A simulation comparison of imputation methods for quantitative data in the presence of multiple data patterns

N. Solaro[a], A. Barbiero[b], G. Manzi[b], and P. A. Ferrari[b]

[a]Department of Statistics and Quantitative Methods, Università degli Studi di Milano-Bicocca, Via Bicocca degli Arcimboldi 8, Milan, Italy; [b]Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, Milan, Italy

Abstract

An extensive investigation via simulation is carried out with the aim of comparing three nonparametric, single imputation methods in the presence of multiple data patterns. The ultimate goal is to provide useful hints for users needing to quickly pick the most effective imputation method among the following: Forward Imputation (*ForImp*), considered in the two variants of *ForImp* with the Principal Component Analysis (PCA), which alternates the use of PCA and the Nearest-Neighbour Imputation (NNI) method in a forward, sequential procedure, and *ForImp* with the Mahalanobis distance, which involves the use of the Mahalanobis distance when performing NNI; the iterative PCA technique, which imputes missing values simultaneously via PCA; the *missForest* method, which is based on random forests and is developed for mixed-type data. Performance of these methods is compared under several data patterns characterized by different levels of kurtosis or skewness and correlation structures.

KEYWORDS

Forward imputation; iterative principal component analysis; Mahalanobis distance; missForest; missing data; Monte Carlo simulation; multivariate exponential power distribution; multivariate skew-normal distribution; nearest-neighbour imputation

AMS CLASSIFICATION

62H25, 62-07, 62-04, 62H99

1. Introduction

'*Thinking the unthinkable*': in this way Bradley Efron in 1979 used to define the new role of simulation techniques in the improvement of statistical science [1]. He reckoned at a very early stage that the advent of high-speed computers would have opened new frontiers for the development of statistical theory. His side message was that, with this possibility, simulated solutions would have been even better than analytical solutions, and the fast spread of Monte Carlo methods subsequent to the availability of increasingly powerful computers would have later testified the rightness of Efron's intuition. Needless to say, in many fields of statistics simulation has become the primary way, or even the unique way, to proceed. For instance, analytical solutions might not exist or be not easily attainable, or, even if available, might not be adequate for application to real situations because grounded on assumptions valid in theory but not exactly in practice.

CONTACT N. Solaro. Email: nadia.solaro@unimib.it

Evaluation of properties of advanced statistical techniques is no exception. One of the problems that usually requires simulation analyses is the appraisal of the performance of missing data imputation methods. A missing datum is missing by definition. It is therefore impossible to assess if a certain method is capable of recovering that very datum well. But it might be possible to simulate data with characteristics similar to the real data from which the missing datum comes and create missingness artificially according to a specific generating mechanism. In this way, evaluation is made by assessing the closeness, according to several criteria, of the imputed data to the simulated ones, which should represent the actual data as best as possible. Moreover, in general, it should be strongly recommended analysing and comparing the performance of the methods under a multitude of different simulation scenarios, and then of data structures, since the goodness of a method may greatly depend on the situation at hand.

Many strategies for dealing with the problem of missing data have been proposed over the years, while research is still advancing on this matter. Statistical literature is incredibly rich of contributions. One needs only think to the recurrent distinctions among different approaches, i.e. between parametric and nonparametric methods, single and multiple imputations, deterministic and random imputations, imputation-based and model-based procedures. Classical theoretical references are provided by Little and Rubin [2] and Schafer [3], while e.g. Molenberghs and Kenward [4] treat missing data handling in clinical studies and Haziza [5] in sample surveys.

Nonetheless, to our knowledge, in literature there are still few contributions addressed to comparing the performance of different imputation methods. A pioneering study was that by Bello [6] who compared five methods, namely the mean substitution method, the EM algorithm, the Dear's principal component method, the general iterative principal component method and the singular value decomposition method. The same author [7] performed a simulation study of imputation techniques in the framework of discriminant analysis. More recently, Marella et al. [8] have evaluated the matching noise, i.e. the discrepancy between the actual and the imputed data, in the important class of K-Nearest Neighbour imputation methods, dealing with different settings of this class. Ning and Cheng [9] have performed a comparison analysis, with an extensive simulation study and an empirical study on real data, on the difference between the nearest-neighbour imputation method and the kernel-weighted regression method in estimating a population mean of incomplete responses and also in classifying the incomplete responses that are missing at random depending on some covariates. In a similar framework, Tutz and Ramzan [10] considered weighted nearest-neighbour imputation methods using distances for selected covariates.

Motivated by several practical problems concerning missing data handling that we faced in a pure nonparametric perspective, we carried out an extensive investigation via simulation for inspecting and comparing the performance of three different nonparametric methods for single imputation of missing data, i.e. the Forward Imputation (*ForImp*) [11,12], the Iterative Principal Component Analysis method (*IPCA*) [13–15] and Stekhoven and Bühlmann's *missForest* method [16]. *ForImp* is a sequential, distance-based, distribution-free imputation procedure that is based on the nearest-neighbour imputation method and can exploit a multivariate data analysis technique to synthesise the information of the complete part of the data [11,12]. The *IPCA* method is an algorithmic-type technique that imputes missing values simultaneously by the iterative use of principal component analysis, recently rearranged by Josse et al. [15] in the more general context of multiple imputation with factorial methods. The *missForest* method is a nonparametric imputation technique for continuous and categorical data based on a random forest, i.e. a random classifier introduced in the context of machine learning [17].

The organization of this paper is as follows. Section 2 summarizes the three compared methods in their main characteristics and algorithmic presentation. Section 3 presents the general simulation settings and data structures considered. Section 4 presents in details four main simulation studies and some other supplementary simulation studies, highlighting the main results from each of them and providing a final discussion with practical hints for users. Section 5 concludes the paper.

2. Imputing quantitative missing data: The considered methods

We focus on nonparametric methods that carry out single imputation of missing data according to different theoretical grounds. The first method is Forward Imputation (*ForImp*) [11,12], considered here in the two variants developed for quantitative data, i.e. *ForImp* with the Mahalanobis distance (*ForImpMahalanobis – FIM* in short) and *ForImp* with the Principal Component Analysis (*ForImpPCA – FIP* in short) [12]. The second method is the Iterative Principal Component Analysis (*IPCA*) [13,14], which has recently been employed in a more general multiple imputation methodology based on factorial methods [15]. The third method is *missForest* [16], which is a random-forests-based imputation method designed specifically for mixed-type data.

In what follows, the main aspects of these methods will be described in short. As previously stated, we confine our investigation to quantitative data because, to our knowledge, in literature there are still few contributions aimed at comparing different imputation approaches for quantitative data. We start by assuming that $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p]$ is a $(n \times p)$-dimensional data matrix, referred to n units and p quantitative variables, which contains missing values in its rows.

2.1. *The ForImp approach with PCA or the Mahalanobis distance*

The *ForImp* approach is a sequential distance-based procedure that exploits the complete part of the data in a step-by-step, forward process involving the Nearest-Neighbour Imputation (NNI) method. Regarding the two *ForImp* methods for quantitative data, the main difference between *FIM* and *FIP* is that *FIM* uses NNI with the Mahalanobis distance to detect donors for incomplete units, while *FIP* alternates the NNI method (applied with a weighted Minkowski distance) with PCA to extract information instrumental in searching for donors from the complete part. The process begins with the complete submatrix $\mathbf{X}_0$ of $\mathbf{X}$ having no missing entry. Initialization of missing data is therefore not required. At each subsequent step, the complete part is updated and enlarged with further completed rows according to the algorithm resumed in Table 1.

2.2. *The IPCA algorithm*

IPCA is an algorithmic-type technique that imputes missing values simultaneously by an iterative use of the PCA method. As such, *IPCA* is part of the iterative imputation methods that involve exploratory multivariate data analysis techniques [13,14]. It is based on the well-known property of minimization in the least-squares sense of PCA. According to this, *IPCA* imputes missing values by minimizing a squared loss function of the difference between the original, centred data matrix $\mathbf{X}$ (previously initialized) and the fitted matrix $\hat{\mathbf{X}}$ obtained as product of the two matrices containing the component scores and the loadings, respectively. As an alternative, the minimization process could be replaced by an iterative use of the singular value decomposition or the alternating least-squares procedure [15]. The *IPCA* algorithm was proved to have a series of good properties, by virtue of which its extension to a multiple imputation perspective by Josse et al. [15] was justified.

IPCA starts from the entire matrix $\mathbf{X}$ fulfilled with initial values and carries out imputation according to the main steps summarized in Table 2.

2.3. *The missForest algorithm*

The *missForest* algorithm [16], developed for mixed-type data (quantitative and categorical variables jointly observed), uses a random forest (RF) [17] trained on the observed part of matrix $\mathbf{X}$ to predict the missing values. Within the scope of our investigation, we confine the use of *missForest* to quantitative data, although we are aware this procedure was developed for a more general usage. Nonetheless, we are interested in exploring how a RF-based imputation method could perform in a situation seemingly less complicated than mixed-type data, as may be the case of quantitative data.

The *missForest* algorithm is described in short in Table 3. For simplicity, the main quantities involved are defined here. Let X_j be a variable with entries in the column-vector $\mathbf{x}_j$ of $\mathbf{X}$ and with missing values corresponding to the set of indices: $\mathbf{i}_{miss}^{(j)} \subseteq \{1, 2, \ldots, n\}$. Then, let the following vectors and matrices be defined:

- $\mathbf{x}_j^{obs}$: the sub-vector of $\mathbf{x}_j$ with the observed values of X_j;
- $\mathbf{x}_j^{miss}$: the sub-vector of $\mathbf{x}_j$ with the missing entries of X_j;
- $\mathbf{X}_{-j}^{obs}$: the matrix of the other variables X_l, $(l \neq j)$, with observations corresponding to the set of indices: $\mathbf{i}_{obs}^{(j)} = \{1, 2, \ldots, n\} \setminus \mathbf{i}_{miss}^{(j)}$. Since this set depends on the observed values of X_j, matrix $\mathbf{X}_{-j}^{obs}$ might be not completely observed;
- $\mathbf{X}_{-j}^{miss}$: the matrix of the other variables X_l, $(l \neq j)$, with observations corresponding to the set of indices $\mathbf{i}_{miss}^{(j)}$. Since this set depends on the missing values of X_j, matrix $\mathbf{X}_{-j}^{miss}$ might indeed not contain missing values.

The algorithm in Table 3 proceeds until the *stopping rule* γ is reached. Confined to quantitative variables, the difference between values $x_{ij,\text{imp}}^{(k)}$ of matrix $\mathbf{X}_{\text{imp}}^{(k)}$ imputed in the current iteration k and values $x_{ij,\text{imp}}^{(k-1)}$ of matrix $\mathbf{X}_{\text{imp}}^{(k-1)}$ imputed in the previous iteration $k-1$ is evaluated through the formula:

$$\frac{\sum_{j=1}^{p} \sum_{i \in \mathbf{i}_{miss}^{(j)}} \left(x_{ij,\text{imp}}^{(k)} - x_{ij,\text{imp}}^{(k-1)} \right)^2}{\sum_{j=1}^{p} \sum_{i \in \mathbf{i}_{miss}^{(j)}} \left(x_{ij,\text{imp}}^{(k)} \right)^2}, \qquad \forall k \geq 1. \tag{1}$$

The algorithm stops when the quantity (1) increases for the first time [16].

3. Method for simulations

Given the algorithmic nature of *FIM*, *FIP*, *IPCA* and *missForest*, assessment of their performance and comparisons among them were carried out through Monte Carlo (MC) simulation. Experimental conditions were fixed such that they reproduced a variety of data patterns frequently encountered in applications. In addition to the number of units and variables, and percentages of MCAR (Missing Completely At Random) values, experimental conditions also concerned the shape of data distributions (kurtosis and skewness) and correlation structures of variables, with the aim of reproducing real data patterns in the most flexible way. To generate data having the desired features we relied on the two families of Multivariate Exponential Power (*MEP*) [18] and Multivariate Skew-Normal (*MSN*) distributions [19,20].

Main objective of the MC simulation studies was to examine the performance of the four imputation methods and detect, if possible, the most effective method regarding the considered data structures. A major aspect of concern was to restrain, as much as possible, the total number of simulation scenarios to be run without losing any meaningful information about the trends. Accordingly, we performed a range of exploratory simulation studies at a first stage, followed by supplementary simulations at a second stage addressed to look more thoroughly into several specific situations suggested by the first kind of studies.

The simulation design is described across its main steps as follows. After briefly mentioning *MEP* and *MSN* main results in Subsect. 3.1, experimental conditions, data patterns and proper simulation settings are sketched in Subsect. 3.2. The simulation procedure is described in Subsect. 3.3, while methods applied to synthesise simulation results are the object of Subsect. 3.4.

3.1. *Multivariate distributions for simulations*

MEP and *MSN* families of distributions play an important role in the description of real data. Their density functions are flexible enough to cover a wide spectrum of real scenarios. In addition, both the families can be regarded as extensions of the multivariate normal (*MVN*) distribution in terms of kurtosis (*MEP*) or skewness (*MSN*) departures. To make clearer the role of their parameters in our simulation study, a synthetic collection of their main theoretical results is given below. Let $\boldsymbol{X}$ be a p-dimensional continuous random vector (r.v.). Then:

- $\boldsymbol{X}$ is $\mathrm{MEP}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \kappa)$ distributed if its density function (d.f.) can be expressed as:

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \kappa) = \frac{p\Gamma(p/2)}{\pi^{p/2}\Gamma(1+p/\kappa)2^{1+p/\kappa}|\boldsymbol{\Sigma}|^{1/2}} \cdot \exp\{-\frac{1}{2}[(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})]^{\kappa/2}\}, \tag{2}$$

  ($\boldsymbol{x} \in \mathbb{R}^p$, $\kappa > 0$), where $\boldsymbol{\mu} \in \mathbb{R}^p$ is the mean vector and $\boldsymbol{\Sigma}$ is the "characteristic matrix", which is square, symmetric and positive-definite. The variance-covariance matrix $\mathrm{V}(\boldsymbol{X})$ is linked to $\boldsymbol{\Sigma}$ by the relation: $\mathrm{V}(\boldsymbol{X}) = c(\kappa, p)\boldsymbol{\Sigma}$, where $c(\kappa, p) = 2^{2/\kappa}\Gamma((p+2)/\kappa)/(p\Gamma(p/\kappa))$. Parameter κ expresses kurtosis departures from the *MVN* distribution and is therefore regarded as the non-normality parameter. Specifically, if $\kappa = 2$ d.f. (2) reduces to the *MVN* distribution, while if $\kappa > 2$ ($\kappa < 2$) a platykurtic (leptokurtic) distribution is obtained. The indicators introduced by Mardia [21] to measure skewness and kurtosis are equal, respectively, to: $\gamma_{1\mathrm{MV}} = 0$ (obviously), and: $\gamma_{2\mathrm{MV}} = \frac{p^2\Gamma(p/\kappa)\Gamma((p+4)/\kappa)}{\Gamma^2((p+2)/\kappa)} - p(p+2)$, where from this latter it is apparent that $\gamma_{2\text{MV}}$ depends only on the κ parameter and the number p of variables [18].

As an illustration, contour plots for the bivariate exponential power distribution with $\boldsymbol{\mu} = \mathbf{0}$, $\sigma_{12} = \sigma_{21} = 0.5$, $\sigma_{11} = \sigma_{22} = 1$, and $\kappa = 1; 2; 14$ are displayed in Figure 1.

