HAL
open science

Nonparametric imputation by data depth

Pavlo Mozharovskyi, Julie Josse, François Husson

To cite this version:

Pavlo Mozharovskyi, Julie Josse, François Husson. Nonparametric imputation by data depth. Journal of the American Statistical Association, 2020, 115 (529), pp.241-253. 10.1080/01621459.2018.1543123 . hal-01440269

HAL Id: hal-01440269

https://hal.science/hal-01440269v1

Submitted on 3 Jun 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Nonparametric imputation by data depth

Pavlo Mozharovskyi*
CREST, Ensai, Universite´ Bretagne Loire
and
Julie Josse
École Polytechnique, CMAP
and
François Husson
IRMAR, Applied Mathematics Unit, Agrocampus Ouest, Rennes

August 6, 2018

Abstract

We present single imputation method for missing values which borrows the idea of data depth—a measure of centrality defined for an arbitrary point of a space with respect to a probability distribution or data cloud. This consists in iterative maximization of the depth of each observation with missing values, and can be employed with any properly defined statistical depth function. For each single iteration, imputation reverts to optimization of quadratic, linear, or quasiconcave functions that are solved analytically by linear programming or the Nelder-Mead method. As it accounts for the underlying data topology, the procedure is distribution free, allows imputation close to the data geometry, can make prediction in situations where local imputation (k-nearest neighbors, random forest) cannot, and has attractive robustness and asymptotic properties under elliptical symmetry. It is shown that a special case—when using the Mahalanobis depth—has direct connection to well-known methods for the multivariate normal model, such as iterated regression and regularized PCA. The methodology is extended to multiple imputation for data stemming from an elliptically symmetric distribution. Simulation and real data studies show good results compared with existing popular alternatives. The method has been implemented as an R-package. Supplementary materials for the article are available online.

Keywords: Elliptical symmetry, Outliers, Tukey depth, Zonoid depth, Local depth, Nonparametric imputation, Convex optimization.

*The major part of this project has been conducted during the postdoc of Pavlo Mozharovskyi at Agrocampus Ouest (Rennes) granted by *Centre Henri Lebesgue* due to program PIA-ANR-11-LABX-0020-01.

1 Introduction

Missing data is a ubiquitous problem in statistics. Non-responses to surveys, machines that break and stop reporting, and data that have not been recorded, impede analysis and threaten the validity of inference. A common strategy (Little and Rubin, 2002) for dealing with missing values is single imputation, replacing missing entries with plausible values to obtain a completed data set, which can then be analyzed.

There are two main families of parametric imputation methods: "joint" and "conditional" modeling, see e.g., Josse and Reiter (2018) for a literature overview. Joint modeling specifies a joint distribution for the data, the most popular being the normal multivariate distribution. The parameters of the distribution, here the mean and the covariance matrix, are then estimated from the incomplete data using an algorithm such as expectation maximization (EM) (Dempster et al., 1977). The missing entries are then imputed with the conditional mean, i.e., the conditional expectation of the missing values, given observed values and the estimated parameters. An alternative is to impute missing values using a principal component analysis (PCA) model which assumes data are generated as a low rank structure corrupted by Gaussian noise. This method is closely connected to the literature on matrix completion Josse and Husson (2012), Hastie et al. (2015), and has shown good imputation capacity due to the plausibility of the low rank assumption (Udell and Townsend, 2017). The conditional modeling approach (van Buuren, 2012) consists in specifying one model for each variable to be imputed, and considers the others variables as explanatory. This procedure is iterated until predictions stabilize. Nonparametric imputation methods have also been developed such as imputation by k-nearest neighbors (kNN) (see Troyanskaya et al., 2001, and references therein) or random forest (Stekhoven and Bühlmann, 2012).

Most imputation methods are defined under the missing (completely) at random (M(C)AR) assumption, which means that the probability of having missing values does not depend on missing data (nor on observed data). Gaussian and PCA imputations are sensitive to outliers and deviations from distributional assumptions, whereas nonparametric methods such as kNN and random forest cannot extrapolate.

Here we propose a family of nonparametric imputation methods based on the notion of a statistical depth function (Tukey, 1975). Data depth is a data-driven multivariate measure of centrality that describes data with respect to location, scale, and shape based on a multivariate ordering. It has been applied in multivariate data analysis (Liu et al., 1999), classification (Jörnsten, 2004, Lange et al., 2014), multivariate risk measurement (Cascos and Molchanov, 2007), and robust linear programming (Bazovkin and Mosler, 2015), but has never been applied in the context of missing data. Depth based imputation provides excellent predictive properties and has the advantages of both global and local imputation methods. It imputes close to the data geometry, while still accounting for global features. In addition, it allows robust imputation in both outliers and heavy-tailed distributions.

Figures 1 and 2 motivate our proposed depth-based imputation by contrasting it to classical methods. First, 150 points are drawn from a bivariate normal distribution with mean $\boldsymbol{\mu}_1 = (1,1)^\top$ and covariance $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}$ and 30% of the entries are removed completely at random in both variables; points with one missing entry are indicated by dotted lines while solid lines provide (oracle) imputation using distribution parameters. The imputation assuming a joint Gaussian distribution using EM estimates is shown by rhombi (Figure 1, left). Zonoid depth-based imputation, represented by filled circles, shows that the sample is not necessarily normal, and that this uncertainty increases as we move to the fringes of the data cloud, where imputed points deviate from the conditional mean towards the unconditional one. Second, the missing

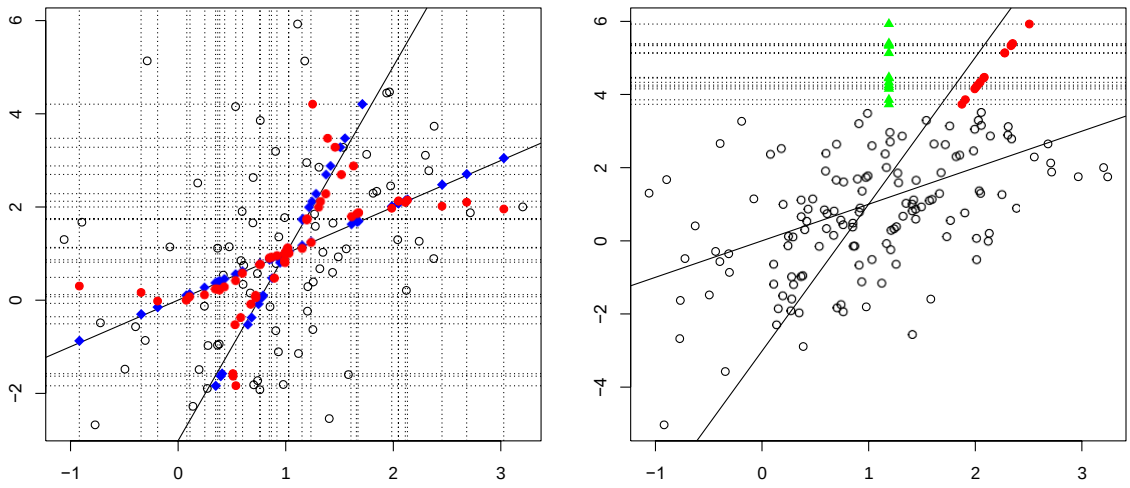

Figure 1: Bivariate normal distribution with 30% MCAR (left) and with MAR in the second coordinate for values > 3.5 (right); imputation using maximum zonoid depth (filled circles), conditional mean imputation using EM estimates (rhombi), and random forest imputation (triangles).

values are generated as follows: the first coordinate is removed when the second coordinate > 3.5 (Figure 1, right). Here, the depth-based imputation allows extrapolation when predicting missing values, while the random forest imputation (triangles) gives, as expected, rather poor results.

In Figure 2 (left), we draw 500 points, 425 from the same normal distribution as above, with 15% of MCAR values and 75 outliers from the Cauchy distribution with the same center and shape matrix and without missing values. In Figure 2 (right), we depict 1000 points drawn from Cauchy distribution with 15% MCAR. As expected, imputation with conditional mean based on EM estimates (rhombi) is rather random. Depth-based imputation with Tukey depth (filled circles) has robust imputed values that are close to the (distribution's) regression lines reflecting data geometry.

The paper is organized as follows. Section 2 describes the algorithm for imputing by data depth and derives its theoretical properties under ellipticity. Section 3 describes the special case of imputation with Mahalanobis depth, emphasizing its relationship to existing imputation methods by regression and PCA, and imputation with zonoid and Tukey depths. For each of them, we suggest an efficient optimization strategy. Next, to go beyond ellipticity, we propose imputation with local depth (Paindaveine and Bever, 2013) appropriate to data with non-convex support. Section 4 provides a comparative simulation and real data study. Section 5 extends the proposed approach to multiple imputation in order to perform statistical inference with missing values. Section 6 concludes the article, gathering together some useful remarks. Proofs are available in the supplementary materials.

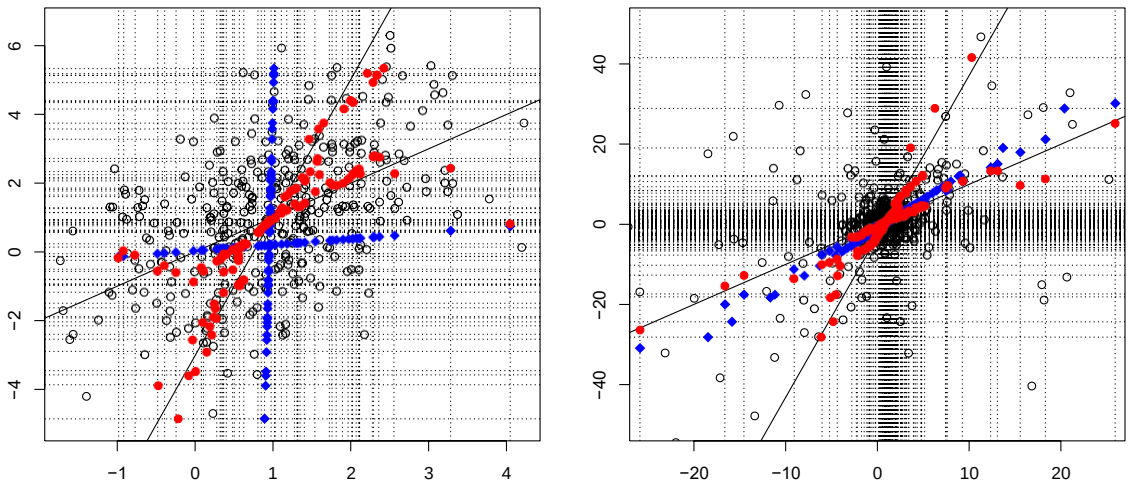

Figure 2: Left: Mixture of normal (425 points, 15% MCAR) and Cauchy (75 points) samples. Right: 1000 Cauchy distributed points with 15% MCAR. Imputation with Tukey depth (filled circles) and conditional mean imputation using EM estimates (rhombi).

2 Imputation by depth maximization

2.1 Imputation by iterative regression

Let X be a random vector in R^d and denote $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ a sample. For a point $\boldsymbol{x}_i \in \boldsymbol{X}$, we denote $miss(i)$ and $obs(i)$ the sets of its coordinates containing missing and observed values, $|miss(i)|$ and $|obs(i)|$ their corresponding cardinalities.

Let the rows $\boldsymbol{x}_i$ be i.i.d. draws from $\mathsf{N}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$. One of the simplest conditional methods for imputing missing values consists in the following iterative regression imputation: (1) initialize missing values arbitrary, using unconditional mean imputation; (2) impute missing values in one variable by the values predicted by the regression model of this variable with the remaining variables taken as explanatory ones, (3) iterate through variables containing missing values until convergence. Here, at each step, each point $\boldsymbol{x}_i$ with missing values at a coordinate j is imputed with the univariate conditional mean $\mathsf{E}[X|X_{\{1,\ldots,d\}\setminus\{j\}} = \boldsymbol{x}_{i,\{1,\ldots,d\}\setminus\{j\}}, \boldsymbol{\mu}_X = \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X = \boldsymbol{\Sigma}_X]$ with the moment estimates $\boldsymbol{\mu}_X = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$ and $\boldsymbol{\Sigma}_X = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu}_X)(\boldsymbol{x}_i - \boldsymbol{\mu}_X)^\top$. After convergence, each point $\boldsymbol{x}_i$ with missing values in $miss(i)$ is imputed with the multivariate conditional mean

$$\begin{aligned} &\mathsf{E}[X|X_{obs(i)} = \boldsymbol{x}_{i,obs(i)}, \boldsymbol{\mu}_X = \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X = \boldsymbol{\Sigma}_X] \\ &= \boldsymbol{\mu}_{\boldsymbol{X}\, miss(i)} + \boldsymbol{\Sigma}_{\boldsymbol{X}\, miss(i),obs(i)} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{X}\, obs(i),obs(i)} \cdot \boldsymbol{x}_{i,obs(i)} - \boldsymbol{\mu}_{\boldsymbol{X}\, obs(i)} \cdot \end{aligned} \tag{1}$$

The last expression is the closed-form solution to

$$\min_{\boldsymbol{z}_{miss(i)} \in \mathsf{R}^{|miss(i)|},\ \boldsymbol{z}_{obs(i)} = \boldsymbol{x}_{obs(i)}} d_M(\boldsymbol{z}, \boldsymbol{\mu}_{\boldsymbol{X}} | \boldsymbol{\Sigma}_{\boldsymbol{X}})$$

with $d_M^2(\boldsymbol{z}, \boldsymbol{\mu}_{\boldsymbol{X}} | \boldsymbol{\Sigma}_{\boldsymbol{X}}) = (\boldsymbol{z} - \boldsymbol{\mu}_{\boldsymbol{X}})^\top \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}_{\boldsymbol{X}})$ being the squared Mahalanobis distance from $\boldsymbol{z}$ to $\boldsymbol{\mu}_{\boldsymbol{X}}$. Minimizing the Mahalanobis distance can be seen as maximizing a centrality measure—the Mahalanobis depth:

$$\max_{\boldsymbol{z}_{miss(i)} \in \mathbb{R}^{|miss(i)|},\, \boldsymbol{z}_{obs(i)} = \boldsymbol{x}_{obs(i)}} D_n^M(\boldsymbol{z} | \boldsymbol{X})$$

where the Manahalobis depth of $\boldsymbol{x} \in \mathbb{R}^d$ w.r.t. X is defined as follows.

