companies with two missing values (#45)	0.225	0.236	0.175	0.065^{*}	0.001^{***}
Companies with three missing values (#100)	0.263	0.074^{*}	0.082^{*}	0.000^{***}	0.000^{***}
Companies with four missing values (#154)	0.054^{*}	0.514	0.468	0.234	0.000^{***}
All companies (#520)	0.021^{**}	0.016^{**}	0.018^{**}	0.000^{***}	0.000^{***}

Table 6.3: p-values for testing $H_0\colon \mu_{\Delta\widehat{\mathrm{CRPS}}(\mathcal{M}_{\mathrm{v}},\mathcal{M})} \geq 0$ against $H_1\colon \mu_{\Delta\widehat{\mathrm{CRPS}}(\mathcal{M}_{\mathrm{v}},\mathcal{M})} < 0$ for the imputation methods $\mathcal{M} \in \{\widetilde{\mathcal{M}}_{\mathrm{v}}, \mathcal{M}_{\mathrm{m}}, \widetilde{\mathcal{M}}_{\mathrm{m}}, \mathcal{M}_{\mathrm{i}}, \mathrm{LSEG}\}$ in groups of companies with the same number of missing values and for all companies with marked significance levels (*: 0.1, **: 0.05, ***: 0.01).

Final Emissions Scores

Let $\mathcal{M} \in \{\mathcal{M}_{\mathrm{v}}, \widetilde{\mathcal{M}}_{\mathrm{v}}, \mathcal{M}_{\mathrm{m}}, \widetilde{\mathcal{M}}_{\mathrm{m}}, \mathcal{M}_{\mathrm{i}}\}$ be an arbitrary but fixed imputation method. So far, for each company $i = 1, \ldots, n$, we only have the collection of Emissions scores $\{s_i^{(\mathcal{M},m)}\}_{m=1,\ldots,M}$ calculated from the single imputed data sets. From the data provider's point of view, however, it could be reasonable to summarize these different scores in a final score $s_i^{(\mathcal{M})}$ per company in order to publish them or use them to continue with the calculation of the ESG score. To generate such final Emissions scores we propose the following procedure which results, as in the original approach of LSEG (2023), in a valid percentile rank as the Emissions score for each company. For all companies $i = 1, \dots, n$, we sum up the single Emissions scores $\{s_i^{(\mathcal{M},m)}\}_{m=1,\dots,M}$ to the values $s_i^{(\mathcal{M},+)} := \sum_{m=1}^{M} s_i^{(\mathcal{M},m)}$ and collect them in the vector

$$\mathbf{s}^{(\mathcal{M},+)} := (s_i^{(\mathcal{M},+)})_{i=1,\dots,n} \in \mathbb{R}^n_{\geq 0}.$$

Then, we apply the percentile ranking giving the final Emissions score for a company i:

$$s_i^{(\mathcal{M})} := \mathrm{PR}^{\leq}_{\mathbf{s}^{(\mathcal{M},+)}}(s_i^{(\mathcal{M},+)}) \in (0,1).$$

The final Emissions scores $s_i^{(\mathcal{M})}$ for all methods $\mathcal{M} \in \{\mathcal{M}_\mathrm{v}, \widetilde{\mathcal{M}}_\mathrm{v}, \mathcal{M}_\mathrm{m}, \widetilde{\mathcal{M}}_\mathrm{m}, \mathcal{M}_\mathrm{i}, \mathrm{LSEG}\}$ and companies $i = 1, \dots, n$ in comparison to the true Emissions scores s_i are given in Figure 6.9. The most important finding is that the LSEG (2023) methodology strongly overestimates the Emissions scores of companies that have no missing values. In a less strong form this trend is also visible for companies with one missing value. On the other hand, for companies with three or four missing values the LSEG (2023) methodology systematically underestimates the Emissions score. For the Emissions scores derived from any real imputation method this is not the case: Especially for companies which reported all scoring variables, the final Emissions scores are almost identical with the true Emissions scores.

For companies without missing values, the independent imputation method $\mathcal{M}_\mathrm{i}$ seems to perform better than the other real imputation methods. Also the methods $\mathcal{M}_\mathrm{m}$ and $\widetilde{\mathcal{M}}_\mathrm{m}$ seem to give more precise Emissions scores than the vine based methods $\mathcal{M}_\mathrm{v}$ and $\widetilde{\mathcal{M}}_\mathrm{v}$. While it is challenging to make a definitive judgment based solely on visual inspection, for the other groups, i. e. for companies with missing values, the vine-based methods appear to outperform the other imputation methods. This suggests that the vine based methods could be superior on the overall level when considering all companies with and without missing values.