- $\boldsymbol{X}$ is $\text{MSN}_p(\boldsymbol{\Omega}, \boldsymbol{\alpha})$ distributed if its d.f. can be expressed as:

$$f(\boldsymbol{x}; \boldsymbol{\Omega}, \boldsymbol{\alpha}) = 2\phi_p(\boldsymbol{x}; \boldsymbol{\Omega})\Phi(\boldsymbol{\alpha}^T\boldsymbol{x}), \qquad (\boldsymbol{x} \in \mathbb{R}^p, \boldsymbol{\alpha} \in \mathbb{R}^p), \tag{3}$$

where: $\phi_p(\boldsymbol{x}; \boldsymbol{\Omega})$ is the $\text{MVN}_p(\mathbf{0}, \boldsymbol{\Omega})$ d.f. with correlation matrix (or “association matrix”) $\boldsymbol{\Omega}$ of full rank, $\Phi(\cdot)$ is the $N(0, 1)$ distribution function, and $\boldsymbol{\alpha}$ is a p-dimensional parameter vector regulating the skewness. In particular, if: $\boldsymbol{\alpha} = \mathbf{0}$, then the d.f. (3) reduces to: $\boldsymbol{X} \sim \text{MVN}_p(\mathbf{0}, \boldsymbol{\Omega})$ with $\boldsymbol{\Omega} \equiv \mathbf{R}$. Expected value is given by: $\text{E}(\boldsymbol{X}) = \boldsymbol{\mu} = \sqrt{2/\pi}\boldsymbol{\delta}$, where: $\boldsymbol{\delta} = \frac{\boldsymbol{\Omega\alpha}}{\sqrt{1+\boldsymbol{\alpha}^T\boldsymbol{\Omega\alpha}}}$, and variance-covariance matrix by: $\text{V}(\boldsymbol{X}) = \boldsymbol{\Sigma} = \boldsymbol{\Omega} - \boldsymbol{\mu\mu}^T$. Correlation matrix $\mathbf{R}$ of $\boldsymbol{X}$ is therefore given by: $\mathbf{R} = \mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1}$, where $\mathbf{D} = \text{diag}\left\{\sqrt{1 - 2\pi^{-1}\delta_j^2}\right\}_{j=1,\dots,p}$. Univariate skewness index γ_1 for variable X_j is defined as: $\gamma_1 = \frac{4-\pi}{2}\frac{\text{E}(X_j)^3}{\text{Var}(X_j)^{3/2}}$, and takes values in $(-0.995, +0.995)$. Multivariate indices of skewness and kurtosis are given, respectively, by: $\gamma_{1\text{MV}} = \left(\frac{4-\pi}{2}\right)^2(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^3$, with values in $(-0.9905, +0.9905)$, and: $\gamma_{2\text{MV}} = 2(\pi - 3)(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2$, with values in $(-0.869, +0.869)$ [19,20]. Scale and location parameters are not comprised in the d.f. (3), but they can be introduced through proper linear transformations [19,20].

Contour plots for the bivariate skew-normal distribution with matrix $\boldsymbol{\Omega}$ having as elements: $\omega_{12} = \omega_{21} = 0.8$, $\omega_{11} = \omega_{22} = 1$, along with $\boldsymbol{\alpha} = \alpha\mathbf{1}_2$ with $\mathbf{1}_2 = (1, 1)^T$ and $\alpha = 1; 4; 10$, are displayed in Figure 2.

3.2. *Experimental conditions, data patterns and simulation settings*

Experimental conditions involved in the MC simulation studies referred to dimensionality of data (i.e. number of units and variables), seriousness of missingness (i.e. percentages of MCAR values) and data patterns. These latter were defined by combining two items of shape, i.e. “Symmetry and Kurtosis” (labelled as SyKu) and “SKewness” (SK), with three structures of correlation of variables, i.e. EquiCorrelations (ECor), Positive-Negative Correlations (PNCor) and Unbalanced Correlations (UnbCor). Data with the desired patterns were generated using the two families of multivariate distributions described in Subsect. 3.1. Their parameters – κ and $\mathbf{R}$ in the *MEP* case, $\boldsymbol{\alpha}$ and $\boldsymbol{\Omega}$ in the *MSN* case –, expressing the experimental conditions linked directly to shape and correlation of the data to be generated, are denoted as *input parameters*. These are distinguished from so-called *output parameters*, which refer instead to some synthesis indices, and then some characteristics, of the generated data. This distinction will help mostly interpret the simulation results obtained in the presence of *MSN* generated data (Subsects. 4.2–4.4). Simulation settings (or scenarios) were ultimately given by the combinations of numbers of units and variables, percentages of MCAR values and input parameters of the multivariate distributions. These scenarios will be described study by study in the next Subsects. 4.1–4.4.

We generated patterns of the SyKu shape (Table 4) through *MEP* distributions using the transformation method described in Gómez et al. [18] and Solaro [22], and fixing the output correlation matrix $\mathbf{R}$ rather than the input characteristic matrix $\mathbf{\Sigma}$ of the *MEP* d.f. (2). For this reason, $\mathbf{R}$ assumes also the role of a matrix of input parameters. On the other hand, patterns related to the SK shape (Table 4) were generated through *MSN* distributions using the method implemented by Azzalini in the R library "sn" [23], which is based on the input (α, ω)-parametrization of the *MSN* d.f. (3). Generating data by fixing appropriate output parameters (e.g. correlation coefficients ρ in $\mathbf{R}$) instead of the input parameters $\boldsymbol{\alpha}$ and $\mathbf{\Omega}$, would have been impracticable in this case because of the various constraints among input–output parameters (Subsect. 3.1). Nonetheless, simulation results can be interpreted all the same in connection with the output parameters, thanks to their strict relationships with the input parameters (Subsects. 4.2–4.4).

After that, we introduced several criteria addressed to describing the generated data patterns in terms of strength and structure of correlations of variables, as well as (symmetric or skew) shape of data distributions. While kurtosis and skewness have homonym univariate and multivariate indices as "natural" syntheses [21], summing up correlation matrices in scalars was less straightforward. After having valued the proposals known in the literature (see e.g. [24]), we relied on the following descriptive indices, which we termed as *correlation indices*. Specifically, we have:

(1) *Eigenvalue-based indices*, used as measures of correlation strength. These are given by the relative eigenvalues (RelEig), i.e. the eigenvalues: $\lambda_{\max} = \lambda_1 \geq \ldots \geq \lambda_s \geq \ldots \geq \lambda_{\min} = \lambda_p \geq 0$ of matrix $\mathbf{R}$, with at least one strict inequality, divided by the number $p = \operatorname{tr}(\mathbf{R})$ of variables:

$$\text{RelEig}_s = \frac{\lambda_s}{p}, \qquad s = 1, \ldots, p, \tag{4}$$

where: $\frac{1}{p} \leq \text{RelEig}_1 \leq 1$ and: $0 \leq \text{RelEig}_s \leq \text{RelEig}_1$, $\forall s \geq 2$. RelEigs are informative about the correlation structure. In case of uncorrelation (i.e. $\mathbf{R} = \mathbf{I}_{(p)}$, with $\mathbf{I}_{(p)}$ the identity matrix of order p), we have: $\text{RelEig}_s = \frac{1}{p}$ for all $s = 1, \ldots, p$. In case of perfect positive correlations (i.e. $\rho_{jl} = 1$ for all $j \neq l$), we have: $\text{RelEig}_1 = 1$ 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. In addition, in case of equicorrelation (i.e. $\rho_{jl} = \rho$, $\forall j, l$), $\mathbf{R}$ admits a unique eigenvalue: $\lambda_* = 1 + (p-1)\rho$, and $(p-1)$ not distinct eigenvalues equal to: $\lambda = 1 - \rho$, [25]. Then, if $\rho > 0$, the maximum eigenvalue is unique and is given by: $\lambda_{\max} = \lambda_* > \lambda$, from which we have: $\text{RelEig}_* > \text{RelEig}$. Otherwise, if $\rho < 0$, the maximum eigenvalue is no more unique and is given by: $\lambda_{\max} = \lambda > \lambda_*$, so that: $\text{RelEig} > \text{RelEig}_*$. Moreover, given that the first $p-1$ eigenvalues λ_s are equal to $\lambda = 1 - \rho$, it holds: $\text{RelEig}_1 = \ldots = \text{RelEig}_{p-1}$;

(2) *Moment-based indices*, which are given by:
   - the minimum and maximum observed correlation coefficients:

$$\rho_{\min} = \min_{l>j=1,\ldots,p} (\rho_{jl}) \quad \text{and:} \quad \rho_{\max} = \max_{l>j=1,\ldots,p} (\rho_{jl}), \tag{5}$$

   where: $-1 \leq \rho_{\min} \leq \rho_{\max} \leq +1$;

- the mean absolute correlation:

$$\bar{\rho}_{\text{abs}} = \frac{2}{p(p-1)} \sum\nolimits_{j=1}^{p} \sum\nolimits_{l>j} |\rho_{jl}| \, , \tag{6}$$

  ($\bar{\rho}_{\text{abs}} \geq 0$), which is a measure of the overall magnitude irrespective of the sign of correlations;
- the absolute skewness index:

$$\text{skew}_{\text{abs}} = \frac{\frac{2}{p(p-1)} \sum_{j=1}^{p} \sum_{l>j} (|\rho_{jl}| - \bar{\rho}_{\text{abs}})^3}{\text{sd}^3_{\text{abs}}}, \qquad \text{with: } \text{sd}_{\text{abs}} > 0, \tag{7}$$

  ($\text{skew}_{\text{abs}} \in (-\infty, +\infty)$), where sd_{abs} at the denominator in (7) is the absolute standard deviation:

$$\text{sd}_{\text{abs}} = \sqrt{\frac{2}{p(p-1)} \sum\nolimits_{j=1}^{p} \sum\nolimits_{l>j} (|\rho_{jl}| - \bar{\rho}_{\text{abs}})^2}. \tag{8}$$

  skew_{abs} is an indicator of the extent of unbalance among absolute correlations, indicating whether absolute correlation coefficients are more concentrated on either lower absolute values (positive skewness), or higher absolute values (negative skewness). Obviously, skew_{abs} is not defined when $\text{sd}_{\text{abs}} = 0$, that is, in the case of absolute equicorrelation (i.e. $|\rho_{jl}| = |\rho|$ for all $j \neq l$).

3.3. *Simulation procedure*

For each simulation setting, we generated a complete $n \times p$ data matrix $\mathbf{X}^*$ ($n > p$) through an *MEP* or *MSN* distribution according to the data shape under study (SyKu or SK, respectively). Subsequently, we set up $T = 1000$ incomplete matrices $\mathbf{X}_t$ from $\mathbf{X}^*$ by deleting 5%, 10%, or 20% of values completely at random, and applied *FIM*, *FIP*, *IPCA* and *missForest* to impute missing values in each $\mathbf{X}_t$, ($t = 1, \ldots, 1000$), using the following options:

- *FIM* and *FIP* (Table 1): We kept the default options of the R library 'GenForImp' [26], which are: (1) a proportion $q = 0.1$ of donors (Table 1, point 4); (2) as for *FIP* only, extraction of the principal components (PCs) from the variance-covariance matrix (i.e. option 'cor=False', see Remark 3, Sect. 2.1 in Solaro et al. [12]), along with the Euclidean distance ($r = 2$ in Table 1, point 4) for donors' detection;
- *IPCA* (Table 2): We used the function 'imputePCA' in the R library 'missMDA' [27] with the default nonparametric 'Regularized method' [15], the maximum number of iterations fixed at 5000, and the number of extracted PCs set at the largest possible value, i.e. $p - 2$, ($p \geq 3$);
- *missForest* (Table 3): We increased the maximum number of iterations from 10 (the default value in the R library 'missForest' [28]) to 50. The other relevant parameters 'ntree' – the number of trees grown in each forest – and 'mtry' – the number of variables randomly sampled at each node of the trees – are kept fixed at their default values, i.e. ntree = 100 and mtry = $\lfloor \sqrt{p} \rfloor$, where $\lfloor x \rfloor$ is the

largest integer not greater than x.

To limit the total number of simulation scenarios to be inspected without losing any meaningful information, we initially tested the imputation performance under the SyKu shape in the presence of the ECor structure only (Table 4). Supplementary MC studies involving the PNCor and the UnbCor structures were performed at a later stage, under the range of simulation settings suggested by the results obtained for the SK shape. These latter were expected to give useful indications about the SyKu shape because, roughly speaking, *MSN* distributions with $\alpha = 1$ (situations of slighter skewness) are quite close to *MEP* distributions with $\kappa = 2$ (normal distribution).

Finally, it is worth remarking that detecting, for each method, the best set of options depending on the types of data structures was beyond our scopes. This would have tremendously enlarged the number of simulation scenarios to be considered. We then applied the four methods at their defaults, excepting the increase of both the number of PCs extracted under *IPCA* (for a better comparability with *FIP*) and the maximum number of iterations under *IPCA* and *missForest*.

3.4. *Methods for summaries and comparisons*

We carried out both descriptive and inferential analyses of simulation results according to a twofold purpose: (1) to inspect the impact of the type of pattern (defined by kurtosis or skewness along with correlation structure) on the imputation performance of *FIM*, *FIP*, *IPCA* and *missForest*, keeping the other experimental conditions fixed (*pattern-impact analysis*); (2) to compare the imputation performance among the four methods *ceteris paribus* (*performance comparison analysis*). As a measure of the imputation performance, we used the Relative Mean Square Error (RMSE), which is similar to the normalized root mean squared error adopted by Stekhoven and Bühlmann [16]. For each method m and under each combination c of the levels of the experimental conditions (said "experimental combination"), RMSE is given by the values $r_{t,c}(\mathsf{m})$:

$$r_{t,c}(\mathsf{m}) = \sum_{j=1}^{p} \tfrac{1}{n\sigma_j^2}(\mathbf{x}_j^* - {}_{\mathsf{m}}\tilde{\mathbf{x}}_{j,t})^T(\mathbf{x}_j^* - {}_{\mathsf{m}}\tilde{\mathbf{x}}_{j,t}), \qquad t = 1, \ldots, 1000, \tag{9}$$

where $\mathbf{x}_j^*$ is column vector j of the known complete matrix $\mathbf{X}^*$, ${}_{\mathsf{m}}\tilde{\mathbf{x}}_{j,t}$ is column vector j of matrix ${}_{\mathsf{m}}\tilde{\mathbf{X}}_t$ with missing values imputed by method m at run t, and σ_j^2 is the variance of variable X_j in $\mathbf{X}^*$, $(j = 1, \ldots, p)$. For the procedure applied to generate the incomplete matrices $\mathbf{X}_t$ (Subsect. 3.3), RMSE values in (9) are independent observations conditionally on the complete matrix $\mathbf{X}^*$.

Descriptive analysis was based on usual synthesis measures (e.g. mean and standard deviation) and graphs, in particular, dot plots of the RMSE mean values: $\bar{r}_c(\mathsf{m}) = \frac{1}{T}\sum_{t=1}^{T} r_{t,c}(\mathsf{m})$, computed for each method m under every experimental combination c, (an ample collection of box plots can be found in [32]).

Methods for inference were chosen consistently with the type of analysis – pattern-impact or performance comparison – to be fulfilled. Pattern-impact analysis was carried out by using the procedure described in Hochberg and Tamhane [29] for computing Tukey uncertainty intervals (U.I.s) around sample means. This procedure, which is based on Tukey's multiple comparison method for balanced data [29], produces simultaneous confidence intervals by which a plurality of means can be compared with each other while preserving the overall nominal $1-\alpha$ confidence level, $(0<\alpha<1)$.