Definition 1. (Mahalanobis, 1936) *$D^M(\boldsymbol{x} | X) = \left(1 + (\boldsymbol{x} - \boldsymbol{\mu}_X)^\top \boldsymbol{\Sigma}_X^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_X)\right)^{-1}$, where $\boldsymbol{\mu}_X$ and $\boldsymbol{\Sigma}_X$ are the location and shape parameters of X.*

In its empirical version $D_n^M(\boldsymbol{x} | \boldsymbol{X})$, the parameters are replaced by their estimates.

The Manahalobis depth is the simplest instance of a statistical depth function. We now generalize the iterative imputation algorithm to other depths.

2.2 Imputation by depth maximization

2.2.1 Definition of data depth

Definition 2. (Zuo and Serfling, 2000*a*) *A bounded non-negative mapping $D(\cdot | X)$ from $\mathbb{R}^d$ to $\mathbb{R}$ is called a* statistical depth function *if it is* (P1) affine invariant, *i.e.*, *$D(\boldsymbol{x} | X) = D(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} | \boldsymbol{A}X + \boldsymbol{b})$ for any invertible $d \times d$ matrix $\boldsymbol{A}$ and any $\boldsymbol{b} \in \mathbb{R}^d$;* (P2) maximal at the symmetry center, *i.e.*, *$D(\boldsymbol{c} | X) = \sup_{\boldsymbol{x} \in \mathbb{R}^d} D(\boldsymbol{x} | X)$ for any X halfspace symmetric around $\boldsymbol{c}$ (A random vector X having distribution P_X is said to be halfspace symmetric around (a center) $\boldsymbol{c} \in \mathbb{R}^d$ if $P_X(H) \geq \frac{1}{2}$ for every halfspace H containing $\boldsymbol{c}$.);* (P3) monotone w.r.t. the deepest point, *i.e.*, *for any X having $\boldsymbol{c}$ as a deepest point, $D(\boldsymbol{x} | X) \leq D(\alpha \boldsymbol{c} + (1 - \alpha) \boldsymbol{x} | X)$ for any $\boldsymbol{x} \in \mathbb{R}^d$ and $\alpha \in [0, 1]$;* (P4) vanishing at infinity, *i.e.*, $\lim_{\|\boldsymbol{x}\| \to 0} D(\boldsymbol{x} | X) = 0$.

Additionally, we require (P5) *quasiconcavity* of $D(\cdot | X)$, upper-level sets (or depth-trimmed regions) $D_\alpha(X) = \{\boldsymbol{x} \in \mathbb{R}^d : D(\boldsymbol{x} | X) \geq \alpha\}$ to be convex, a useful property for optimization. We denote $D_n(\cdot | \boldsymbol{X})$ the corresponding empirical depth. See also Zuo and Serfling (2000*b*) for a reference on depth contours.

2.2.2 Imputation by depth maximization

We suggest a unified framework to impute missing values by depth maximization, which extends iterative regression imputation. More precisely, consider the following iterative scheme: (1) initialize missing values arbitrarily using unconditional mean imputation; (2) impute a point $\boldsymbol{x}$ containing missing coordinates with the point $\boldsymbol{y}$ maximizing data depth conditioned on observed values $\boldsymbol{x}_{obs}$:

$$\boldsymbol{y} = \operatorname*{argmax}_{\boldsymbol{z}_{miss} \in \mathbb{R}^{|miss|},\, \boldsymbol{z}_{obs} = \boldsymbol{x}_{obs}} D_n(\boldsymbol{z} | \boldsymbol{X}) \,; \tag{2}$$

(3) iterate until convergence.

The solution of (2) can be non-unique (see Figure 1 in the supplementary materials for an illustration) and the depth value may become zero immediately beyond the convex hull of the

support of the distribution. To avoid these problems, we suggest *imputation by depth* (ID) of an $\boldsymbol{x}$ which has missing values with $\boldsymbol{y} = ID\left(\boldsymbol{x}, D_n(\cdot|\boldsymbol{X})\right)$:

$$ID\left(\boldsymbol{x}, D_n(\cdot|\boldsymbol{X})\right) = \text{ave}\left(\underset{\boldsymbol{u}\in\mathbb{R}^d,\, \boldsymbol{u}_{obs}=\boldsymbol{x}_{obs}}{\arg\min} \{\|\boldsymbol{u}-\boldsymbol{v}\| \,|\, \boldsymbol{v}\in D_{n,a^*}(\boldsymbol{X})\}\right), \tag{3}$$

$$\text{with} \quad a^* = \inf_{a\in(0;1)}\left(a \,|\, D_{n,a}(\boldsymbol{X}) \cap \{\boldsymbol{z} \,|\, \boldsymbol{z}\in\mathbb{R}^d,\, \boldsymbol{z}_{obs}=\boldsymbol{x}_{obs}\} = \emptyset\right),$$

where ave is the averaging operator. The imputation by iterative maximization of depth is summarized in Algorithm 1. The complexity of Algorithm 1 is $O\left(N_s n_{miss}\Omega(D)\right)$. It depends on the data geometry and on the missing values (through the number of outer-loop iterations N_s necessary to achieve s-convergence), the number of points containing missing values n_{miss}, and the depth-specific complexities for solving (3) $\Omega(D)$ are detailed in subsections of Section 3.

Algorithm 1 Single imputation

```
function IMPUTE.DEPTH.SINGLE(X)
    Y ← X
    μ ← μ̂^(obs)(X)                          ▷ Calculate mean, ignoring missing values
    for i = 1 : n do
        if miss(i) ≠ ∅ then
            y_{i,miss(i)} ← μ_{miss(i)}      ▷ Impute with unconditional mean
    I ← 0
    repeat                                   ▷ Iterate until convergence or maximal iteration
        I ← I + 1
        Z ← Y
        for i = 1 : n do
            if miss(i) ≠ ∅ then
                y_i ← ID(y_i, D_n(·|Z))      ▷ Impute with maximum depth
    until max_{i∈{1,...,n},j∈{1,...,d}} |y_{i,j} − z_{i,j}| < s or I = I_max
    return Y
```

2.2.3 Theoretical properties for elliptical distributions

An elliptical distribution is defined as follows (see Fang et al. (1990), and Liu and Singh (1993) in the data depth context).

Definition 3. *A random vector X in $\mathbb{R}^d$ is elliptical if and only if there exists a vector $\boldsymbol{\mu}_X \in \mathbb{R}^d$ and $d \times d$ symmetric and positive semi-definite invertible matrix $\boldsymbol{\Sigma}_X = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$ such that for a random vector U uniformly distributed on the unit sphere $\mathbb{S}^{d-1}$ and a non-negative random variable R, it holds that $X \overset{D}{=} \boldsymbol{\mu}_X + R\boldsymbol{\Lambda}U$. We then say that $X \sim \mathbb{E}_d(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X, F_R)$, where F_R is the cumulative distribution function of the generating variate R.*

Theorem 1 shows that for an elliptical distribution, imputation of one point with a quasi-concave uniformly consistent depth converges to the center of the conditional distribution when conditioning on the observed values. Theorem 1 is illustrated in Figure 2 in the supplementary materials.

Theorem 1 (One row consistency)**.** *Let* $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ *be a data set in* $\mathbb{R}^d$ *drawn i.i.d. from* $X \sim \mathcal{E}_d(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X, F_R)$ *with* $d \geq 2$, F_R *absolutely continuous with strictly decreasing density, and let* $\boldsymbol{x} = (\boldsymbol{x}_{obs}, \boldsymbol{x}_{miss}) \in \mathbb{R}^d$ *with* $|obs(\boldsymbol{x})| \geq 1$. *Further, let* $D(\cdot|X)$ *satisfy (P1)–(P5) and* $D_{n,a}(\boldsymbol{X}) \xrightarrow[n\to\infty]{a.s.} D_a(X)$. *Then for* $\boldsymbol{y} = ID_{\boldsymbol{x}, D_n(\cdot|\boldsymbol{X})}(\cdot|\boldsymbol{X})$,

$$\cdot \boldsymbol{y}_{miss} - \boldsymbol{\mu}_{X\,miss} - \boldsymbol{\Sigma}_{X\,miss,obs}\boldsymbol{\Sigma}^{-1}_{X\,obs,obs}(\boldsymbol{x}_{obs} - \boldsymbol{\mu}_{X\,obs}) \cdot \xrightarrow[n\to\infty]{a.s.} 0\,.$$

Theorem 2 states that if missing values constitute a portion of the sample but are in a single variable, the imputed values converge to the center of the conditional distribution when conditioning on the observed values.

Theorem 2 (One column consistency)**.** *Let* $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ *be a data set in* $\mathbb{R}^d$ *drawn i.i.d. from* $X \sim \mathcal{E}_d(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X, F_R)$ *with* $d \geq 2$, F_R *absolutely continuous with strictly decreasing density, and let* $miss(i) = \{j\}$ *with probability* $p \in (0, 1)$ *for a fixed* $j \in \{1, \ldots, d\}$. *Let* $D(\cdot|Z)$ *satisfy (P1)–(P5) and* $D_{n,a}(\boldsymbol{Z}) \xrightarrow[n\to\infty]{a.s.} D_a(Z)$ *for* $Z = (1-p)X + pZ^j$ *with* $Z^j = \boldsymbol{\mu}_{X\,miss(i)} + \{\boldsymbol{\Sigma}^{-1}_{X_{-j},-j}(X_{-j} - \boldsymbol{\mu}_{X_{-j}})$ *if* $miss(i) = \{j\}$ *otherwise. Further, let* $\boldsymbol{Y}$ *exist such that* $\boldsymbol{y}_i = ID_{\boldsymbol{x}_i, D_n(\cdot|\boldsymbol{Y})}(\cdot|\boldsymbol{Y})$ *and denoting* $-j$ *for* $\{1, \ldots, d\} \setminus \{j\}$, *and* $\boldsymbol{y}_i = \boldsymbol{x}_i$

$$\cdot \boldsymbol{y}_{i,j} - \boldsymbol{\mu}_{Xj} - \boldsymbol{\Sigma}_{Xj,-j}\boldsymbol{\Sigma}^{-1}_{X_{-j},-j}(\boldsymbol{x}_{i,-j} - \boldsymbol{\mu}_{X_{-j}}) \cdot \xrightarrow[n\to\infty]{a.s.} 0\,.$$

3 Which depth to use?

The generality of the proposed methodology lies in the possibility of using any notion of depth which defines imputation properties. We focus here on imputation with Manahalobis, zonoid, and Tukey depths. These are of particular interest because they are quasiconcave and require two, one, and zero first moments of the underlying probability measure, respectively.

Corollary 1. *Theorems 1 and 2 hold for the Tukey depth, for the zonoid depth if* $\mathbb{E}[\|X\|] < \infty$, *and for the Mahalanobis depth if* $\mathbb{E}[\|X\|^2] < \infty$.

In addition, the function $f(\boldsymbol{z}_{miss}) = D_n(\boldsymbol{z}|\boldsymbol{X})$ subject to $\boldsymbol{z}_{obs} = \boldsymbol{x}_{obs}$ in equation (2), iteratively optimized in Algorithm 1, is quadratic for the Mahalanobis depth, continuous inside $\mathrm{conv}(\boldsymbol{X})$ (the smallest convex set containing $\boldsymbol{X}$) for the zonoid depth, and stepwise discrete for the Tukey depth, which in all cases leads to efficient implementations. For a trivariate Gaussian sample, $f(\boldsymbol{z}_{miss})$ is depicted in Figure 1 in the supplementary materials.

The use of a non-quasiconcave depth (e.g., simplicial, spatial (Nagy, 2017), etc.) results in non-convex optimization when maximizing depth, and this non-stability impedes numerical convergence of the algorithm.