To statistically compare the final Emissions scores resulting from the vine based method $\mathcal{M}_\mathrm{v}$ to the final Emissions scores from the other methods, we perform paired t-tests. More precise, for any two methods $\mathcal{M}_1, \mathcal{M}_2 \in \{\mathcal{M}_\mathrm{v}, \widetilde{\mathcal{M}}_\mathrm{v}, \mathcal{M}_\mathrm{m}, \widetilde{\mathcal{M}}_\mathrm{m}, \mathcal{M}_\mathrm{i}, \mathrm{LSEG}\}$ let

$$\Delta s_i(\mathcal{M}_1, \mathcal{M}_2) := |s_i^{(\mathcal{M}_1)} - s_i| - |s_i^{(\mathcal{M}_2)} - s_i|$$

be the difference of the absolute errors of Emissions scores derived from methods $\mathcal{M}_1$ and $\mathcal{M}_2$ for company i. We are testing the hypotheses

$$H_0\colon \mu_{\Delta s(\mathcal{M}_\mathrm{v},\mathcal{M})} \geq 0 \quad \text{against} \quad H_1\colon \mu_{\Delta s(\mathcal{M}_\mathrm{v},\mathcal{M})} < 0,$$

or, formulated less formally, regarding the final Emissions score

$$H_0\colon \mathcal{M}_\mathrm{v} \text{ is worse than } \mathcal{M} \quad \text{against} \quad H_1\colon \mathcal{M}_\mathrm{v} \text{ is better than } \mathcal{M}.$$