With the aim of comparing the RMSE means referred to the same imputation method across different experimental combinations, we re-adapted the above procedure as follows. Let $E=\{E_1,\dots E_h\}$ be the set of h *active* experimental conditions, i.e. conditions in regard to which we want to perform comparisons within each given method. Set E depends on the type of simulation study considered. For instance, under the SyKu-ECor pattern (Table 4), we have $h=2$ active conditions given by kurtosis (E_1) and correlation structure (E_2). The other $\bar{h}$ conditions not included in E, and forming the set $\bar{E}$, are the *inactive* conditions. In all the considered simulation studies, inactive conditions are given by the number of variables, the number of units and percentage of MCAR values (so that: $\bar{h}=3$), and are kept fixed at specific levels, say, $\bar{E}=\bar{e}^*$. Then, for every set: $E_*=\{E\mid\bar{E}=\bar{e}^*\}$ of h active conditions considered at the fixed levels $\bar{e}^*$ of $\bar{E}$, we set up a h-way full ANOVA model for explaining RMSE with the effects of the levels of the h conditions E_j in E_*:

$$R^*_{i_1i_2\dots i_h,t}(\mathsf{m})={}_{\mathsf{m}}\theta^*+{}_{\mathsf{m}}\tau^*_{1i_1}+\dots+{}_{\mathsf{m}}\tau^*_{hi_h}+\dots+{}_{\mathsf{m}}(\tau_1\tau_2\dots\tau_h)^*_{i_1i_2\dots i_h}+\varepsilon^*_{i_1i_2\dots i_h,t}(\mathsf{m}),\quad(10)$$

$(i_j=1,\dots,I_j,\ j=1,\dots,h,\ t=1,\dots,1000)$, where $R^*_{i_1i_2\dots i_h,t}(\mathsf{m})$ is RMSE of method m under the combination of levels $(i_1,i_2,\dots,i_h)$ of the h active conditions in the set E_* at the t-th simulation run, ${}_{\mathsf{m}}\theta^*$ is the overall RMSE mean of method m conditional on the levels $\bar{e}^*$ of the $\bar{h}$ inactive conditions in $\bar{E}$, parameters ${}_{\mathsf{m}}\tau^*$ are fixed effects associated to the single levels, or their combinations, of the h active conditions in E_*, and $\varepsilon_{i_1i_2\dots i_h,t}(\mathsf{m})$ are i.i.d. $N(0,{}_{\mathsf{m}}\sigma^2)$ random errors, with ${}_{\mathsf{m}}\sigma^2$ unknown.

Next, let $\mathcal{I}^*_k=(i_1,i_2,\dots,i_h\,|\,\bar{e}^*)$ be the k-th combination of levels of the h active conditions in E_*, with: $k=1,\dots,K$, and: $K=\prod_{j=1}^h I_j$ expressing the total number of these combinations. As said before, we have $T=1000$ observations for each $\mathcal{I}^*_k$. The RMSE sample mean of method m under $\mathcal{I}^*_k$ is given by: $\bar{R}_{\mathcal{I}^*_k}(\mathsf{m})=\frac{1}{T}\sum_{t=1}^T R^*_{i_1i_2\dots i_h,t}(\mathsf{m})$, $\forall k$. We have: $\mathrm{E}\big[\bar{R}_{\mathcal{I}^*_k}(\mathsf{m})\big]=\theta_{\mathcal{I}^*_k}(\mathsf{m})$, and: $\mathrm{Var}\big[\bar{R}_{\mathcal{I}^*_k}(\mathsf{m})\big]=\frac{{}_{\mathsf{m}}\sigma^2}{T}$, $\forall k$. Then, Tukey U.I. around $\bar{R}_{\mathcal{I}^*_k}(\mathsf{m})$ at the overall $1-\alpha$ confidence level is given by:

$$\bar{R}_{\mathcal{I}^*_k}(\mathsf{m})\ \pm Q^{(\alpha)}_{K,\nu}\frac{S(\mathsf{m})}{2\sqrt{T}},\qquad(11)$$

where $Q^{(\alpha)}_{K,\nu}$ is the Studentized range, $\nu=K(T-1)$ expresses the degrees of freedom of the residual deviance $ESS(\mathsf{m})$ of model (10), and $S^2(\mathsf{m})$ is the unbiased sample variance estimator for ${}_{\mathsf{m}}\sigma^2$ given by: $S^2(\mathsf{m})=\frac{ESS(\mathsf{m})}{K(T-1)}$.

Performance comparison analysis was carried out by means of the Jonckheere–Terpstra (JT) test [30] for ordered alternatives, with the aim of appraising which among the four imputation methods proves to have smaller imputation errors under the various simulation scenarios. We regarded the best performing method as the one that, in case of a significant result, has the smallest *p-value*, or, equivalently, the highest absolute value on the reference asymptotic normal distribution of the JT test.

4. Simulation studies and their results

In the ensuing subsections, simulation studies are described in terms of data patterns (Table 4), simulation settings and main findings. Simulation results are discussed taking into account the relationships between input and output parameters of the *MEP* and *MSN* distributions used for data generation (Subsects. 3.1 and 3.2). As already pointed out, these relationships will help mostly understand the results related to the SK shape.

Given the great number of the considered scenarios, simulation results presented here exclusively concern 20% of MCAR values, since it better emphasizes differences among *FIM*, *FIP*, *IPCA* and *missForest* [31], and mostly pertain to the case of $p = 5$ variables and $n = 1000$ units. Besides being tested in all the studies, this experimental setting is representative of the trends observed under most of the other scenarios. An ample collection of the omitted results can be found in the supplemental material (SM) and in [32].

Results are displayed through graphs and numeric tables. Dot plots and tables of RMSE mean values with 95%-Tukey U.I.s are part of the pattern-impact analysis (Subsect. 3.4), addressed to discovering, for each method, potential effects of the active conditions (i.e. correlation structure and shape of data distribution) on the imputation performance under fixed levels of inactive conditions (i.e. number of variables, number of units and percentage of MCAR values).

Figures and tables are built with a similar structure over all the simulation studies. Figures are set up as matrices of panels of dot plots, whose column headers contain the name of the imputation methods, and row headers the number of variables (e.g. Figure 3) or the input parameters of the correlation structure (e.g. Figure 4). In each panel, input kurtosis/skewness parameters are put on the horizontal axis, while the vertical axis reports RMSE mean values. According to the procedure adopted to compute Tukey U.I.s (Subsect. 3.4), comparisons among RMSE means can be made simultaneously, at fixed levels of the inactive conditions, within every column in the graphs (and in the tables), that is, with regard to a same method. Tukey U.I.s that do not overlap in such comparisons pick out the RMSE means that are significantly different at 0.05 level, as well as the active conditions (and their levels) under which significant differences are observed for a specific imputation method.

JT test results pertain to the performance comparison analysis, addressed here to detecting the most effective method from among *FIM*, *FIP* and *IPCA* under the various simulation settings. Since by the descriptive analysis *missForest* showed a poor or not readily understandable performance in many of the considered situations, we decided to confine these comparisons to *FIM*, *FIP* and *IPCA*. In line with this, to detect the best method from among these three, we carried out six separate one-sided JT tests at the 0.05 significance level to test the null hypothesis:

$$H_0 : F_{FIM}(x) = F_{FIP}(x) = F_{IPCA}(x), \qquad \forall x \geq 0, \tag{12}$$

where $F_{\mathsf{m}}(\cdot)$ in (12) denotes the empirical RMSE distribution function of method m, against each of the following six ordered alternatives:

$$\begin{aligned}
1 &= H_1 : F_{FIM}(x) \leq F_{FIP}(x) \leq F_{IPCA}(x) \\
2 &= H_1 : F_{FIP}(x) \leq F_{FIM}(x) \leq F_{IPCA}(x) \\
3 &= H_1 : F_{IPCA}(x) \leq F_{FIP}(x) \leq F_{FIM}(x) \\
4 &= H_1 : F_{FIM}(x) \leq F_{IPCA}(x) \leq F_{FIP}(x) \\
5 &= H_1 : F_{FIP}(x) \leq F_{IPCA}(x) \leq F_{FIM}(x) \\
6 &= H_1 : F_{IPCA}(x) \leq F_{FIM}(x) \leq F_{FIP}(x),
\end{aligned} \tag{13}$$

with at least a strict inequality for any x. Tables regarding the JT test contain the number of the "most significant" ranking according to the numbering of hypotheses (13) if the result is significant, or a "n.s." label if the result is not significant. In addition, cells in the tables are coloured differently depending on which method appears as the first in the significant ranking, thus proving to have smaller imputation errors. Grey cells denote rankings 1 and 4, where *FIM* proves to have the best performance. Light-grey cells indicate rankings 2 and 5, where *FIP* is the best. Blank cells stand for rankings 3 and 6, with *IPCA* as the best.

The four exploratory studies listed in Table 4 are covered in Subsects. 4.1–4.4, respectively, while results of supplementary studies are summarised in Subsect. 4.5.

4.1. *Simulation study 1: Symmetry and kurtosis with equicorrelation*

Table 5 reports experimental conditions and correlation indices for the SyKu-ECor study (complete table in SM), whose data patterns were generated through *MEP* distributions in the presence of equicorrelation. Three different types of distributions were involved, i.e. leptokurtic ($\kappa = 1$), normal ($\kappa = 2$) and platykurtic ($\kappa = 14$), in order to study the effect of kurtosis on the imputation performance. Regarding the correlation structure, equicorrelation matrices were considered with three different levels of magnitude for ρ, i.e. uncorrelation ($\rho = 0$), positive low ($\rho = 0.3$) and positive high ($\rho = 0.7$) correlations. In this study, correlation indices (Subsect. 3.2) are given basically by the relative eigenvalues (4), the first two and the last of which are displayed in Table 5. For the fact that the ρs coincide, we have: $\mathrm{sd}_{\mathrm{abs}} = 0$, so that the absolute skewness index in (7) is undefined. Moreover, since the ρs are also positive, the maximum eigenvalue is unique, while the other $p - 1$ eigenvalues assume the same value (Subsect. 3.2).

Figure 3 displays dot plots of RMSE mean values (with 95%-Tukey U.I.s) of *FIM*, *FIP*, *IPCA* and *missForest* for each combination of levels (ρ, κ), with $p = 5$ (first row of panels) and $p = 10$ variables (second row), and $n = 1000$ units, while Table 6 contains the corresponding numeric results (all the omitted tables and figures are provided in SM). *IPCA* shows the best performance because it has the lowest RMSE mean values under almost all the simulation scenarios. Overall, *FIM* and *FIP* tend to have a similar and intermediate performance between *IPCA* and *missForest* for small ρs ($\rho = 0; 0.3$). When $\rho = 0.7$, their performance tends however to worsen, especially *FIM* with $p = 10$.

Pattern-impact analysis is based on the 2-way full ANOVA model that derives from equation (10) by including correlation structure and kurtosis as active conditions, and with the levels specified in Table 5 (input parameters). RMSE means can be compared each other within every panel in Figure 3. The main results are summed up as follows:

- *Correlation structure effect*: It can be noticed that, for every value of κ with fixed

p, Tukey U.I.s do not overlap in any of the panels in Figure 3. It means that the RMSE mean values of all the methods tend to decrease as the correlation level becomes stronger, or equivalently, as the first relative eigenvalue increases (Table 5). This occurs whatever the value of the kurtosis parameter is, as the number of variables and units varies (see SM also).

- *Kurtosis effect*: The four methods behave slightly differently. Specifically,
  - when $\rho = 0$, *FIM* and *FIP* tend to have significantly smaller errors in the presence of leptokurtic ($\kappa = 1$) rather than normal ($\kappa = 2$) or platykurtic ($\kappa = 14$) data. With higher values of ρ, *FIP* tends to perform better under normal data (although sometimes $\kappa = 2$ is not significantly different to $\kappa = 14$), and *FIM* under platykurtic data, while both *FIM* and *FIP* seem to perform worst in the presence of leptokurtic data (especially *FIM* when $\rho = 0.7$). However, when either $p = 3$ or $n = 500$, other kinds of trends occur. For instance, with $p = 3$ and $n = 500$, *FIM* and *FIP* always perform better for $\kappa = 14$ (see SM).
  - *IPCA* tends to have smaller RMSE mean values when data are normal, although it proves to be less sensitive than the other methods to kurtosis of data distribution when ρ is low. For instance, in the panel in the first row ($p = 5$) and third column of Figure 3, Tukey U.I.s overlap when $\rho = 0$ and $\rho = 0.3$ (see also Table 6). Other trends can however be observed in the presence of lower dimensionality of data (e.g. when $p = 3$ and $n = 500$ with $\rho \geq 0.3$, results are better for platikurtic data).
  - *missForest* has a less clear trend over the various simulation settings. Nonetheless, mostly it appears to perform better with normal data as well as to be more severely affected by platykurtic data ($\kappa = 14$) for higher values of ρ and p.

Regarding the performance comparison analysis confined to *FIM*, *FIP* and *IPCA*, JT test results are displayed, for all p and n with 20% of MCAR values, in Table 7. With only two exceptions occurring when $\rho = 0$ and $p = 10$, *IPCA* is always the best imputation method compared to *FIM* and *FIP*. In addition, the prevalence of ranking 3, instead of ranking 6, reveals that *FIM* more frequently performs worse than the other two methods. In the other omitted tables with 5% and 10% of MCAR values reported in SM, *IPCA* proves to be the best method in almost all the scenarios.

4.2. *Simulation study 2: Skewness and equicorrelation*

Table 8 summarises experimental conditions and correlation indices of the SK-ECor study (skewness with equicorrelation; complete table in SM). It is the first of the three exploratory studies that involve data patterns generated by *MSN* distributions. These studies mostly differ for the structure assigned to the input correlation matrix $\mathbf{\Omega}$ (Subsect. 3.1), and consequently for the structure taken over by the output correlation matrix $\mathbf{R}$. Regarding the shape of data distribution, the same four levels of skewness parameter α are considered for every variable (i.e. $\alpha_j = \alpha$, for all $j = 1, \ldots, p$), and range from slight skewness ($\alpha = 1$) to stronger skewness ($\alpha = 30$), passing through two intermediate levels ($\alpha = 4$ and $\alpha = 10$).

As already pointed out (Subsect. 3.2), input and output parameters in the *MSN* case are linked together by complex patterns of variations. For instance, values of the output correlation coefficient ρ depend on both the input parameters ω and α along with the number p of variables. With ω and p fixed, ρ varies, therefore, with

α, although in general, such a variation is of small magnitude [32]. For this reason, in the correspondence between the input and output correlations reported in Table 8, an approximate value of ρ is provided for each pair (ω, p), thus meaning that a small range of values of ρ corresponds to the set of values of α chosen for the study.

The SK-ECor pattern is characterised by having matrix $\mathbf{\Omega}$ with the same non-negative ω for all pairs of variables, and matrix $\mathbf{R}$ preserving the same equicorrelation structure of $\mathbf{\Omega}$ although with different values (Table 8). For α in the range $[1, 30]$ and $p = 5$, we thus have three levels of magnitude of ρ: Negative low ρs (≈ -0.1), resulting from $\omega = 0$; positive low ρs (≈ 0.2), corresponding to $\omega = 0.5$; positive moderate ρs (≈ 0.6), related to $\omega = 0.8$, (Table 8). Moreover, in this case also, correlation indices are substantially given by the relative eigenvalues (the first two and the last provided in Table 8) because of the equicorrelation structure. In particular, in the presence of negative low ρs ($\omega = 0$), the maximum eigenvalue is no more unique (Subsect. 3.2), and it coincides with the subsequent first $p-2$ eigenvalues (Table 8). This is the main difference between data built up with negative low ρs and those ones with positive ρs.

Dot plots of RMSE mean values with 95%-Tukey U.I.s are displayed in Figure 4, with $p = 5$ variables, $n = 1000$ units and 20% of MCAR values. Panels in the rows correspond to the three levels of input/output correlation described above. According to the descriptive analysis, two different trends can be noticed. Whatever the value of α, in the presence of negative low ρs ($\omega = 0$), *FIM* has the smallest RMSE means, followed by *FIP*, while *IPCA* and *missForest* perform poorly. In the case of positive low ρs ($\omega = 0.5$), *FIM* and *FIP* perform similarly, although *FIP* is slightly better ($\alpha = 1$ excepted, where *IPCA* is better). When ρ assumes moderate values ($\omega = 0.8$), the trend becomes opposite. *FIM* produces the worst results (apart from *missForest* with $\alpha = 10$), *IPCA* is the best, and *FIP* performs very similarly to *IPCA*.