3.1 Mahalanobis depth

Imputation with the Mahalanobis depth is related to existing methods. First, we show the link with the minimization of the covariance determinant.

Proposition 1 (Covariance determinant is quadratic in a point's missing entries)**.** *Let* $\boldsymbol{X}(\boldsymbol{y}) = \boldsymbol{x}_1, \ldots, (\boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,|obs(i)|}, \boldsymbol{y}^\top)^\top, \ldots, \boldsymbol{x}_n^\top$ *be a* $n \times d$ *matrix with* $\boldsymbol{\Sigma}_X(\boldsymbol{y})$ *invertible for all* $\boldsymbol{y} \in \mathbb{R}^{|miss(i)|}$. *Then* $|\boldsymbol{\Sigma}_X(\boldsymbol{y})|$ *is quadratic and globally minimized in* $\boldsymbol{y} = \boldsymbol{\mu}_{X\,miss(i)}(\boldsymbol{y}) + \boldsymbol{\Sigma}_{X\,miss(i),obs(i)}(\boldsymbol{y})\boldsymbol{\Sigma}^{-1}_{X\,obs(i),obs(i)}(\boldsymbol{y}) \cdot (\boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,|obs(i)|}) - \boldsymbol{\mu}_{X\,obs(i)}$.

From Proposition 1 it follows that the minimum of the covariance determinant is unique and the determinant itself decreases at each iteration. Thus, to impute points with missing coordinates one-by-one and iterate until convergence constitutes the block coordinate descent method, which can be proved to numerically converge due to Proposition 2.7.1 from Bertsekas (1999) (as long as $\boldsymbol{\Sigma}_X$ is invertible).

Further, Theorem 3 states that imputation using the maximum Mahalanobis depth, iterative (multiple-output) regression, and regularized PCA (Josse and Husson, 2012) with $S = d - 1$ dimensions, all converge to the same imputed sample.

Theorem 3. *Suppose that we impute $\boldsymbol{X} = (\boldsymbol{X}_{miss}, \boldsymbol{X}_{obs})$ in $\mathbb{R}^d$ with $\boldsymbol{Y}$ so that $\boldsymbol{y}_i = \operatorname{argmax}_{\boldsymbol{z}_{obs(i)} = \boldsymbol{y}_{obs(i)}} D_n^M(\boldsymbol{z}|\boldsymbol{Y})$ for each i with $|miss(i)| > 0$ and $\boldsymbol{y}_i = \boldsymbol{x}_i$ otherwise. Then for each such $\boldsymbol{y}_i$, it also holds that:*

- *$\boldsymbol{x}_i$ is imputed with the* **conditional mean***:*

$$\boldsymbol{y}_{i,miss(i)} = \boldsymbol{\mu}_{\boldsymbol{Y}\, miss(i)} + \boldsymbol{\Sigma}_{\boldsymbol{Y}\, miss(i),obs(i)} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{Y}\, obs(i),obs(i)} (\boldsymbol{x}_{obs(i)} - \boldsymbol{\mu}_{\boldsymbol{Y}\, obs(i)})$$

  which is equivalent to **single**- *and* **multiple**-*output* **regression**,
- *$\boldsymbol{Y}$ is a* **stationary point** *of $|\boldsymbol{\Sigma}_{\boldsymbol{X}}(\boldsymbol{X}_{miss})|$: $\frac{\partial |\boldsymbol{\Sigma}_{\boldsymbol{X}}|}{\partial \boldsymbol{X}_{miss}}(\boldsymbol{Y}_{miss}) = \boldsymbol{0}$, and*
- *each missing coordinate j of $\boldsymbol{x}_i$ is imputed with* **regularized PCA** *as in Josse & Husson (2012) with any $0 < \sigma^2 \leq \lambda_d$ and with $\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^\top$ the singular value decomposition (SVD): $\boldsymbol{y}_{i,j} = \sum_{s=1}^{d} \boldsymbol{U}_{i,s} \frac{\lambda_s - \sigma^2}{\lambda_s} \boldsymbol{V}_{j,s} + \boldsymbol{\mu}_{\boldsymbol{Y}\, j}$.*

The first point of the theorem sheds light on the connection between imputation by Mahalanobis depth and the iterative regression imputation of Section 2.1. When the Mahalanobis depth is used in Algorithm 1, each $\boldsymbol{x}_i$ with missingness in $miss(i)$ is imputed by the multivariate conditional mean as in equation (1), and thus lies in the $d - |miss(i)|$-dimensional multiple-output regression subspace of $\boldsymbol{X}_{\cdot,miss(i)}$ on $\boldsymbol{X}_{\cdot,obs(i)}$. This subspace is obtained as the intersection of the single-output regression hyperplanes $\boldsymbol{X}_{\cdot,j}$ on $\boldsymbol{X}_{\cdot,\{1,\ldots,d\}\setminus\{j\}}$ for all $j \in miss(i)$ corresponding to missing coordinates. The third point strengthens the method as imputation with regularized PCA has proved to be highly efficient in practice due to its sticking to low-rank structure of importance and ignoring noise.

The complexity of imputing a single point with the Mahalanobis depth is $O(nd^2 + d^3)$. Despite its good properties, it is not robust to outliers. However, robust estimates for $\boldsymbol{\mu}_{\boldsymbol{X}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ can be used, e.g., the minimum covariance determinant ones (MCD, see Rousseeuw and Van Driessen, 1999).

3.2 Zonoid depth

Koshevoy and Mosler (1997) define a zonoid trimmed region, with $a \in (0, 1]$, as

$$D_a^z(X) = \left\{ \int_{\mathbb{R}^d} \boldsymbol{x} g(\boldsymbol{x}) dP_X(\boldsymbol{x}) : g : \mathbb{R}^d \to \left[0, \frac{1}{a}\right] \text{ measurable and } \int_{\mathbb{R}^d} g(\boldsymbol{x}) dP_X(\boldsymbol{x}) = 1 \right\},$$

and for $a = 0$ as $D_0^z(X) = \mathrm{cl} \cup_{a \in (0,1]} D_a^z(X)$, where cl denotes the closure. Its empirical version can be defined as

$$D_{n,a}^z(\boldsymbol{X}) = \left\{ \sum_{i=1}^{n} \lambda_i \boldsymbol{x}_i : \sum_{i=1}^{n} \lambda_i = 1,\ \lambda_i \geq 0,\ a\lambda_i \leq \frac{1}{n}\ \forall\ i \in \{1, \ldots, n\} \right\}.$$

Definition 4. (Koshevoy and Mosler, 1997) *The zonoid depth of* $\boldsymbol{x}$ *w.r.t.* X *is defined as*

$$D^z(\boldsymbol{x}|X) = \begin{cases} \sup\{\alpha : \boldsymbol{x} \in D^z_\alpha(X)\} & \text{if } \boldsymbol{x} \in \operatorname{conv}\operatorname{supp}(X) \\ 0 & \text{otherwise.} \end{cases}$$

For a comprehensive reference on the zonoid depth, the reader is referred to Mosler (2002).

Imputation of a point $\boldsymbol{x}_i$ in Algorithm 1 is then performed by a slight modification of the linear programming for computation of zonoid depth with variables γ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^\top$:

$$\min \gamma \quad \text{s.t.} \quad \boldsymbol{X}^\top_{\cdot,obs(i)}\boldsymbol{\lambda} = \boldsymbol{x}_{i,obs(i)}\,,\ \boldsymbol{\lambda}^\top\mathbf{1}_n = 1\,,\ \gamma\mathbf{1}_n - \boldsymbol{\lambda} \geq \mathbf{0}_n\,,\ \boldsymbol{\lambda} \geq \mathbf{0}_n\,.$$

Here $\boldsymbol{X}_{\cdot,obs(i)}$ stands for the completed $n \times |obs(i)|$ data matrix containing columns corresponding only to non-missing coordinates of $\boldsymbol{x}_i$, and $\mathbf{1}_n$ (respectively $\mathbf{0}_n$) is a vector of ones (respectively zeros) of length n. In the implementation, we use the simplex method, which is known for being fast despite its exponential complexity. This implies that, for each point $\boldsymbol{x}_i$, imputation is performed by the weighted mean:

$$\boldsymbol{y}_{i,miss(i)} = \boldsymbol{X}^\top_{\cdot,miss(i)}\boldsymbol{\lambda}\,,$$

the average of the maximum number of equally weighted points. Additional insight on the position of imputed points with respect to the sample can be gained by inspecting the optimal weights λ_i. Zonoid imputation is related to local methods such as as kNN imputation, as only some of the weights are positive.

3.3 Tukey depth

Definition 5. (Tukey, 1975) *The Tukey depth of* $\boldsymbol{x}$ *w.r.t.* X *is defined as* $D^T(\boldsymbol{x}|X) = \inf\{P_X(H) : H \text{ a closed halfspace}, \boldsymbol{x} \in H\}$.

In the empirical version, the probability is substituted by the portion of $\boldsymbol{X}$ giving $D^T(\boldsymbol{x}|\boldsymbol{X}) = \min_{\boldsymbol{u}\in S^{d-1}}\{\frac{1}{n}\cdot|\{i : \boldsymbol{x}_i^\top\boldsymbol{u} \geq \boldsymbol{x}^\top\boldsymbol{u}\,,\ i = 1, \ldots, n\}|\}$. For more information on Tukey depth see Donoho and Gasko (1992).

With nonparametric imputation by Tukey depth, one can expect that after convergence of Algorithm 1, for each point initially containing missing values, it holds that $\boldsymbol{y}_i = \operatorname{argmax}_{\boldsymbol{z}_{obs}=\boldsymbol{x}_{obs}} \min_{\boldsymbol{u}\in S^{d-1}}\{|\{k : \boldsymbol{y}_k^\top\boldsymbol{u} \geq \boldsymbol{z}^\top\boldsymbol{u},\ k \in \{1, \ldots, n\}\}|\}$. Thus, imputation is performed according to the maximin principle based on criteria involving indicator functions, which implies robustness of the solution. Note that as the Tukey depth is not continuous, the searched-for maximum (2) may be non-unique (see Figure 1 (top right) in the supplementary materials), and we impute with the barycenter of the maximizing arguments (3). Due to the combinatorial nature of the Tukey depth, to speed up implementation, we run $2d$ times the Nelder-Mead downhill-simplex algorithm, and take the average over the solutions. The imputation is illustrated in Figure 3 in the supplementary materials.

The Tukey depth can be computed exactly (Dyckerhoff and Mozharovskyi, 2016) with complexity $O(n^{d-1}\log n)$, although to avoid computational burden we also implement its approximation with random directions (Dyckerhoff, 2004) having complexity $O(kn)$, with k denoting the number of random directions. All of the experiments are performed with exactly computed Tukey depth, unless stated otherwise.

3.4 Beyond ellipticity: local depth

Imputation with the so-called "global depth" from Definition 2 may be appropriate in applications even if the data moderately deviate from ellipticity (see Section 4.2.6). However, it can fail when the distribution has non-convex support or several modes. A solution is to use the local depth in Algorithm 1.

Definition 6. (Paindaveine and Bever, 2013) *For a depth $D(\cdot|X)$, the β-local depth is defined as $LD^{\beta}(\cdot, X) : \mathbb{R}^d \to \mathbb{R}^+ : \boldsymbol{x} \mapsto LD^{\beta}(\boldsymbol{x}, X) = D(\boldsymbol{x}|X^{\beta,\boldsymbol{x}})$ with $X^{\beta,\boldsymbol{x}}$ the conditional distribution of X conditioned on $\bigcap_{a\geq 0,\, P_Y(D_a(Y))\geq\beta} D_a(Y)$, where Y has the distribution $P_Y = \frac{1}{2} P_X + \frac{1}{2} P_{2\boldsymbol{x}-}$*

The locality level β should be chosen in a data-driven way, for instance by cross-validation. An important advantage of this approach is that any depth satisfying Definition 2 can be plugged in to the local depth. We suggest using the Nelder-Mead algorithm to enable imputation with maximum local depth regardless of the chosen depth notion.

3.5 Dealing with outsiders

A number of depths that exploit the geometry of the data are equal to zero beyond $\mathrm{conv}(\boldsymbol{X})$, including the zonoid and Tukey depths. Although (3) deals with this situation, for a finite sample it means that points with missing values having the maximal value in at least one of the observed coordinates will never move from the initial imputation because they will become vertices of the $\mathrm{conv}(\boldsymbol{X})$. For the same reason, other points to be imputed and lying exactly on the $\mathrm{conv}(\boldsymbol{X})$ will not move much during imputation iterations. As such points are not numerous and would need to move quite substantially to influence imputation quality, we impute them—during the initial iterations—using the spatial depth function (Vardi and Zhang, 2000), which is everywhere non-negative. This resembles the so-called "outsider treatment" introduced by Lange et al. (2014). Another possibility is to extend the depth beyond $\mathrm{conv}(\boldsymbol{X})$, see e.g., Einmahl et al. (2015) for the Tukey depth.

