A COMPREHENSIVE SIMULATION STUDY ON THE FORWARD IMPUTATION

NADIA SOLARO ALESSANDRO BARBIERO GIANCARLO MANZI PIER ALDA FERRARI

Working Paper n. 2015-04

FEBBRAIO 2015

DIPARTIMENTO DI ECONOMIA, MANAGEMENT E METODI QUANTITATIVI

Via Conservatorio 7
20122 Milano
tel. ++39 02 503 21501 (21522) - fax ++39 02 503 21450 (21505)
http://www.economia.unimi.it
E Mail: dipeco@unimi.it

Università degli Studi di Milano

A comprehensive simulation study on the Forward Imputation

Nadia Solaro[1], **Alessandro Barbiero**[2], **Giancarlo Manzi**[2], **and Pier Alda Ferrari**[2]

[1]Nadia Solaro
Department of Statistics and Quantitative Methods
Università degli Studi di Milano-Bicocca
Via Bicocca degli Arcimboldi, 8 – 20126 Milano, Italy
E-mail address: nadia.solaro@unimib.it

[2]Alessandro Barbiero, Giancarlo Manzi, and Pier Alda Ferrari
Department of Economics, Management and Quantitative Methods
Università degli Studi di Milano
Via Conservatorio, 7 – 20122 Milano, Italy
E-mail addresses: alessandro.barbiero@unimi.it, giancarlo.manzi@unimi.it, pieralda.ferrari@unimi.it

23rd February 2015

Abstract

The Nearest Neighbour Imputation (NNI) method has a long history in missing data imputation. Likewise, multivariate dimensional reduction techniques allow for preserving the maximum information from the data. Recently, the combined use of these methodologies has been proposed to solve data imputation problems and exploit as much as information from the complete part of the data. In this paper we perform an extensive simulation study to test the performance of this new imputation approach (called "Forward Imputation" - *ForImp*). We compare the two *ForImp* methods developed for missing quantitative data (the first one called *ForImpPCA* involving the NNI method and the Principal Component Analysis (PCA) as a multivariate data analysis technique, and the second one called *ForImpMahalanobis*, which involves the Mahalanobis distance for NNI) with other two imputation techniques regarded as benchmark, namely Stekhoven and Bühlmann's *missForest* method, which is a nonparametric imputation technique for continuous and/or categorical data based on a random forest, and the *Iterative PCA*, which is an algorithmic-type technique that imputes missing values simultaneously by an iterative use of PCA. The simulation study is based on constructing simulated data with different levels of kurtosis or skewness and strength of linear relationship of variables, so that the performance of the four methods can be compared on various data patterns. Distributions used for these simulated data belong to the families of Multivariate Exponential Power and Multivariate Skew-Normal distributions, respectively. Results tend to favour *ForImpMahalanobis* especially in the presence of skew data with small or negative correlations of a same magnitude, or a mix of negative and positive correlations of low level, whereas *ForImpPCA* works better than it when a slightly higher level of correlations is present in the data.

Keywords: Correlation, Data patterns, Kurtosis, Mahalanobis distance, MissForest, Nearest Neighbour Imputation, Principal Component Analysis, Skewness.

JEL classifications: C15, C18, C38, C63.

1 Introduction

Missing data are a recurring problem in almost every field of quantitative research. Working with large datasets inevitably means dealing with incompleteness, hence the need to find a solution to this problem easy to implement. Very different approaches have been proposed through the years, like, for example, deletion methods, model-based procedures, nonparametric or distribution-free procedures, single/multiple imputation, and so on. For a thorough review and discussion of missing data techniques see Little and Rubin (2002) and Rässler et al. (2013).

Within the nonparametric framework, distance-based methods expressly consider distances between complete and incomplete units to impute missing values. The nearest-neighbour imputation (NNI) is a prominent case of distance-based method, since imputation is performed by relying on donors, which are the complete units nearest to the incomplete ones detected according to a specific, pre-chosen measure of "closeness". Recently, a forward imputation (*ForImp*) approach was introduced by Solaro et al. (2015) as a sequential distance-based, distribution-free imputation procedure (see also Solaro et al. (2014)). *ForImp* applies the NNI method *forward* and exclusively to the complete part of data that updates further to every step of an iteration procedure, and *possibly* in alternation with a multivariate data analysis (MVDA) technique, which is introduced to synthesize the information of the complete part of data. In Solaro et al. (2015), two alternative *ForImp* methods were specifically proposed for imputing missing quantitative data: (A) *ForImp* with the Principal Component Analysis (*ForImpPCA – FIP*), which involves the NNI method and PCA as the MVDA technique; (B) *ForImp* with the Mahalanobis distance (*ForImpMahalanobis – FIM*), which involves only NNI applied to the original variables.

This work is addressed to inspect the quality of performance of the *FIP* and *FIM* methods by means of an extensive Monte Carlo simulation study that involves a multitude of different data patterns. This is the objective of Section 2. Alternative imputation methods were also considered for benchmark, namely: (i) Stekhoven and Bühlmann's *missForest* method (2012); (ii) the Iterative PCA (Nora-Chouteau (1974); Greenacre (1984), *IPCA*). The first is a nonparametric imputation technique for continuous and/or categorical data based on a random forest, i.e. a random classifier introduced in the context of machine learning (Breiman, 2001). The second is an algorithmic-type technique that imputes missing values simultaneously by the iterative use of PCA. *IPCA* is at the core of the multiple imputation method with PCA, recently introduced by Josse et al. (2011) as a part of a more general methodology with principal component methods (*missMDA*). Both *missForest* and *IPCA* are nonparametric methods, suitable in the exploratory framework, and available in the R environment (R Development Core Team, 2014). In particular, *IPCA* is implemented in the R library "missMDA" by Josse et al. (2011), and *missForest* in the homonymous R library "missForest" by Stekhoven and Bühlmann (2012).

Another main aspect of concern, which is strictly linked to the performance assessment carried out in this context, is how to choose the "ideal" imputation method consistently with the types of data structures at hand. This point is particularly crucial in an exploratory framework, like the one we refer to, where no distribution assumption is made on data. A contribution on this matter is proposed in Section 3, in which descriptive indices are introduced to synthesize the considered correlation structures and thus recognize, also by means of kurtosis and skewness indices, the typology of data pattern. Some conclusions are finally given in Section 4.

2 Simulation study

The two *ForImp* methods *FIP* and *FIM* perform the imputation of missing quantitative data in a step-by-step process that is carried out forward and sequentially by starting from the complete part of data. No initialization of missing data is required. Imputation is fulfilled by exploiting the covariance/correlation structure inherent in the complete part of data, which is sequentially updated at every step of the process. The main difference between the two methods is that *FIP* involves both the NNI method (applied with a weighted Minkowski distance of order r, $r \geq 1$) and a MVDA technique, i.e. PCA, while *FIM* uses only NNI (applied with the Mahalanobis distance). All the methodological details concerning *FIP* and *FIM* are given in Solaro et al. (2015).

With the purpose of assessing and comparing the performance of *FIM* and *FIP*, an extensive Monte Carlo simulation study was undertaken in the presence of data having different shapes, as given by kurtosis or skewness, and correlation structures. These characteristics, opportunely combined together to produce different data patterns, were treated, along with data dimensionality (i.e. number of units and variables), as *exogenous factors*, in that they are relevant to the data, and not to the methods. On the other hand, options more closely pertaining to the methods, i.e. donors' quantiles and metrics (only for *FIP*), were regarded as *endogenous factors*. The main objective of the simulation study was then to examine the performance of *FIM* and *FIP* with respect to both exogenous and endogenous factors, and detect, if possible, the most effective method according to specific features of data patterns. In order to generate data having the desired shapes, heavy/thin-tailed symmetric or skew data respectively, which are very common in real situations, we relied on two families of multivariate distributions. The first is the Multivariate Exponential Power (*MEP*) family (Gómez et al., 1998), which belongs to the class of elliptical symmetric multivariate distributions (Fang et al., 1990). The second is the Multivariate Skew-Normal (*MSN*) family of distributions (Azzalini and Dalla Valle, 1996; Azzalini and Capitanio, 1998).

In Subsect. 2.1, after briefly mentioning *MEP* and *MSN* main results, the simulation design is described across its main steps, i.e. the definition of data patterns and the pertinent simulation settings, the simulation procedure and summary of results. These latter are then presented and analysed in Subsect. 2.2.

2.1 Simulation design

One aspect of main concern was to fix experimental conditions, namely the exogenous factors, in a way that they would reproduce a variety of data patterns we often encountered in applications. On this point, besides considering "dimensionality of data" (number of units and variables) and "seriousness of missingness" (i.e. percentages of missing values), correlations of variables, kurtosis and skewness of the data distribution were more closely taken into account, for they were expected to greatly affect the imputation performance. Specific data patterns were then defined by combining different items of "shape" and "linear relationship" together. Regarding shape, we considered the two forms: SyKu (i.e. symmetry and kurtosis), and SK (i.e. skewness). For the linear relationship, we relied on the structures: ECor (equal correlations, or equicorrelations), PNCor (positive-negative correlations), and UnbCor (unbalanced correlations), each of which was modulated at three different levels of correlation strength, e.g. absent/low, moderate, or medium/high.

Another aspect of concern was to restrain, as much as possible, the total number of simulation scenarios to be run without loosing any meaningful information about the trends. This is the reason why we performed, at a first stage, an exploratory simulation

Table 1: Formal definitions of *MEP* and *MSN* family of distributions

MEP family of distributions: $\boldsymbol{X} \sim \text{MEP}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \kappa)$	
Density function	$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \kappa) = \frac{p\Gamma(p/2)}{\pi^{p/2}\Gamma(1+p/\kappa)2^{1+p/\kappa}\lvert\boldsymbol{\Sigma}\rvert^{1/2}} \cdot \exp\{-\frac{1}{2}[(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})]^{\kappa/2}\},$ $\boldsymbol{\mu} \in \mathbb{R}^p,\ \kappa > 0,\ \boldsymbol{\Sigma}_{p\times p}$ ("characteristic matrix", positive-definite)
Meaning of parameters	– Mean vector: $\text{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ – Variance-covariance matrix: $\text{V}(\boldsymbol{X}) = c(\kappa, p)\boldsymbol{\Sigma}$, with: $c(\kappa, p) = 2^{2/\kappa}\Gamma((p+2)/\kappa)/(p\Gamma(p/\kappa))$ – Kurtosis parameter: $\kappa = 2 \rightarrow$ normal distribution $\kappa < 2$ ($\kappa > 2$) $\rightarrow$ leptokurtic (platykurtic) distribution
Skewness indices	Univariate: $\gamma_1 = 0$; Multivariate (MV): $\gamma_{1\text{MV}} = 0$
MV kurtosis index	$\gamma_{2\text{MV}} = \frac{p^2\Gamma(p/\kappa)\Gamma((p+4)/\kappa)}{\Gamma^2((p+2)/\kappa)} - p(p+2)$
MSN family of distributions: $\boldsymbol{X} \sim \text{MSN}_p(\boldsymbol{\Omega}, \boldsymbol{\alpha})$	
Density function	$f(\mathbf{x}; \boldsymbol{\Omega}, \boldsymbol{\alpha}) = 2\phi_p(\mathbf{x}; \boldsymbol{\Omega})\Phi(\boldsymbol{\alpha}^t\mathbf{x})$, where: – $\phi_p(\mathbf{x}; \boldsymbol{\Omega})$ is the $\text{MVN}_p(\mathbf{0}, \boldsymbol{\Omega})$ d.f. with correlation matrix $\boldsymbol{\Omega}$ – $\Phi(\cdot)$ is the $N(0,1)$ distribution function, and $\boldsymbol{\alpha} \in \mathbb{R}^p$
Meaning of parameters	– Mean vector: $\text{E}(\boldsymbol{X}) = \boldsymbol{\mu} = \sqrt{2/\pi}\boldsymbol{\delta}$, with: $\boldsymbol{\delta} = \frac{\boldsymbol{\Omega}\boldsymbol{\alpha}}{\sqrt{1+\boldsymbol{\alpha}^t\boldsymbol{\Omega}\boldsymbol{\alpha}}}$ – Variance-covariance matrix: $\text{V}(\boldsymbol{X}) = \boldsymbol{\Sigma} = \boldsymbol{\Omega} - \boldsymbol{\mu}\boldsymbol{\mu}^t$ – Correlation matrix: $\mathbf{R} = \mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1}$, with: $\mathbf{D} = \text{diag}\left\{\sqrt{1 - 2\pi^{-1}\delta_j^2}\right\}_{j=1,\ldots,p}$ – Parameter related to the skewness: $\boldsymbol{\alpha} \in \mathbb{R}^p$. If: $\boldsymbol{\alpha} = \mathbf{0}$, then: $\boldsymbol{X} \sim \text{MVN}_p(\mathbf{0}, \boldsymbol{\Omega})$
Skewness indices	– Univariate: $\gamma_1 = \frac{4-\pi}{2}\frac{\text{E}(X_j)^3}{\text{Var}(X_j)^{3/2}} \in (-0.995, +0.995)$ – MV: $\gamma_{1\text{MV}} = \left(\frac{4-\pi}{2}\right)^2(\boldsymbol{\mu}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^3 \in (-0.9905, +0.9905)$
MV kurtosis index	$\gamma_{2\text{MV}} = 2(\pi-3)(\boldsymbol{\mu}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2 \in (-0.869, +0.869)$

study, followed by, at a second stage, supplementary simulations addressed to look more thoroughly into specific situations.