Pattern-impact analysis is based on the 2-way full ANOVA model (10) with input correlation ω and skewness α as active conditions, whose levels are specified in Table 8 (input parameters). RMSE means can be compared across the panels inside a same column of Figure 4, or equivalently within the columns of Table 9. Regarding the correlation structure effect, once again RMSE means tend to shrink in value as the correlation level increases. Notably, for each α and each method, Tukey U.I.s of $\omega = 0$ never overlap with those of $\omega = 0.8$. Regarding the skewness effect, a sort of dichotomy between $\alpha = 1$ and $\alpha \geq 4$ appears in Figure 4, especially for *FIM*, *FIP* and *IPCA*. In particular, when $\omega = 0$ and $\omega = 0.8$, Tukey U.I.s referred to $\alpha = 1$ do not overlap with those of $\alpha \geq 4$. In the case of negative low ρs ($\omega = 0$), *FIM*, *FIP* and *IPCA* perform better when data are more skew ($\alpha \geq 4$), while in the presence of higher values of ρ ($\omega = 0.8$), they tend to perform better for less skew data ($\alpha = 1$).

The case of positive low ρs ($\omega = 0.5$) represents an intermediate situation. *FIM* and *FIP* prove to be less sensitive to skewness – their U.I.s overlap for every ω –, while *IPCA* performs clearly better for $\alpha = 1$. On the other hand, *missForest* has a relatively strange performance. The skewness effect is evidently not monotone, especially for $\omega = 0$ and $\omega = 0.8$.

Lastly, JT test results in Table 10 concern the performance comparison analysis carried out for detecting the best method from among *FIM*, *FIP* and *IPCA*. Clearly separated trends can be read. In the presence of negative low ρs ($\omega = 0$), the best method is *FIM*, followed by *FIP* (ranking 1 in (13)). Few exceptions are observed when $\alpha = 1$ and data dimensionality is low ($p = 3$ with $n = 500; 1000$, and $p = 5$ with $n = 500$), where the best method is *IPCA*. As for positive low ρs ($\omega = 0.5$), *IPCA* has the best performance, although *FIP* turns out to be the best with $p = 5$ and more skew data ($\alpha \geq 4$). Finally, in the presence of higher values of ρs ($\omega = 0.8$), *IPCA* has always the best performance (tables concerning 5% and 10% of MCAR values in SM).

4.3. *Simulation study 3: Skewness and positive–negative correlations*

Experimental conditions, correspondence of input–output correlations and correlation indices regarding the SK-PNCor study (skewness with positive and negative correlations; complete table in SM) are reported in Table 11. Data under this pattern are generated with matrix $\mathbf{\Omega}$ containing the same ω in absolute value but with alternating sign. This produces an output matrix $\mathbf{R}$ having positive and negative correlations (PNCor) of a very similar magnitude, according to the formal structure described in Table 11. For example, with $p = 5$ and $\omega = 0.2$ matrix $\mathbf{\Omega}$ is:

$$\mathbf{\Omega} = \begin{pmatrix} 1 & 0.2 & -0.2 & 0.2 & -0.2 \\ 0.2 & 1 & -0.2 & 0.2 & -0.2 \\ -0.2 & -0.2 & 1 & -0.2 & 0.2 \\ 0.2 & 0.2 & -0.2 & 1 & -0.2 \\ -0.2 & -0.2 & 0.2 & -0.2 & 1 \end{pmatrix}.$$

Next, by setting: $\alpha = 1$ the following matrix $\mathbf{R}$ is obtained:

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

where, according to the notation of Table 11, $\rho_1 = -0.299$, $\rho_2 = 0.088$, and $\rho_3 = 0.163$.

Overall, values of ρ remain quite stable as p and/or α vary, while they are mainly sensitive to variations of ω. When $p = 5$, levels of output ρ derived from the values chosen here for ω, i.e. $\omega = 0.2; 0.5; 0.8$, are reported in Table 11 under the correspondence of input-output correlations. The three correlation structures thus obtained (i.e. PN low, PN moderate and PN high) differ both in magnitude and in the extent of unbalance among absolute correlations. In particular, in the case of PN moderate ρs, correlation coefficients are more concentrated on higher absolute values ($\text{skew}_{\text{abs}} \approx -0.75$) than PN low ρs ($\text{skew}_{\text{abs}} \approx -0.53$) and PN high ρs ($\text{skew}_{\text{abs}} \approx 0$).

An excerpt of the results concerning the SK-PNCor study with $p = 5$ variables, $n = 1000$ units and 20% of MCAR values is provided in Figure 5 and Table 12. The three rows of panels in Figure 5 correspond to the three correlation structures described above. Although most remarks would be very similar to those advanced for the SK-ECor pattern (Subsect. 4.2), it is worth pointing out that: (a) in the presence of PN low ρs ($\omega = 0.2$), *FIM* has the smallest RMSE means, and *missForest* the highest. On the other hand, *IPCA* shows the best performances with PN moderate-high ρs ($\omega = 0.5; 0.8$), while *FIM* performs very poorly in the presence of PN high 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.

Pattern-impact analysis is based again on the 2-way full ANOVA model (10) having input correlation ω and skewness α as active conditions with the levels given in Table 11. Once again, as the correlation becomes stronger, or as the first relative eigenvalue of $\mathbf{R}$ increases, RMSE means decrease for all the methods. In particular, for each α, Tukey U.I.s of every method do not overlap over the different values of ω. As for the skewness effect, again we have the dichotomy: $\alpha = 1$ vs. $\alpha \geq 4$, just observed in the SK-ECor study (Subsect. 4.2), and it is shared, more or less, by all the methods. Specifically, (a) 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 ($\alpha \geq 4$); (b) in the presence of PN high ρs ($\omega = 0.8$), they tend to perform better for less skew data ($\alpha = 1$). In most situations, with ω fixed, Tukey U.I.s overlap for $\alpha \geq 4$, but they do not in the comparison between $\alpha = 1$ and $\alpha \geq 4$.

JT test results referred to the performance comparison analysis are provided in Table 13. In all the considered scenarios with PN moderate or PN high ρs ($\omega = 0.5; 0.8$), *IPCA* performs better than *FIM* and *FIP*. In the presence of low ρs ($\omega = 0.2$) with $p = 5$, *FIM* is the best method, followed by *FIP*, whereas for $p = 10$, *IPCA* proves to have the best performances (tables with 5% and 10% of MCAR values in SM).

4.4. *Simulation study 4: Skewness and unbalanced correlations*

Table 14 reports the basic features of the SK-UnbCor study (skewness and unbalanced correlations), which is treated in short also in [12]. Input matrix $\boldsymbol{\Omega}$ contains a same negative value $\omega_1 = -\omega$ in the first column/row, and a same positive value $\omega_2 = \omega/c$ in the other columns/rows ($\omega = 0.2; 0.5; 0.8$, $c = 1; 1.25; 1.5$). Coefficient c thus regulates the extent of unbalance among input correlations in $\boldsymbol{\Omega}$, which also reflects, yet not in a linear way, among the output correlations in $\mathbf{R}$. In this way, two distinct values of ρ, i.e. ρ_1 and ρ_2, are obtained in connection with ω_1 and ω_2, respectively. As before, values of ρ mostly vary with ω rather than α, so that from the various combinations of values of ω and c we obtain the six different correlation structures of $\mathbf{R}$ provided in Table 14 (see also [12]). They mainly differ in the overall magnitude of the absolute correlations (measured by $\bar{\rho}_{\text{abs}}$), and in the extent of unbalance between $\rho_{\min}$ and $\rho_{\max}$, while, overall, correlation coefficients are more concentrated on low absolute correlations ($\text{skew}_{\text{abs}} > 0$).

From Figure 6 it can be observed that: (a) In the presence of the "negative low and nearly null" ρs ($\omega = 0.2$, first three rows of panels), *FIM* shows the smallest RMSE mean values, followed by *FIP*, while *IPCA* and *missForest* have a worse performance; (b) in the presence of the structures "negative moderate ρs" with positive low ρs ($\omega = 0.5$, $c = 1; 1.25$) and nearly null ρs ($\omega = 0.5$, $c = 1.5$), *FIP* has the lowest RMSE mean values with few exceptions (i.e. $\alpha = 1$, $c = 1$, where *IPCA* performs better, and: $\alpha \geq 4$, $c = 1.5$, where *FIM* performs better); (c) the three structures with negative high ρs ($\omega = 0.8$, last three rows of panels) highlight the main differences among the methods. In particular, there is an inversion of trend moving from the "negative high and positive high" ρs ($c = 1$), where *IPCA* has the smallest RMSE mean values and *FIM* the highest, to the "negative high and positive low" ρs ($c = 1.5$), where *FIM* performs best and *IPCA* worst. The row of panels related to the "negative high and positive moderate" ρs ($c = 1.25$) displays an intermediate situation, where *missForest*, *FIM* and *FIP* have a very similar good performance, while *IPCA* performs worst.

Pattern-impact analysis is based on the 3-way full ANOVA model derived from equation (10) by including input correlation ω, coefficient c and skewness α as active conditions with the levels specified in Table 14 (input parameters). Results of this analysis are shown in Figure 6 in the form of 95%-Tukey U.I.s, and in Table 15. Comparisons among RMSE mean values can be made within a same column of Figure 6 (and Table 15), i.e. concerning a same method. As for the correlation structure effect, once again smaller imputation errors are observed in the presence of higher levels of correlation. Notably, Tukey U.I.s of every method do not overlap over the different values of ω, for fixed c and α. Regarding the skewness effect, the dichotomy $\alpha = 1$ vs. $\alpha \geq 4$ is still 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, (its Tukey U.I.s overlap for every c). On the other hand, when $\omega = 0.8$ all the methods tend to perform better for less skew ($\alpha = 1$) than more skew ($\alpha \geq 4$) data. The case $\omega = 0.5$ looks intermediate. Both the trends are observed, in particular the first (better for more skew data) when $c = 1.5$, and the second (better for less skew data) when $c = 1$.

Finally, JT test results related to the performance comparison analysis are provided in Table 16. *FIM* performs better than the other methods in almost all the considered scenarios, especially when data are more skew-distributed and correlation coefficients are more unbalanced ($c = 1.5$). *IPCA* works better with more balanced moderate or high ρs (i.e. $\omega = 0.5; 0.8$ with $c = 1$), with the only exceptions of $n = 1000$, $\alpha \geq 4$, and $\omega = 0.5$, where *FIP*, followed by *FIM*, performs better than *IPCA*.

4.5. *Supplementary simulation studies and computational efficiency*

As sketched before in Table 4, several supplementary studies were undertaken in order to examine additional simulation scenarios. These latter were planned *a posteriori*, in the light of the main findings resulted from the exploratory studies of Subsects. 4.1–4.4.

The SyKu shape was also considered in the presence of negative equicorrelations (SyKu-NegECor pattern) with magnitude similar to the study 1 of the SK-ECor pattern (Subsect. 4.2). Moreover, we 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 (Tables 11 and 14). Finally, the UnbCor structure was also set up with all positive correlations for both the SyKu and SK shapes (i.e. SyKu-PosUnbCor and SK-PosUnbCor patterns). All the tables and figures concerning these further experimental conditions and simulation results are provided in SM. The main findings are summed up as follows:

- *SyKu-NegECor pattern.* To have consistent matrices $\mathbf{R}$ such that they were positive-definite, values of ρ could not be less than nearly -0.2. We considered, therefore, two levels of ρ, i.e. $\rho = -0.1; -0.2$. Once again, *FIM* and *FIP* tend to have smaller errors in the presence of leptokurtic data ($\kappa = 1$) when absolute correlation values are lower (i.e. when $\rho = -0.1$), otherwise imputation errors are smaller when data are platykurtic ($\kappa = 14$). This same trend is shared also by *missForest*, while *IPCA* performs better in both the cases when data are leptokurtic. Finally, by the JT test, *FIM* always proves to be the best method in all these additional scenarios, followed by *FIP* (see SM).
- *SyKu-PNCor pattern.* As regards the correlation effect, results obtained are very similar to the SK-PNCor pattern. RMSE mean values tend to decrease as the mean absolute correlation (or as the first relative eigenvalue) increases. Regarding the kurtosis effect, *FIM* and *FIP* perform better in the presence of platykurtic data, while *IPCA* with normal data. Depending on the case, *missForest* may perform better in the presence of either normal or platykurtic data (see SM). Concerning the performance comparison analysis, JT test proves that *FIM* is the best method in the presence of low mean absolute correlation values (i.e.

$\bar{\rho}_{\text{abs}} < 0.3$), otherwise *IPCA* performs best.

- *SyKu-UnbCor pattern.* The correlation effect is similar as described above. As for the kurtosis effect, results obtained evidence that *FIM* and *FIP* tend to perform better in the presence of platykurtic data, especially for higher mean absolute correlation values. On the other hand, *IPCA* performs better in the presence of leptokurtic data, while *missForest* tends to perform better when data are platykurtic with low mean absolute correlation values, or when data are normal with higher mean absolute correlation values. Once again, JT test reveals that overall *FIM* is the best method in the presence of lower correlation values, while *IPCA* is the best in the presence of higher correlation values.
- *SyKu-PosUnbCor pattern.* Apart from few exceptions, both *FIM* and *FIP* tend to perform better in the presence of normal or platykurtic data, *missForest* with normal data and *IPCA* with leptokurtic data. Overall, by the JT test, *FIM* confirms as the best method in the presence of lower correlation values, otherwise *IPCA* is the best method.
- *SK-PosUnbCor pattern.* Results are very similar to those obtained under the corresponding SK-UnbCor pattern (Subsect. 4.4). Therefore, it seems that it is the extent of unbalance among correlations along with the magnitude of absolute correlations, rather than the sign, to be the most important discriminant elements among the imputation methods.

As a final result, Table 17 reports a general indication of the average times (in seconds) and standard deviations of the four methods under each combination of the number of variables (p) and units (n), and percentages of MCAR values. Times were recorded in 1000 simulation runs carried out by a PC with Windows 8.1 Pro 64-bit operating system, i7-4500U CPU processor, 1.80 GHz clock frequency, and 8.00 GB RAM. The type of distribution (*MEP* or *MSN*) and their parameters (e.g., correlations of variables) turned out to be substantially uninfluential. For simplicity, we have then collapsed all the times recorded in the various exploratory studies by computing overall averages within each triple ($p, n,$ % MCAR values). The fastest method is *IPCA*, while the slowest is *missForest*. As expected, execution times tend to grow with the increase of dimensionality of data and percentages of MCAR values, but for *missForest* this trend appears to be not monotonic.

4.6. *Discussion*

Taking into account the previous descriptive and inferential analyses that concerned *FIM*, *FIP* and *IPCA*, the main impressive findings of the simulation studies are now provided in the form of practical hints for users pertaining to the considered data structures:

- *Equicorrelation patterns.* We have seen that if data have a symmetric distribution and correlation coefficients are positive, *IPCA* is the method with the smallest imputation errors. Otherwise, if correlation coefficients are negative, *FIM* turns out to be the best one. On the other hand, if data are skew-distributed, *FIM* tends to perform better in the presence of negative low correlations and *IPCA* in the presence of positive moderate-high correlations, while *FIP* has an intermediate performance. Besides the mean absolute correlation $\bar{\rho}_{\text{abs}}$, the relative eigenvalues are the other measures used to synthesize the output correlation matrices $\mathbf{R}$. Regardless the symmetric or skew shape of the data, we have noticed that the second relative eigenvalue RelEig_2 can be used as a criteria helpful in choosing a method rather than another. In line with this, *FIM* has proved overall to perform better than *FIP* and *IPCA* when: $\text{RelEig}_2 > \frac{1}{p}$, where p is the number of variables. 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 these two, while *IPCA* is the best method when $\text{RelEig}_2 < \frac{1}{p}$. Nonetheless, if data are skew, the above condition concerning RelEig_2 may be relaxed, e.g. *FIP* might again be the best method even if RelEig_2 is slightly under the threshold of $\frac{1}{p}$. This means that for *FIM* and *FIP* skewness of data implies better results also in the presence of slightly higher positive correlations. Finally, as RelEig_1 approaches to 1 and RelEig_2 to 0, that is, with the increasing of the correlation magnitude, *IPCA* tends to produce the best results.