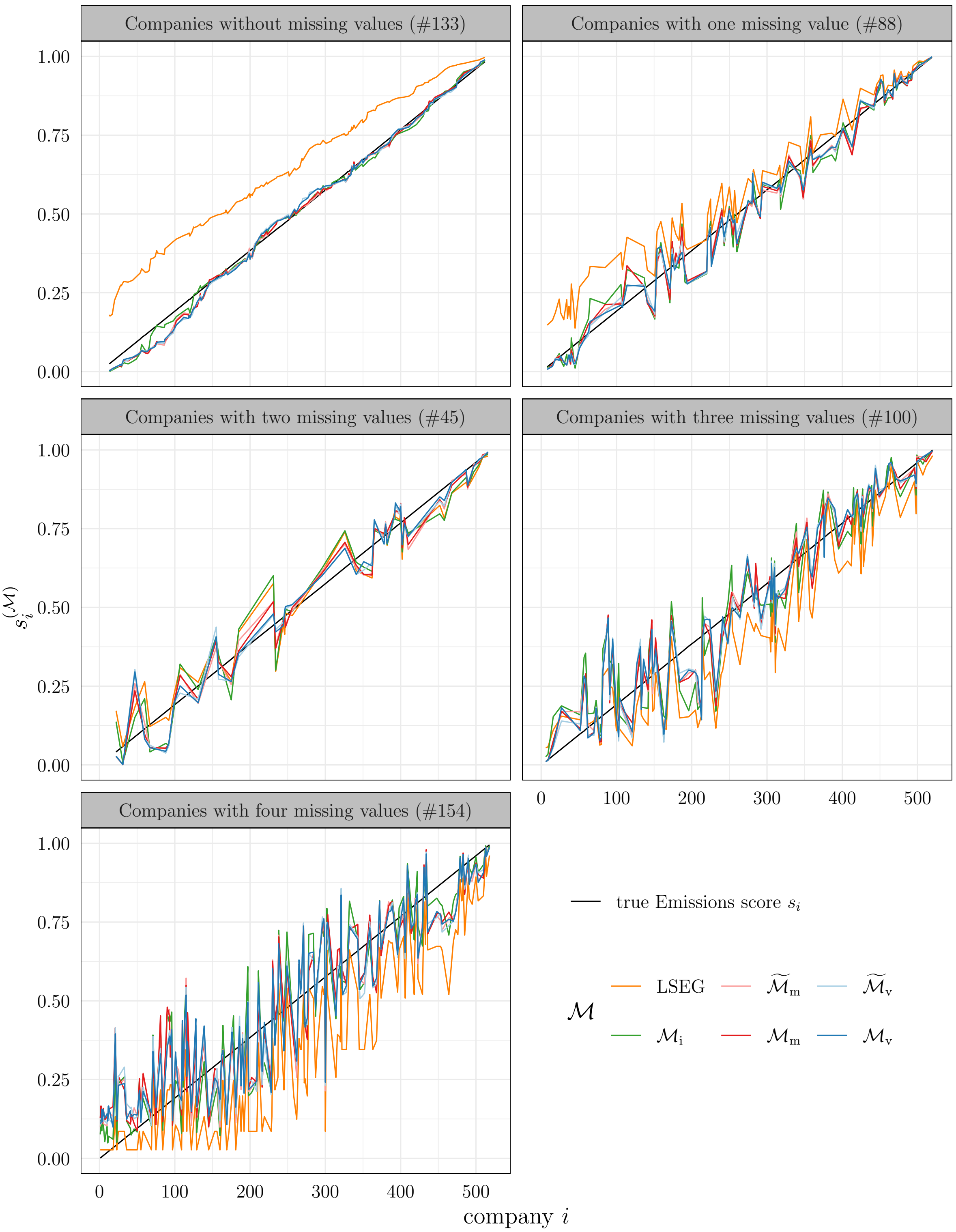


Figure 6.9: Emissions scores $s_i^{(\mathcal{M})}$ per company i calculated by imputation method $\mathcal{M} \in \{\mathcal{M}_{\mathrm{v}}, \widetilde{\mathcal{M}}_{\mathrm{v}}, \mathcal{M}_{\mathrm{m}}, \widetilde{\mathcal{M}}_{\mathrm{m}}, \mathcal{M}_{\mathrm{i}}, \mathrm{LSEG}\}$ compared to the true Emissions scores s_i grouped by companies with the same number of missing values.

	$\widetilde{\mathcal{M}}_{\mathrm{v}}$	$\mathcal{M}_{\mathrm{m}}$	$\widetilde{\mathcal{M}}_{\mathrm{m}}$	$\mathcal{M}_{\mathrm{i}}$	LSEG
Companies without missing values (#133)	0.007^{***}	0.993	0.999	1.000	0.000^{***}
Companies with one missing value (#88)	0.683	0.005^{***}	0.040^{**}	0.000^{***}	0.000^{***}
Companies with two missing values (#45)	0.338	0.345	0.179	0.059^{*}	0.113
Companies with three missing values (#100)	0.414	0.125	0.068^{*}	0.000^{***}	0.000^{***}
Companies with four missing values (#154)	0.086^{*}	0.378	0.383	0.277	0.000^{***}
All companies (#520)	0.071^{*}	0.096^{*}	0.121	0.000^{***}	0.000^{***}

Table 6.4: p-values for testing $H_0\colon \mu_{\Delta s(\mathcal{M}_{\mathrm{v}},\mathcal{M})} \geq 0$ against $H_1\colon \mu_{\Delta s(\mathcal{M}_{\mathrm{v}},\mathcal{M})} < 0$ for the imputation methods $\mathcal{M} \in \{\widetilde{\mathcal{M}}_{\mathrm{v}}, \mathcal{M}_{\mathrm{m}}, \widetilde{\mathcal{M}}_{\mathrm{m}}, \mathcal{M}_{\mathrm{i}}, \mathrm{LSEG}\}$ in groups of companies with the same number of missing values and for all companies with marked significance levels (*: 0.1, **: 0.05, ***: 0.01).

The results of the paired t-tests are given in Table 6.4. They confirm our conclusions from the visual analysis. For companies without missing values, method $\mathcal{M}_{\mathrm{v}}$ is outperformed by the methods $\mathcal{M}_{\mathrm{m}}$, $\widetilde{\mathcal{M}}_{\mathrm{m}}$ and $\mathcal{M}_{\mathrm{i}}$. However, if one considers all companies together, the vine based method $\mathcal{M}_{\mathrm{v}}$ with natural imputation order is statistically superior to the other methods at significance level 0.1, except to method $\widetilde{\mathcal{M}}_{\mathrm{m}}$. In comparison to the original LSEG (2023) methodology and the independent imputation $\mathcal{M}_{\mathrm{i}}$, method $\mathcal{M}_{\mathrm{v}}$ even performs better for the majority of categories at significance level 0.01.