4 Experimental study

4.1 Choice of competitors

We assess the prediction abilities of Tukey, zonoid, and Mahalanobis depth imputation, and the robust Mahalanobis depth imputation using MCD mean and covariance estimates, with the robustness parameter chosen in an optimal way due to knowledge of the simulation setting. We measure their performance against the competitors: conditional mean imputation based on EM estimates of the mean and covariance matrix; regularized PCA imputation with rank 1 and 2; two nonparametric imputation methods: random forest (using the default implementation in the `R`-package `missForest`), and kNN imputation choosing k from $\{1, \ldots, 15\}$, minimizing the imputation error over 10 validation sets as in Stekhoven and Bühlmann (2012). Mean and oracle (if possible) imputations are used to benchmark the results.

4.2 Simulated data

4.2.1 Elliptical setting with Student-t distribution

We generate 100 points according to an elliptical distribution (Definition 3) with $\boldsymbol{\mu}_2 = (1, 1, 1)^\top$ and the shape $\boldsymbol{\Sigma}_2 = (1, 1, 1)^\top, (1, 4, 4)^\top, (1, 4, 8)^\top$, where F is the univariate Student-t distribution ranging in number of degrees of freedom (d.f.) from the Gaussian to the Cauchy: $t = \infty, 10, 5, 3, 2, 1$. For each of the 1000 simulations, we remove 5%, 15% and 25% of values completely at random (MCAR), and compute the median and the median absolute deviation from the median (MAD) of the root mean square error (RMSE) of each imputation method. Table 1 presents the results for 25% missing values. The conclusions with other percentages (see the supplementary materials) are the same, but as expected, performances decrease with increasing percentage of missing data.

Distr.	D^{Tuk}	D^{zon}	D^{Mah}	$D^{Mah}_{MCD,\mathbf{75}}$	EM	regPCA1	regPCA2	kNN	RF	mean	oracle
$t\ \infty$	1.675	*1.609*	1.613	1.991	**1.575**	1.65	1.613	1.732	1.763	2.053	1.536
	(0.205)	(*0.1893*)	(0.1851)	(0.291)	(**0.1766**)	(0.1846)	(0.1856)	(0.2066)	(0.2101)	(0.2345)	(0.1772)
$t\ 10$	1.871	1.81	*1.801*	2.214	**1.755**	1.836	*1.801*	1.923	1.96	2.292	1.703
	(0.2445)	(0.2395)	(*0.2439*)	(0.3467)	(**0.2379**)	(0.2512)	(*0.2433*)	(0.2647)	(0.2759)	(0.2936)	(0.2206)
$t\ 5$	2.143	2.089	*2.079*	2.462	**2.026**	2.108	*2.08*	2.235	2.259	2.612	1.949
	(0.3313)	(0.3331)	(*0.3306*)	(0.4323)	(**0.3144**)	(0.3431)	(*0.3307*)	(0.3812)	(0.3656)	(0.3896)	(0.3044)
$t\ 3$	2.636	2.603	2.62	2.946	**2.516**	*2.593*	2.619	2.757	2.79	3.165	2.384
	(0.5775)	(0.5774)	(0.5745)	(0.6575)	(**0.5537**)	(*0.561*)	(0.5741)	(0.5874)	(0.5856)	(0.6042)	(0.5214)
$t\ 2$	**3.563**	3.73	3.738	3.989	*3.567*	3.692	3.738	3.798	3.849	4.341	3.175
	(**1.09**)	(1.236)	(1.183)	(1.287)	(*1.146*)	(1.186)	(1.19)	(1.133)	(1.19)	(1.252)	(0.9555)
$t\ 1$	*16.58*	19.48	19.64	**16.03**	18.5	18.22	19.61	17.59	17.48	20.32	13.55
	(*13.71*)	(16.03)	(16.2)	(**12.4**)	(15.46)	(15.02)	(16.1)	(14.59)	(14.33)	(16.36)	(10.71)

Table 1: Median and MAD of the RMSEs of the imputation for a sample of 100 points drawn from elliptically symmetric Student-t distributions with $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_2$ with 25% of MCAR values, over 1000 repetitions. Bold values indicate the best results, italics the second best.

As expected, the behavior of the different imputation methods changes with the number of d.f., as does the overall leadership trend. For the Cauchy distribution, robust methods perform best: Mahalanobis depth-based imputation using MCD estimates, followed closely by the one using Tukey depth. For 2 d.f., when the first moment exists but not the second, EM- and Tukey-depth-based imputations perform similarly, with a slight advantage to the Tukey depth in terms of MAD. For larger numbers of d.f., when two first moments exist, EM takes the lead. It is followed by the group of regularized PCA methods, and Mahalanobis- and zonoid-depth-based imputation. Note that the Mahalanobis depth and regularized PCA with rank two perform similarly (the small difference can be explained by numerical precision considerations), see Theorem 3. Both nonparametric methods perform poorly, being "unaware" of the ellipticity of the underlying distribution, but give reasonable results for the Cauchy distribution because of insensitivity to correlation. By default, we present the results obtained with spatial depth for the outsiders. For the Tukey depth, implementation is also available using the extension by Einmahl et al. (2015).

4.2.2 Contaminated elliptical setting

We then modify the above setting by adding 15% of outliers (which do not contain missing values) that stem from the Cauchy distribution with the same parameters $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_2$. As expected, Table 2 shows that the best RMSEs are obtained by the robust imputation methods: Tukey depth and Mahalanobis depth with MCD estimates. Being restricted to a neighborhood, nonparametric methods often impute based on non-outlying points, and thus perform less well as the preceding

Distr.	D^{Tuk}	D^{zon}	D^{Mah}	$D^{Mah}_{MCD,\mathbf{.75}}$	EM	regPCA1	regPCA2	kNN	RF	mean	oracle
$t\ \infty$	**1.751**	1.86	1.945	*1.81*	1.896	1.958	1.945	1.859	1.86	2.23	1.563
	(**0.2317**)	(0.3181)	(0.4299)	(*0.239*)	(0.3987)	(0.4495)	(0.4328)	(0.2602)	(0.2332)	(0.3304)	(0.1849)
$t\ 10$	**1.942**	2.087	2.165	*2.022*	2.112	2.196	2.165	2.051	2.047	2.48	1.733
	(**0.2976**)	(0.4295)	(0.5473)	(*0.3128*)	(0.5226)	(0.5729)	(0.5479)	(0.3143)	(0.3043)	(0.4163)	(0.2266)
$t\ 5$	**2.178**	2.333	2.421	*2.231*	2.376	2.398	2.421	2.315	2.325	2.766	1.939
	(**0.3556**)	(0.4924)	(0.6026)	(*0.381*)	(0.5715)	(0.6035)	(0.5985)	(0.3809)	(0.3946)	(0.528)	(0.2979)
$t\ 3$	**2.635**	2.864	2.935	*2.664*	2.828	2.916	2.93	2.797	2.838	3.34	2.356
	(**0.6029**)	(0.7819)	(0.8393)	(*0.5877*)	(0.7773)	(0.8221)	(0.8384)	(0.6045)	(0.6228)	(0.7721)	(0.4946)
$t\ 2$	**3.763**	4.082	4.136	*3.783*	4.036	4.09	4.14	3.955	4.026	4.623	3.323
	(**1.17**)	(1.535)	(1.501)	(*1.224*)	(1.518)	(1.585)	(1.503)	(1.265)	(1.354)	(1.561)	(1.04)
$t\ 1$	*17.17*	20.43	20.27	**16.46**	19.01	19.81	20.53	18.96	19.04	21.04	14.44
	(*13.27*)	(15.99)	(15.91)	(**12.94**)	(15.21)	(16.15)	(16.28)	(14.73)	(14.62)	(15.56)	(11.33)

Table 2: Median and MAD of the RMSEs of the imputation for 100 points drawn from elliptically symmetric Student-t distributions, with $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_2$ contaminated with 15% of outliers, and 25% of MCAR values on non-contaminated data, repeated 1000 times.

group. The rest of the included imputation methods cannot deal with the contaminated data and perform rather poorly.

4.2.3 The MAR setting

We next generate highly correlated Gaussian data by setting $\boldsymbol{\mu}_3 = (1,\ 1,\ 1)$ and the covariance matrix to $\boldsymbol{\Sigma}_3 = (1,\ 1.75,\ 2)^\top,\ (1.75,\ 4,\ 4)^\top,\ (2,\ 4,\ 8)^\top$. We insert missing values according to the MAR mechanism: the first and third variables are missing depending on the value of the second variable. Figure 3 (left) shows the boxplots of the RMSEs. As we expected, semiparametric

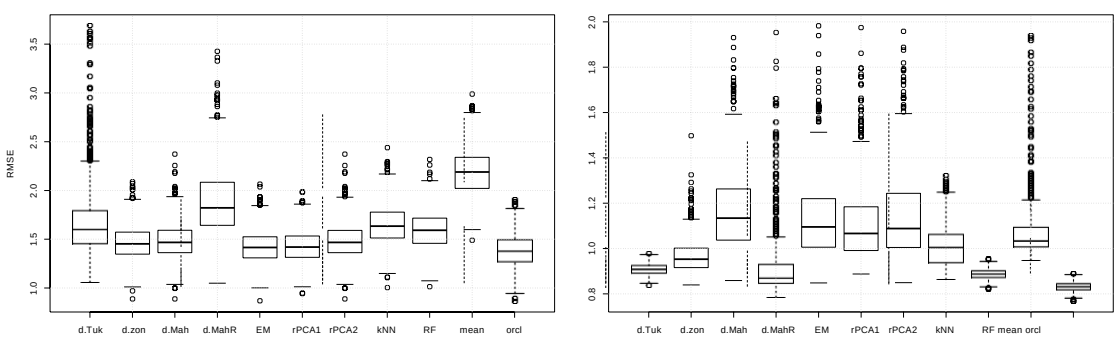

Figure 3: Left: RMSEs for different imputation methods for 100 points drawn from a correlated 3-dimensional Gaussian distribution with $\boldsymbol{\mu}_3$ and $\boldsymbol{\Sigma}_3$ with MAR values (see implementation for details), over 1000 repetitions. Right: 1000 points drawn from a 6-dimensional Gaussian distribution with $\boldsymbol{\mu}_4$ and $\boldsymbol{\Sigma}_4$ contaminated with 15% of outliers, and 15% of MCAR values on non-contaminated data, over 500 repetitions.

methods (EM, regularized PCA and Mahalanobis depth) perform close to the oracle imputation. The good performance of the rank 1 regularized PCA can be explained by the high correlation between variables. The zonoid depth imputes well despite having no parametric knowledge. Nonparametric methods are unable to capture the correlation, while robust methods "throw away" points possibly containing valuable information.

4.2.4 The low-rank model

We consider as a stress-test an extremely contaminated low-rank model by adding to a 2-dimensional low-rank structure a Cauchy-distributed noise. Generally, while capturing any structure is rather meaningless in this setting (confirmed by the high MADs in Table 3), the performance of the methods is "proportional to the way they ignore" dependency information. For this reason, mean imputation as well as nonparametric methods perform best. The Tukey and zonoid depths perform second best by accounting only for fundamental features of the data. This can be also said about the regularized PCA when keeping the first principal component only. The remaining methods try to reflect the data structure, but are distracted either by the low rank or the heavy-tailed noise.

	D^{Tuk}	D^{zon}	D^{Mah}	$D^{Mah}_{MCD.75}$	EM	regPCA1	regPCA2	kNN	RF	mean
Median RMSE	0.4511	0.4536	0.4795	0.5621	0.4709	0.4533	0.4664	**0.4409**	0.4444	*0.4430*
Mad of RMSE	0.3313	0.3411	0.3628	0.4355	0.3595	0.3461	0.3554	**0.3302**	0.3389	*0.3307*

Table 3: Medians and MADs of the RMSE for a rank-two model in R^4 of 50 points with Cauchy noise and 20% of missing values according to MCAR, over 1000 repetitions.

4.2.5 Contamination in higher dimensions

To check the resistance to outliers in higher dimensions, we consider a simulation setting similar to that of Section 4.2.2, in dimension 6, with a normal multivariate distribution with $\boldsymbol{\mu}_4 = (0, \ldots, 0)^\top$ and a Toeplitz covariance matrix $\boldsymbol{\Sigma}_4$ (having $\sigma_{i,j} = 2^{-|i-j|}$ as entries). The data are contaminated with 15% of outliers and have 15% of MCAR values on non-contaminated data. The Tukey depth is approximated using 1000 random directions. Figure 3 (right) shows that the Tukey depth imputation has high predictive quality, comparable to that of the random forest imputation even with only 1000 random directions.

4.2.6 Skewed distributions and distributions with non-convex support

First, let us consider only a slight deviation from ellipticity. We simulate 150 points from a skewed normal distribution (Azzalini and Capitanio, 1999), insert 15% MCAR values, and impute them with global (Tukey, zonoid and Mahalanobis) depths and their local versions (see Section 3.4). This is shown in Figure 4. In this setting, both global and local imputation perform similarly.