- *Positive-negative correlation patterns.* Data generated with these patterns present output correlation matrices having both positive and negative entries of a similar magnitude along with a more or less marked extent of unbalance among them. The performance of *FIM*, *FIP* and *IPCA* have now to be related to a plurality of correlation indices, in particular RelEig_2 along with $\bar{\rho}_{\text{abs}}$ and skew_{abs}. Overall, *FIM* has shown better performances than *FIP* and *IPCA* when $\text{RelEig}_2 \geq \frac{1}{p}$ and $\bar{\rho}_{\text{abs}} \leq 0.2$. In any case, even if RelEig_2 is slightly below the threshold of $\frac{1}{p}$ and $\bar{\rho}_{\text{abs}} > 0.2$, *FIM* might still be the best method, provided that the correlation coefficients are not too much unbalanced towards higher absolute values. As a rule of thumb we refer to: $\text{skew}_{\text{abs}} > -0.6$, although it would require to be validated in more general contexts. In its turn, *FIP* has proved in most scenarios to have an intermediate performance between *FIM* and *IPCA* (e.g. rankings 1 and 3 of the hypotheses (13)). We argue, therefore, that under a wider spectrum of graduated levels of the correlation coefficients *FIP* might result as the best performing method. On the other hand, *IPCA* confirms to certainly have the best performances in the presence of lower values of RelEig_2 ($< \frac{1}{p}$), especially when $\bar{\rho}_{\text{abs}} \geq 0.4$ and $\text{skew}_{\text{abs}} \leq -0.6$ (i.e. stronger unbalance towards higher absolute correlations).
- *Unbalanced correlation patterns.* Remarks similar to the positive-negative correlation patterns can be advanced. Once again, the performance of the methods seems mostly tied to the values assumed by RelEig_2, $\bar{\rho}_{\text{abs}}$ and skew_{abs} jointly considered. In particular, *FIM* performs better than *FIP* and *IPCA* when $\text{RelEig}_2 \geq \frac{1}{p}$, or when: $\bar{\rho}_{\text{abs}} < 0.4$ with correlation coefficients unbalanced towards lower absolute values ($\text{skew}_{\text{abs}} > 0$). That is to say that a low/moderate magnitude of the mean absolute correlation is not sufficient for *FIM* to perform better than the other two methods, because unbalance among the correlation coefficients has to be towards lower absolute correlations. A counterexample is provided by the scenario with: $\rho_1 = 0.18$ and $\rho_2 = 0.50$ under both the SyKu-PosUnbCor (Table S81 in SM) and SK-PosUnbCor (Table S84 in SM) patterns. Here we have: $\bar{\rho}_{\text{abs}} = 0.37$, 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 to perform better than the others when the above considered indices are very close to those thresholds, or slightly overcome them. Finally, *IPCA* shows a performance better than *FIM* and *FIP* in the presence of higher magnitudes of correlations ($\bar{\rho}_{\text{abs}} \geq 0.6$, say), with a stronger unbalance towards absolute larger values ($\text{skew}_{\text{abs}} < 0$).

As a final remark, under the considered experimental conditions we have noticed that the reference values given for RelEig_2 and $\bar{\rho}_{\text{abs}}$ might be relaxed depending on the shape of the data (i.e. symmetric or skew distributions). For instance, as just observed, *FIM* and *FIP* tend to improve their performance in the presence of skew data, while *IPCA* with symmetric data.

5. Conclusions

In the presence of such a vast literature concerning the imputation methods, one of the main problem to cope with is how to choose the best performing method depending on the real situation at hand. Within the scope of our work, we have tried to address this point by comparing, under several distinct data patterns, the three nonparametric methods: *ForImp* (in the two variants *FIM* and *FIP*), *IPCA* and *missForest*, which we have described in short in Sect. 2. We have then confined the performance comparison analysis to the methods that have shown the most convincing results in the descriptive analyses, namely *FIM*, *FIP* and *IPCA*. In this regard, we have decided not to take *missForest* into account because it did not show satisfactory or readily understandable results in many of the considered simulation scenarios. However, it is worth pointing out that *missForest* is an imputation method designed for mixed-type data, and this aspect could explain its lacking effectiveness for quantitative data under the experimental conditions considered here.

A crucial point in our simulation comparison was the choice of how to simulate artificial data such that they were representative of real situations as much as possible. To this end, we set up a variety of data patterns that differed in shape as well as in the correlation structure of variables. To obtain different items of shape we relied on two families of multivariate distributions,i.e. the multivariate exponential power (*MEP*) and the multivariate skew-normal (*MSN*), respectively, both of which include the multivariate normal distribution as a special case. In such a way, we could set up either symmetric leptokurtic or platykurtic data (the SyKu shape, generated with *MEP*) or asymmetric data (the SK shape, generated with *MSN*). As for the correlation of variables, we considered the three different structures given by equicorrelation (the ECor structure), positive-negative correlations (the PNCor structure) and unbalanced correlations (the UnbCor structure). Data patterns were then provided by the combinations of the two items of shape with these three correlation structures.

We argue that the simulation method we have designed to analyse the performance in an imputation problem could have a more general application scope, e.g. concerning other kinds of statistical techniques and investigations. In our case, the extensive simulation comparison studies we have undertaken permitted us to give some useful indications to the potential users about the choice of the most performing imputation method, among the ones considered, in a nonparametric, single imputation perspective and with respect to multiple data patterns close to real situations.

References

[1] Efron B. Bootstrap methods: another look at the jackknife. Ann Stat. 1979;7(1):1-26.

[2] Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed. New York: Wiley; 2002.

[3] Schafer JL. Analysis of incomplete multivariate data. London: Chapman and Hall/CRC; 1997.

[4] Molenberghs G, Kenward MG. Missing data in clinical studies. Chichester: Wiley; 2007.

[5] Haziza D. Imputation and inference in the presence of missing data. In: Pfeffermann D, Rao CR, editors. Sample surveys: design, methods and applications. Handbook of Statistics; 29A. Amsterdam: North Holland; 2009. p. 215-246.

[6] Bello AL. Choosing among imputation techniques for incomplete multivariate data: a simulation study. Commun Stat-Theor M. 1993;22(3):853-877.

[7] Bello AL. A simulation study of imputation techniques in linear quadratic and kernel discriminant analyses. J Stat Comput Sim. 1993;48(3-4):167-180.

[8] Marella D, Scanu M, Conti PL. On the matching noise of some nonparametric imputation procedures. Stat Probabil Lett. 2008;78(12):1593-1600.

[9] Ning J, Cheng PE. A comparison study of nonparametric imputation methods. Stat Comput. 2012;22(1):273-285.

[10] Tutz G, Ramzan S. Improved methods for the imputation of missing data by nearest neighbor methods. Comput Stat Data Anal. 2015;90:84-99.

[11] Ferrari PA, Annoni P, Barbiero A, Manzi G. An imputation method for categorical variables with application to nonlinear principal component analysis. Comput Stat Data Anal. 2011;55:2410-2420.

[12] Solaro N, Barbiero A, Manzi G, Ferrari PA. A sequential distance-based approach for imputing missing data: Forward Imputation. Adv Data Anal Classi. 2017;11:395-414.

[13] Nora-Chouteau C. Une méthode de reconstitution et d'analyse de données incomplètes [dissertation]. Paris: Université Pierre et Marie Curie; 1974.

[14] Greenacre M. Theory and applications of correspondence analysis. London: Academic Press; 1984.

[15] Josse J, Pagès J, Husson F. Multiple imputation in principal component analysis. Adv Data Anal Classi. 2011;5:231-246.

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

[17] Breiman L. Random forests. Mach Learn. 2001;45:5-32.

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

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

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

[21] Mardia, KV. Measures of multivariate skewness and kurtosis with applications. Biometrika. 1970;57(3):519-530.

[22] Solaro N. Random variate generation from multivariate exponential power distribution. Statistica & Applicazioni. 2004;II(2):25-44.

[23] Azzalini A. Package 'sn': The skew-normal and related distributions, such as the skew-t. 2017 - [R package version 1.5-0]. Available from: https://CRAN.R-project.org/package=sn

[24] Seber GAF. Multivariate observations. New York: Wiley; 1984.

[25] Kaiser HF. A measure of the average intercorrelation. Educ Psychol Meas. 1968;28:245-247.

[26] Solaro N, Barbiero A, Manzi G, Ferrari PA. Package 'GenForImp': The Forward Imputation - a sequential distance-based approach for imputing missing data. 2015 - [R package version 1.0.0]. Available from: http://CRAN.R-project.org/package=GenForImp

[27] Husson F, Josse J. Package 'missMDA': Handling missing values with multivariate data analysis. 2017 - [R package version 1.11]. Available from: http://CRAN.R-project.org/package=missMDA

[28] Stekhoven DJ. Package 'missForest': Nonparametric missing value imputation using random forest. 2016 - [R package version 1.4]. Available from: http://CRAN.R-project.org/package=missForest

[29] Hochberg Y, Tamhane AC. Multiple comparison procedures. New York: John Wiley & Sons; 1987.

[30] Hollander M, Wolfe DA. Nonparametric statistical methods. 2nd ed. New York: Wiley-Interscience; 1999.

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

[32] Solaro N, Barbiero A, Manzi G, Ferrari PA. A comprehensive simulation study on the Forward Imputation. Milan (IT): Università degli Studi di Milano; 2015. (DEMM working paper; no. 2015-04). Available from: https://ideas.repec.org/p/mil/wpdepa/2015-04.html

Table 1. The *ForImp* algorithm with the *FIM* and *FIP* methods.

1. *Step* $k = 0$: Split $\mathbf{X}$ into a $(n_0 \times p)$-dimensional submatrix $\mathbf{X}_0$ free of missing data ($p \leq n_0 < n$), and K submatrices $\mathbf{X}_k$ of dimension $(n_k \times p)$, with $k < p$ the number of missing values in each row (it is not necessary that $n_k > 0$ for all $k = 1, \ldots, K$, and $n_k > p$ for all $k = 0, 1, \ldots, K$. See [12]).

2. *Step* $k \geq 1$:
If *FIP* = *True*: Perform a PCA on $\mathbf{X}_{k-1}$ (i.e. the complete submatrix available up to the $(k-1)$-th step) to obtain eigenvalues $\lambda_s^{(k-1)}$ and eigenvectors $\boldsymbol{\omega}_s^{(k-1)}$ with generic loading $\omega_{js}^{(k-1)}$, $(j, s = 1, \ldots, p)$.
PCA input matrix can be either the variance-covariance matrix or the correlation matrix of $\mathbf{X}_{k-1}$.

3. If *FIP* = *True*: Compute Pseudo-Principal Components (PPCs), denoted with $\tilde{C}_s$, for the incomplete units in $\mathbf{X}_k$ (i.e. the submatrix with k missing entries in its row) and the complete units in $\mathbf{X}_{k-1}$ involving only common complete variables X_l, with $l \notin \iota_k$ and ι_k the set formed by the k-combinations of the p indices of variables that present missing values in the rows of $\mathbf{X}_k$. PPCs are then given by:
$\tilde{C}_{s(\iota_k)}^{(k)} = \sum_{\substack{l=1 \\ l \notin \iota_k}}^{p} \omega_{ls}^{(k-1)} X_l^{(k)}$ for the incomplete units in $\mathbf{X}_k$, and:
$\tilde{C}_{s(\iota_k)}^{(k-1)} = \sum_{\substack{l=1 \\ l \notin \iota_k}}^{p} \omega_{ls}^{(k-1)} X_l^{(k-1)}$ for the complete units in $\mathbf{X}_{k-1}$, $(s = 1, \ldots, p)$.

4a. If *FIP* = *True*: Compute the weighted Minkowski distance d_r of order r, $r \geq 1$, (with weights given by the square root of the eigenvalues $\lambda_s^{(k-1)}$ divided by their total sum), between each incomplete unit in $\mathbf{X}_k$ and each complete unit in $\mathbf{X}_{k-1}$ using their PPC scores computed as above.

4b. Else if *FIM* = *True*: Compute the Mahalanobis distance d_M between each incomplete unit in $\mathbf{X}_k$ and each complete unit in $\mathbf{X}_{k-1}$ using the values of the common complete variables X_l, $(l \notin \iota_k)$. The variance-covariance matrix involved in the formula of d_M is computed using the common complete variables X_l.

4c. If *FIP* or *FIM* = *True*: The set of donors for the incomplete unit $u_i^{(k)}$ is formed by the first $q100\%$ of the complete units, available up to the $(k-1)$-th step, that correspond to the q-th quantile $d_{q,i}$ of the Minkowski (*FIP*) or Mahalanobis (*FIM*) distances $(0 < q < 1;\ i = 1, \ldots, n_k)$.

5. For each incomplete unit $u_i^{(k)}$, the missing value x_{ij} on variable X_j is imputed with the weighted mean:
$\tilde{x}_{ij}^{(k)} = \frac{\sum_{\delta=1}^{n_\delta} x_{\delta j}^{(k-1)} \frac{1}{d_{\delta i}}}{\sum_{\delta=1}^{n_\delta} \frac{1}{d_{\delta i}}}$, $\forall j \in \iota_k$, where n_δ is the total number of donors for $u_i^{(k)}$ and $d_{\delta i}$ is the distance between the δ-th donor and unit $u_i^{(k)}$.
Repeat the imputation for each $i = 1, \ldots, n_k$ to obtain the imputed matrix $\tilde{\mathbf{X}}_k$.

6. Set up the new complete matrix $\mathbf{X}_k$ by stacking $\mathbf{X}_{k-1}$ with the imputed $\tilde{\mathbf{X}}_k$.

⟶ Repeat points 2 to 6 for *FIP* (4b excepted), or points 4b to 6 for *FIM*, until matrix $\mathbf{X}$ is completely imputed.

Table 2. The *IPCA* imputation algorithm.

1.	*Step* $k = 0$: Initialize matrix $\mathbf{X}$ by substituting the missing values with values opportunely assigned (e.g. variable means). Call $\mathbf{X}^{(0)}$ the matrix imputed in this way and set up matrix $\mathbf{M}^{(0)}$ with the mean vector in its rows.
2.	*Step* $k \geq 1$: An $(n \times S)$ matrix of PC scores $\hat{\mathbf{F}}^{(k)}$ along with a $(p \times S)$ loading matrix $\hat{\mathbf{U}}^{(k)}$ (with columns orthogonal and of unit norm), $S < p < n$, are found such that the reconstruction error: $\mathcal{E} = \Vert\mathbf{X}^{(k-1)} - \mathbf{M}^{(k-1)} - \mathbf{F}\mathbf{U}^T\Vert_F^2$ is minimized, where $\Vert\cdot\Vert_F$ is the Frobenius norm of a matrix (i.e. $\Vert\mathbf{A}\Vert_F = \sqrt{\mathrm{tr}\,(\mathbf{A}\mathbf{A}^T)}$, with $\mathbf{A}$ an $(n \times p)$-dimensional matrix).
3.	Compute the matrix: $\hat{\mathbf{X}}^{(k)} = \hat{\mathbf{F}}^{(k)}\hat{\mathbf{U}}^{(k)^T} + \mathbf{M}^{(k-1)}$, and replace the missing entries in $\mathbf{X}$ with the corresponding fitted values in $\hat{\mathbf{X}}^{(k)}$ to obtain the completed imputed matrix: $\mathbf{X}^{(k)} = \mathbf{W} * \mathbf{X} + (\mathbf{J} - \mathbf{W}) * \hat{\mathbf{X}}^{(k)}$, where $*$ is the element-wise (or Hadamard) product, $\mathbf{J}$ is the unit matrix containing all 1s, and $\mathbf{W}$ is the matrix with elements $w_{ij} = 0$ if x_{ij} is missing and $w_{ij} = 1$ otherwise.
4.	Compute matrix $\mathbf{M}^{(k)}$ with the updated mean vector in its rows using the completed imputed $\mathbf{X}^{(k)}$.
$\longrightarrow$	Repeat points 2 to 4 until a convergence criterion is met.

Table 3. The *missForest* imputation algorithm (case of quantitative variables only).