Summary

Multiple imputation of the incomplete scoring variable data using any of the considered real imputation methods yields Emissions scores which are notably closer to the true Emissions scores compared to those derived from the original LSEG (2023) methodology. It has been shown that the LSEG (2023) methodology systematically overestimates the true Emissions scores for companies without missing values, while the scores for companies with several missing values are systematically underestimated. Both effects do not occur if multiple imputation is applied. Among the real imputation methods, the vine based method with a natural imputation order statistically outperforms the other methods in both the sample continuous ranked probability scores and the final Emissions scores.

7 Conclusion

This thesis is concerned with multiple imputation using vine copulas. We presented a novel imputation method that generates the imputations via sampling from a sequence of D-vine quantile regression models based on the fully conditional specification approach. In a simulation study, the comparison to other, well established methods with regard to randomization valid inference as well as with regard to the interval score and the continuous ranked probability score has shown that our vine based method can improve the quality of the imputations in certain situations. While it performed similarly to the imputation with `mice()`, it outperformed the imputation with `amelia()` in view of the mixture of continuous and discrete variables. The other considered copula-based method, the imputation with `CoImp()`, was inferior to all other methods. Nevertheless, the vine based imputation is computationally much more complex than the comparable performing imputation method using `mice()`. Thus, for high-dimensional data our approach is not yet practicable.

Further, we applied our imputation method on real ESG-data to recalculate the Emissions scores of companies in the machinery sector. We compared the resulting Emissions scores to the scores derived from other imputation methods and to the original approach of LSEG (2023) where missing values are not imputed. Our method significantly improved the Emissions scores and especially overcame two shortfalls of the LSEG (2023) methodology, namely the systematic overestimation of the Emissions scores for companies which reported values for all scoring variables and the systematic underestimation of the Emissions scores for companies for which several values were missing.

Future work could examine the performance of the presented approach when nonparametric pair-copulas are allowed. Additionally the effect of replacing the D-vine quantile regression models with C-vine quantile regression models could be investigated. To speed up the vine based imputation we suggest to truncate the D-vine quantile regression models to a smaller number of allowed covariates. For a general speed up an implementation on GPU's would be useful.

Bibliography

Albert, J. H. and S. Chib (1993). “Bayesian Analysis of Binary and Polychotomous Response Data”. In: *Journal of the American Statistical Association* 88.422, pp. 669–679. DOI: 10.1080/01621459.1993.10476321.

Audigier, V., F. Husson, and J. Josse (2016). “Multiple imputation for continuous variables using a Bayesian principal component analysis”. In: *Journal of Statistical Computation and Simulation* 86.11, pp. 2140–2156. DOI: 10.1080/00949655.2015.1104683.

Bai, F., M. Shang, and Y. Huang (2024). “Corporate culture and ESG performance: Empirical evidence from China”. In: *Journal of Cleaner Production* 437. DOI: 10.1016/j.jclepro.2024.140732.

Barnard, J. and D. B. Rubin (1999). “Miscellanea. Small-sample degrees of freedom with multiple imputation”. In: *Biometrika* 86.4, pp. 948–955. DOI: 10.1093/biomet/86.4.948.

Bedford, T. and R. M. Cooke (2002). “Vines–a new graphical model for dependent random variables”. In: *The Annals of Statistics* 30.4, pp. 1031–1068. DOI: 10.1214/aos/1031689016.

Benuzzi, M., K. Bax, S. Paterlini, and E. Taufer (2024). *Chasing ESG performance: How Methodologies Shape Outcomes.* DOI: 10.2139/ssrn.4662257.

Bosse, N. I., H. Gruson, A. Cori, E. van Leeuwen, S. Funk, and S. Abbott (2024). *Evaluating Forecasts with scoringutils in R.* DOI: 10.48550/arXiv.2205.07090.

Brand, J. P. L. (1999). “Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets”. PhD thesis. Erasmus University Rotterdam.

Breiman, L., J. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification and Regression Trees.* DOI: 10.1201/9781315139470.

Burgette, L. F. and J. P. Reiter (2010). “Multiple Imputation for Missing Data via Sequential Regression Trees”. In: *American Journal of Epidemiology* 172.9, pp. 1070–1076. DOI: 10.1093/aje/kwq260.

Buuren, S. van (2007). “Multiple imputation of discrete and continuous data by fully conditional specification”. In: *Statistical Methods in Medical Research* 16.3, pp. 219–242. DOI: 10.1177/0962280206074463.