Further, let us consider an extreme departure from elliplicity with the moon-shaped example from Paindaveine and Bever (2013). We generate 150 bivariate observations from $(X_1, X_2)^\top$ with $X_1 \sim U(-1, 1)$ and $X_2 | X_1 = x_1 \sim U\left(1.5(1 - x_1^2), 2(1 - x_1^2)\right)$, and introduce 15% of MCAR values on X_2, see Figure 5 (left). Figure 5 (right) shows boxplots of the RMSE for single imputation using local Tukey, zonoid and Mahalanobis depths. If the depth and value of β are properly chosen (this can be achieved by cross-validation), the local-depth imputation considerably outperforms the classical methods as well as the global depth.

4.3 Real data

We validate the proposed methodology on three real data sets taken from the UCI Machine Learning Repository (Dua and Karra Taniskidou, 2017) and on the Cows data set. We thus consider

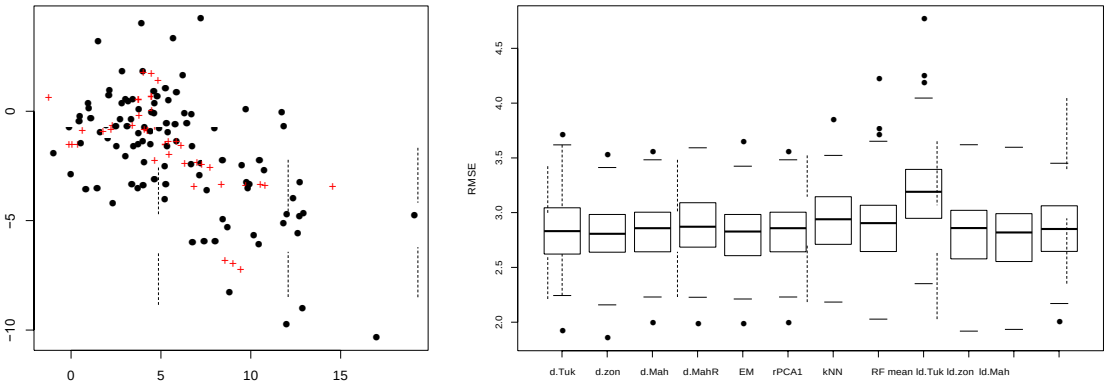

Figure 4: Left: An example of Tukey depth imputation (pluses). Right: boxplots of RMSEs of the prediction for 150 points drawn from a skewed distribution with 15% MCAR, over 100 repetitions; ld.* stands for the local depth with $\beta = 0.8$.

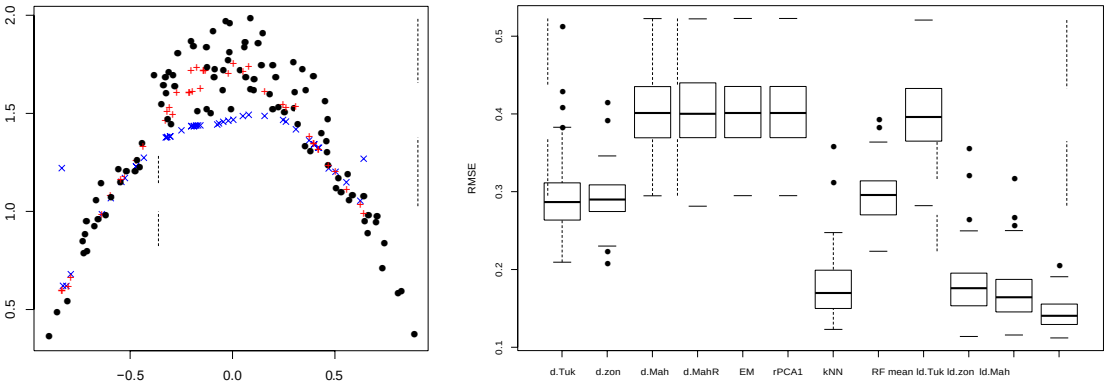

Figure 5: Left: Comparison of global (crosses) and local (pluses) Tukey depth imputation. Right: boxplots of RMSEs of predictions for 150 points drawn from the moon-shaped distribution with 15% MCAR values in the second coordinate, over 100 repetitions; ld.* stands for the local depth with $\beta = 0.2$.

Banknotes ($n = 100$, $d = 3$), Glass ($n = 76$, $d = 3$), Blood Transfusion ($n = 502$, $d = 3$, Yeh et al., 2009), and Cows ($n = 3454$, $d = 6$). For details on the experimental design, see the implementation. Figure 6 shows boxplots of the RMSEs for the ten imputation methods considered. The zonoid depth is stable across data sets and provides the best results.

Observations in Banknotes are clustered in two groups, which explains the poor performance of the mean and one-dimensional regularized PCA imputation. The zonoid depth searches for a compromise between local and global features and performs the best. The Tukey depth captures the data geometry, but under-exploits information on points' location. Methods imputing by conditional mean (Mahalanobis depth, EM-based, and regularized PCA imputation) perform similarly and reasonably well while imputing in two-dimensional affine subspaces. The Glass data is challenging as it highly deviates from ellipticity, and part of the data lie sparsely in part

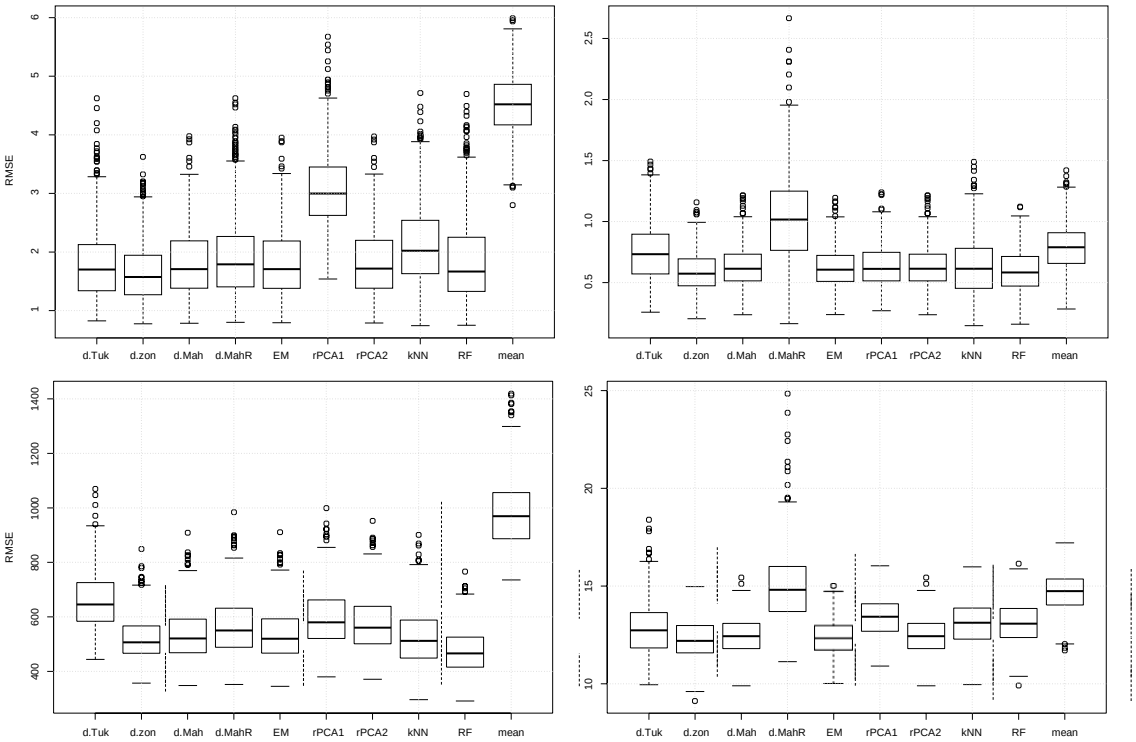

Figure 6: RMSEs for the Banknotes (top, left), Glass (top, right), Cows (bottom, left), and Blood Transfusion (bottom, right) data sets with 15% (5% for Cows) of MCAR values over 500 repetitions.

of the space, but do not seem to be outlying. Thus, the mean, and robust Mahalanobis and Tukey depth imputation perform poorly. Accounting for local geometry, random forest and zonoid depth perform slightly better. For the Cows data, which is larger-dimensional, the best results are obtained with random forest, but followed closely by zonoid depth imputation which reflects the data structure. The Tukey depth with 1000 directions struggles, while the satisfactory results of EM suggest that the data are close to elliptical. The Blood Transfusion data visually resemble a tetrahedron dispersed from one of its vertices. Thus, mean imputation can be substantially improved. Nonparametric methods and rank one regularized PCA perform poorly because they disregard dependency between dimensions. Better imputation is delivered by those capturing correlation: the depth- and EM-based methods.

Table 4 shows the time taken by different imputation methods. Zonoid imputation is very fast, and the approximation scheme by Dyckerhoff (2004) allows for a scalable application of the Tukey depth.

5 Multiple imputation for the elliptical family

When the objective is to predict missing entries as well as possible, single imputation is well suited. When analyzing complete data, it is important to go further, so as to better reflect the uncertainty in predicting missing values. This can be done with multiple imputation (MI)(Little

Data set	D^{Tuk}	D^{zon}	D^{Mah}	D^{Mah}_{MCD}
Higher dimension (Section 4.2.5) ($n = 1000$, $d = 6$)	3230*	1210	0.102	1.160
Banknotes ($n = 100$, $d = 3$)	81.2	0.376	0.010	0.126
Glass ($n = 76$, $d = 3$)	12.6	0.143	0.008	0.085
Cows ($n = 3454$, $d = 6$)	4490*	14300	0.212	2.37
Blood transfusion ($n = 502$, $d = 3$)	26400	51.3	0.46	0.775

Table 4: Median (in seconds, over 35 runs) execution time for depth-based imputation. * indicates approximate Tukey depth with 1000 random directions.

and Rubin, 2002) where several plausible values are generated for each missing entry, leading to several imputed data sets. MI then applies a statistical method to each imputed data set, and aggregates the results for inference. Under the Gaussian assumption, the generation of several imputed data sets is achieved by drawing missing values from the Gaussian conditional distribution of the missing entries, e.g., imputing $\boldsymbol{x}_{miss}$ by draws from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ conditional on $\boldsymbol{x}_{obs}$, with the mean and covariance matrix estimated by EM. This method is called stochastic EM. The objective is to impute close to the underlying distribution. However, this is not enough to perform *proper* (Little and Rubin, 2002) multiple imputation, since uncertainty in the imputation model's parameters must also be reflected. This is usually obtained either using a bootstrap or Bayesian approach, see e.g., Schafer (1997), Efron (1994), van Buuren (2012) for more details.

The generic framework of depth-based single imputation developed above allows for multiple imputation to be extended to the more general elliptical framework. We first show how to reflect the uncertainty due to the distribution (Section 5.1), then apply bootstrap to reflect model uncertainty, and state the complete algorithm (Section 5.2).

5.1 Stochastic single depth-based imputation

The extension of stochastic EM to the elliptically symmetric distribution consists in drawing from a conditional distribution that is also elliptical. For this we design a Monte Carlo Markov chain (MCMC), see Figure 7 for an illustration of a single iteration. First, starting with a point with missing values and observed values $\boldsymbol{x}_{obs}$, we impute it with $\boldsymbol{\mu}^*$ by maximizing its depth (3), see Figure 7 (right). Then, for each $\boldsymbol{y}$ with $\boldsymbol{y}_{obs} = \boldsymbol{x}_{obs}$ it holds that $D(\boldsymbol{y}|\boldsymbol{X}) \leq D(\boldsymbol{\mu}^*|\boldsymbol{X})$. The cumulative distribution function (with the normalization constant omitted as it is used, exceptionally, for drawing random variables) of the depth of the random vector Y corresponding to $\boldsymbol{y}$ can be written as

$$F_{D(Y|X)}(y) = \int_0^y f_{D(X|X)}(z) \frac{d_M^{d-1}(z)}{\sqrt{d_M^2(z) - d_M^2 \cdot D(\boldsymbol{\mu}^*|X)}^{\Sigma_{|miss(x)|-1}}} \times \frac{d_M(z)}{\sqrt{d_M^2(z) - d_M^2 \cdot D(\boldsymbol{\mu}^*|X)}^{\Sigma}} dz, \tag{4}$$

where $f_{D(X|X)}$ denotes the density of the depth for a random vector X w.r.t. itself, and $d_M(z)$ is the Mahalanobis distance to the center as a function of depth (see the supplementary materials for the derivation of this). For the specific case of the Mahalanobis depth, $d_M(x) = \sqrt{1/x - 1}$. Then, we draw a quantile Q uniformly on $[0, F_{D(Y|X)}(D(\boldsymbol{\mu}^*|\boldsymbol{X}))]$ that gives the value of

the depth of $\boldsymbol{y}$ as $a = F^{-1}_{D(Y|X)}(Q)$, see Figure 7 (left). a defines a depth contour, which is depicted as an ellipsoid in Figure 7 (right). Finally, we draw $\boldsymbol{y}$ uniformly in the intersection of this contour with the hyperplane of missing coordinates: $\boldsymbol{y} \in \partial D_a(\boldsymbol{X}) \cap \{\boldsymbol{z} \in \mathsf{R}^d \mid \boldsymbol{z}_{obs(\boldsymbol{x})} = \boldsymbol{x}_{obs}\}$. This is done by drawing $\boldsymbol{u}$ uniformly on $\mathsf{S}^{|miss(\boldsymbol{x})|-1}$ and transform- ing it using a conditional scatter matrix, obtaining $\boldsymbol{u}^* \in \mathsf{R}^d$, where $\boldsymbol{u}^*_{miss(\boldsymbol{x})} = \boldsymbol{\Lambda u}$ (with $\boldsymbol{\Lambda}(\boldsymbol{\Lambda})^\mathsf{T} = \boldsymbol{\Sigma}_{miss(\boldsymbol{x}),miss(\boldsymbol{x})} - \boldsymbol{\Sigma}_{miss(\boldsymbol{x}),obs(\boldsymbol{x})}\boldsymbol{\Sigma}^{-1}_{obs(\boldsymbol{x}),obs(\boldsymbol{x})}\boldsymbol{\Sigma}_{obs(\boldsymbol{x}),miss(\boldsymbol{x})}$) and $\boldsymbol{u}^*_{obs(\boldsymbol{x})} = \boldsymbol{0}$. Such a $\boldsymbol{u}^*$ is uniformly distributed on the conditional depth contour. Then $\boldsymbol{x}$ is imputed as $\boldsymbol{y} = \boldsymbol{\mu}^* + \beta\boldsymbol{u}^*$, where β is a scalar obtained as the positive solution of $\boldsymbol{\mu}^* + \beta\boldsymbol{u}^* \in \partial D_a(\boldsymbol{X})$ (e.g., the quadratic equation $(\boldsymbol{\mu}^* + \beta\boldsymbol{u}^* - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^* + \beta\boldsymbol{u}^* - \boldsymbol{\mu}) = d^2_M(a)$ in the case of the Mahalanobis depth), see Figure 7 (right).