1.	*Step* $k = 0$: Initialize matrix $\mathbf{X}$ by substituting the missing values with values opportunely assigned (e.g. variable means). Set up vector $\mathbf{s}$ with the p indices of variables sorted according to the increasing number of missing values.
2.	*Step* $k \geq 1$, with previous results stored in matrix $\mathbf{X}_{\text{imp}}^{(k-1)}$: Fit a random forest (RF) using $\mathbf{X}_{-j}^{obs}$ as matrix of predictors and $\mathbf{x}_{j}^{obs}$ as response to have a trained RF.
3.	Predict $\mathbf{x}_{j}^{miss}$ by using the trained RF on $\mathbf{X}_{-j}^{miss}$.
4.	Update matrix $\mathbf{X}_{\text{imp}}^{(k)}$ with the new predictions of missing values in $\mathbf{x}_{j}^{miss}$.
5.	Repeat points 2 to 4 for all j in $\mathbf{s}$.
⟶	Restart from point 2 until the stopping rule γ in (1) is reached.

Table 4. Synoptic table of the data patterns considered in both the exploratory and supplementary simulation studies.

Pattern label	Description	MV distrib.	Type of study
SyKu-ECor	Symmetry and kurtosis with equicorrelations	*MEP*	exploratory
SyKu-NegECor	Symmetry and kurtosis with negative equicorrelations	*MEP*	supplementary
SyKu-PNCor	Symmetry and kurtosis with positive-negative correlations	*MEP*	supplementary
SyKu-UnbCor	Symmetry and kurtosis with unbalanced correlations	*MEP*	supplementary
SyKu-PosUnbCor	Symmetry and kurtosis with positive unbalanced correlations	*MEP*	supplementary
SK-ECor	Skewness with equicorrelations	*MSN*	exploratory
SK-PNCor	Skewness with positive-negative correlations	*MSN*	exploratory
SK-UnbCor	Skewness with unbalanced correlations	*MSN*	exploratory
SK-PosUnbCor	Skewness with positive unbalanced correlations	*MSN*	supplementary

Table 5. SyKu-ECor pattern: Experimental conditions, input–output parameters and correlation indices.

Data dimensionality and percentage of MCAR values:

− Number of variables in $\mathbf{X}^*$	$p = 3; 5; 10$
− Number of units in $\mathbf{X}^*$	$n = 500; 1000$
− Percentage of MCAR values	5%; 10%; 20%

Generation of SyKu-ECor pattern from $\text{MEP}_p(\mathbf{0}, \mathbf{\Sigma}, \kappa)$ with $\mathbf{\Sigma} = c^{-1}(\kappa, p)\mathbf{R}$ and $c(\kappa, p) = 2^{2/\kappa}\Gamma((p+2)/\kappa)/(p\Gamma(p/\kappa))$:

→ *Input parameters:*	→ *Output parameters:*
− Kurtosis parameter: $\kappa = 1; 2; 14$	− Output correlation coefficients in $\mathbf{R}$:
− Input correlation coefficients in $\mathbf{R}$: $\rho_{lj} = \rho = 0; 0.3; 0.7$ for $l \neq j = 1, \ldots, p$	$\rho_{lj} = \rho = 0; 0.3; 0.7 \quad$ for $l \neq j$

Correspondence between input and output kurtosis for each $p = 5; 10$:

Input kurtosis:	Output MV kurtosis index:	Type:
$\kappa = 1$	→ $\gamma_{2\text{MV}} = 11.67$ $(p = 5)$, 21.82 $(p = 10)$	positive excess kurtosis
$\kappa = 2$	→ $\gamma_{2\text{MV}} = 0$ for all p	mesokurtic
$\kappa = 14$	→ $\gamma_{2\text{MV}} = -7.25$ $(p = 5)$, -15.65 $(p = 10)$	negative excess kurtosis

Correlation indices computed for the output $\mathbf{R}$ matrices with $p = 5; 10$:

− $p = 5$ variables

Output ρ	RelEig_1	RelEig_2	RelEig_5	$\rho_{\min}$	$\rho_{\max}$	$\bar{\rho}_{\text{abs}}$
$\rho = 0$	0.20	0.20	0.20	0	0	0
$\rho = 0.3$	0.44	0.14	0.14	0.3	0.3	0.3
$\rho = 0.7$	0.76	0.06	0.06	0.7	0.7	0.7

− $p = 10$ variables

Output ρ	RelEig_1	RelEig_2	RelEig_{10}	$\rho_{\min}$	$\rho_{\max}$	$\bar{\rho}_{\text{abs}}$
$\rho = 0$	0.10	0.10	0.10	0	0	0
$\rho = 0.3$	0.37	0.07	0.07	0.3	0.3	0.3
$\rho = 0.7$	0.73	0.03	0.03	0.7	0.7	0.7

Table 6. SyKu-ECor pattern: 95%-Tukey uncertainty intervals around RMSE mean values of *FIM*, *FIP*, *IPCA* and *missForest*, with $p = 5; 10$ variables, $n = 1000$ units and 20% of MCAR values.

		ForImpMahalanobis			*ForImpPCA*			*IPCA*			*missForest*		
ρ	κ	mean	lower	upper	mean	lower	upper	mean	lower	upper	mean	lower	upper
$p = 5$ variables													
$\rho = 0$	$\kappa = 1$	1.024	1.021	1.026	1.024	1.021	1.026	1.012	1.010	1.015	1.093	1.091	1.096
	$\kappa = 2$	1.033	1.031	1.036	1.033	1.031	1.035	1.014	1.012	1.016	1.095	1.093	1.098
	$\kappa = 14$	1.037	1.034	1.039	1.037	1.034	1.039	1.012	1.010	1.014	1.125	1.123	1.128
$\rho = 0.3$	$\kappa = 1$	0.869	0.867	0.872	0.859	0.857	0.862	0.834	0.831	0.836	0.907	0.904	0.909
	$\kappa = 2$	0.862	0.859	0.864	0.853	0.850	0.855	0.829	0.827	0.831	0.901	0.898	0.903
	$\kappa = 14$	0.863	0.861	0.865	0.857	0.855	0.859	0.831	0.829	0.833	0.913	0.910	0.916
$\rho = 0.7$	$\kappa = 1$	0.486	0.483	0.488	0.434	0.432	0.437	0.379	0.377	0.381	0.414	0.411	0.416
	$\kappa = 2$	0.447	0.445	0.450	0.406	0.404	0.408	0.373	0.371	0.376	0.406	0.404	0.409
	$\kappa = 14$	0.436	0.434	0.438	0.409	0.407	0.412	0.387	0.385	0.390	0.425	0.423	0.428
$p = 10$ variables													
$\rho = 0$	$\kappa = 1$	2.047	2.044	2.051	2.047	2.043	2.050	2.055	2.052	2.058	2.146	2.143	2.150
	$\kappa = 2$	2.057	2.054	2.061	2.058	2.054	2.061	2.040	2.037	2.043	2.148	2.145	2.152
	$\kappa = 14$	2.069	2.066	2.073	2.070	2.066	2.073	2.045	2.042	2.048	2.140	2.136	2.143
$\rho = 0.3$	$\kappa = 1$	1.747	1.743	1.750	1.650	1.647	1.653	1.594	1.591	1.598	1.668	1.665	1.671
	$\kappa = 2$	1.727	1.723	1.730	1.637	1.634	1.640	1.581	1.578	1.584	1.671	1.667	1.674
	$\kappa = 14$	1.704	1.701	1.708	1.631	1.628	1.634	1.590	1.587	1.593	1.682	1.679	1.685
$\rho = 0.7$	$\kappa = 1$	1.161	1.157	1.164	0.783	0.780	0.787	0.682	0.679	0.685	0.724	0.720	0.727
	$\kappa = 2$	1.089	1.086	1.093	0.753	0.750	0.756	0.682	0.679	0.685	0.723	0.720	0.726
	$\kappa = 14$	1.034	1.031	1.038	0.763	0.760	0.766	0.713	0.710	0.716	0.760	0.756	0.763

Table 7. SyKu-ECor pattern: JT test for detection of the best imputation method among *FIM*, *FIP* and *IPCA*, with 20% of MCAR values.

		$p = 3$		$p = 5$		$p = 10$	
	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

Legend. Numbers in the cells refer to the hypotheses (13): 1: *FIM* the best; 2: *FIP* the best; 3, 6: *IPCA* the best.

Table 8. SK-ECor pattern: Experimental conditions, correspondence of input–output correlations and correlation indices.

Data dimensionality and percentage of MCAR values:	
– Number of variables in $\mathbf{X}^*$	$p = 3; 5; 10$
– Number of units in $\mathbf{X}^*$	$n = 500; 1000$
– Percentage of MCAR values	5%; 10%; 20%

Generation of SK-ECor pattern from $\text{MSN}_p(\boldsymbol{\Omega}, \boldsymbol{\alpha})$ with $\boldsymbol{\Omega} = \left[\omega_{lj}\right]_{l \neq j=1,\ldots,p}$ and $\boldsymbol{\alpha} = [\alpha_j]_{j=1,\ldots,p}$:

→ *Input parameters:*	→ *Output parameters:*
– Skewness parameter: $\alpha_j = \alpha = 1; 4; 10; 30$, $\forall j$	– Output correlation coefficients in $\mathbf{R}$:
– Input correlation coefficients in $\boldsymbol{\Omega}$: $\omega = 0; 0.5; 0.8$ for $l \neq j = 1, \ldots, p$	$\rho_{lj} = \rho$, for $l \neq j = 1, \ldots, p$ with approximate ρ values given below

Correspondence between input and output skewness for $p = 5$:

Input skewness:	Output MV skewness index:	Strength:
$\alpha = 1$ (with: $\omega = 0; 0.5; 0.8$)	$\rightarrow \gamma_{1\text{MV}} \in (0.26, 0.68)$	moderate-medium skewness
$\alpha \geq 4$ (with: $\omega = 0; 0.5; 0.8$)	$\rightarrow \gamma_{1\text{MV}} \in (0.89, 0.99)$	strong skewness

Input–output correlation correspondence for $p = 5$ variables and $\alpha \in [1, 30]$:

Input correlations in $\boldsymbol{\Omega}$:	Output correlations in $\mathbf{R}$:	Correlation structure:
(1) $\omega = 0$ and $p = 5$	$\rightarrow \rho \approx -0.14$	negative low ρs
(2) $\omega = 0.5$ and $p = 5$	$\rightarrow \rho \approx 0.20$	positive low ρs
(3) $\omega = 0.8$ and $p = 5$	$\rightarrow \rho \approx 0.58$	positive moderate ρs

Correlation indices computed for the output $\mathbf{R}$ matrices with the structures (1)–(3):

	RelEig_1	RelEig_2	RelEig_5	$\rho_{\min}$	$\rho_{\max}$	$\bar{\rho}_{\text{abs}}$	skew_{abs}
(1)	≈ 0.23	≈ 0.23	≈ 0.09	≈ -0.14	≈ -0.14	≈ 0.14	—
(2)	≈ 0.36	≈ 0.16	≈ 0.16	≈ 0.20	≈ 0.20	≈ 0.20	—
(3)	≈ 0.66	≈ 0.09	≈ 0.09	≈ 0.58	≈ 0.58	≈ 0.58	—

Table 9. SK-ECor pattern: 95%-Tukey uncertainty intervals around RMSE mean values of *FIM*, *FIP*, *IPCA* and *missForest*, with $p = 5$ variables, $n = 1000$ units and 20% of MCAR values.

		ForImpMahalanobis			*ForImpPCA*			*IPCA*			*missForest*		
ω	α	mean	lower	upper	mean	lower	upper	mean	lower	upper	mean	lower	upper
$\omega = 0$	$\alpha = 1$	0.948	0.946	0.951	0.956	0.953	0.958	0.993	0.991	0.996	1.023	1.020	1.025
	$\alpha = 4$	0.904	0.902	0.907	0.922	0.920	0.925	0.988	0.985	0.990	0.980	0.977	0.983
	$\alpha = 10$	0.899	0.897	0.902	0.918	0.916	0.921	0.985	0.983	0.988	0.974	0.971	0.977
	$\alpha = 30$	0.898	0.896	0.901	0.918	0.916	0.921	0.986	0.984	0.989	1.016	1.013	1.019
$\omega = 0.5$	$\alpha = 1$	0.941	0.938	0.944	0.936	0.934	0.939	0.931	0.929	0.934	0.992	0.990	0.995
	$\alpha = 4$	0.944	0.942	0.947	0.938	0.936	0.941	0.954	0.952	0.957	0.990	0.987	0.993
	$\alpha = 10$	0.944	0.942	0.947	0.938	0.936	0.941	0.956	0.954	0.959	0.992	0.989	0.995
	$\alpha = 30$	0.942	0.939	0.945	0.935	0.933	0.938	0.953	0.951	0.956	1.023	1.021	1.026
$\omega = 0.8$	$\alpha = 1$	0.611	0.608	0.614	0.582	0.579	0.585	0.559	0.556	0.561	0.603	0.600	0.605
	$\alpha = 4$	0.620	0.617	0.622	0.596	0.593	0.598	0.592	0.590	0.595	0.616	0.613	0.619
	$\alpha = 10$	0.620	0.618	0.623	0.596	0.593	0.598	0.594	0.591	0.596	0.641	0.638	0.644
	$\alpha = 30$	0.619	0.616	0.622	0.595	0.593	0.598	0.593	0.591	0.596	0.614	0.612	0.617

Table 10. SK-ECor pattern: JT test for detection of the best imputation method among *FIM*, *FIP* and *IPCA*, with 20% of MCAR values.

		$p = 3$		$p = 5$		$p = 10$	
	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

Legend. Numbers in the cells refer to the hypotheses (13): 1: *FIM* the best; 2, 5: *FIP* the best; 3, 6: *IPCA* the best.

Table 11. SK-PNCor pattern: Experimental conditions, correspondence of input–output correlations and correlation indices.

Data dimensionality and percentage of MCAR values:

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

Generation of SK-PNCor pattern from $\mathrm{MSN}_p(\boldsymbol{\Omega}, \boldsymbol{\alpha})$ with $\boldsymbol{\Omega} = \left[\omega_{lj}\right]_{l \neq j = 1,\ldots,p}$ and $\boldsymbol{\alpha} = [\alpha_j]_{j=1,\ldots,p}$:

→ *Input parameters:*

– Skewness parameter: $\alpha_j = \alpha = 1; 4; 10; 30, \ \forall j$

– For $j = 2, \ldots, p$ and $\omega = 0.2; 0.5; 0.8$:

$$\begin{cases} \omega_{1j} = \omega_{j1} = (-1)^j \omega \\ \omega_{jv} = \omega \qquad \text{if } \operatorname{sign}(\omega_{lj}) = \operatorname{sign}(\omega_{lv}) \\ \omega_{jv} = -\omega \quad \text{ if } \operatorname{sign}(\omega_{lj}) \neq \operatorname{sign}(\omega_{lv}), \\ \qquad\qquad (l, v = 1, \ldots, p, \ l \neq v \neq j) \end{cases}$$

→ *Output parameters:*

– Output correlation coefficients in $\mathbf{R}$:
For odd (even) p, set: $m = p - 1$, $(m = p - 2)$.
Then, for each j (# is “number”):

$$\begin{cases} \rho_{jv} = \rho_1 \text{ if } \omega_{jv} = -\omega \text{ and \#neg. } \omega_{jl} = \frac{m}{2} \\ \rho_{jv} = \rho_2 \text{ if } \omega_{jv} = \omega \text{ and:} \\ \qquad\qquad \text{for odd } p: \text{\#pos. } \omega_{jl} = \frac{m}{2} \\ \qquad\qquad \text{for even } p: \text{\#pos. } \omega_{jl} = \frac{m}{2} + 1 \\ \rho_{jv} = \rho_3 \text{ otherwise,} \\ (l, v = 1, \ldots, p, \ l \neq v \neq j) \end{cases}$$

Correspondence between input and output skewness for $p = 5$:

Input skewness:	Output MV skewness index:	Strength:
$\alpha = 1$ (with: $\omega = 0.2; 0.5; 0.8$)	→ $\gamma_{1\mathrm{MV}} \in (0.06, 0.22)$	moderate skewness
$\alpha \geq 4$ (with: $\omega = 0.2; 0.5; 0.8$)	→ $\gamma_{1\mathrm{MV}} \in (0.98, 0.99)$	strong skewness

Input–output correlation correspondence for $p = 5$ variables and $\alpha \in [1, 30]$:

Input corr. in $\boldsymbol{\Omega}$:	Output correlations in $\mathbf{R}$:	Correlation structure:
(1) $\omega = 0.2$ and $p = 5$	→ $\rho_1 \approx -0.30$, $\rho_2 \approx 0.06$, $\rho_3 \approx 0.15$	positive-negative low ρs
(2) $\omega = 0.5$ and $p = 5$	→ $\rho_1 \approx -0.56$, $\rho_2 \approx 0.37$, $\rho_3 \approx 0.50$	positive-negative moderate ρs
(3) $\omega = 0.8$ and $p = 5$	→ $\rho_1 \approx -0.78$, $\rho_2 \approx 0.70$, $\rho_3 \approx 0.77$	positive-negative high ρs

Correlation indices computed for the output $\mathbf{R}$ matrices with the structures (1)–(3):

	RelEig_1	RelEig_2	RelEig_5	$\rho_{\min}$	$\rho_{\max}$	$\bar{\rho}_{\mathrm{abs}}$	$\mathrm{skew}_{\mathrm{abs}}$
(1)	≈ 0.38	≈ 0.19	≈ 0.07	≈ -0.32	≈ 0.16	≈ 0.23	≈ -0.53
(2)	≈ 0.60	≈ 0.13	≈ 0.05	≈ -0.56	≈ 0.50	≈ 0.50	≈ -0.75
(3)	≈ 0.81	≈ 0.06	≈ 0.03	≈ -0.78	≈ 0.78	≈ 0.76	≈ 0

Table 12. SK-PNCor pattern: 95%-Tukey uncertainty intervals around RMSE mean values of *FIM*, *FIP*, *IPCA* and *missForest*, with $p = 5$ variables, $n = 1000$ units and 20% of MCAR values.