Buuren, S. van (2018). *Flexible Imputation of Missing Data.* 2nd ed. DOI: 10.1201/9780429492259.

Buuren, S. van and K. Groothuis-Oudshoorn (2011). “mice: Multivariate Imputation by Chained Equations in R”. In: *Journal of Statistical Software* 45.3, pp. 1–67. DOI: 10.18637/jss.v045.i03.

Chapon, A., T. B. M. J. Ouarda, and Y. Hamdi (2023). “Imputation of missing values in environmental time series by D-vine copulas”. In: *Weather and Climate Extremes* 41, p. 100591. DOI: 10.1016/j.wace.2023.100591.

Czado, C. and T. Nagler (2022). “Vine Copula Based Modeling”. In: *Annual Review of Statistics and Its Application* 9, pp. 453–477. DOI: 10.1146/annurev-statistics-040220-101153.

D’Orazio, M. (2024). *StatMatch: Statistical Matching or Data Fusion.* R package version 1.4.2. URL: https://CRAN.R-project.org/package=StatMatch.

DeYoreo, M., J. P. Reiter, and D. S. Hillygus (2017). “Bayesian Mixture Models with Focused Clustering for Mixed Ordinal and Nominal Data”. In: *Bayesian Analysis* 12.3, pp. 679–703. DOI: 10.1214/16-BA1020.

Di Lascio, F. M. L., S. Giannerini, and A. Reale (2015). “Exploring copulas for the imputation of complex dependent data”. In: *Statistical Methods & Applications* 24.1, pp. 159–175. DOI: 10.1007/s10260-014-0287-2.

Di Lascio, F. M. L., A. Gatto, and S. Giannerini (2024). *CoImp: Parametric and Non-Parametric Copula-Based Imputation Methods.* R package version 2.0.0. URL: https://CRAN.R-project.org/package=CoImp.

Drempetic, S., C. Klein, and B. Zwergel (2020). “The Influence of Firm Size on the ESG Score: Corporate Sustainability Ratings Under Review”. In: *Journal of Business Ethics* 167.2, pp. 333–360. DOI: 10.1007/s10551-019-04164-1.

Gebregziabher, M. and S. M. DeSantis (2010). “Latent class based multiple imputation approach for missing categorical data”. In: *Journal of Statistical Planning and Inference* 140.11, pp. 3252–3262. DOI: 10.1016/j.jspi.2010.04.020.

Gelfand, A. E. and A. F. M. Smith (1990). “Sampling-Based Approaches to Calculating Marginal Densities”. In: *Journal of the American Statistical Association* 85.410, pp. 398–409. DOI: 10.1080/01621459.1990.10476213.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis.* 3rd ed. DOI: 10.1201/b16018.

Gelman, A., A. Jakulin, M. G. Pittau, and Y.-S. Su (2008). “A weakly informative default prior distribution for logistic and other regression models”. In: *The Annals of Applied Statistics* 2.4, pp. 1360–1383. DOI: 10.1214/08-AOAS191.

Gelman, A. and D. B. Rubin (1992). “Inference from Iterative Simulation Using Multiple Sequences”. In: *Statistical Science* 7.4, pp. 457–472. DOI: 10.1214/ss/1177011136.

Gneiting, T. and A. E. Raftery (2007). “Strictly Proper Scoring Rules, Prediction, and Estimation”. In: *Journal of the American Statistical Association* 102.477, pp. 359–378. DOI: 10.1198/016214506000001437.

Gower, J. C. (1971). “A General Coefficient of Similarity and Some of Its Properties”. In: *Biometrics* 27.4, pp. 857–871. DOI: https://doi.org/10.2307/2528823.

Hasler, C., R. V. Craiu, and L.-P. Rivest (2018). “Vine Copulas for Imputation of Monotone Non-response”. In: *International Statistical Review* 86.3, pp. 488–511. DOI: 10.1111/insr.12263.

Heitjan, D. F. and R. J. A. Little (1991). “Multiple Imputation for the Fatal Accident Reporting System”. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 40.1, pp. 13–29. DOI: 10.2307/2347902.

Hollenbach, F. M., I. Bojinov, S. Minhas, N. W. Metternich, M. D. Ward, and A. Volfovsky (2021). “Multiple Imputation Using Gaussian Copulas”. In: *Sociological Methods & Research* 50.3, pp. 1259–1283. DOI: 10.1177/0049124118799381.