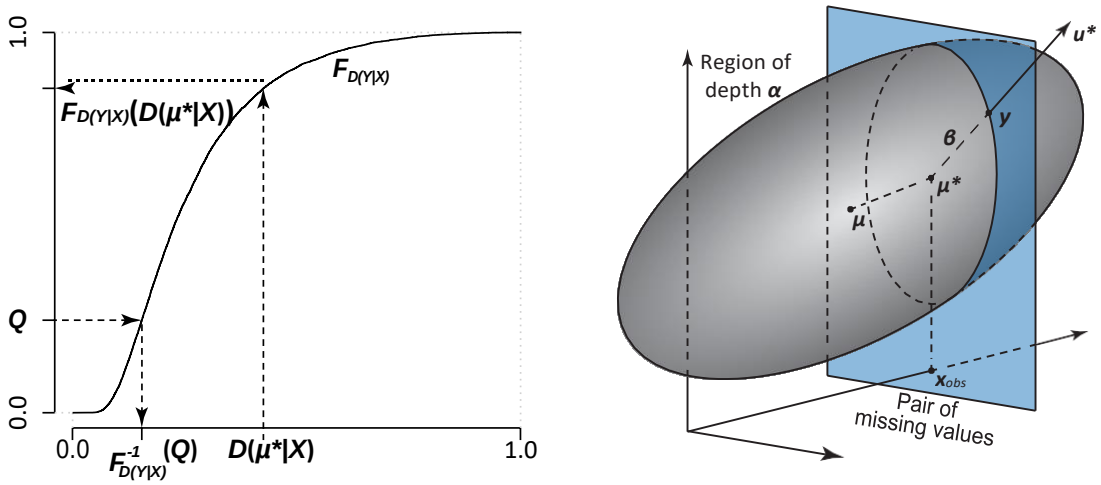

Figure 7: Illustration of an application of (4) to impute by drawing from the conditional distribution of an elliptical distribution. Drawing the depth $D = F^{-1}_{D(Y|X)}(Q)$ via the depth cumulative distribution function $F_{D(Y|X)}$ (left) and locating the corresponding imputed point $\boldsymbol{y}$ (right).

5.2 Depth-based multiple imputation

We use a bootstrap approach to reflect uncertainty due to the estimation of the underlying semi-parametric model. The depth-based procedure for multiple imputation (called DMI), detailed in Algorithm 2 consists of the following steps: first, a sequence of indices $\boldsymbol{b} = (b_1, \ldots, b_n)$ is drawn from $b_i \sim U(1, \ldots, n)$ for $i = 1, \ldots, n$, and this sequence is used to obtain new incomplete data set $\boldsymbol{X}_{\boldsymbol{b},\cdot} = (\boldsymbol{x}_{b_1}, \ldots, \boldsymbol{x}_{b_n})$. Then, on each incomplete data set, the stochastic single depth imputation method described in Section 5.1 is applied.

Algorithm 2 Depth-based multiple imputation

function IMPUTE.DEPTH.MULTIPLE($\boldsymbol{X}$, num.burnin, num.sets)
  for $m = 1 : num.sets$ **do**
    $\boldsymbol{Y}^{(m)} \leftarrow$ IMPUTE.DEPTH.SINGLE($\boldsymbol{X}$) ▷ Start MCMC with a single imputation
    $\boldsymbol{b} = (b_1, \ldots, b_n) = U(1, \ldots, n), \ldots, U(1, \ldots, n)$ ▷ Draw bootstrap sequence
    for $k = 1 : (num.burnin + 1)$ **do**
      $\boldsymbol{\Sigma} \leftarrow \hat{\boldsymbol{\Sigma}}(\boldsymbol{Y}^{(m)}_{\boldsymbol{b},\cdot})$
      Estimate $f_{D(X|X)}$ using $\boldsymbol{Y}^{(m)}$.
      for $i = 1 : n$ **do**
        if $miss(i) \neq \emptyset$ **then**
          $\boldsymbol{\mu}^* \leftarrow$ IMPUTE.DEPTH.SINGLE($\boldsymbol{x}_i, \boldsymbol{Y}^{(m)}_{\boldsymbol{b},\cdot}$) ▷ Single-impute point
          $\boldsymbol{u} \leftarrow U(\mathbb{S}^{|miss(i)|-1})$
          $\boldsymbol{u}^*_{miss(i)} \leftarrow \boldsymbol{u}\boldsymbol{\Lambda}$ ▷ Calculate random direction
          $\boldsymbol{u}^*_{obs(i)} \leftarrow 0$
          Calculate $F_{D(Y|X)}$
          $Q \leftarrow U[0, F_{D(Y|X)}(D(\boldsymbol{\mu}^*|\boldsymbol{Y}^{(m)}_{\boldsymbol{b},\cdot}))]$ ▷ Draw depth
          $a \leftarrow F^{-1}_{D(Y|X)}(Q)$
          $\beta \leftarrow$ positive solution of $\boldsymbol{\mu}^* + \beta\boldsymbol{u}^* \in \partial D_a(\boldsymbol{Y}^{(m)}_{\boldsymbol{b},\cdot})$.
          $\boldsymbol{y}^{(m)}_{i,miss(i)} \leftarrow \boldsymbol{\mu}^*_{miss(i)} + \beta\boldsymbol{u}^*_{miss(i)}$ ▷ Impute one point
  return $\boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(num.sets)}$

5.3 Experiments

5.3.1 Stochastic single depth-based imputation preserves quantiles

We generate 500 points from an elliptical Student-t distribution with degrees of freedom, with $\boldsymbol{\mu}_4 = (-1, -1, -1, -1)^\top$ and $\boldsymbol{\Sigma}_4 = (0.5, 0.5, 1, 1)^\top, (0.5, 1, 1, 1)^\top, (1, 1, 4, 4)^\top, (1, 1, 4, 10)^\top$, adding 30% of MCAR values, and compare the imputation with stochastic EM, stochastic PCA (Josse and Husson, 2012), and stochastic depth imputation (Section 5.1). On the completed data, we calculate quantiles for each variable and compare them with those obtained for the initial complete data. Table 5 shows the medians of the results over 2000 simulations for the first variable; the results are the same for the other variables. The stochastic EM and PCA methods, which generate noise from the normal model, do not lead to accurate quantile estimates. The proposed method gives excellent results with only slight deviations in the tail of the distribution due to difficulties in reflecting the density's shape. Although such an outcome is expected, it considerably broadens the scope of practice in comparison to the deep-rooted Gaussian imputation.

Quantile:	0.5	0.75	0.85	0.9	0.95	0.975	0.99	0.995
complete	-1.0013	-0.4632	-0.1231	0.1446	0.6398	1.2017	2.0661	2.8253
stoch. EM	-1.0008	-0.4225	-0.0649	0.2114	0.6902	1.2022	1.9782	2.6593
stoch. PCA	-0.9999	-0.4248	-0.0650	0.2121	0.7048	1.2398	2.0417	2.7481
depth	-0.9996	-0.4643	-0.1232	0.1491	0.6509	1.2142	2.0827	2.8965

Table 5: Median (over 2000 repetitions) quantiles of the imputed variable X_1 obtained from an elliptical sample of 500 points drawn from the Student-t distribution with 3 d.f. with 30% of MCAR values.

5.3.2 Inference with missing values

We explore the performance of DMI for inference with missing values by estimating coefficients of a regression model. Data are generated according to the following model: $Y = \boldsymbol{\beta}^\top(1, X^\top)^\top + s$, with $\boldsymbol{\beta} = (0.5, 1, 3)^\top$ and $X \sim N\left((1, 1)^\top, \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}\right)$, then 30% of MCAR values are introduced. We employ DMI and perform multiple imputation using the `R`-packages `Amelia` and `mice` under their default settings, generating 20 imputed data sets. For each, we run the regression model to estimate the parameters and their variance, and combine the results according to Rubin's rules (Little and Rubin, 2002). Here competitors are in a favourable setting as they are based on Gaussian distribution assumptions. We indicate the medians, the coverage of the 95% confidence interval, and the width of this interval, for the estimates of $\boldsymbol{\beta}$ in Table 6 with a sample size of 500, over 2000 simulations. In addition to the missing data, one difficulty comes from the high correlation ($\approx$0.988) between two of the variables.

	β_0			β_1			β_2		
	med	cov	width	med	cov	width	med	cov	width
	20 multiply-imputed data sets								
`Amelia`	0.487	0.931	0.489	1.01	0.941	0.399	2.998	0.929	0.206
`mice`	0.519	0.984	1.6	1.081	0.98	1.807	2.881	0.982	1.502
regPCA	0.495	0.971	0.853	1.04	0.964	0.751	2.957	0.936	0.334
DMI	0.504	0.971	0.613	0.989	0.979	0.519	3.003	0.97	0.26

Table 6: Medians (med), 95% coverage (cov), and width of the confidence intervals (width) for the regression parameters based on 20 imputed data sets over 2000 repetitions, for a sample of 500 observations from a regression model with 30% MCAR values.

`Amelia` has minor under-coverage problems, and `mice` provides biased coefficients and has large over-coverage issues as it is based on regression imputations that are unstable in the presence of high correlation. DMI, on the other hand, suffers from a slight amount of over-coverage, but in general provides valid inference.

6 Conclusions

The depth imputation framework we propose here fills the gap between global imputation of regression- and PCA-based methods, and the local imputation of methods such as random forest and kNN. It reflects uncertainty in the distribution assumption by imputing data close to the data geometry, is robust in the sense of the distribution and outliers, and still functions with MAR data. When used with the Mahalanobis depth, using data depth as a concept, the link between iterative regression, regularized PCA, and imputation with values that minimize the determinant of the covariance matrix, was established. Our empirical study shows the effectiveness of the suggested methodology for various elliptic distributions and real data. In addition, the method has been naturally extended to multiple imputation for the elliptical family, broadening the scope of existing tools for multiple imputation.

The methodology is general, i.e., any reasonable notion of data depth can be used, which then determines the imputation properties. In the empirical study, the zonoid depth behaves well in general, and for real data in particular. However if robustness is an issue, the Tukey depth may be preferable. The projection depth (Zuo and Serfling, 2000*a*) is an appropriate choice if only a few points contain missing values in a data set that is substantially outlier-contaminated. This specific case is not included in the article, but imputation based on projection depth is implemented in the associated `R` package. To reflect multimodality of the data, the suggested framework has been used with localized depths, see e.g. Paindaveine and Bever (2013).

A serious issue with data depths is their computation. Using approximate versions of data depths (which can also be found in the implementation) is a first step to handling larger data sets. Our methodology has been implemented as the `R`-package `imputeDepth`. Source code of the package and of the experiment-reproducing files can be downloaded from `https://github.com/julierennes/imputeDepth`.

Acknowledgments

The authors thank co-editor in chief Regina Liu, the associate editor, and three anonymous reviewers for their insightful remarks. The authors also gratefully thank Davy Paindaveine and Bernard Delyon for fruitful discussions as well as Germain Van Bever for providing the original code for the local depth, which served as a starting point for its implementation in the `R`-package `imputeDepth`.

Supplementary materials

Supplementary materials These contain additional figures, results on experiments with other percentages of missing values, as well as proofs. (`NIDD_Supplement.pdf`)

`R`-package and codes for reproducing experiments The archive contains the source code of the accompanying `R`-package `imputeDepth` as well as scripts for reproducing the experiments and figures of the article. The most up-to-date content of this archive can be found online using the link `https://github.com/julierennes/imputeDepth`. (`NIDD_codes.zip`).