		ForImpMahalanobis			*ForImpPCA*			*IPCA*			*missForest*		
ω	α	mean	lower	upper	mean	lower	upper	mean	lower	upper	mean	lower	upper
$\omega = 0.2$	$\alpha = 1$	0.875	0.873	0.878	0.882	0.880	0.884	0.892	0.890	0.894	0.932	0.930	0.934
	$\alpha = 4$	0.829	0.826	0.831	0.846	0.844	0.848	0.875	0.873	0.877	0.888	0.886	0.891
	$\alpha = 10$	0.822	0.820	0.824	0.841	0.839	0.843	0.870	0.868	0.872	0.882	0.880	0.885
	$\alpha = 30$	0.821	0.819	0.823	0.841	0.839	0.843	0.871	0.869	0.873	0.879	0.877	0.881
$\omega = 0.5$	$\alpha = 1$	0.665	0.663	0.667	0.649	0.647	0.651	0.626	0.624	0.628	0.695	0.693	0.697
	$\alpha = 4$	0.634	0.632	0.636	0.631	0.629	0.633	0.612	0.610	0.614	0.663	0.661	0.665
	$\alpha = 10$	0.631	0.629	0.633	0.629	0.627	0.631	0.611	0.609	0.613	0.662	0.660	0.665
	$\alpha = 30$	0.630	0.628	0.632	0.629	0.627	0.631	0.610	0.608	0.612	0.660	0.658	0.663
$\omega = 0.8$	$\alpha = 1$	0.387	0.384	0.389	0.339	0.337	0.341	0.301	0.299	0.303	0.342	0.340	0.344
	$\alpha = 4$	0.396	0.394	0.398	0.357	0.355	0.359	0.321	0.319	0.323	0.355	0.353	0.358
	$\alpha = 10$	0.395	0.393	0.397	0.358	0.355	0.360	0.323	0.321	0.325	0.355	0.353	0.357
	$\alpha = 30$	0.395	0.393	0.397	0.359	0.356	0.361	0.323	0.321	0.325	0.355	0.352	0.357

Table 13. SK-PNCor pattern: JT test for detection of the best imputation method among *FIM*, *FIP* and *IPCA*, with 20% of MCAR values.

			$p = 5$		$p = 10$	
		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

Legend. Numbers in the cells refer to the system of hypotheses (13): 1, 4: *FIM* the best; 3, 6: *IPCA* the best.

Table 14. SK-UnbCor pattern: Experimental conditions, correspondence of input–output correlations and correlation indices.

Data dimensionality and percentage of MCAR values:

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

Generation of SK-UnbCor pattern from $\mathrm{MSN}_p(\boldsymbol{\Omega}, \boldsymbol{\alpha})$ with $\boldsymbol{\Omega} = \left[\omega_{lj}\right]_{l \neq j=1,\ldots,p}$ and $\boldsymbol{\alpha} = \left[\alpha_j\right]_{j=1,\ldots,p}$:

→ Input parameters:

– Skewness parameter: $\alpha_j = \alpha = 1; 4; 10; 30, \forall j$

– Input correlation coefficients in $\boldsymbol{\Omega}$:

$\omega_{1j} = \omega_1 = -\omega, \quad \omega > 0, \; j \neq 1$

$\omega_{lj} = \omega_2 = \omega/c, \qquad \text{for } l, j \neq 1, \; l \neq j,$

with $\omega = 0.2; 0.5; 0.8$ and $c = 1; 1.25; 1.5$

→ Output parameters:

– Output correlation coefficients in $\mathbf{R}$:

$\rho_{1j} = \rho_1, \quad j \neq 1$

$\rho_{lj} = \rho_2, \quad \text{for } l, j \neq 1, \; l \neq j$

Correspondence between input and output skewness for each $p = 5$:

Input skewness:	Output MV skewness index:	Strength:
$\alpha = 1$ (with: $\omega = 0.2; 0.5; 0.8$)	→ $\gamma_{1\mathrm{MV}} \in (0.27, 0.42)$	moderate skewness
$\alpha \geq 4$ (with: $\omega = 0.2; 0.5; 0.8$)	→ $\gamma_{1\mathrm{MV}} \in (0.89, 0.99)$	strong skewness

Input–output correlation correspondence for $p = 5$ variables and $\alpha \in [1, 30]$:

Input correlations in $\boldsymbol{\Omega}$:	Output correlations in $\mathbf{R}$:	Correlation structure:
(1) $\omega = 0.2$ and $c = 1; 1.25; 1.5$	→ $\rho_1 \approx -0.25$, $\rho_2 \approx 0$	negative low and nearly null ρs
(2) $\omega = 0.5$ and $c = 1; 1.25$	→ $\rho_1 \approx -0.4$, $\rho_2 \approx 0.2$	neg. moderate and pos. low ρs
(3) $\omega = 0.5$ and $c = 1.5$	→ $\rho_1 \approx -0.4$, $\rho_2 \approx 0.1$	neg. moderate and nearly null ρs
(4) $\omega = 0.8$ and $c = 1$	→ $\rho_1 \approx -0.7$, $\rho_2 \approx 0.6$	neg. high and pos. high ρs
(5) $\omega = 0.8$ and $c = 1.25$	→ $\rho_1 \approx -0.7$, $\rho_2 \approx 0.35$	neg. high and pos. moderate ρs
(6) $\omega = 0.8$ and $c = 1.5$	→ $\rho_1 \approx -0.7$, $\rho_2 \approx 0.25$	neg. high and pos. low ρs

Correlation indices computed for the output $\mathbf{R}$ matrices with the structures (1)–(6):

	RelEig_1	RelEig_2	RelEig_5	$\rho_{\min}$	$\rho_{\max}$	$\bar{\rho}_{\mathrm{abs}}$	$\mathrm{skew}_{\mathrm{abs}}$
(1)	≈ 0.30	≈ 0.20	≈ 0.09	≈ -0.25	≈ -0.05	≈ 0.12	≈ 0.41
(2)	≈ 0.43	≈ 0.16	≈ 0.08	≈ -0.42	≈ 0.19	≈ 0.26	≈ 0.41
(3)	≈ 0.38	≈ 0.18	≈ 0.06	≈ -0.40	≈ 0.08	≈ 0.20	≈ 0.41
(4)	≈ 0.65	≈ 0.09	≈ 0.04	≈ -0.66	≈ 0.58	≈ 0.62	≈ 0.20
(5)	≈ 0.58	≈ 0.12	≈ 0.03	≈ -0.63	≈ 0.35	≈ 0.46	≈ 0.41
(6)	≈ 0.53	≈ 0.16	≈ 0.01	≈ -0.63	≈ 0.22	≈ 0.39	≈ 0.41

Table 15. SK-UnbCor pattern: 95%-Tukey uncertainty intervals around RMSE mean values of *FIM*, *FIP*, *IPCA* and *missForest*, with $p = 5$ variables, $n = 1000$ units and 20% of MCAR values.

		ForImpMahalanobis			*ForImpPCA*			*IPCA*			*missForest*		
ρ, c	α	mean	lower	upper	mean	lower	upper	mean	lower	upper	mean	lower	upper
$\omega = 0.2$, $c = 1$	$\alpha = 1$	0.939	0.936	0.942	0.942	0.939	0.946	0.956	0.953	0.959	1.005	1.002	1.008
	$\alpha = 4$	0.911	0.908	0.914	0.923	0.919	0.926	0.960	0.957	0.963	1.015	1.012	1.018
	$\alpha = 10$	0.909	0.906	0.912	0.921	0.918	0.925	0.962	0.959	0.965	0.977	0.974	0.981
	$\alpha = 30$	0.910	0.907	0.913	0.922	0.919	0.925	0.963	0.960	0.966	0.977	0.974	0.980
$\omega = 0.2$, $c = 1.25$	$\alpha = 1$	0.936	0.933	0.939	0.941	0.938	0.945	0.962	0.959	0.965	1.002	0.999	1.006
	$\alpha = 4$	0.898	0.895	0.901	0.913	0.909	0.916	0.961	0.958	0.964	1.004	1.000	1.007
	$\alpha = 10$	0.898	0.895	0.901	0.914	0.910	0.917	0.964	0.960	0.967	0.966	0.963	0.969
	$\alpha = 30$	0.896	0.893	0.899	0.912	0.909	0.915	0.963	0.960	0.966	0.963	0.960	0.966
$\omega = 0.2$, $c = 1.5$	$\alpha = 1$	0.932	0.929	0.934	0.938	0.935	0.942	0.965	0.962	0.969	0.999	0.995	1.002
	$\alpha = 4$	0.892	0.889	0.895	0.910	0.906	0.913	0.964	0.961	0.967	1.000	0.997	1.004
	$\alpha = 10$	0.889	0.886	0.892	0.908	0.904	0.911	0.966	0.962	0.969	0.958	0.955	0.961
	$\alpha = 30$	0.886	0.883	0.889	0.905	0.901	0.908	0.963	0.960	0.966	0.956	0.953	0.959
$\omega = 0.5$, $c = 1$	$\alpha = 1$	0.836	0.833	0.839	0.825	0.822	0.828	0.818	0.815	0.821	0.895	0.892	0.899
	$\alpha = 4$	0.839	0.836	0.842	0.834	0.830	0.837	0.847	0.844	0.850	0.899	0.896	0.902
	$\alpha = 10$	0.841	0.838	0.844	0.837	0.834	0.840	0.853	0.850	0.856	0.904	0.901	0.907
	$\alpha = 30$	0.840	0.837	0.843	0.836	0.833	0.840	0.853	0.850	0.856	0.901	0.898	0.904
$\omega = 0.5$, $c = 1.25$	$\alpha = 1$	0.843	0.840	0.846	0.840	0.837	0.843	0.853	0.850	0.856	0.907	0.904	0.911
	$\alpha = 4$	0.839	0.836	0.842	0.844	0.841	0.848	0.881	0.878	0.884	0.912	0.908	0.915
	$\alpha = 10$	0.838	0.835	0.841	0.846	0.842	0.849	0.884	0.881	0.887	0.911	0.908	0.914
	$\alpha = 30$	0.839	0.836	0.842	0.846	0.842	0.849	0.884	0.881	0.887	0.911	0.907	0.914
$\omega = 0.5$, $c = 1.5$	$\alpha = 1$	0.837	0.834	0.840	0.842	0.839	0.846	0.872	0.869	0.875	0.908	0.905	0.911
	$\alpha = 4$	0.830	0.827	0.833	0.846	0.843	0.849	0.899	0.896	0.902	0.907	0.904	0.910
	$\alpha = 10$	0.826	0.823	0.829	0.843	0.840	0.846	0.900	0.897	0.903	0.904	0.901	0.908
	$\alpha = 30$	0.826	0.823	0.829	0.842	0.839	0.845	0.901	0.898	0.904	0.903	0.900	0.906
$\omega = 0.8$, $c = 1$	$\alpha = 1$	0.552	0.549	0.555	0.515	0.512	0.518	0.482	0.479	0.485	0.520	0.517	0.524
	$\alpha = 4$	0.578	0.575	0.581	0.548	0.544	0.551	0.531	0.528	0.534	0.558	0.554	0.561
	$\alpha = 10$	0.579	0.576	0.582	0.550	0.547	0.553	0.536	0.533	0.539	0.558	0.555	0.562
	$\alpha = 30$	0.580	0.577	0.583	0.551	0.548	0.554	0.538	0.535	0.541	0.559	0.556	0.562
$\omega = 0.8$, $c = 1.25$	$\alpha = 1$	0.621	0.618	0.624	0.628	0.625	0.632	0.648	0.644	0.651	0.615	0.611	0.618
	$\alpha = 4$	0.649	0.646	0.652	0.667	0.664	0.670	0.716	0.713	0.719	0.656	0.653	0.659
	$\alpha = 10$	0.650	0.647	0.653	0.668	0.664	0.671	0.723	0.720	0.726	0.657	0.654	0.660
	$\alpha = 30$	0.650	0.647	0.653	0.667	0.664	0.670	0.722	0.719	0.725	0.658	0.654	0.661
$\omega = 0.8$, $c = 1.5$	$\alpha = 1$	0.482	0.479	0.485	0.644	0.640	0.647	0.755	0.752	0.758	0.551	0.548	0.554
	$\alpha = 4$	0.512	0.509	0.515	0.688	0.685	0.692	0.843	0.840	0.846	0.603	0.600	0.607
	$\alpha = 10$	0.515	0.512	0.518	0.689	0.686	0.692	0.855	0.852	0.858	0.607	0.604	0.610
	$\alpha = 30$	0.516	0.513	0.519	0.689	0.686	0.692	0.855	0.852	0.858	0.606	0.603	0.609

Table 16. SK-UnbCor pattern: JT test for detection of the best imputation method among *FIM*, *FIP* and *IPCA*, with 20% of MCAR values.

		$p = 5$					
		$n = 500$			$n = 1000$		
	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

Legend. Numbers in the cells refer to the hypotheses (13): 1, 4: *FIM* the best; 2: *FIP* the best; 3, 6: *IPCA* the best.