Honaker, J. and G. King (2010). “What to Do about Missing Values in Time-Series Cross-Section Data”. In: *American Journal of Political Science* 54.2, pp. 561–581. DOI: 10.1111/j.1540-5907.2010.00447.x.

Honaker, J., G. King, and M. Blackwell (2011). “Amelia II: A Program for Missing Data”. In: *Journal of Statistical Software* 45.7, pp. 1–47. DOI: 10.18637/jss.v045.i07.

Joe, H. (1996). “Families of m-Variate Distributions With Given Margins and $m(m-1)/2$ Bivariate Dependence Parameters”. In: *Distributions with fixed marginals and related topics*. Ed. by L. Rüschendorf, B. Schweizer, and M. D. Taylor. DOI: 10.1214/lnms/1215452614.

Jordan, A., F. Krüger, and S. Lerch (2019). “Evaluating Probabilistic Forecasts with scoringRules”. In: *Journal of Statistical Software* 90.12, pp. 1–37. DOI: 10.18637/jss.v090.i12.

Kaufman, L. and P. J. Rousseeuw (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. DOI: 10.1002/9780470316801.

Kim, H. J., J. P. Reiter, Q. Wang, L. H. Cox, and A. F. Karr (2014). “Multiple Imputation of Missing or Faulty Values Under Linear Constraints”. In: *Journal of Business & Economic Statistics* 32.3, pp. 375–386. DOI: 10.1080/07350015.2014.885435.

King, G., J. Honaker, A. Joseph, and K. Scheve (2001). “Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation”. In: *American Political Science Review* 95.1, pp. 49–69. DOI: 10.1017/S0003055401000235.

Kotsantonis, S. and G. Serafeim (2019). “Four Things No One Will Tell You About ESG Data”. In: *Journal of Applied Corporate Finance* 31.2, pp. 50–58. DOI: 10.1111/jacf.12346.

Kraus, D. and C. Czado (2017). “D-vine copula based quantile regression”. In: *Computational Statistics & Data Analysis* 110, pp. 1–18. DOI: 10.1016/j.csda.2016.12.009.

Little, R. J. A. and D. B. Rubin (2020). *Statistical Analysis with Missing Data*. 3rd ed. DOI: 10.1002/9781119482260.

Little, R. J. A. and M. D. Schluchter (1985). “Maximum likelihood estimation for mixed continuous and categorical data with missing values”. In: *Biometrika* 72.3, pp. 497–512. DOI: 10.1093/biomet/72.3.497.

Liu, C. and D. B. Rubin (1998). “Ellipsoidally symmetric extensions of the general location model for mixed categorical and continuous data”. In: *Biometrika* 85.3, pp. 673–688. DOI: 10.1093/biomet/85.3.673.

LSEG (2023). *Environmental, social and governance scores from LSEG*. URL: https://www.lseg.com/content/dam/data-analytics/en_us/documents/methodology/lseg-esg-scores-methodology.pdf (visited on 06/19/2024).

Manrique-Vallier, D. and J. P. Reiter (2014). “Bayesian multiple imputation for large-scale categorical data with structural zeros”. In: *Survey Methodology* 40.1, pp. 125–134.

Min, A. and C. Czado (2011). “Bayesian model selection for D-vine pair-copula constructions”. In: *Canadian Journal of Statistics* 39.2, pp. 239–258. DOI: 10.1002/cjs.10098.

Murray, J. S. (2018). “Multiple Imputation: A Review of Practical and Theoretical Findings”. In: *Statistical Science* 33.2, pp. 142–159. DOI: 10.1214/18-STS644.

Murray, J. S. and J. P. Reiter (2016). “Multiple Imputation of Missing Categorical and Continuous Values via Bayesian Mixture Models With Local Dependence”. In: *Journal of the American Statistical Association* 111.516, pp. 1466–1479. DOI: 10.1080/01621459.2016.1174132.

Nagler, T. and D. Kraus (2024). *vinereg: D-Vine Quantile Regression.* R package version 0.10.0. URL: https://CRAN.R-project.org/package=vinereg.

Nagler, T. and T. Vatter (2022). *kde1d: Univariate Kernel Density Estimation.* R package version 1.0.5. URL: https://CRAN.R-project.org/package=kde1d.