Multivariate distributions for simulation. Data with the desired patterns were generated by relying on *MEP* (Gómez et al., 1998) and *MSN* (Azzalini and Dalla Valle, 1996; Azzalini and Capitanio, 1998) families of distributions. They can be regarded as extensions of the multivariate normal (*MVN*) distribution in terms of kurtosis or skewness departures, respectively. To make clearer the role of parameters in our simulation study, and the data patterns deriving from them, a synthetic collection of the main theoretical results is given in Table 1. Given the complex patterns of variations among parameters of the *MSN* distributions, a few remarks are worth making about the link between the (α, ω)-parametrization, on one hand, and the correlation coefficient ρ as well as the skewness indices γ_1 (univariate) and $\gamma_{1\text{MV}}$ (multivariate), on the other hand. Table 2 reports an instance of the computations performed for different *MSN* distributions with $p = 3, 5, 10$

Table 2: Values of output correlation coefficient and skewness indices for $p = 3, 5$ and 10 variables and different values of ω and $\alpha = 1, 4, 10, 30$

Output		$p = 3$			$p = 5$			$p = 10$		
Input	α	ρ	γ_1	$\gamma_{1\text{MV}}$	ρ	γ_1	$\gamma_{1\text{MV}}$	ρ	γ_1	$\gamma_{1\text{MV}}$
$\omega = 0$	1	-0.189	0.035	0.141	-0.119	0.018	0.266	-0.061	0.007	0.478
	4	-0.262	0.058	0.838	-0.144	0.023	0.895	-0.067	0.007	0.941
	10	-0.268	0.060	0.964	-0.146	0.024	0.974	-0.068	0.008	0.982
	30	-0.269	0.060	0.988	-0.146	0.024	0.989	-0.068	0.008	0.990
$\omega = 0.3$	1	0.026	0.105	0.254	0.058	0.087	0.507	0.092	0.070	0.799
	4	-0.053	0.154	0.891	0.030	0.103	0.945	0.085	0.073	0.977
	10	-0.059	0.157	0.974	0.028	0.104	0.983	0.084	0.073	0.988
	30	-0.060	0.158	0.989	0.028	0.104	0.990	0.084	0.073	0.990
$\omega = 0.5$	1	0.214	0.186	0.319	0.221	0.179	0.598	0.238	0.163	0.856
	4	0.138	0.264	0.910	0.193	0.206	0.957	0.231	0.169	0.981
	10	0.132	0.271	0.977	0.191	0.208	0.985	0.231	0.170	0.989
	30	0.131	0.272	0.989	0.191	0.209	0.990	0.231	0.170	0.990
$\omega = 0.8$	1	0.609	0.402	0.400	0.592	0.457	0.685	0.587	0.472	0.897
	4	0.558	0.571	0.928	0.572	0.524	0.967	0.582	0.489	0.984
	10	0.554	0.584	0.980	0.570	0.528	0.987	0.582	0.490	0.990
	30	0.554	0.586	0.989	0.570	0.529	0.990	0.582	0.490	0.991

variables, respectively, with the components of the r.v. $\boldsymbol{X}$ sharing the same $\alpha_j = \alpha$ and $\omega_{lj} = \omega$ parameters, $(j, l = 1, \ldots, p,\ j \neq l)$. The grey lines are introduced to ease reading and comparing the results.

As expected, with ω and p kept fixed, the skewness indices γ_1 and $\gamma_{1\text{MV}}$ increase as α increases. The same occurs for ω growing, with α and p fixed. On the contrary, by keeping ω and α fixed, γ_1 decreases while $\gamma_{1\text{MV}}$ increases when p increases. Similarly, *ceteris paribus*, ρ increases as ω grows, while it tends to decrease with α growing, although mostly in a subtle way. It is interesting to note, for the following discussion, that the value of ρ is always lower than ω. In conclusion, while $\gamma_{1\text{MV}}$ is sensitive to variation of ω and α together, ρ and γ_1 are mainly sensitive to ω. As for the number p of variables, ρ and $\gamma_{1\text{MV}}$ increase as p increases, while γ_1 decreases. The only exception is the case $(\alpha = 1, \omega = 0.8)$, where both ρ and γ_1 decrease as p increases. However, in the range of the considered values, the most appreciable numerical variations are observed as ω varies, and for α moving from 1 (a situation of "slighter" skewness) to 4 (situations of "manifest" skewness). For $p \geq 5$ and $\alpha \geq 4$ the value of ρ, γ_1 and $\gamma_{1\text{MV}}$ keep quite stable in magnitude as α varies.

Data patterns and simulation settings. Table 3 collects all the data patterns and the pertinent simulation settings considered in the exploratory simulation study. Apart from dimensionality of data and percentage of missing values, *input parameters* specified in the table concern the multivariate distributions used for random data generation (Table 1). *Output parameters* play a more important role, especially for the subsequent interpretation of results, since they pertain to characteristics of the final distribution of data. As such, then, they could be easily ascertained on any real dataset, by computing the usual summary descriptive statistics.

The SyKu shape was artificially generated through *MEP* distributions (Table 1), by using the transformation method described in Gómez et al. (1998) and Solaro (2004), and fixing the output correlation matrix $\mathbf{R}$ rather than the input characteristic matrix $\boldsymbol{\Sigma}$ in the *MEP* density function (d.f.). Attention was then focused on three different types of distributions, i.e. leptokurtic ($\kappa = 1$), normal ($\kappa = 2$), and platykurtic ($\kappa = 14$). As

apparent from Table 1, kurtosis depending on the κ parameter only for fixed p, its effect on imputation performance can be studied straightforwardly, irrespective of the strength of correlations. The SyKu shape was initially considered in the presence of a non-negative equicorrelation structure with three different levels of magnitude for ρ, (Table 3).

The SK shape was generated through *MSN* distributions (Table 1). As before noticed, output skewness and strength of linear relationship among variables are strictly related to both vector $\boldsymbol{\alpha}$ and, more importantly, matrix $\boldsymbol{\Omega}$ (Table 2). Given the various constraints among the parameters, random variate generation by fixing the output ρ and/or γ_1 and $\gamma_{1\text{MV}}$ would have been impracticable. We relied therefore on the method implemented by Azzalini in the R library "sn" (Azzalini, 2013), which is based on the (α, ω)-parametrization of the *MSN* d.f.. Interpretation to simulation results was then based on the relationships between input and output parameters, (e.g. Table 2).

Examinations concerning the SK shape were performed according to the three studies indicated in Table 3, which essentially differ for the structure assigned to $\boldsymbol{\Omega}$, and consequently the output correlation matrix $\mathbf{R}$. The 1$^{\text{st}}$ study refers to the SK-ECor pattern (skewness with equicorrelations). Matrix $\boldsymbol{\Omega}$ contains the same non-negative ω for all the pairs of the p components. Matrix $\mathbf{R}$ preserves the same equicorrelation structure (ECor) with values of its entries given in Table 2. In particular, three levels of magnitude of ρs are considered in the simulations: Low-negative ρs (≈ -0.1), resulting from $\omega = 0$; low-positive ρs (≈ 0.2), corresponding to $\omega = 0.5$; medium-positive ρs (≈ 0.6), given by $\omega = 0.8$, (Table 2).

The 2$^{\text{nd}}$ study concerns the SK-PNCor pattern (skewness with positive and negative correlations). Matrix $\boldsymbol{\Omega}$ contains the same ω in absolute value but with alternating sign. This produces an output matrix $\mathbf{R}$ having positive and negative correlations of a very similar magnitude (PNCor), according to the structure formally described in Table 3, and with values in the ranges reported in Table 4. Ranges of ρ coefficients remain quite stable as p and/or α vary, (punctual values of output parameters are omitted). They are mainly sensitive to variations of ω parameter. In the simulation study, we considered therefore the following levels: For $\omega = 0.2$, positive-negative low ρs ($\rho_1 \approx -0.3$, $\rho_2 \approx 0.1$, $\rho_3 \approx 0.2$); for $\omega = 0.5$, positive-negative moderate ρs ($\rho_1 \approx -0.5$, $\rho_2 \approx 0.4$, $\rho_3 \approx 0.5$); for $\omega = 0.8$, positive-negative high ρs ($\rho_1 \approx -0.75$, $\rho_2 \approx 0.7$, $\rho_3 \approx 0.75$).

As an instance of the structure of $\boldsymbol{\Omega}$ and $\mathbf{R}$ regarded in this 2$^{\text{nd}}$ study, if $p = 5$ and $\omega = 0.2$, matrix $\boldsymbol{\Omega}$ is equal to:

$$\boldsymbol{\Omega} = \begin{pmatrix} 1 & 0.2 & -0.2 & 0.2 & -0.2 \\ 0.2 & 1 & -0.2 & 0.2 & -0.2 \\ -0.2 & -0.2 & 1 & -0.2 & 0.2 \\ 0.2 & 0.2 & -0.2 & 1 & -0.2 \\ -0.2 & -0.2 & 0.2 & -0.2 & 1 \end{pmatrix},$$

while if $\alpha = 1$ then matrix $\mathbf{R}$ is given by:

$$\mathbf{R} = \begin{pmatrix} 1 & 0.088 & -0.299 & 0.088 & -0.299 \\ 0.088 & 1 & -0.299 & 0.088 & -0.299 \\ -0.299 & -0.299 & 1 & -0.299 & 0.163 \\ 0.088 & 0.088 & -0.299 & 1 & -0.299 \\ -0.299 & -0.299 & 0.163 & -0.299 & 1 \end{pmatrix},$$

where: $\rho_1 = -0.299$, $\rho_2 = 0.088$, and $\rho_3 = 0.163$ according to the notation of Table 3.

The 3$^{\text{rd}}$ study pertains to the SK-UnbCor pattern (skewness with unbalanced correlations). We assume that the first component in $\boldsymbol{\Omega}$ is negatively correlated with all the

Table 3: Data patterns and experimental conditions in the simulation study

– Number of units in $\mathbf{X}$	$n = 500; 1000$
– Percentage of MCAR missing values	$5\%; 10\%; 20\%$

• *The SyKu shape* – Symmetry and Kurtosis
⇒ Data generation from $\text{MEP}_p(\mathbf{0}, \mathbf{\Sigma}, \kappa)$, with $\mathbf{\Sigma} = c^{-1}(\kappa, p)\mathbf{R}$, (Table 1):

→ *Input parameters*	→ *Output parameters*
– Number of variables in $\mathbf{X}$ $p = 3; 5; 10$	– Kurtosis index $\gamma_{2\text{MV}} = \gamma_{2\text{MV}}(\kappa, p)$ for $\kappa = 1$: $\gamma_{2\text{MV}} = 7.50, 11.67, 21.82$
– Kurtosis parameter $\kappa = 1; 2; 14$	for $\kappa = 2$: $\gamma_{2\text{MV}} = 0$, $\forall p$ for $\kappa = 14$: $\gamma_{2\text{MV}} = -4.05, -7.25, -15.64$

⊞ *ECor* structure → *The SyKu-ECor pattern*:

– Correlation coefficient $\rho = 0; 0.3; 0.7$	– Correlation coefficient $\rho = 0; 0.3; 0.7$

• *The SK shape* – Skewness
⇒ Data generation from $\text{MSN}_p(\mathbf{\Omega}, \boldsymbol{\alpha})$, with $\mathbf{\Omega} = [\omega_{lj}]_{l,j=1,\ldots,p}$ and $\boldsymbol{\alpha} = [\alpha_j]_{j=1,\ldots,p}$, (Table 1):

→ *Input parameters*	→ *Output parameters*
– Skewness parameter $\alpha_j = \alpha = 1; 4; 10; 30$, $\forall j$	

⊞ *ECor* structure (1st study) → *The SK-ECor pattern*:

∗ Number of variables and input correlation:	∗ Skewness and output correlation:
– $p = 3; 5; 10$ – $\omega_{lj} = \omega = 0; 0.5; 0.8$	– γ_1 and $\gamma_{1\text{MV}}$ given in Table 2 – $\rho_{lj} = \rho$, $\forall l, j$, with values given in Table 2

⊞ *PNCor* structure (2nd study) → *The SK-PNCor pattern*:

∗ Number of variables and input correlation:	∗ Skewness and output correlation:
– $p = 5; 10$	– γ_1 and $\gamma_{1\text{MV}}$ with range given in Table 4 – For odd (even) p, set: $m = p - 1$ $(= p - 2)$.
– For $j = 2, \ldots, p$:	Then, for each j (# is number):
$\begin{cases} \omega_{1j} = \omega_{j1} = (-1)^j \omega \\ \omega_{jv} = \omega \quad \text{if } \text{sign}(\omega_{lj}) = \text{sign}(\omega_{lv}) \\ \omega_{jv} = -\omega \quad \text{if } \text{sign}(\omega_{lj}) \neq \text{sign}(\omega_{lv}), \\ \qquad (l, v = 1, \ldots, p,\ l \neq v \neq j) \end{cases}$	$\begin{cases} \rho_{jv} = \rho_1 \text{ if } \omega_{jv} = -\omega \text{ and } \#\text{neg. } \omega_{jl} = \frac{m}{2} \\ \rho_{jv} = \rho_2 \text{ if } \omega_{jv} = \omega \text{ and:} \\ \qquad \text{for odd } p : \#\text{pos. } \omega_{jl} = \frac{m}{2} \\ \qquad \text{for even } p : \#\text{pos. } \omega_{jl} = \frac{m}{2} + 1 \\ \rho_{jv} = \rho_3 \text{ otherwise,} \\ (l, v = 1, \ldots, p,\ l \neq v \neq j) \end{cases}$
with $\omega = 0.2; 0.5; 0.8$	with ρ_1, ρ_2, and ρ_3 given in Table 4