References

Azzalini, A. and Capitanio, A. (1999), 'Statistical applications of the multivariate skew normal distribution', *Journal of the Royal Statistical Society: Series B* **61**(3), 579–602.

Bazovkin, P. and Mosler, K. (2015), 'A general solution for robust linear programs with distortion risk constraints', *Annals of Operations Research* **229**(1), 103–120.

Bertsekas, P. D. (1999), *Nonlinear programming. Second edition*, MIT Press.

Cascos, I. and Molchanov, I. (2007), 'Multivariate risks and depth-trimmed regions', *Finance and Stochastics* **11**(3), 373–397.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society. Series B (Methodological)* **39**(1), 1–38.

Donoho, D. L. and Gasko, M. (1992), 'Breakdown properties of location estimates based on halfspace depth and projected outlyingness', *The Annals of Statistics* **20**(4), 1803–1827.

Dua, D. and Karra Taniskidou, E. (2017), 'UCI machine learning repository'.
URL: *http://archive.ics.uci.edu/ml*

Dyckerhoff, R. (2004), 'Data depths satisfying the projection property', *Advances in Statistical Analysis* **88**(2), 163–190.

Dyckerhoff, R. and Mozharovskyi, P. (2016), 'Exact computation of the halfspace depth', *Computational Statistics and Data Analysis* **98**, 19–30.

Efron, B. (1994), 'Missing data, imputation, and the bootstrap', *Journal of the American Statistical Association* **89**(426), 463–475.

Einmahl, J. H. J., Li, J. and Liu, R. Y. (2015), 'Bridging centrality and extremity: Refining empirical data depth using extreme value statistics', *Ann. Statist.* **43**(6), 2738–2765.

Fang, K., Kotz, S. and Ng, K. (1990), *Symmetric multivariate and related distributions*, Monographs on statistics and applied probability, Chapman and Hall.

Hastie, T., Mazumder, R., Lee, D. J. and Zadeh, R. (2015), 'Matrix completion and low-rank svd via fast alternating least squares', *Journal of Machine Learning Research* **16**, 3367–3402.

Jörnsten, R. (2004), 'Clustering and classification based on the L_1 data depth', *Journal of Multivariate Analysis* **90**(1), 67–89. Special Issue on Multivariate Methods in Genomic Data Analysis.

Josse, J. and Husson, F. (2012), 'Handling missing values in exploratory multivariate data analysis methods', *Journal de la Société Française de Statistique* **153**(2), 79–99.

Josse, J. and Reiter, J. P. (2018), 'Introduction to the special section on missing data', *Statist. Sci.* **33**(2), 139–141.

Koshevoy, G. and Mosler, K. (1997), 'Zonoid trimming for multivariate distributions', *The Annals of Statistics* **25**(5), 1998–2017.

Lange, T., Mosler, K. and Mozharovskyi, P. (2014), 'Fast nonparametric classification based on data depth', *Statistical Papers* **55**(1), 49–69.

Little, R. and Rubin, D. (2002), *Statistical analysis with missing data*, Wiley series in probability and mathematical statistics. Probability and mathematical statistics, Wiley.

Liu, R. Y., Parelius, J. M. and Singh, K. (1999), 'Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by liu and singh)', *The Annals of Statistics* **27**(3), 783–858.

Liu, R. Y. and Singh, K. (1993), 'A quality index based on data depth and multivariate rank tests', *Journal of the American Statistical Association* **88**(401), 252–260.

Mahalanobis, P. C. (1936), 'On the generalised distance in statistics', **2**(1), 49–55.

Mosler, K. (2002), *Multivariate Dispersion, Central Regions, and Depth: The Lift Zonoid Approach*, Lecture Notes in Statistics, Springer New York.

Nagy, S. (2017), 'Monotonicity properties of spatial depth', *Statistics and Probability Letters* **129**, 373 – 378.

Paindaveine, D. and Bever, G. V. (2013), 'From depth to local depth: A focus on centrality', *Journal of the American Statistical Association* **108**(503), 1105–1119.

Rousseeuw, P. J. and Van Driessen, K. (1999), 'A fast algorithm for the minimum covariance determinant estimator', *Technometrics* **41**(3), 212–223.

Schafer, J. (1997), *Analysis of Incomplete Multivariate Data*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, CRC Press.

Stekhoven, D. J. and Bühlmann, P. (2012), 'MissForest – non-parametric missing value imputation for mixed-type data.', *Bioinformatics* **28**(1), 112–118.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B. (2001), 'Missing value estimation methods for dna microarrays', *Bioinformatics* **17**(6), 520–525.

Tukey, J. W. (1975), Mathematics and the Picturing of Data, *in* R. D. James, ed., 'International Congress of Mathematicians 1974', Vol. 2, pp. 523–532.

Udell, M. and Townsend, A. (2017), 'Nice latent variable models have log-rank', *arXiv:1705.07474* .

van Buuren, S. (2012), *Flexible Imputation of Missing Data (Chapman & Hall/CRC Interdisciplinary Statistics)*, Chapman and Hall/CRC.

Vardi, Y. and Zhang, C.-H. (2000), 'The multivariate L1-median and associated data depth', *Proceedings of the National Academy of Sciences* **97**(4), 1423–1426.

Yeh, I.-C., Yang, K.-J. and Ting, T.-M. (2009), 'Knowledge discovery on RFM model using bernoulli sequence', *Expert Systems with Applications* **36**(3, Part 2), 5866–5871.

Zuo, Y. and Serfling, R. (2000*a*), 'General notions of statistical depth function', *The Annals of Statistics* **28**(2), 461–482.

Zuo, Y. and Serfling, R. (2000*b*), 'Structural properties and convergence results for contours of sample statistical depth functions', *The Annals of Statistics* **28**(2), 483–499.

Supplementary Materials to the article "Nonparametric imputation by data depth"

by Pavlo Mozharovskyi, Julie Josse and François Husson

1 Additional figures

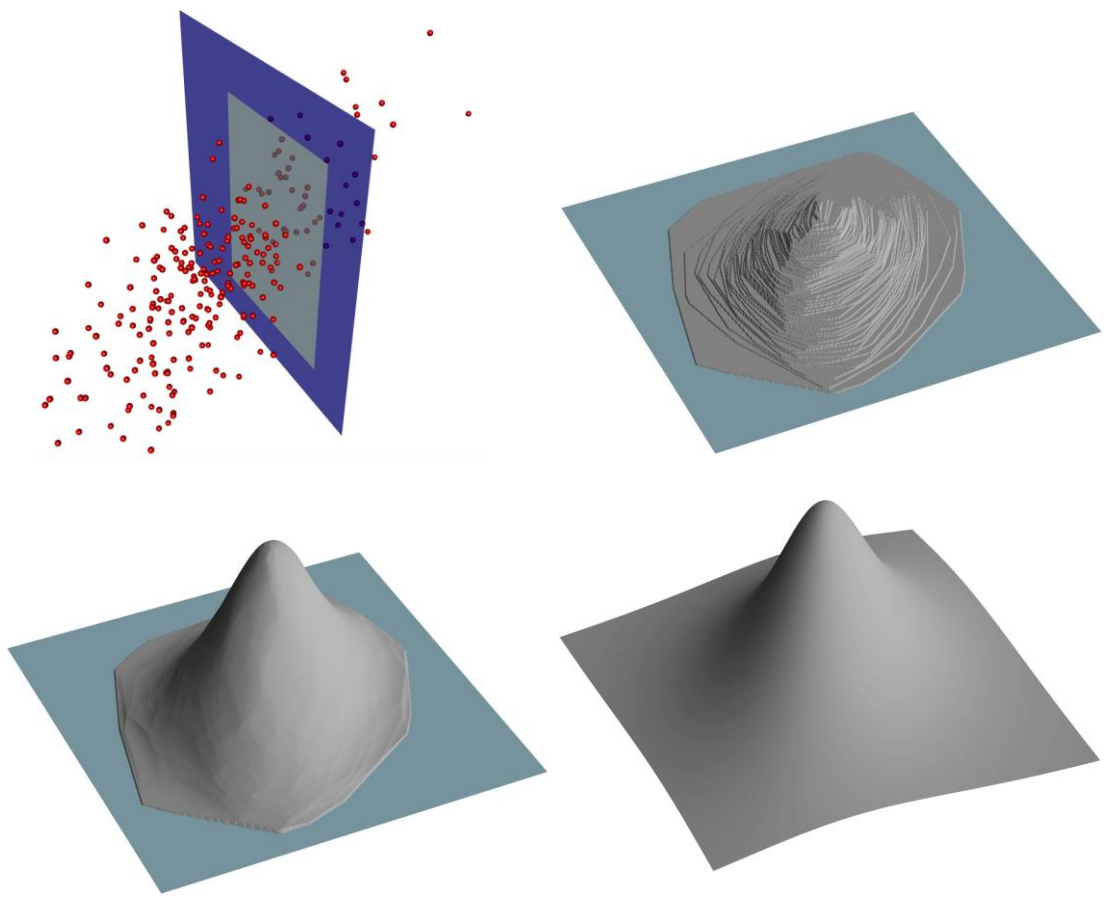

Figure 8: A Gaussian sample consisting of 250 points and a hyperplane of two missing coordinates (top, left), and the function $f(\boldsymbol{z}_{miss})$ to be optimized on each single iteration of Algorithm 1, for the smaller rectangle, for Tukey (top, right), zonoid (bottom, left), and Mahalanobis (bottom, right) depth. For the Tukey depth the maximum is not unique, and forms a polygon.

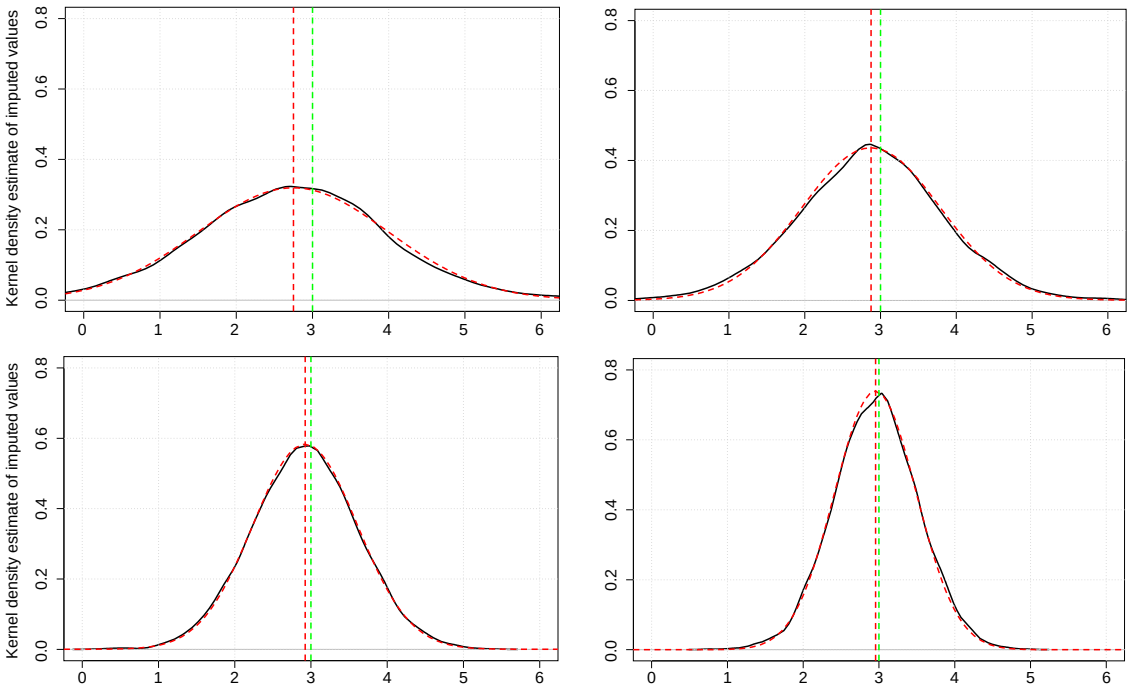

Figure 9: Samples of size 100 (top, left), 200 (top, right), 500 (bottom, left), and 1000 (bottom, right) are drawn from the bivariate Cauchy distribution with the location and scatter parameters $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$ from the introduction. Single point with one missing coordinate is imputed with the Tukey depth. Its kernel density estimate (solid) and the best approximating Gaussian curve (dashed) over $10,000$ repetitions are plotted. The population's conditional center given the observed value equals 3.