Nagler, T. and T. Vatter (2023). *rvinecopulib: High Performance Algorithms for Vine Copula Modeling.* R package version 0.6.3.1.1. URL: https://CRAN.R-project.org/package=rvinecopulib.

Novo, A. A. and J. L. Schafer (2023). *norm: Analysis of Multivariate Normal Datasets with Missing Values.* R package version 1.0-11.1. URL: https://CRAN.R-project.org/package=norm.

Olkin, I. and R. F. Tate (1961). “Multivariate Correlation Models with Mixed Discrete and Continuous Variables”. In: *The Annals of Mathematical Statistics* 32.2, pp. 448–465. DOI: 10.1214/aoms/1177705052.

Paddock, S. M. (2002). “Bayesian nonparametric multiple imputation of partially observed data with ignorable nonresponse”. In: *Biometrika* 89.3, pp. 529–538. DOI: 10.1093/biomet/89.3.529.

Parzen, E. (1962). “On Estimation of a Probability Density Function and Mode”. In: *The Annals of Mathematical Statistics* 33.3, pp. 1065–1076. DOI: 10.1214/aoms/1177704472.

R Core Team (2023). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Raghunathan, T. E., J. M. Lepkowski, J. Van Hoewyk, and P. Solenberger (2001). “A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models”. In: *Survey Methodology* 27.1, pp. 85–95.

Reiter, J. P. (2005). “Using CART to Generate Partially Synthetic Public Use Microdata”. In: *Journal of Official Statistics* 21.3, pp. 441–462.

Rubin, D. B. (1978). “Multliple Imputations in Sample Surveys – A Phemomenoligical Bayesian Approach to Nonresponse”. In: *Proceedings of the Survey Research Methods Section.* American Statistical Association, pp. 20–29.

Rubin, D. B. (1981). “The Bayesian Bootstrap”. In: *The Annals of Statistics* 9.1, pp. 130–134. URL: https://www.jstor.org/stable/2240875.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* DOI: 10.1002/9780470316696.

Rubin, D. B. (1996). “Multiple Imputation After 18+ Years”. In: *Journal of the American Statistical Association* 91.434, pp. 473–489. DOI: 10.2307/2291635.

Rubin, D. B. and N. Schenker (1986). “Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse”. In: *Journal of the American Statistical Association* 81.394, pp. 366–374. DOI: 10.2307/2289225.

Sahin, Ö., K. Bax, C. Czado, and S. Paterlini (2022). “Environmental, Social, Governance scores and the Missing pillar—Why does missing information matter?” In: *Corporate Social Responsibility and Environmental Management* 29.5, pp. 1782–1798. DOI: 10.1002/csr.2326.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.*

Schallhorn, N., D. Kraus, T. Nagler, and C. Czado (2017). *D-vine quantile regression with discrete variables.* DOI: 10.48550/arXiv.1705.08310.

Schenker, N. and J. M. Taylor (1996). “Partially parametric techniques for multiple imputation”. In: *Computational Statistics & Data Analysis* 22.4, pp. 425–446. DOI: 10.1016/0167-9473(95)00057-7.

Si, Y. and J. P. Reiter (2013). “Nonparametric Bayesian Multiple Imputation for Incomplete Categorical Variables in Large-Scale Assessment Surveys”. In: *Journal of Educational and Behavioral Statistics* 38.5, pp. 499–521. DOI: 10.3102/1076998613480394.

Sklar, M. (1959). “Fonctions de répartition à N dimensions et leurs marges”. In: *Annales de l’ISUP* 8.3, pp. 229–231. URL: https://hal.science/hal-04094463.

Su, Y.-S., A. Gelman, J. Hill, and M. Yajima (2011). “Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box”. In: *Journal of Statistical Software* 45.2, pp. 1–31. DOI: 10.18637/jss.v045.i02.

Tanner, M. A. and W. H. Wong (1987). “The Calculation of Posterior Distributions by Data Augmentation”. In: *Journal of the American Statistical Association* 82.398, pp. 528–540. DOI: 10.1080/01621459.1987.10478458.

Vermunt, J. K., J. R. van Ginkel, L. A. van der Ark, and K. Sijtsma (2008). “MULTIPLE IMPUTATION OF INCOMPLETE CATEGORICAL DATA USING LATENT CLASS ANALYSIS”. In: *Sociological Methodology* 38.1, pp. 369–397. DOI: 10.1111/j.1467-9531.2008.00202.x.