⊞ *UnbCor* structure (3rd study) → *The SK-UnbCor pattern*:

∗ Number of variables and input correlation:	∗ Skewness and output correlation:
– $p = 5$	– γ_1 and $\gamma_{1\text{MV}}$ with range given in Table 5
– $\begin{cases} \omega_{1j} = \omega_{j1} = -\omega, \\ \omega_{lj} = \omega / c, \quad \text{for } l \neq 1, \end{cases}$	– $\begin{cases} \rho_{1j} = \rho_1, \\ \rho_{lj} = \rho_2, \quad \text{for } l \neq 1, \end{cases}$
with $\omega = 0.2; 0.5; 0.8$ and $c = 1; 1.25; 1.5$	with ρ_1 and ρ_2 given in Table 5

others by a same value $-\omega$, while the other components are positively correlated to each other by a common parameter equal to ω/c, where c regulates the extent of unbalancing in

Table 4: The SK-PNCor pattern. Range of variation of output parameters (*MSN* data, 2nd study)

Input	Output parameters				
	ρ_1	ρ_2	ρ_3	γ_1	$\gamma_{1\text{MV}}$
$\omega = 0.2$	$(-0.33, -0.24)$	$(0.05, 0.12)$	$(0.15, 0.20)$	$(\ 0.0, 0.4)$	$(0.20, 0.99)$
	(positive-negative low ρs)				
$\omega = 0.5$	$(-0.57, -0.48)$	$(0.36, 0.41)$	$(0.48, 0.50)$	$(\ 0.0, 0.6)$	$(0.14, 0.99)$
	(positive-negative moderate ρs)				
$\omega = 0.8$	$(-0.79, -0.72)$	$(0.66, 0.75)$	$(0.73, 0.78)$	$(-0.1, 0.3)$	$(0.06, 0.99)$
	(positive-negative high ρs)				

Table 5: The SK-UnbCor pattern. Range of variation of output parameters (*MSN* data, 3rd study)

Input parameters		Output parameters			
		ρ_1	ρ_2	γ_1	$\gamma_{1\text{MV}}$
$\omega = 0.2$	$c = 1$	$(-0.27, -0.25)$	$(-0.02, \ 0.03)$	$(\ 0.0, 0.1)$	$(0.30, 0.99)$
	$c = 1.25$	$(-0.26, -0.24)$	$(-0.05, \ 0.00)$	$(\ 0.0, 0.1)$	$(0.28, 0.99)$
	$c = 1.5$	$(-0.26, -0.24)$	$(-0.07, -0.03)$	$(\ 0.0, 0.1)$	$(0.26, 0.99)$
		(negative low and nearly null ρs)			
$\omega = 0.5$	$c = 1$	$(-0.43, -0.41)$	$(\ 0.21, \ 0.27)$	$(-0.1, 0.2)$	$(0.36, 0.99)$
	$c = 1.25$	$(-0.42, -0.40)$	$(\ 0.12, \ 0.18)$	$(-0.1, 0.2)$	$(0.30, 0.99)$
		(negative moderate and low ρs)			
	$c = 1.5$	$(-0.42, -0.39)$	$(\ 0.06, \ 0.13)$	$(-0.1, 0.2)$	$(0.26, 0.99)$
		(negative moderate and nearly null ρs)			
$\omega = 0.8$	$c = 1$	$(-0.68, -0.65)$	$(\ 0.57, \ 0.63)$	$(-0.2, 0.5)$	$(0.41, 0.99)$
		(negative high and high ρs)			
	$c = 1.25$	$(-0.68, -0.62)$	$(\ 0.33, \ 0.41)$	$(-0.4, 0.3)$	$(0.33, 0.99)$
		(negative high and moderate ρs)			
	$c = 1.5$	$(-0.68, -0.62)$	$(\ 0.20, \ 0.29)$	$(-0.9, 0.3)$	$(0.26, 0.99)$
		(negative high and low ρs)			

the correlation matrix $\mathbf{\Omega}$, (Table 3). This produces an output matrix $\mathbf{R}$ having the same structure of $\mathbf{\Omega}$ (UnbCor) and entries with values in the ranges given in Table 5. Relations between input and output correlations are now much more sophisticated. In particular, for $\omega = 0.2$, as c varies, we have always a negative low ρ for the first variable ($\rho_1 \approx -0.25$) and nearly null values for the other variables ($\rho_2 \approx 0$). This case is therefore denoted as "negative low and nearly null" level. For $\omega = 0.5$, there are two instances of "negative moderate and low" level (i.e. $c = 1; 1.25$, where: $\rho_1 \approx -0.4$ and $\rho_2 \approx 0.15; 0.25$), and one "negative moderate and nearly null" level ($c = 1.5$, where: $\rho_1 \approx -0.4$ and $\rho_2 \approx 0.1$). For $\omega = 0.8$, there are: one "negative high and high" level ($c = 1$, where: $\rho_1 \approx -0.7$ and $\rho_2 \approx 0.6$), one "negative high and moderate" level ($c = 1.25$, where: $\rho_1 \approx -0.7$ and $\rho_2 \approx 0.4$), and one "negative high and low" level ($c = 1.5$, where: $\rho_1 \approx -0.7$ and $\rho_2 \approx 0.2$).

In conclusion, by combining together the considered number n of units and p of variables, percentages of missing values (generated through a MCAR mechanism), and the input parameters related to the *MEP* or *MSN* distributions, we examined a total number of artificial scenarios equal to: 162, in the case of the SyKu shape (*MEP* data), and: 216 (1st study)+ 144 (2nd study) + 216 (3rd study) = 576, in the case of the SK shape (*MSN* data).

It is worth remarking that imputation under the SyKu shape was initially tested in the presence of the ECor structure only. Supplementary simulations also considering the PNCor and the UnbCor structures were subsequently performed with additional simulation settings suggested by the results obtained under the SK shape. The reason is that *MSN*

data with $\alpha = 1$, accounting for situations of slighter skewness, are then quite close to $\kappa = 2$ of the *MEP* family. Their results were therefore expected to give useful indications on how further inspecting patterns under the SyKu shape. In addiction, a modified version of the 3$^{\text{rd}}$ study was also introduced for both the SyKu and the SK shapes to examine the role of the sign of the correlation coefficients concerning the first component.

Simulation procedure. Under each scenario a complete $n \times p$ data matrix $\mathbf{X}^*$ $(n > p)$ was first generated according to a specific *MEP* or *MSN* distribution. Then, 1,000 incomplete matrices $\mathbf{X}_t$ were derived from it by deleting a 5%, 10%, or 20% percentage, respectively, of MCAR values ($t = 1, \ldots,$ 1,000). Subsequently, *missForest*, *IPCA*, *FIP*, and *FIM* were applied to each $\mathbf{X}_t$ with the following options. As for *missForest*, the maximum number of iterations was increased from 10 (the default value in the R library "missForest" by Stekhoven and Bühlmann (2012)) to 50. Regarding *IPCA*, we used the function "imputePCA" in the R library "missMDA" with the default "Regularized method" (Josse et al., 2011), the maximum number of iterations fixed at 5,000, and the number of extracted PCs set at the largest possible value, i.e. $p - 2$ $(p \geq 3)$. Since for *FIM* (and *FIP*) it is necessary to make a choice for the number of donors (and the metric) the simulation study was carried out this way. In the case of *FIP*, PCs were extracted from the variance-covariance matrix (i.e. option "cor=False", see Remark 3, Sect. 2.1 in Solaro et al. (2015)), and donors' detection was carried out with, respectively, city-block, Euclidean, and Lagrange distance ($r = 1; 2; \infty$, respectively, in formula (3), Solaro et al. (2015)). Furthermore, for both *FIP* and *FIM* we considered four different quantiles q of distances, i.e. $q = 0.05; 0.1; 0.15;$ 0.2. Therefore, to each incomplete $\mathbf{X}_t$ the method *FIP* was applied in 12 variants (given by the combinations of the three distances with the four quantiles), while *FIM* in 4 variants (given by the four quantiles).

Summary of simulation results. The performance of the considered methods was evaluated and compared by means of the Relative Mean Square Error (RMSE):

$$_{\mathsf{m}}\text{RMSE}_t = \sum_{j=1}^{p} \frac{1}{n\sigma_j^2} (\mathbf{x}_j^* - {}_{\mathsf{m}}\tilde{\mathbf{x}}_{j,t})^t (\mathbf{x}_j^* - {}_{\mathsf{m}}\tilde{\mathbf{x}}_{j,t}), \tag{1}$$

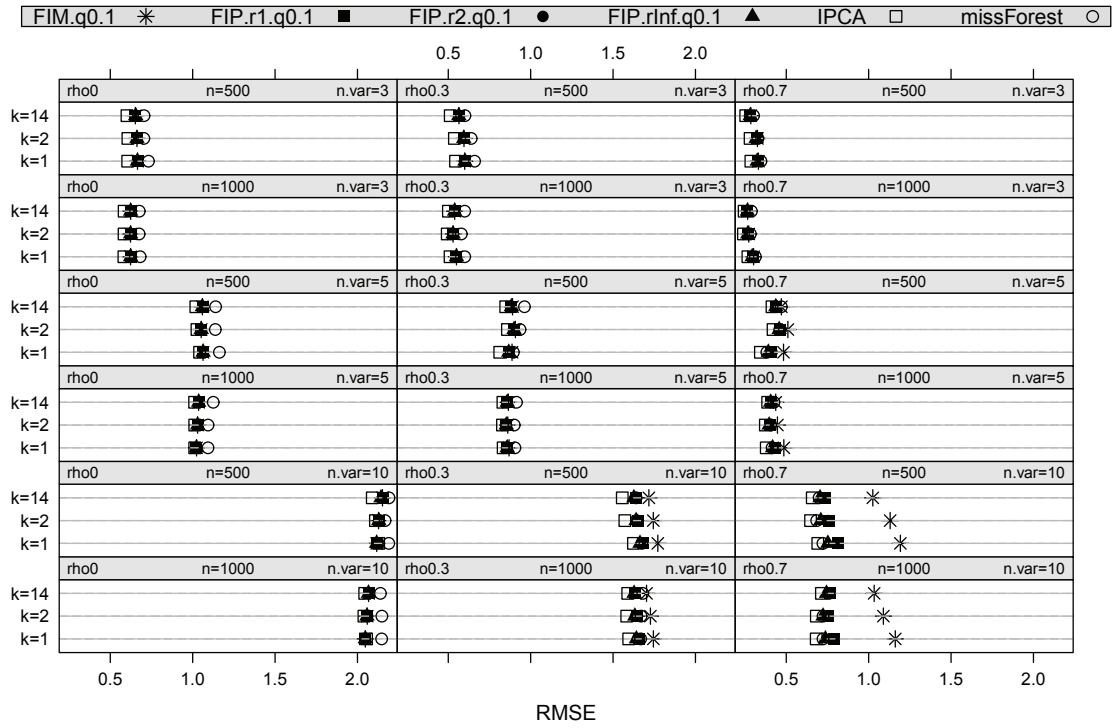
$(t = 1, \ldots, 1000)$, where $\mathbf{x}_j^*$ is the j-th column vector of the complete matrix $\mathbf{X}^*$, ${}_{\mathsf{m}}\tilde{\mathbf{x}}_{j,t}$ is the j-th column vector of the matrix ${}_{\mathsf{m}}\tilde{\mathbf{X}}_t$ imputed with method m at the t-th simulation run, and σ_j^2 is the variance of the j-th variable in $\mathbf{X}^*$.

After that, we carried out descriptive and inferential analyses of RMSE values. Descriptive analyses were performed by computing usual synthesis measures (mean, standard deviation, and quartiles), and displaying results in graphical form, i.e. dotplots of RMSE median values along with boxplots of RMSE distributions, in order to ease the comparison. With the specific purpose of an inferential analysis, the Jonckheere-Terpstra (J-T) test (Hollander and Wolfe, 1999) was used to compare the RMSE distributions of the examined methods obtained under a same simulation scenario against different ordered alternative hypotheses, and test if these methods produce significantly different results. These are explained case by case in the next Subsect. 2.2.