Table 17. Computational efficiency: Times in seconds (mean ± standard deviation) for each method computed as an overall average over the exploratory studies with *MEP* and *MSN* distributions

p	n	% MCAR values	*ForImpMahalanobis*	*ForImpPCA*	*IPCA*	*missForest*
3	500	5%	0.44 ± 0.10	0.49 ± 0.12	0.01 ± 0.01	2.26 ± 0.61
		10%	0.45 ± 0.09	0.56 ± 0.11	0.01 ± 0.01	2.05 ± 0.61
		20%	0.78 ± 0.31	1.00 ± 0.39	0.02 ± 0.02	2.80 ± 1.47
	1000	5%	1.69 ± 0.25	2.25 ± 0.38	0.02 ± 0.01	14.86 ± 4.71
		10%	1.45 ± 0.60	2.16 ± 0.87	0.02 ± 0.02	9.57 ± 5.14
		20%	1.67 ± 0.71	2.66 ± 1.12	0.04 ± 0.02	8.06 ± 4.18
5	500	5%	0.58 ± 0.19	0.68 ± 0.22	0.02 ± 0.01	5.48 ± 2.05
		10%	0.77 ± 0.36	0.92 ± 0.46	0.02 ± 0.01	5.88 ± 3.78
		20%	0.84 ± 0.42	1.02 ± 0.52	0.03 ± 0.02	4.98 ± 3.59
	1000	5%	2.01 ± 0.70	2.76 ± 0.92	0.04 ± 0.03	31.94 ± 14.06
		10%	2.30 ± 0.98	3.30 ± 1.33	0.05 ± 0.03	27.73 ± 13.91
		20%	2.94 ± 0.93	4.26 ± 1.30	0.10 ± 0.06	28.28 ± 11.60
10	500	5%	1.31 ± 0.40	1.55 ± 0.46	0.06 ± 0.03	22.46 ± 8.44
		10%	1.61 ± 0.61	1.87 ± 0.71	0.08 ± 0.04	20.34 ± 9.31
		20%	3.15 ± 0.38	3.57 ± 0.44	0.27 ± 0.16	27.72 ± 5.93
	1000	5%	3.17 ± 0.88	4.37 ± 1.17	0.10 ± 0.06	80.25 ± 29.72
		10%	3.93 ± 1.26	5.36 ± 1.64	0.14 ± 0.09	74.42 ± 29.99
		20%	5.68 ± 1.42	7.34 ± 1.78	0.37 ± 0.21	75.12 ± 25.98

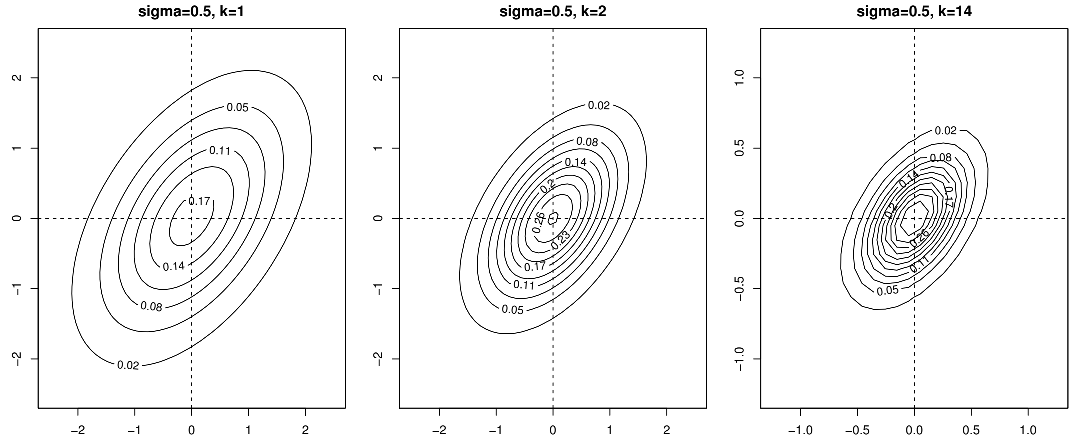

Figure 1. Contour plots for *MEP* distribution.

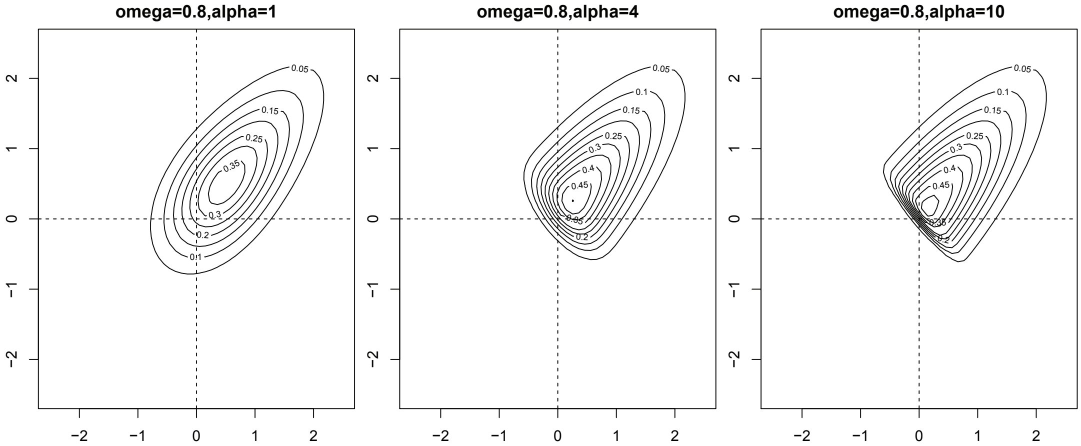

Figure 2. Contour plots for *MSN* distribution.

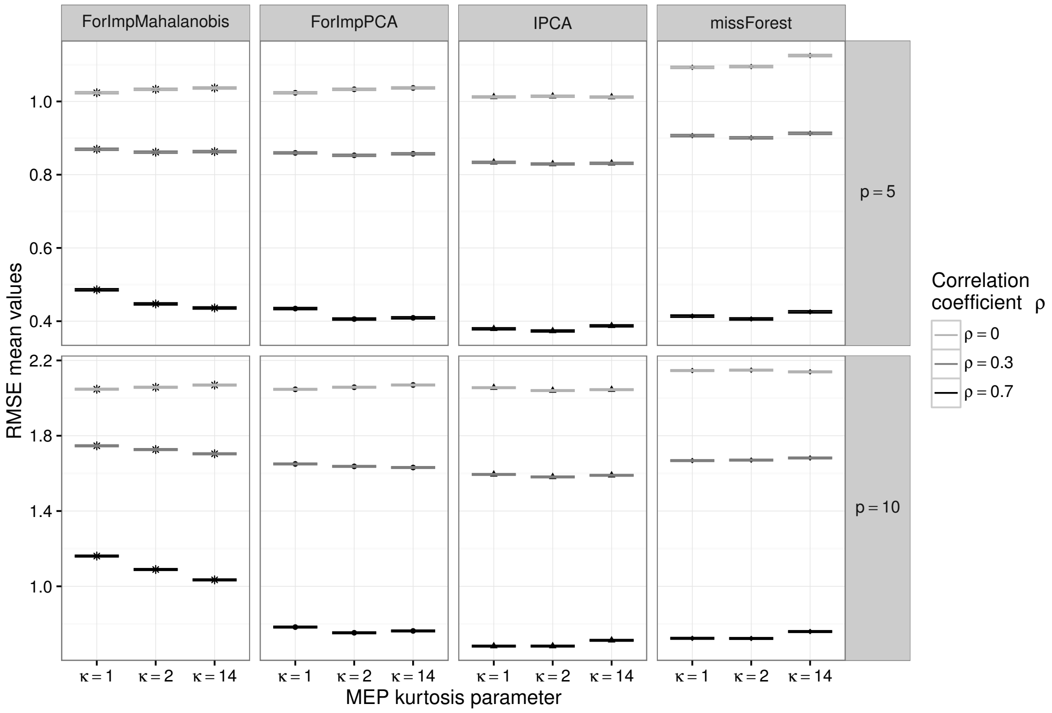

Figure 3. SyKu-ECor pattern with $p = 5; 10$ variables, $n = 1000$ units and 20% of MCAR values – Dot plots of RMSE mean values with 95%-Tukey uncertainty intervals concerning *FIM*, *FIP*, *IPCA* and *missForest*.

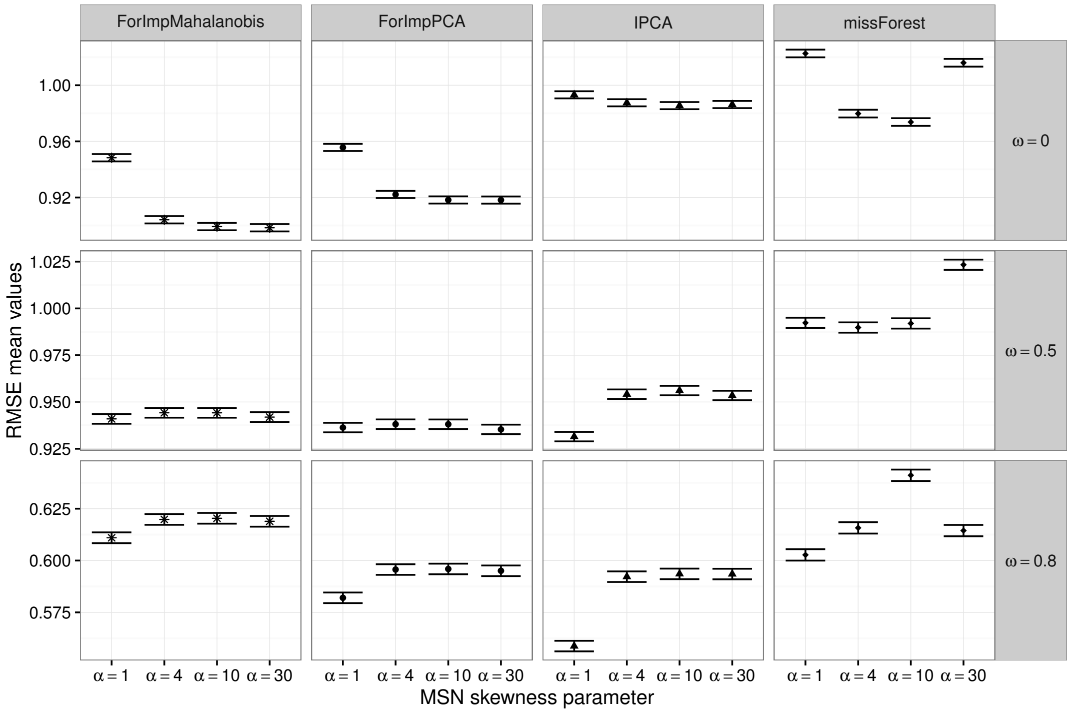

Figure 4. SK-ECor pattern with $p = 5$ variables, $n = 1000$ units and 20% of MCAR values – Dot plots of RMSE mean values with 95%-Tukey uncertainty intervals concerning *FIM*, *FIP*, *IPCA* and *missForest*.

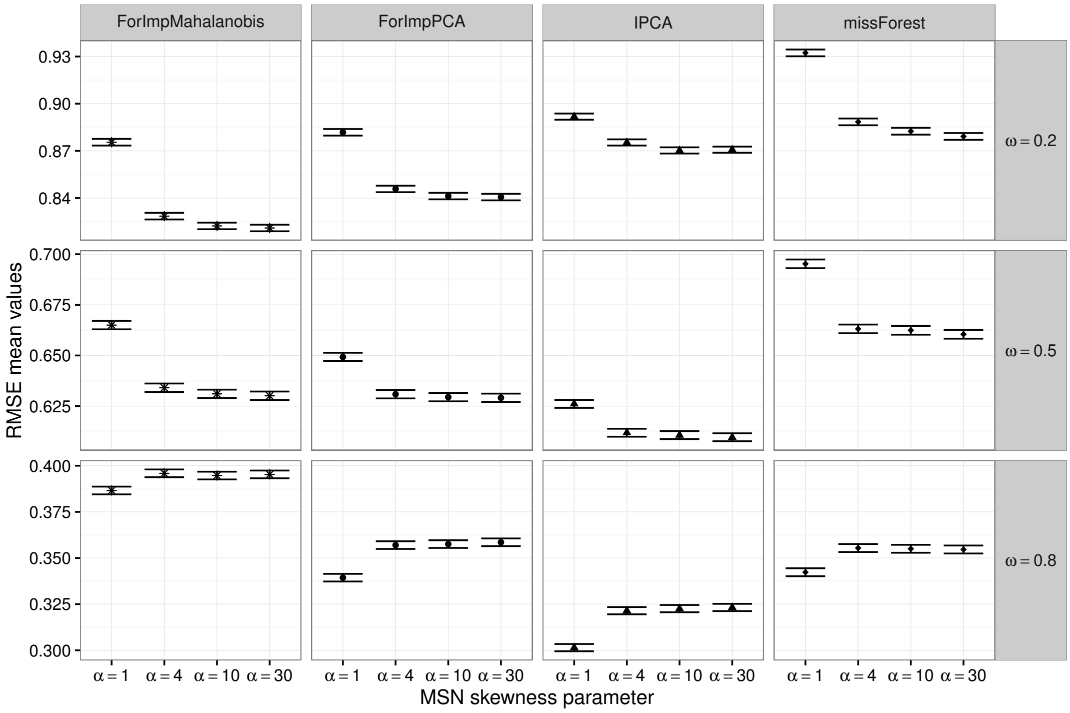

Figure 5. SK-PNCor pattern with $p = 5$ variables, $n = 1000$ units and 20% of MCAR values – Dot plots of RMSE mean values with 95%-Tukey uncertainty intervals concerning *FIM*, *FIP*, *IPCA* and *missForest*.

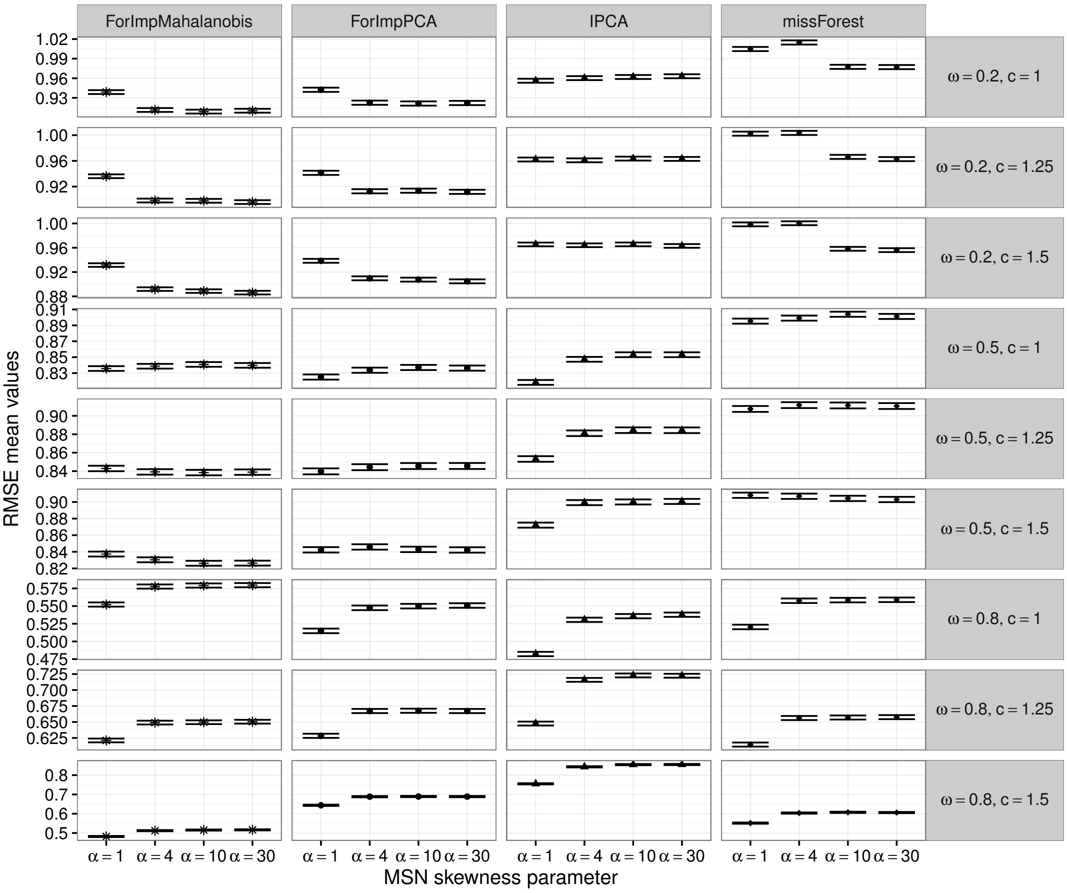

Figure 6. SK-UnbCor pattern with $p = 5$ variables, $n = 1000$ units and 20% of MCAR values – Dot plots of RMSE mean values with 95%-Tukey uncertainty intervals concerning *FIM*, *FIP*, *IPCA* and *missForest*.