2.2 Simulation results

Simulation results here presented concern the scenarios with 20% of missing values, since they better emphasize differences among the four methods *missForest*, *IPCA*, *FIM*, and *FIP* (Solaro et al., 2014). An ample collection of the omitted results can be found in the

Figure 1: SyKu-ECor pattern (*MEP* data) – Dotplots of RMSE median values of *missForest*, *IPCA*, *FIM* and *FIP* with $q = 0.1$ donor quantile, and 20% of missing values

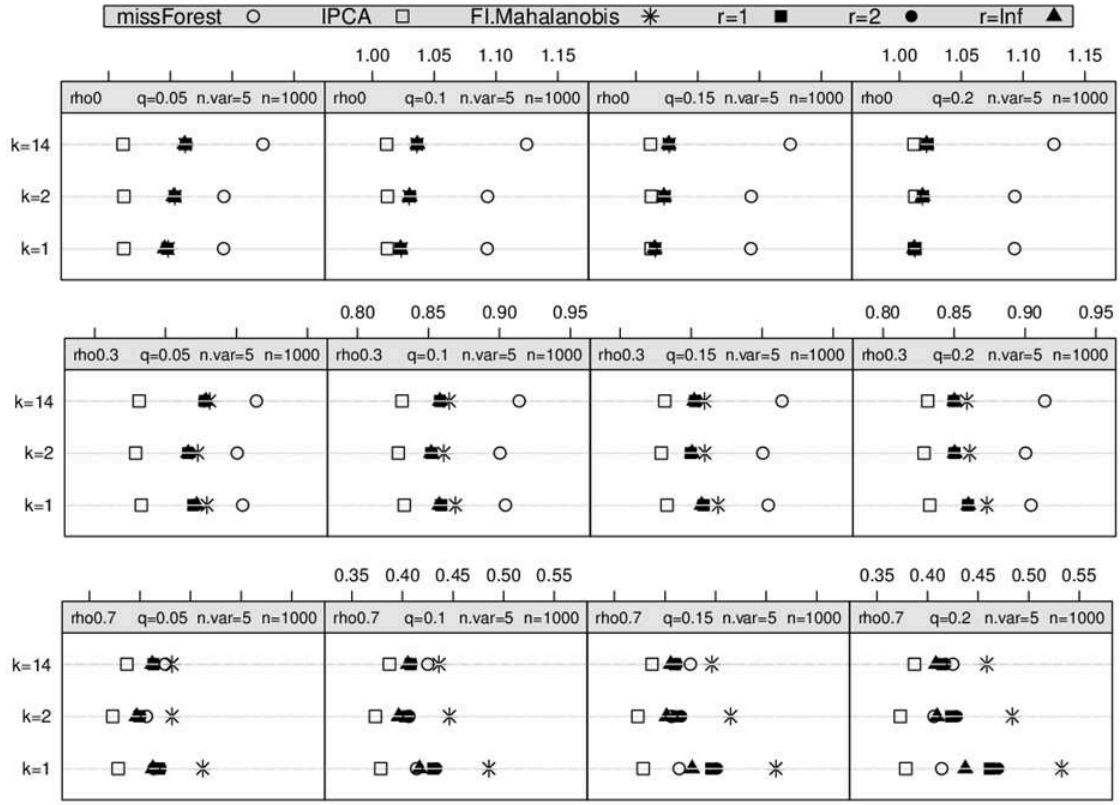


electronic supplementary material (ESM). Results are here displayed through dotplots of RMSE median values. Inspection of boxplots of the various RMSE distributions (in the ESM) revealed in fact that the median is highly representative for all the methods, since imputation errors are symmetrically distributed around it with a similar extent of dispersion (i.e. the boxes have a similar length). Moreover, the graphs here displayed mostly regard the case of $p = 5$ variables and $n = 1000$ units. Besides being common to all the studies, such a combination of experimental conditions well represents the overall observed trends.

Descriptive analysis. A few general considerations are worth making before examining results from each data pattern in detail. As expected, regarding the exogenous factors "dimensionality" and "correlation", the four methods share the fact that, on the whole, their RMSE tends to increase with the dimensionality of data, especially the number of variables, and to decrease as the value of (both input and output) correlation parameters increase. This is consistent with the fact that if variables are medium/highly correlated then imputation is generally subject to smaller errors. Figure 1, which concerns results obtained under the SyKu shape with the ECor structure (SyKu-ECor pattern), displays an instance of such a trend. In addiction, the most important factor that seems to discriminate, on the whole, between a "good" and a "less good" imputation method reveals to be the type of correlation structure along with the magnitude of correlation coefficients. As it will be seen soon, their impact can get even stronger if data are skew.

Regarding the endogenous factors, proper of *FIM* and *FIP*, what can be observed in general is that if correlations are low then selecting a higher proportion of donors (e.g. $q = 0.2$) leads to smaller errors, while if correlations are high, having few donors (e.g.

Figure 2: SyKu-ECor pattern – Dotplots of RMSE median values of *missForest*, *IPCA*, *FIM* and *FIP* with $q = 0.05, 0.1, 0.15, 0.2$ donor quantiles, 20% of missing values, $p = 5$ variables, and $n = 1000$ units



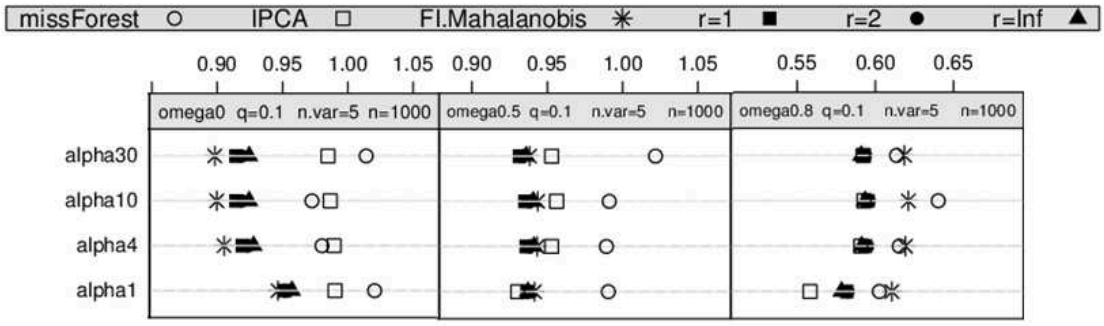
$q = 0.05$) implies better results. As for the choice of the distance in *FIP*, if correlations are low, then in most of the scenarios the city-block distance ($r = 1$) turns out to be an optimal choice, while if correlations are high, Lagrange distance ($r = \infty$) generally leads to better results.

To contain the exposition as much as possible, the analysis of the endogenous factors is exclusively carried out from an inferential point of view, and is reported later on. In what follows, attention will be paid to the main peculiarities found in each data pattern.

► *The SyKu-ECor pattern* (symmetry with equicorrelations, derived from *MEP* data, Table 3). As it can be clearly noticed in Figure 2, *IPCA* proves to be the best imputation method. *FIM* and *FIP* tend to have an overlapping and intermediate performance between *IPCA* and *missForest* for small ρs ($\rho = 0; 0.3$) and low data dimensionality (Figure 1). For high ρ and high dimensionality, their performance tend however to worsen, especially *FIM*, (see e.g. the last panel in Figure 1).

- ⊵ *Kurtosis effect. IPCA* proves to be less sensitive than the other methods to kurtosis of data distribution, especially when ρ is small. *missForest* tends to produce smaller imputation errors under normal data, while it seems more badly affected by platykurtic data. Regarding *FIM* and *FIP*, when $\rho = 0$ they tend to have smaller errors for leptokurtic data. As ρ increases, *FIP* tends to perform better under normally distributed data, and *FIM* under platykurtic data, while both the methods seem to perform worst in the presence of leptokurtic data, especially *FIM* for higher ρs (e.g. the last row of panels in Figure 2).

Figure 3: SK-ECor pattern (*MSN* data, 1st study) – Dotplots of RMSE median values of *missForest*, *IPCA*, *FIM* and *FIP* with $q = 0.1$ donor quantile, 20% of missing values, $p = 5$ variables, and $n = 1000$ units



→ *SK shape* (derived from *MSN* data). Interpretation of results in this case is complicated by the little intuitive relationship between input (matrix $\mathbf{\Omega}$) and output (matrix $\mathbf{R}$) correlation structures. Nonetheless, to simplify the graphical layout, dotplots preserve the indication of the input ω parameters, while results are mainly read by taking into account the output ρ parameters described in Subsect. 2.1 along with Tables 2, 4, and 5.

► *The SK-ECor pattern* (skewness with equicorrelations, 1st study, Table 3). Figure 3, related to $p = 5$, $n = 1000$, and $q = 0.1$ for donor quantile, well represents the trends observed in the other scenarios. The panels correspond to the three levels of output correlation described in Subsect. 2.1, i.e. low-negative ($\omega = 0$), low-positive ($\omega = 0.5$), and medium ($\omega = 0.8$) values of ρ (Table 2).

Three different trends can be noticed. Whichever the value of α is, in the presence of low-negative ρs ($\omega = 0$) *FIM* is the best imputation method, followed by *FIP*. *IPCA* and *missForest* have a clear bad performance (1st panel). In the case of low-positive ρs ($\omega = 0.5$), the points of *FIM* and *FIP* tend to overlap, although *FIP* is slightly better, thus denoting a similar performance. They are again the best imputation methods, with the only exception of $\alpha = 1$, where skewness is less strong (2nd panel). When ρ assumes medium values ($\omega = 0.8$), the trend becomes opposite. *FIM* gives the worst results (apart from an isolated point of *missForest* for $\alpha = 10$), while *FIP* performance is very similar to *IPCA*.

⊵ *Skewness effect.* By looking at the performance of *FIM* and *FIP* over the three panels in Figure 3, a sort of dichotomy between $\alpha = 1$ and $\alpha \geq 4$ seems apparent. This is particularly marked when $\omega = 0$, since in this case $\alpha = 1$ represents a situation much closer to the symmetry than the other values of ω, (see γ_1 and $\gamma_{1\text{MV}}$ in Table 2). Substantially, for low-negative ρs ($\omega = 0$), *FIM* and *FIP* perform better when data are more asymmetrically distributed ($\alpha \geq 4$), while in the presence of higher values of ρ ($\omega = 0.8$), they tend to perform better for data less asymmetrically distributed ($\alpha = 1$). The case of low-positive ρs ($\omega = 0.5$) represents an intermediate situation, since *FIM* and *FIP* seem less sensitive to skewness. *missForest* also proves to be fairly sensitive to variations of α, but the effect is evidently not monotone, especially for $\omega = 0$. On the contrary, *IPCA* looks less sensitive than the other methods to variations of parameter α.

► *The SK-PNCor pattern* (skewness with positive and negative correlations, 2nd study, Table 3). An extract of the results concerning the second study (Table 3) is provided in

Figure 4: SK-PNCor pattern (*MSN* data, 2nd study) – Dotplots of RMSE median values of *missForest*, *IPCA*, *FIM* and *FIP* with $q = 0.1$ donor quantile, 20% of missing values, $p = 5$ variables, and $n = 1000$ units

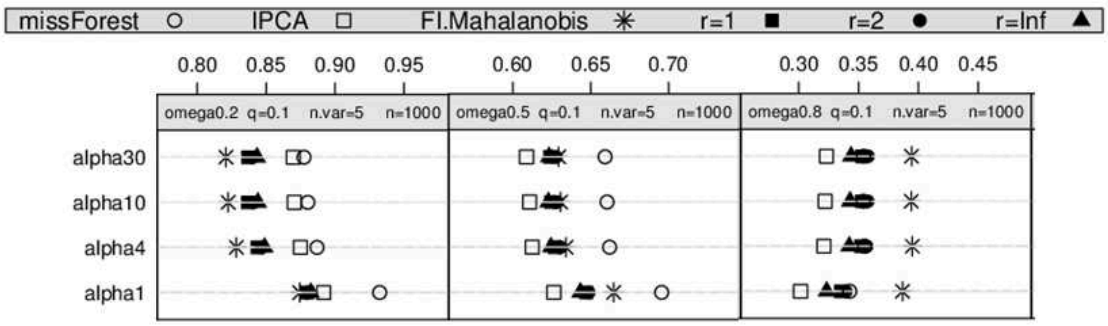


Figure 4, with $p = 5$ variables, $n = 1000$ units, and $q = 0.1$ for *FIM* and *FIP*. This plot well represents the other scenarios in this study. The three panels concern the three levels of output correlations, given by three distinct values of ρ described in Subsect. 2.1 along with Table 4, i.e. positive-negative (PN) low ($\omega = 0.2$), PN moderate ($\omega = 0.5$), and PN high ($\omega = 0.8$) values of ρ.

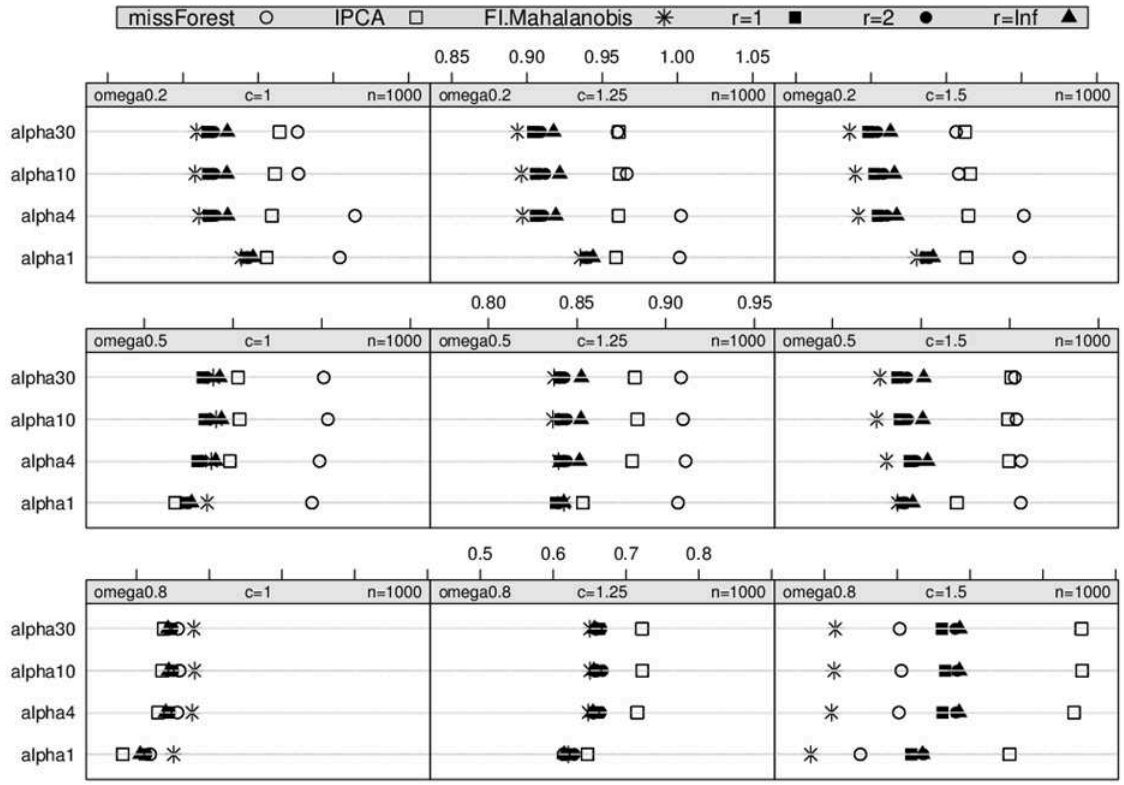
Most remarks would be very similar to the above described SK-ECor pattern. We limit therefore to point out that: (a) in the presence of PN low ρs ($\omega = 0.2$), *FIM* is the best method, and *missForest* the worst. On the contrary, with higher PN ρs ($\omega = 0.5; 0.8$) *IPCA* is the best method, while *FIM* turns out to be the worst for higher PN ρs ($\omega = 0.8$); (b) *FIP* always shows intermediate performances between the best and the worst methods; (c) *missForest* tends to improve its performance with the increasing of correlation levels.

- $\trianglerighteq$ *Skewness.* Again, it can be observed the dichotomy: $\alpha = 1$ vs. $\alpha \geq 4$ just pointed out in the previous study (Figure 3). This is common, more or less, to all the methods, so it can be summed up as follows: (1) in the presence of PN low or moderate ρs ($\omega = 0.2; 0.5$), all the methods tend to produce smaller errors for more skew data ($\alpha \geq 4$); (2) in the presence of PN high ρs ($\omega = 0.8$), they tend to perform better for less skew data ($\alpha = 1$).

▶ *The SK-UnbCor pattern* (skewness with unbalanced correlations, 3rd study, Table 3). As regards the third study (Table 3), Figure 5 shows the results obtained for $p = 5$ and $n = 1000$, with $q = 0.1$ for *FIM* and *FIP*. The full description of input-output correlations is provided in Subsect. 2.1 along with Table 5.

Overall, it is worth pointing out that: (a) In the presence of the "negative low and nearly null" level ($\omega = 0.2$), *FIM* confirms to be the best imputation method, followed by *FIP* (Figure 5, 1st row of panels). *IPCA* and *missForest* have a worse performance; (b) in the presence of moderate values of ρ for the first variable ($\omega = 0.5$, 2nd row of panels), *FIP* proves mostly to be the best method, (with few exceptions given by: $\alpha = 1$, $c = 1$, where *IPCA* is better, and: $\alpha \geq 4$, $c = 1.5$, where the best is *FIM*); (c) the case in which the first variable has high ρs ($\omega = 0.8$, last row of panels) highlights the main differences among the methods. In particular, there is an inversion of trend moving from the "negative high and high" level ($c = 1$, 1st column, last row), where *IPCA* is the best method and *FIM* the worst, to the "negative high and low" level ($c = 1.5$, last column, last row), where *FIM* is the best method and *IPCA* the worst. The panel related to the "negative high and moderate" level ($c = 1.25$; 2nd column, last row) displays an intermediate situation, where

Figure 5: SK-UnbCor pattern (*MSN* data, 3rd study) – Dotplots of RMSE median values of *missForest*, *IPCA*, *FIM* and *FIP* with $q = 0.1$ donor quantile, 20% of missing values, $p = 5$ variables, and $n = 1000$ units



missForest, *FIM* and *FIP* have a very similar good performance, while *IPCA* is the worst method.

- *Skewness.* Once again, the dichotomy $\alpha = 1$ vs. $\alpha \geq 4$ is clearly visible. In particular, when $\omega = 0.2$, *missForest*, *FIM*, and *FIP* perform better for more skew data ($\alpha \geq 4$) while *IPCA* seems less sensitive to variations of α. On the contrary, when $\omega = 0.8$ all the methods tend to perform better for less skew data ($\alpha = 1$). The case $\omega = 0.5$ is intermediate, showing both the trends, in particular the first for $c = 1.5$ (better for more skew), and the second for $c = 1$ (better for less skew).

Inferential analysis. J-T test is applied (in all cases at the significance level of 0.05) to verify whether assumed conjectures about specific aspects find support in the simulation results. Given the copious number of observations (1,000 values in all) for each RMSE distribution, J-T test has been applied with the asymptotic normal distribution as test statistic. Once again, results here provided refer to the case of the strongest missingness we considered, i.e. the 20% of missing values.

One of the main aspects of concern was testing the effects of the endogenous factors, i.e. (a) donor quantile, and (b) distance in selecting donors, on imputation performance of *FIM* ((a) only) and *FIP* (both (a) and (b)). Another major aspect of concern regards: (c) comparisons among the performances of *IPCA*, *FIM*, and *FIP*. We have decided to discard *missForest* from this analysis because *IPCA* has proved to perform better than *missForest* in almost all the considered scenarios. We have then focused on comparisons between the best method within those two, and *FIM* and *FIP*.

(a) Effect of donor quantile. Two separate one-sided J-T tests have been carried out to test the null hypothesis:

$$H_0 : F_{q=0.05}(x) = F_{q=0.1}(x) = F_{q=0.15}(x) = F_{q=0.2}(x) \tag{2}$$

against the two ordered alternatives:

$$H_1' : F_{q=0.05}(x) \leq F_{q=0.1}(x) \leq F_{q=0.15}(x) \leq F_{q=0.2}(x) \tag{3}$$

and:

$$H_1'' : F_{q=0.05}(x) \geq F_{q=0.1}(x) \geq F_{q=0.15}(x) \geq F_{q=0.2}(x) \tag{4}$$

with at least a strict inequality in the ordered alternatives (3)-(4), where $F_{q=q*}(\cdot)$ denotes the distribution function of RMSE obtained with *FIP* or *FIM* run at a fixed $q = q*$ quantile of donors.

It is worth noting that rejecting the null hypothesis (2) in favour of the alternative (3) gives empirical support to the fact that $q = 0.05$ produces better results than $q = 0.2$, but it does not exclude that $q = 0.1$ (or also $q = 0.15$) could lead to results as good as $q = 0.05$ (analogously for the alternative (4)). Nonetheless, we are mainly interested in appraising differences in performance between extreme values of donor quantiles, rather than intermediate quantiles.

Test results are provided separately for *FIM* and *FIP* (with the Euclidean distance) in Tables 6–13. Given the presence of similar trends between the two methods, results can be summarized as follows:

(1) *SyKu-ECor pattern* (Tables 6 and 7). Overall, $q = 0.2$ proves to be better than $q = 0.05$ for $\rho = 0$ and $\rho = 0.3$. There are only few exceptions when $\rho = 0.3$, ($p = 10$, $n = 1000$, $\kappa = 1$ – for both *FIM* and *FIP* –, and $\kappa = 2$ – for *FIM* only –, in which $q = 0.05$ is the best choice). On the other hand, when $\rho = 0.7$, $q = 0.05$ produces smaller errors than $q = 0.2$ (with few exceptions when $\kappa = 14$, where $q = 0.2$ is better for both *FIM* and *FIP*).

(2) *SK-ECor pattern* (Tables 8 and 9). Apart from few exceptions, when $p = 3$, $q = 0.2$ produces better results in both *FIM* and *FIP*, regardless of correlation strength. When $p = 5; 10$, *FIM* performs better with $q = 0.2$ for low ρs ($\omega = 0; 0.5$), and with $q = 0.05$ for higher values of ρ ($\omega = 0.8$). *FIP* has a similar performance with few exceptions (i.e. when $\omega = 0.8$ with $p = 5; 10$ variables and $n = 500$ units, where $q = 0.2$ leads to better results than $q = 0.05$).

(3) *SK-PNCor pattern* (Tables 10 and 11). *FIM* requires more donors ($q = 0.2$) in the presence of PN low ρs ($\omega = 0.2$) with a smaller number of units ($n = 500$) for each α, or with more units ($n = 1000$) for less skew data ($\alpha = 1$). In all the other considered scenarios (i.e. $\omega = 0.5; 0.8$), *FIM* requires smaller proportions of donors ($q = 0.05$). *FIP* shares the same trend of *FIM*, except in the presence of PN moderate ρs ($\omega = 0.5$) with $n = 500$ units, where $q = 0.2$ seems more effective.

(4) *SK-UnbCor pattern* (Tables 12 and 13). *FIM* and *FIP* perform very similarly. In the case of "negative low and nearly null" ρs ($\omega = 0.2$ and $c = 1; 1.25; 1.5$), $q = 0.2$ is to be preferred, while for high ρs ($\omega = 0.8$ and $c = 1; 1.25; 1.5$), $q = 0.05$ gives rise to better results. In the presence of "negative moderate and low or nearly null" ρs ($\omega = 0.5$ and $c = 1; 1.25; 1.5$), it is the number of units that makes the difference. If $n = 500$, more donors are necessary ($q = 0.2$), while if $n = 1000$, fewer donors ($q = 0.05$) are better (with only few exceptions in the case of *FIP*).

Table 6: SyKu-ECor pattern. J-T test for the effect of donor quantile in *FIM*

		$p=3$		$p=5$		$p=10$	
	n	500	1000	500	1000	500	1000
$\rho=0$	$\kappa=1,2,14$	II	II	II	II	II	II
$\rho=0.3$	$\kappa=1,2$	II	II	II	II	II	I
	$\kappa=14$	II	II	II	II	II	II
$\rho=0.7$	$\kappa=1,2$	I	I	I	I	I	I
	$\kappa=14$	II	II	I	I	I	I

Table 7: SyKu-ECor pattern. J-T test for the effect of donor quantile in *FIP* (with $r=2$)

		$p=3$		$p=5$		$p=10$	
	n	500	1000	500	1000	500	1000
$\rho=0$	$\kappa=1,2,14$	II	II	II	II	II	II
$\rho=0.3$	$\kappa=1$	II	II	II	II	II	I
	$\kappa=2,14$	II	II	II	II	II	II
$\rho=0.7$	$\kappa=1,2$	ns	I	I	I	I	I
	$\kappa=14$	II	II	II	I	I	I

Table 8: SK-ECor pattern. J-T test for the effect of donor quantile in *FIM*

		$p=3$		$p=5$		$p=10$	
	n	500	1000	500	1000	500	1000
$\omega=0$	$\alpha=1,4,10,30$	II	II	II	II	II	II
$\omega=0.5$	$\alpha=1,4,10,30$	II	II	II	II	II	II
$\omega=0.8$	$\alpha=1,4$	II	ns	I	I	I	I
	$\alpha=10$	II	I	I	I	I	I
	$\alpha=30$	II	ns	I	I	I	I

Table 9: SK-ECor pattern. J-T test for the effect of donor quantile in *FIP* (with $r=2$)

		$p=3$		$p=5$		$p=10$	
	n	500	1000	500	1000	500	1000
$\omega=0$	$\alpha=1,4,10,30$	II	II	II	II	II	II
$\omega=0.5$	$\alpha=1,4,10,30$	II	II	II	II	II	II
$\omega=0.8$	$\alpha=1,4,10,30$	II	II	II	I	II	I

Table 10: SK-PNCor pattern. J-T test for the effect of donor quantile in *FIM*

		$p=5$		$p=10$	
	n	500	1000	500	1000
$\omega=0.2$	$\alpha=1$	II	II	II	II
	$\alpha=4$	II	ns	II	ns
	$\alpha=10,30$	II	I	II	ns
$\omega=0.5$	$\alpha=1$	II	I	I	I
	$\alpha=4,10,30$	I	I	I	I
$\omega=0.8$	$\alpha=1,4,10,30$	I	I	I	I

Legend in Tables 6–10. I: $q=0.05 \leq q=0.1 \leq q=0.15 \leq q=0.2$, and
II: $q=0.05 \geq q=0.1 \geq q=0.15 \geq q=0.2$ (at least a strict inequality); ns: not significant.

Table 11: SK-PNCor pattern. J-T test for the effect of donor quantile in *FIP* (with $r = 2$)

		$p = 5$		$p = 10$	
	n	500	1000	500	1000
$\omega = 0.2$	$\alpha = 1$	II	II	II	II
	$\alpha = 4, 10, 30$	II	I	II	II
$\omega = 0.5$	$\alpha = 1, 4, 10, 30$	II	I	II	I
$\omega = 0.8$	$\alpha = 1, 4, 10, 30$	I	I	I	I

Table 12: SK-UnbCor pattern. J-T test for the effect of donor quantile in *FIM*

		$p = 5$					
		$n = 500$			$n = 1000$		
	c	1	1.25	1.5	1	1.25	1.5
$\omega = 0.2$	$\alpha = 1, 4, 10, 30$	II	II	II	II	II	II
$\omega = 0.5$	$\alpha = 1$	II	II	II	ns	ns	ns
	$\alpha = 4, 10, 30$	II	II	II	I	I	I
$\omega = 0.8$	$\alpha = 1, 4, 10, 30$	I	I	I	I	I	I

Table 13: SK-UnbCor pattern. J-T test for the effect of donor quantile in *FIP* (with $r = 2$)

		$p = 5$					
		$n = 500$			$n = 1000$		
	c	1	1.25	1.5	1	1.25	1.5
$\omega = 0.2$	$\alpha = 1, 4, 10, 30$	II	II	II	II	II	II
$\omega = 0.5$	$\alpha = 1$	II	II	II	II	II	I
	$\alpha = 4, 10, 30$	II	II	II	I	I	I
$\omega = 0.8$	$\alpha = 1, 4, 10, 30$	I	I	I	I	I	I

Legend in Tables 11–13. I: $q = 0.05 \leq q = 0.1 \leq q = 0.15 \leq q = 0.2$, and
II: $q = 0.05 \geq q = 0.1 \geq q = 0.15 \geq q = 0.2$ (at least a strict inequality); ns: not significant.

(b) Effect of distance. Once again, two separate one-sided J-T tests have been carried out to test the null hypothesis:

$$H_0 : F_{r=1}(x) = F_{r=2}(x) = F_{r=\infty}(x) \tag{5}$$

against the two ordered alternatives:

$$H_1' : F_{r=1}(x) \leq F_{r=2}(x) \leq F_{r=\infty}(x) \tag{6}$$

and:

$$H_1'' : F_{r=1}(x) \geq F_{r=2}(x) \geq F_{r=\infty}(x) \tag{7}$$

(with at least a strict inequality), where $F_{r=r*}(\cdot)$ denotes the distribution function of RMSE obtained from *FIP* with distance $r = r*$.

As before, it is worth noting that if the null hypothesis (5) is rejected in favour of the alternative (6), thus judging the city-block distance as better than Lagrange, nothing excludes that the Euclidean distance can be as good as the city-block (an analogous remark holds for the alternative (7)). While reading the J-T test results, this point has then to be taken into account.

Results are displayed in Tables 14 to 17. Specifically,

(1) *SyKu-ECor pattern* (Table 14). The choice of the metric is significant essentially for $\rho = 0.7$ (apart from $p = 3$ with $\kappa = 14$), and $p = 10$ (excepted $\rho = 0$ with $\kappa = 1; 2$), in which cases the Lagrange distance leads to better results than city-block. The choice thus seems mostly tied to the presence of a strong linear relationship between variables and/or a high dimensionality of data.

(2) *SK-ECor pattern* (Table 15). With the exception of $p = 3$, for low-negative ρs ($\omega = 0$) the best metric turns out the city-block distance, while for moderate ρs ($\omega = 0.8$) it is Lagrange. The case of low-positive ρs ($\omega = 0.5$) gives rise to less well-framed results.

(3) *SK-PNCor pattern* (Table 16). Most of the results concerning PN low ρs ($\omega = 0.2$) are characterized by the city-block distance as the best one. For PN moderate and PN high ρs ($\omega = 0.5; 0.8$), the Lagrange distance proves to be the best choice (with the exception of $p = 5$, $n = 500$, and $p = 5$, $n = 1000$, $\alpha = 30$, where test results are not significant).

(4) *SK-UnbCor pattern* (Table 17). Apart from several not significant results occurring for less skew data ($\alpha = 1$) with low and/or moderate ρs ($\omega = 0.2; 0.5$) and a smaller number of units ($n = 500$), most of the times the city-block distance is the best metric, while Lagrange should be preferred for higher values of ρ ($\omega = 0.8$ and $c = 1; 1.25$).

(c) Comparison among IPCA, FIM, and FIP. In order to detect the best method among the three under the various scenarios, six separate one-sided J-T tests have been applied at the 0.05 significance level to test the null hypothesis:

$$H_0 : F_{\mathsf{m}=FIM}(x) = F_{\mathsf{m}=FIP}(x) = F_{\mathsf{m}=IPCA}(x), \tag{8}$$

where $F_{\mathsf{m}=\mathsf{m}*}(\cdot)$ in (8) denotes the empirical distribution function of RMSE of method $\mathsf{m}*$, against each of the following six ordered alternatives:

$$\begin{aligned} 1 &= FIM < FIP < IPCA, \\ 2 &= FIP < FIM < IPCA, \\ 3 &= IPCA < FIP < FIM, \\ 4 &= FIM < IPCA < FIP, \\ 5 &= FIP < IPCA < FIM, \\ 6 &= IPCA < FIM < FIP. \end{aligned} \tag{9}$$

Comparisons of *FIM* and *FIP* with *IPCA* are performed twice. The first time, by considering *FIM* and *FIP* at their "intermediate" options (donor quantile $q = 0.1$ – *FIM* and *FIP* –, and Euclidean distance – $r = 2$, only for *FIP* –), which are regarded as possible candidates to be the default options. The second time, at their best endogenous factor levels detected by the previous J-T test analyses (a) and (b).

For each scenario the best method is then judged as the one that, in case of a significant result, has associated the smallest *p-value*, or, equivalently, the highest absolute value on the reference asymptotic normal distribution.

Results are displayed in Tables 18 to 21. In case of a significant result, the number of the "most significant" ranking is given according to the numbering of system (9). Moreover, the cells are differently coloured depending on which method appears as the first in the

Table 14: SyK-ECor pattern. J-T test for the effect of metrics in *FIP* (with $q = 0.1$)

		$p=3$		$p=5$		$p=10$	
	n	500	1000	500	1000	500	1000
$\rho=0$	$\kappa=1,2$	ns	ns	ns	ns	ns	ns
	$\kappa=14$	ns	ns	ns	ns	B	B
$\rho=0.3$	$\kappa=1,2,14$	ns	ns	ns	ns	B	B
$\rho=0.7$	$\kappa=1,2$	B	B	B	B	B	B
	$\kappa=14$	ns	ns	B	B	B	B

Table 15: SK-ECor pattern. J-T test for the effect of metrics in *FIP* (with $q = 0.1$)

		$p=3$		$p=5$		$p=10$	
	n	500	1000	500	1000	500	1000
$\omega=0$	$\alpha=1$	ns	ns	ns	A	A	A
	$\alpha=4,10,30$	ns	ns	A	A	A	A
$\omega=0.5$	$\alpha=1$	ns	ns	ns	ns	B	B
	$\alpha=4,10,30$	ns	ns	ns	A	ns	ns
$\omega=0.8$	$\alpha=1$	ns	ns	B	B	B	B
	$\alpha=4,10,30$	ns	ns	ns	B	B	B

Table 16: SK-PNCor pattern. J-T test for the effect of metrics in *FIP* (with $q = 0.1$)

		$p=5$		$p=10$	
	n	500	1000	500	1000
$\omega=0.2$	$\alpha=1$	ns	ns	ns	A
	$\alpha=4,10,30$	A	A	ns	A
$\omega=0.5$	$\alpha=1,4,10$	ns	B	B	B
	$\alpha=30$	ns	ns	B	B
$\omega=0.8$	$\alpha=1,4,10,30$	B	B	B	B

Table 17: SK-UnbCor pattern. J-T test for the effect of metrics in *FIP* (with $q = 0.1$)

		$p=5$					
		$n=500$			$n=1000$		
	c	1	1.25	1.5	1	1.25	1.5
$\omega=0.2$	$\alpha=1$	ns	ns	ns	A	A	A
	$\alpha=4,10,30$	A	A	A	A	A	A
$\omega=0.5$	$\alpha=1$	ns	ns	ns	A	A	A
	$\alpha=4,10,30$	A	A	A	A	A	A
$\omega=0.8$	$\alpha=1,4,10,30$	B	B	A	B	B	A

Legend in Tables 14–17. A: $r = 1 \leq r = 2 \leq r = \infty$, and B: $r = \infty \leq r = 2 \leq r = 1$ (at least a strict inequality); ns: not significant.

significant ranking. Grey cells denote rankings 1 and 4, where *FIM* is the best. Light-grey cells denote ranking 2 and 5, where *FIP* is the best. A blank background in the cells denotes rankings 3 and 6, with *IPCA* as the best. Comparisons there reported are made with *FIM* and *FIP* applied at their default options. Some further remarks about *FIM* and *FIP* tested at their best endogenous factor levels (not shown in the tables) are given in the text if they produce significant changes among the rankings in (9).

Regarding the main achieved results:

(1) *SyKu-ECor pattern* (Table 18). With only two exceptions when $\rho = 0$, *IPCA* is always the best imputation method. In addition, the prevalence of ranking 3, instead of ranking 6, reveals that *FIM* more frequently performs worse than the other two methods. Switching the options of *FIM* and *FIP* to their best endogenous factor levels does not produce compelling results, apart from the scenarios with $\rho = 0$ and $p = 10$, where with $q = 0.2$ and $r = \infty$, *FIP* performs better than *FIM* and *IPCA*, in the order.

(2) *SK-ECor pattern* (1^{st} study, Table 19). Clear separated trends can be read. For low-negative ρs ($\omega = 0$) the best method is *FIM*, followed by *FIP* (ranking 1). As for the endogenous factors, the three cases in which ranking 6 (*IPCA* the best) prevails for $\alpha = 1$ turn, respectively: (i) to not significant differences when $p = 3$, if $q = 0.2$ (and $r = 2$) is fixed for *FIM* (and *FIP*); (ii) to *FIM* as the best method when $p = 5$ and $n = 500$, if $q = 0.2$ is considered. For low-positive ρs ($\omega = 0.5$), *IPCA* has the best performance for $p = 3$, but with a higher number of variables and more skew data ($\alpha \geq 4$) *FIP* shows a good performance. In particular, by fixing the endogenous factors at their best levels, *FIP* overcomes *IPCA* by using more donors ($q = 0.2$) along with the Euclidean distance ($r = 2$) when $p = 5$, or the Lagrange distance ($r = \infty$) when $p = 10$ and $n = 500$. However, when $p = 10$ with $n = 1000$, *IPCA* proves again to be the best method, although the observed previous trend seems to suggest that in the presence of a wider proportion of donors ($q > 0.2$) *FIP* could further improve its performance. Finally, for higher values of ρs ($\omega = 0.8$) *IPCA* is always the best method.

(3) *SK-PNCor pattern* (2^{nd} study, Table 20). In all the considered scenarios with PN moderate or PN high values of ρ ($\omega = 0.5; 0.8$), *IPCA* perform better than *FIM* and *FIP*. For low ρs ($\omega = 0.2$) with $p = 5$, *FIM* is the best method, followed by *FIP*, whereas for $p = 10$, *IPCA* proves to have better performances. Anyway, as before, the trend observed for $p = 5$ seems to suggest that a wider proportion of donors could lead to improve the performance of both *FIM* and *FIP*.

(4) *SK-UnbCor pattern* (3^{rd} study, Table 21). This is the study where our methods exhibit the best results, especially *FIM*, which performs better than the others in almost all the considered scenarios. *IPCA* works better with balanced moderate or high ρs ($c = 1$ with $\omega = 0.5; 0.8$), with the only exceptions of $n = 1000$, $\alpha \geq 4$, and $\omega = 0.5$, where *FIP*, followed by *FIM*, is better than *IPCA*.

Supplementary simulations. As before mentioned, a supplementary study was undertaken in order to examine additional scenarios, not comprised among those in Table 3 and formulated *a posteriori* by taking into account the indications provided by the exploratory study. The SyKu shape was also considered in the presence of negative equicorrelations, with magnitude similar to the 1^{st} study of the SK-ECor pattern. Moreover, we also introduced the SyKu-PNCor and SyKu-UnbCor patterns, with correlation matrices $\mathbf{R}$ having entries of magnitude similar to the output $\mathbf{R}$ of the SK-PNCor and SK-UnbCor patterns, respectively (Table 3). Finally, the UnbCor structure was also set up by inserting all positive correlations. Results are here omitted (a part of them is given in ESM), but the main indications can be summed up as follows:

- *SyKu-ECor, negative equicorrelations.* To have consistent matrices $\mathbf{R}$, such that they were positive-definite, ρ could not be less than nearly -0.2. In all these additional scenarios, *FIM* always proves to be the best method, followed by *FIP*.

Table 18: SyKu-ECor pattern. J-T test for the best imputation method

		$p=3$		$p=5$		$p=10$	
	n	500	1000	500	1000	500	1000
$\rho=0$	$\kappa=1$	3	6	6	3	1	2
	$\kappa=2$	3	3	6	3	3	6
	$\kappa=14$	6	6	3	6	3	6
$\rho=0.3$	$\kappa=1$	6	3	3	3	3	3
	$\kappa=2$	3	3	3	3	3	3
	$\kappa=14$	3	6	3	3	3	3
$\rho=0.7$	$\kappa=1,2,14$	3	3	3	3	3	3

Table 19: SK-ECor pattern. J-T test for the best imputation method

		$p=3$		$p=5$		$p=10$	
	n	500	1000	500	1000	500	1000
$\omega=0$	$\alpha=1$	6	6	6	1	1	1
	$\alpha=4,10,30$	1	1	1	1	1	1
$\omega=0.5$	$\alpha=1$	3	3	3	3	3	3
	$\alpha=4,10,30$	3	3	5	2	3	3
$\omega=0.8$	$\alpha=1,4,10,30$	3	3	3	3	3	3

Table 20: SK-PNCor pattern. J-T test for the best imputation method

		$p=5$		$p=10$	
	n	500	1000	500	1000
$\omega=0.2$	$\alpha=1$	3	1	3	3
	$\alpha=4,10$	4	1	3	6
	$\alpha=30$	1	1	3	6
$\omega=0.5$	$\alpha=1,4,10,30$	3	3	3	3
$\omega=0.8$	$\alpha=1,4,10,30$	3	3	3	3

Table 21: SK-UnbCor pattern. J-T test for the best imputation method

		$p=5$					
		$n=500$			$n=1000$		
	c	1	1.25	1.5	1	1.25	1.5
$\omega=0.2$	$\alpha=1$	6	6	6	1	1	1
	$\alpha=4,10,30$	4	1	1	1	1	1
$\omega=0.5$	$\alpha=1$	3	3	6	3	2	1
	$\alpha=4$	3	6	1	2	1	1
	$\alpha=10$	3	4	1	2	1	1
	$\alpha=30$	3	1	1	2	1	1
$\omega=0.8$	$\alpha=1,4,10,30$	3	1	1	3	1	1

- *SyKu-PNCor and SyKu-UnbCor patterns.* Results obtained are very similar to the SK-PNCor and SK-UnbCor patterns, although platykurtic data seem indicate a better performance of *FIM* and *FIP*.

- *SyKu-UnbCor and SK-UnbCor patterns with all positive correlations.* Results are very similar to those obtained under the corresponding patterns obtained with one variable negatively correlated with all the others. Therefore, it seems that it is the

extent of unbalancing among correlations plus the magnitude of absolute correlations, rather than the sign, to be the most important discriminant elements among the imputation methods.

3 Descriptive criteria for the choice of the imputation method

Simulations performed in this work confirm the idea that which the best imputation method is, it ultimately depends on the pattern of data. In our intentions, the experimental conditions in Table 3, along with those considered in the supplementary simulations, were chosen to interpret, as best as possible, data patterns that might be encountered in practice.

The main concern is then how a specific data pattern (for instance, one of the kinds considered in this work) could be recognized in a real dataset. In this regard, we have seen that strength and structure of correlations of variables, along with symmetric or skew nature of data distributions, are the elements to be considered in the choice of the "most suitable" imputation method.

After having valued the range of statistics known in the literature for summing up variance-covariance or correlation matrices in a scalar (see e.g. Seber (1984)), we have introduced the criteria defined in Table 22, called correlation indices:

1. *Eigenvalue-based indices*, given by the relative eigenvalues (RelEig), used for measuring the strength of correlations. In particular, if: $\mathbf{R} = \mathbf{I}_{(p)}$, then: $\text{RelEig}_s = \frac{1}{p}$ for all $s = 1, \ldots, p$. If: $\rho_{jl} = 1$ for all $j \neq l$, then: $\text{RelEig}_1 = 1$ and $\text{RelEig}_s = 0$ for all $s \geq 2$. The same occurs if $\mathbf{R}$ contains any $\rho_{jl} = -1$ in a consistent manner. Moreover, in the equicorrelation case (i.e. $\rho_{jl} = \rho$, $\forall j, l$), it was proved (see e.g. Kaiser (1968)) that $\mathbf{R}$ has a unique eigenvalue: $\lambda_* = 1 + (p-1)\rho$, and $(p-1)$ eigenvalues: $\lambda = 1 - \rho$ such that: $\lambda_{\max} = \lambda_* > \lambda$ (and then: $\text{RelEig}_* > \text{RelEig}$) if $\rho > 0$, while: $\lambda_* < \lambda = \lambda_{\max}$ (or $\text{RelEig}_* < \text{RelEig}$) if $\rho < 0$. In this latter case, the first two eigenvalues λ_1 and λ_2 would be both equal to $1 - \rho$, so that: $\text{RelEig}_1 = \text{RelEig}_2$.

Table 22: Definition of correlation indices

Definition	Range
Eigenvalue-based indices:	
$\text{RelEig}_s = \frac{\lambda_s}{p}$, with: $p = \text{tr}(\mathbf{R})$ and: $\lambda_1 \geq \ldots \geq \lambda_p$, $(s = 1, \ldots, p)$	$\frac{1}{p} \leq \text{RelEig}_1 \leq 1$ $0 \leq \text{RelEig}_s < 1$, $s \geq 2$
Moment-based indices:	
$\bar{\rho}_{\text{abs}} = \frac{2}{p(p-1)} \sum_{j=1}^{p} \sum_{l>j} \lvert\rho_{jl}\rvert$	$\bar{\rho}_{\text{abs}} \geq 0$
$\text{sd}_{\text{abs}} = \sqrt{\frac{2}{p(p-1)} \sum_{j=1}^{p} \sum_{l>j} (\lvert\rho_{jl}\rvert - \bar{\rho}_{\text{abs}})^2}$	$\text{sd}_{\text{abs}} \geq 0$
$\text{skew}_{\text{abs}} = \frac{\frac{2}{p(p-1)} \sum_{j=1}^{p} \sum_{l>j} (\lvert\rho_{jl}\rvert - \bar{\rho}_{\text{abs}})^3}{\text{sd}_{\text{abs}}^3}$	$\text{skew}_{\text{abs}} \in (-\infty, +\infty)$
Ratio-based indices:	
$\text{rrho} = \frac{\rho_{\max}}{\rho_{\min}}$, where: $\rho_{\max} = \max_{j,l}(\rho_{jl})$, and: $\rho_{\min} = \min_{j,l}(\rho_{jl})$, $\rho_{\min} \neq 0$	If $\lvert\rho_{\max}\rvert \leq \lvert\rho_{\min}\rvert$, rrho $\in [-1, +1]$
$\text{UnbI} = \text{sign}(\text{rrho}) \frac{\sum_{j=1}^{p} \sum_{l>j} \lvert\rho_{jl}\rvert_{\in \iota_{\max}} / N(\iota_{\max})}{\sum_{j=1}^{p} \sum_{l>j} \lvert\rho_{jl}\rvert_{\in \iota_{\min}} / N(\iota_{\min})}$ with: $N(\iota_{\min})$ and $N(\iota_{\max})$ the occurrences of values in the minimum $\iota_{\min}$ or the maximum $\iota_{\max}$ interval, resp.	If $\lvert\rho_{\max}\rvert > \lvert\rho_{\min}\rvert$, $\lvert\text{rrho}\rvert > 1$ $\lvert\text{UnbI}\rvert \geq 1$

2. *Moment-based indices*, given by the average of correlation coefficients in absolute value ($\bar{\rho}_{\text{abs}}$, absolute mean correlation), which measures the overall magnitude irrespective of the sign of correlations, along with the absolute standard deviation (sd_{abs}), which is an indicator of the unbalancing among the correlations, in that: $\text{sd}_{\text{abs}} = 0$ if $|\rho_{jl}| = |\rho|$ for all $j \neq l$, plus the absolute skewness index (skew_{abs}), which indicates if the correlation distribution is more peaked on either smaller (positive skewness) or greater values (negative skewness). In this sense, the absolute skewness index gives the shape of the unbalancing among correlations.

3. *Ratio-based indices*. One of this is given by the ratio of the largest to the smallest correlation coefficients (rrho). A more refined version we propose is the "unbalancing index" (UnbI). It is set up by, first, aggregating the correlation coefficients in absolute value of the upper (lower) triangular part of $\mathbf{R}$ in intervals. Here we considered the five intervals: $[0, 0.15), [0.15, 0.3), [0.3, 0.5),\ [0.5, 0.7),\ [0.7, 1]$. Second, by considering the two extreme observed intervals, the one, $\iota_{\min}$, with a number $N(\iota_{\min})$ of the smallest coefficients, and the other, $\iota_{\max}$, with a number $N(\iota_{\max})$ of the largest. Finally, by setting up the adjusted ratio reported in Table 22. The UnbI index then gives the ratio of the mean of the largest absolute correlations to the mean of the smallest ones. By definition, it is provided with the same sign of rrho in order to indicate whether the maximum correlation coefficient has discordant sign with respect to the minimum. These two indices, with ranges of variation reported in Table 22, aim at reflecting the presence of unbalancing among the coefficients in $\mathbf{R}$. In particular, the UnbI index should help understand the extent of the unbalancing among the correlations, in that rrho depends on two single values only (i.e. the observed minimum and maximum), and does not take into account potential coefficients in $\mathbf{R}$ that could be close to the minimum and/or the maximum (in absolute value). Hence, to a same value of rrho different values of UnbI could correspond. Much depends on how absolute correlations fall into the minimum and the maximum observed intervals. Moreover, we have that: rrho = UnbI if, and only if, the mean of the largest and the smallest absolute correlations coincide with the maximum and the minimum observed correlations, respectively.

Table 23 contains an instance of computations of the above indices with respect to a subset of the correlation matrices considered for each data pattern in both the exploratory and supplementary simulation studies with $p = 5$ variables. Discussion mainly involves the first two relative eigenvalues, and especially the second one, since it proved to be greatly informative about the magnitude and the structure of $\mathbf{R}$. *FIM* and *FIP* are considered at their default options, (i.e. $q = 0.1$ donor quantile, and Euclidean distance, $r = 2$, only for *FIP*). Once again, *IPCA* is used as a benchmark for comparisons, whose outcomes are indicated in the table as coloured rows. Grey-coloured rows refer to experimental situations where *FIM* proved to perform better than all the other methods. Light-grey rows denote the situations in which *FIP* is either the best method, or shares the best performance with another method. Blank rows correspond to *IPCA* as the best method.

Empirical evidence has given support to the following fact:

- *Equicorrelation patterns.* By construction, equicorrelation matrices always have: $\text{sd}_{\text{abs}} = 0$, an undefined skew_{abs}, and rrho = UnbI = 1 (with the only exception of the undefined rrho and UnbI for $\rho = 0$). Evaluation is then entirely based on the relative eigenvalues and the absolute mean correlation. In particular, by the second relative eigenvalue it is apparent that *FIM* performs better than the other considered methods when: $\text{RelEig}_2 > \frac{1}{p}$. This is the situation in which $\mathbf{R}$ contains a same negative

Table 23: Values of correlation indices computed for $p = 5$ variables

Indices	Eigenvalue-based		Moment-based			Ratio-based	
Data patterns	RelEig_1	RelEig_2	$\bar{\rho}_{\text{abs}}$	sd_{abs}	skew_{abs}	rrho	UnbI
Equicorrelation Patterns							
SyKu-ECor (for each κ)							
$\rho = -0.2$	0.240	0.240	0.2	0	–	1	1
$\rho = -0.1$	0.220	0.220	0.1	0	–	1	1
$\rho = 0$	0.200	0.200	0	0	–	–	–
$\rho = 0.3$	0.440	0.140	0.3	0	–	1	1
$\rho = 0.7$	0.760	0.060	0.7	0	–	1	1
SK-ECor ($\alpha = 4$)							
$\rho = -0.144$	0.229	0.229	0.144	0	–	1	1
$\rho = 0.030$	0.224	0.194	0.030	0	–	1	1
$\rho = 0.193$	0.354	0.161	0.193	0	–	1	1
$\rho = 0.572$	0.657	0.086	0.572	0	–	1	1
Positive-Negative Correlation Patterns							
SyKu-PNCor (for each κ)							
$\rho_1 = -0.20, \rho_2 = -0.05, \rho_3 = 0.02$	0.291	0.210	0.137	0.078	−0.444	−0.100	−4.709
$\rho_1 = -0.41, \rho_2 = 0.16, \rho_3 = 0.28$	0.461	0.168	0.322	0.113	−0.611	−0.683	−2.158
$\rho_1 = -0.50, \rho_2 = 0.30, \rho_3 = 0.40$	0.546	0.140	0.430	0.090	−0.626	−0.800	−1.619
$\rho_1 = -0.50, \rho_2 = 0.40, \rho_3 = 0.50$	0.577	0.120	0.470	0.046	0	−1.000	−1.000
$\rho_1 = -0.80, \rho_2 = 0.70, \rho_3 = 0.80$	0.816	0.060	0.770	0.046	0	−1.000	−1.143
SK-PNCor ($\alpha = 4$)							
$\rho_1 = -0.24, \rho_2 = -0.04, \rho_3 = 0.01$	0.309	0.208	0.156	0.098	−0.428	−0.059	−6.697
$\rho_1 = -0.41, \rho_2 = 0.16, \rho_3 = 0.28$	0.459	0.168	0.320	0.110	−0.617	−0.693	−2.117
$\rho_1 = -0.49, \rho_2 = 0.26, \rho_3 = 0.40$	0.531	0.147	0.410	0.100	−0.694	−0.693	−1.784
$\rho_1 = -0.56, \rho_2 = 0.37, \rho_3 = 0.50$	0.600	0.126	0.498	0.086	−0.754	−0.890	−2.098
$\rho_1 = -0.78, \rho_2 = 0.70, \rho_3 = 0.77$	0.805	0.061	0.760	0.039	−0.852	−0.986	−1.123
Unbalanced Correlation Patterns							
SyKu-UnbCor (for each κ)							
• One var. negatively correlated							
$\rho_1 = -0.60, \rho_2 = 0.60$	0.680	0.080	0.600	0	–	−1.000	−1.000
$\rho_1 = -0.60, \rho_2 = 0.40$ ($\kappa = 14$)	0.588	0.120	0.480	0.098	0.408	−0.667	−1.500
$\rho_1 = -0.60, \rho_2 = 0.20$	0.507	0.160	0.360	0.196	0.408	−0.333	−3.000
$\rho_1 = -0.40, \rho_2 = 0.20$ ($\kappa = 14$)	0.431	0.160	0.280	0.098	0.408	−0.500	−2.000
$\rho_1 = -0.30, \rho_2 = 0$	0.320	0.200	0.120	0.147	0.408	0	–
• All positive correlations							
$\rho_1 = 0.18, \rho_2 = 0.50$	0.518	0.184	0.373	0.161	−0.408	2.860	2.860
$\rho_1 = 0.40, \rho_2 = 0.05$	0.376	0.190	0.190	0.171	0.408	8.000	8.000
$\rho_1 = 0.40, \rho_2 = 0.20$ ($\kappa = 2, 14$)	0.431	0.160	0.280	0.098	0.408	2.000	2.000
$\rho_1 = 0.60, \rho_2 = 0.20$	0.507	0.160	0.360	0.196	0.408	3.000	3.000
$\rho_1 = 0.60, \rho_2 = 0.40$ ($\kappa = 2, 14$)	0.588	0.120	0.480	0.098	0.408	1.500	1.500
$\rho_1 = 0.60, \rho_2 = 0.60$	0.680	0.080	0.600	0	–	1.000	1.000
SK-UnbCor ($\alpha = 4$)							
• One var. negatively correlated							
$\rho_1 = -0.66, \rho_2 = 0.58$	0.690	0.083	0.612	0.036	0.408	−0.889	−1.000
$\rho_1 = -0.63, \rho_2 = 0.34$	0.574	0.131	0.456	0.137	0.408	−0.551	−1.816
$\rho_1 = -0.63, \rho_2 = 0.21$	0.523	0.157	0.378	0.203	0.408	−0.339	−2.952
$\rho_1 = -0.42, \rho_2 = 0.22$	0.445	0.156	0.298	0.098	0.408	−0.520	−1.922
$\rho_1 = -0.39, \rho_2 = 0.07$	0.380	0.186	0.199	0.158	0.408	−0.179	−5.596
$\rho_1 = -0.26, \rho_2 = -0.04$	0.291	0.208	0.128	0.105	0.408	0.163	6.135
• All positive correlations							
$\rho_1 = 0.18, \rho_2 = 0.50$	0.518	0.184	0.373	0.161	−0.408	2.860	2.860
$\rho_1 = 0.20, \rho_2 = 0.10$	0.317	0.180	0.141	0.052	0.408	2.090	2.090
$\rho_1 = 0.44, \rho_2 = 0.24$	0.463	0.151	0.321	0.096	0.408	1.802	1.782
$\rho_1 = 0.59, \rho_2 = 0.33$	0.557	0.133	0.437	0.128	0.408	1.782	1.782
$\rho_1 = 0.63, \rho_2 = 0.21$	0.519	0.159	0.373	0.204	0.408	3.019	3.019

ρ. On the other hand, if $\text{RelEig}_2 \approx \frac{1}{p}$, *FIP* tends to perform better than *FIM* and *IPCA*, or as well as the best method between *FIM* and *IPCA*. The two conditions concerning RelEig_2 can be relaxed if data are skew. In this sense, *FIM* can be again the best method even if $\text{RelEig}_2 < \frac{1}{p}$ (see e.g. the second and third rows under the SK-ECor pattern, where *FIM* and *FIP* are jointly the best methods). This means that for *FIM* and *FIP* skewness of data implies better results than symmetry also in the presence of slightly higher positive correlations. On the other hand, if data are (nearly) symmetric *FIM* and *FIP* proves to perform better than *IPCA* only for negative ρs. Finally, as RelEig_1 approaches to 1 and RelEig_2 to 0, that is with the increasing of the correlation magnitude, *IPCA* produces the best results.

- *Positive-negative correlation patterns.* By the last two columns in Table 23 (rrho and UnbI), it is apparent that PN-Cor patterns can present a more or less marked unbalancing among correlations. Nonetheless, this is an indirect consequence of the way in which matrices $\mathbf{R}$ were obtained from the (α, ω)-parameterization of the *MSN* studies (Subsect. 2.1), whilst the main objective was to define correlation matrices having both positive and negative entries of a similar magnitude. The performance of the methods is now linked to a plurality of indices, in particular RelEig_2 along with $\bar{\rho}_{\text{abs}}$, skew_{abs}, rrho, and UnbI. Regarding these last two, under the considered scenarios we always have: rrho $\in [-1, 0)$ and UnbI ≤ -1. Once again, *FIM* performs better than the other methods if $\text{RelEig}_2 \geq \frac{1}{p}$. Otherwise, in the presence of skew data *FIM* is still better if these following conditions jointly occur: $\bar{\rho}_{\text{abs}} \leq 0.3$ (low/moderate absolute mean correlation), $\text{skew}_{\text{abs}} > -0.65$ (not too marked unbalancing towards the largest absolute correlations), rrho > -0.7 (maximum correlation far smaller than the absolute minimum correlation), and UnbI ≤ -2 (mean of the largest absolute correlations more than twice the mean of the smallest absolute correlations). To its turn, *FIP* is the best method when such indices are very close to those thresholds, or slightly overcome them, especially $\bar{\rho}_{\text{abs}}$, which can be a little higher. On the other hand, if data are symmetric, the above thresholds become smaller in absolute value, especially for $\bar{\rho}_{\text{abs}}$, skew_{abs}, and rrho. Finally, *IPCA* confirms to have the best performances for lower values of RelEig_2 ($< \frac{1}{p}$), with $\bar{\rho}_{\text{abs}} \geq 0.45$, $\text{skew}_{\text{abs}} < -0.7$, rrho < -0.8, and UnbI > -2, for both skew and symmetric data.

- *Unbalanced correlation patterns.* Remarks similar to the PN-Cor patterns can be advanced. Once again, the performance of the methods seems mostly tied to the values assumed by RelEig_2, $\bar{\rho}_{\text{abs}}$, skew_{abs}, rrho, and UnbI, jointly considered. Under the considered patterns, we have: (i) $-1 \leq$ rrho < 0 and UnbI ≤ -1, in the case of a unique variable negatively correlated with all the others, (with the only exception in the last row of SK-UnbCor); (ii) rrho = UnbI ≥ 1, in the case of all positive correlations. Moreover, skew_{abs} assumes only the two values: -0.408 or 0.408, i.e. unbalancing towards larger or smaller absolute correlations, respectively, (with a unique exception of non-definiteness). Again, *FIM* performs better than the others when $\text{RelEig}_2 \geq \frac{1}{p}$. Otherwise, as before, the absolute mean correlation has to be: (i) $\bar{\rho}_{\text{abs}} < 0.4$ under these additional conditions: $\text{skew}_{\text{abs}} > 0$ (unbalancing towards lower values), $-0.4 <$ rrho < 0, and UnbI < -2, in the case of one negatively correlated variable, or: (ii) $\text{skew}_{\text{abs}} > 0$ and rrho = UnbI > 3, in the case of all positive correlations. Hence, a low/moderate average magnitude of absolute correlations does not suffice for *FIM* to perform at best. Unbalancing has to be towards lower correlations. An empirical counterexample is given by the case: $\rho_1 = 0.18$ and $\rho_2 = 0.50$ in Table 23, under both SyKu-UnbCor and SK-UnbCor patterns. Here we have:

$\bar{\rho}_{\text{abs}} = 0.373$, i.e. a moderate average magnitude, but correlations are unbalanced towards the highest value of 0.5 ($\text{skew}_{\text{abs}} < 0$). In this case, neither *FIM* nor *FIP* perform well. Moreover, *FIP* proves again to perform better than the others when the above considered indices are very close to their thresholds, or slightly overcome them, especially rrho, which can be around -0.5 in the case of one negatively correlated variable, or close to 2 when correlations are all positive. Otherwise, *IPCA* proves to perform better than *FIM* and *FIP* in the presence of higher magnitudes of correlations ($\bar{\rho}_{\text{abs}} \geq 0.6$, say), with a stronger unbalancing towards absolute larger values (mainly captured by $\text{skew}_{\text{abs}} < 0$). Overall, these results seem sharper for *FIM* and *FIP* if data are skew, thus supporting the fact that they tend to perform better here than in the presence of symmetric data. Moreover, under the SyKu-UnbCor patterns light-grey-coloured rows in Table 23, indicating *FIP* as the best, refer to platykurtic data ($\kappa = 14$), and sometimes also to normal data ($\kappa = 2$), supporting the idea of a slight kurtosis effect.

4 Discussion and conclusions

By taking into account the above illustration, along with the previous descriptive and inferential analyses, the main impressive findings of the work can now be given in the form of practical hints for users. In particular, with regard to the methods *FIM* and *FIP* it can be concluded that:

1. *FIM* works well especially in the presence of data with small or negative correlations of a same magnitude (ECor patterns, Table 23), or a mix of negative and positive correlations (PNCor and UnbCor patterns, Table 23), provided that such correlation coefficients are more or less strongly unbalanced towards lower absolute values. Such considerations hold particularly for skew data;

2. *FIP* has characteristics similar to *FIM*, but tends to perform best with a slightly higher level of correlations, i.e. small/medium correlations, according to the findings obtained under the PNCor patterns and UnbCor patterns, (Table 23).

Otherwise, in the presence of either symmetric or skew data with medium/high correlations, especially unbalanced towards higher values, other imputation methods such as *IPCA* could give better results. To this regard, it is worth remarking that, while not showing satisfactory results in most scenarios of the present study, *missForest* (not inspected by the J-T test) was expressly designed for imputation in the case of mixed-type data. This could explain its lacking effectiveness for quantitative data.

Finally, as concerns the analysis of the endogenous factors of *FIM* and *FIP*:

i. the choice concerning donor quantiles appears to be much linked to the magnitude of correlation of variables. In particular, very small correlations would require a higher number of donors, while very high correlations a smaller number. However, it could be argued that in general a good choice is to fix the percentage of donors equal to 10% (which is settled as default option in *FIM* and *FIP*) or 15%.

ii. Overall, as concerns *FIP*, differences among the various Minkowski distances here considered (i.e. city-block, Euclidean, and Lagrange distances) did not appear too substantial in the considered comparisons among the methods. Euclidean distance could hence be used as a default metric. Nonetheless, the performance of *FIP* can be improved by taking into account the data structure more carefully, for instance by

considering that in the presence of higher levels of correlations the Lagrange distance seems a better choice than the city-block distance.

References

Azzalini A (2013) R package "sn": The skew-normal and skew-t distributions (version 0.4-18). `http://azzalini.stat.unipd.it/SN`

Azzalini A, Capitanio A (1999) Statistical applications of the multivariate skew normal distribution. J R Stat Soc B 61(3):579–602

Azzalini A, Dalla Valle A (1996) The multivariate skew-normal distribution. Biometrika 83(4):715–726

Breiman L (2001) Random forests. Mach Learn 45:5–32

Fang KT, Kotz S, Ng KW (1990) Symmetric multivariate and related distributions. Monographs on Statistics and Applied Probability, 36. Chapman & Hall, New York

Gómez E, Gómez-Villegas MA, Marin JM (1998) A multivariate generalization of the power exponential family of distributions. Commun Stat-Theor M 27(3):589–600

Greenacre M (1984) Theory and applications of correspondence analysis. Academic Press, London

Hollander M, Wolfe DA (1999) Nonparametric statistical methods, 2nd edn. Wiley-Interscience, New York

Josse J, Pagès J, Husson F (2011) Multiple imputation in principal component analysis. Adv Data Anal Classif 5:231–246

Kaiser HF (1968) A measure of the average intercorrelation. Educ Psychol Meas 28:245–247

Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York

Mardia KV (1970) Measures of multivariate skewness and kurtosis with applications. Biometrika 57(3):519–530

Nora-Chouteau C (1974) Une méthode de reconstitution et d'analyse de données incomplètes. Ph.D. thesis, Université Pierre et Marie Curie

R Foundation for Statistical Computing, A Language and Environment for Statistical Computing, Vienna (2013). `http://www.R-project.org`

Rässler S, Rubin DB, Zell ER (2013) Imputation. Wiley Interdisciplinary Reviews: Computational Statistics 5(1):20–29. doi: 10.1002/wics.1240

Seber GAF (1984) Multivariate observations. John Wiley & Sons, New York

Solaro N (2004) Random variate generation from Multivariate Exponential Power distribution. Statistica & Applicazioni II(2):25–44

Solaro N, Barbiero A, Manzi G, Ferrari PA (2014) Algorithmic-type imputation techniques with different data structures: Alternative approaches in comparison. In: Vicari D, Okada A, Ragozini G, Weihs C (eds) Analysis and modeling of complex data in behavioural and social sciences, Studies in Classification, Data Analysis, and Knowledge Organization. Springer International Publishing, Cham (CH):253–261

Solaro N, Barbiero A, Manzi G, Ferrari PA (2015) A sequential distance-based approach for imputing missing data: The Forward Imputation. *submitted*

Stekhoven DJ, Bühlmann P (2012) MissForest - nonparametric missing value imputation for mixed-type data. Bioinformatics 28(1):112–118