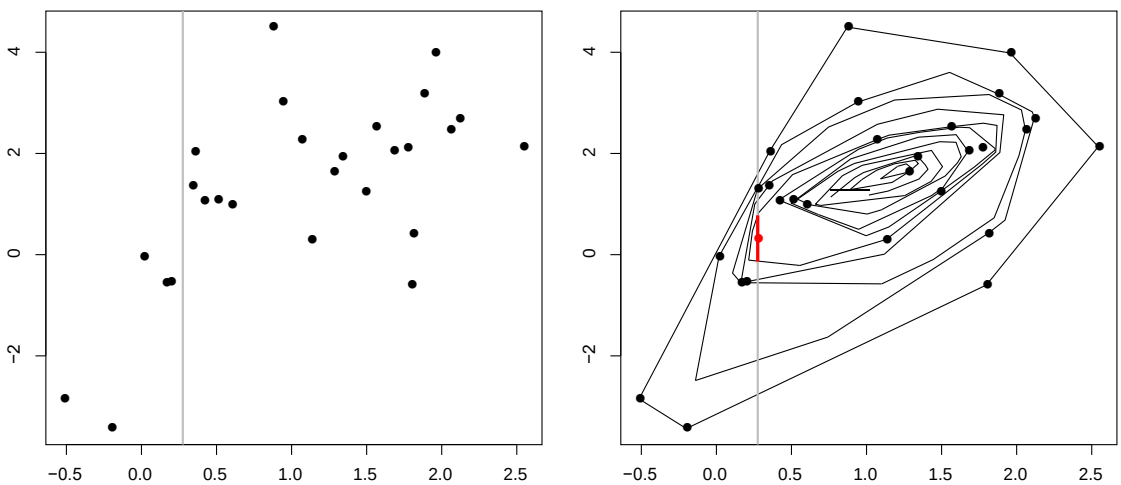

Figure 10: Illustration of imputation with the Tukey depth. When imputing the point with a missing second coordinate (left), the maximum of the constrained Tukey depth is non-unique (the red line segment), and an average over the optimal arguments (the red point) is used in equation (3) (right).

2 Simulation results with other percentages of missing values

When varying the percentage of missing values, the general trend remains unchanged. The small differences seen can be summarized as follows: with a decreasing percentage of missing values, the difference between EM and Mahalanobis depth imputation (and thus also the rank two PCA one) shrinks, and indeed the latter performs comparably to EM for 5% missingness. For the same percentage and the Cauchy distribution, nonparametric methods (kNN and random forest) perform comparably to the Tukey depth due to a sufficient quantity of available observations and an absence of correlation structure (outliers are generated from Cauchy distribution as well).

Distr.	D^{Tuk}	D^{zon}	D^{Mah}	$D^{Mah}_{MCD.75}$	EM	regPCA1	regPCA2	kNN	RF	mean	oracle
$t\ \infty$	1.577	1.547	*1.532*	1.537	**1.518**	1.596	*1.532*	1.684	1.681	2.058	1.487
	(0.2345)	(0.2128)	*(0.216)*	(0.2199)	**(0.2129)**	(0.2327)	*(0.2159)*	(0.2422)	(0.2445)	(0.2774)	(0.2004)
$t\ 10$	1.748	1.718	*1.693*	1.709	**1.69**	1.769	*1.692*	1.853	1.871	2.275	1.642
	(0.287)	(0.2838)	*(0.2737)*	(0.2827)	**(0.2826)**	(0.3039)	*(0.2741)*	(0.3168)	(0.3085)	(0.3757)	(0.2771)
$t\ 5$	1.993	1.971	*1.956*	*1.956*	**1.933**	2.017	*1.956*	2.125	2.134	2.565	1.874
	(0.378)	(0.3602)	*(0.3799)*	*(0.361)*	**(0.361)**	(0.3759)	*(0.3796)*	(0.4126)	(0.3976)	(0.4732)	(0.3492)
$t\ 3$	2.417	2.434	2.39	**2.333**	*2.362*	2.431	2.39	2.55	2.592	3.045	2.235
	(0.5996)	(0.6032)	(0.5792)	**(0.5571)**	*(0.5808)*	(0.5734)	(0.5793)	(0.6154)	(0.612)	(0.6943)	(0.5319)
$t\ 2$	*3.31*	3.373	3.431	**3.192**	3.366	3.437	3.422	3.538	3.555	4.155	2.986
	(1.191)	(1.273)	(1.314)	**(1.148)**	(1.249)	(1.343)	(1.289)	(1.33)	(1.321)	(1.466)	(1.063)
$t\ 1$	**13.19**	15.13	15.17	*13.39*	14.86	14.82	15.22	14.09	13.91	16.77	11.17
	(10.83)	(12.06)	(11.74)	*(10.32)*	(11.57)	(11.64)	(11.94)	(11.28)	(11.06)	(13.22)	(8.901)

Table 7: Median and MAD of the RMSEs of the imputation for a sample of 100 points drawn from elliptically symmetric Student-t distributions with $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_2$ having 15% of MCAR values, over 1000 repetitions.

Distr.	D^{Tuk}	D^{zon}	D^{Mah}	$D^{Mah}_{MCD.75}$	EM	regPCA1	regPCA2	kNN	RF	mean	oracle
$t\ \infty$	**1.656**	1.754	1.853	*1.671*	1.817	1.855	1.853	1.793	1.762	2.182	1.514
	(0.2523)	(0.3142)	(0.4062)	*(0.2906)*	(0.3937)	(0.4158)	(0.4071)	(0.2974)	(0.282)	(0.3821)	(0.2121)
$t\ 10$	**1.859**	1.973	2.048	*1.865*	2.027	2.05	2.044	1.995	1.968	2.45	1.677
	(0.3062)	(0.4031)	(0.511)	*(0.3917)*	(0.4933)	(0.5069)	(0.5119)	(0.3861)	(0.3402)	(0.4778)	(0.279)
$t\ 5$	**2.09**	2.23	2.31	*2.109*	2.267	2.348	2.31	2.255	2.233	2.749	1.91
	(0.4275)	(0.5122)	(0.6217)	*(0.4841)*	(0.6006)	(0.6504)	(0.6219)	(0.4543)	(0.4476)	(0.6089)	(0.3742)
$t\ 3$	**2.507**	2.697	2.772	*2.541*	2.737	2.791	2.779	2.707	2.699	3.32	2.239
	(0.6389)	(0.7977)	(0.8516)	*(0.7133)*	(0.8243)	(0.9306)	(0.8495)	(0.6946)	(0.7254)	(0.964)	(0.5497)
$t\ 2$	**3.462**	3.68	3.733	*3.517*	3.669	3.807	3.736	3.709	3.762	4.476	3.061
	(1.35)	(1.577)	(1.6)	*(1.39)*	(1.589)	(1.648)	(1.601)	(1.413)	(1.444)	(1.794)	(1.136)
$t\ 1$	**11.81**	14.12	14.22	*12.34*	13.78	13.73	14.31	12.58	13.64	15.73	10.37
	(9.738)	(12.09)	(12.05)	*(9.631)*	(11.48)	(11.01)	(12.12)	(10.36)	(11.44)	(12.65)	(8.249)

Table 8: Median and MAD of the RMSEs of the imputation for 100 points drawn from elliptically symmetric Student-t distributions with $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_2$ contaminated with 15% outliers, and 15% of MCAR values on non-contaminated data, over 1000 repetitions.

Distr.	D^{Tuk}	D^{zon}	D^{Mah}	$D^{Mah}_{MCD.75}$	EM	regPCA1	regPCA2	kNN	RF	mean	oracle
$t\ \infty$	1.464	1.454	**1.447**	1.453	*1.449*	1.529	**1.447**	1.571	1.581	2.009	1.404
	(0.3694)	(0.3713)	**(0.3595)**	(0.3663)	*(0.3593)*	(0.401)	**(0.3594)**	(0.3892)	(0.3946)	(0.486)	(0.3399)
$t\ 10$	1.649	1.597	**1.57**	*1.572*	**1.57**	1.665	**1.57**	1.755	1.754	2.2	1.529
	(0.4316)	(0.4285)	**(0.4163)**	*(0.4203)*	**(0.4206)**	(0.4502)	**(0.4163)**	(0.4565)	(0.46)	(0.5737)	(0.4278)
$t\ 5$	1.816	1.799	**1.757**	*1.758*	**1.757**	1.876	**1.757**	1.955	1.972	2.402	1.712
	(0.5134)	(0.5129)	**(0.49)**	*(0.4991)*	**(0.4899)**	(0.5499)	**(0.4901)**	(0.555)	(0.5345)	(0.7318)	(0.4869)
$t\ 3$	2.213	2.184	2.147	**2.101**	*2.139*	2.242	2.147	2.37	2.343	2.844	2.054
	(0.7882)	(0.8159)	(0.8016)	**(0.7618)**	*(0.8)*	(0.7782)	(0.801)	(0.8563)	(0.8357)	(1.011)	(0.7649)
$t\ 2$	2.837	2.919	2.813	**2.68**	*2.8*	2.911	2.813	3.03	2.99	3.578	2.529
	(1.249)	(1.342)	(1.309)	**(1.196)**	*(1.287)*	(1.311)	(1.31)	(1.325)	(1.331)	(1.554)	(1.133)
$t\ 1$	**7.806**	8.718	8.911	8.286	8.9	9.118	8.935	*8.135*	8.138	10.99	6.367
	(6.351)	(7.135)	(7.127)	(6.602)	(7.124)	(7.334)	(7.137)	*(6.605)*	(6.563)	(8.952)	(5.12)

Table 9: Median and MAD of the RMSEs of the imputation for a sample of 100 points drawn from elliptically symmetric Student-t distributions, with $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_2$ having 5% of MCAR values, over 1000 repetitions.

Distr.	D^{Tuk}	D^{zon}	D^{Mah}	$D^{Mah}_{MCD.75}$	EM	regPCA1	regPCA2	kNN	RF	mean	oracle
$t\ \infty$	**1.552**	1.613	1.709	*1.553*	1.701	1.769	1.709	1.695	1.603	2.167	1.406
	(0.3693)	(0.4107)	(0.4867)	*(0.4379)*	(0.4788)	(0.5248)	(0.4877)	(0.407)	(0.3924)	(0.5981)	(0.3171)
$t\ 10$	**1.706**	1.778	1.874	*1.73*	1.861	1.906	1.875	1.884	1.823	2.398	1.564
	(0.4415)	(0.5106)	(0.6104)	*(0.4912)*	(0.6032)	(0.6053)	(0.6111)	(0.5182)	(0.4797)	(0.7059)	(0.3938)
$t\ 5$	**1.868**	1.951	2.038	*1.877*	2.027	2.172	2.039	2.077	1.995	2.57	1.698
	(0.5565)	(0.5843)	(0.6859)	*(0.5679)*	(0.6806)	(0.7747)	(0.6819)	(0.6256)	(0.6102)	(0.8625)	(0.491)
$t\ 3$	*2.243*	2.348	2.421	**2.226**	2.42	2.525	2.421	2.429	2.392	3.05	2.016
	(0.8064)	(0.8694)	(0.9166)	**(0.8258)**	(0.9345)	(1.019)	(0.9237)	(0.8484)	(0.8521)	(1.171)	(0.7047)
$t\ 2$	**2.902**	3.032	3.183	*2.933*	3.163	3.196	3.188	3.142	3.071	4.073	2.55
	(1.375)	(1.498)	(1.566)	*(1.421)*	(1.558)	(1.565)	(1.582)	(1.472)	(1.43)	(2.007)	(1.129)
$t\ 1$	**7.464**	8.487	8.531	8.334	8.5	8.675	8.541	*7.958*	8.1	10.82	6.245
	(5.916)	(6.869)	(7.081)	(6.867)	(6.988)	(7.261)	(7.117)	*(6.509)*	(6.922)	(8.802)	(4.874)

Table 10: Median and MAD of the RMSEs of the imputation for 100 points drawn from elliptically symmetric Student-t distributions with $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_2$ contaminated with 15% of outliers, with 5% MCAR values on non-contaminated data, over 1000 repetitions.

3 Proofs

Proof of Theorem 1:
Due to the fact that $D_{n,a}(\boldsymbol{X}) \xrightarrow[n\to\infty]{a.s.} D_a(X)$, in what follows we focus on the population version only. For $X \sim \mathsf{E}_d(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X, F_R)$ allow the transform $X \mapsto Z = \boldsymbol{R}\boldsymbol{\Sigma}^{-1/2}(X - \boldsymbol{\mu})$, with $\boldsymbol{R}$ being a rotation operator such that w.l.o.g. $\boldsymbol{x} \mapsto \boldsymbol{z}$, such that missing values still constitute a $|miss(\boldsymbol{x})|$-dimensional affine space parallel to missing coordinates' axes. Since contours $D_a(Z)$ are concentric spheres centered at the origin, $D^*_a(Z)$ in (3) is of the form $\{\boldsymbol{v} \mid \boldsymbol{v} = \boldsymbol{z}^{\jmath} + \beta\boldsymbol{r}, \beta \geq 0\}$ with $\boldsymbol{z}^{\jmath}_{obs(\boldsymbol{z})} = \boldsymbol{z}_{obs}$ and $\boldsymbol{z}^{\jmath}_{miss(\boldsymbol{z})} = \boldsymbol{0}_{|miss(\boldsymbol{x})|}$, and $\boldsymbol{r} \in \mathsf{S}^{|miss(\boldsymbol{x})|-1}$, a unit sphere in the

linear span of $miss(\boldsymbol{z})$. Because of the fact that $P\{\boldsymbol{x} \in \mathbb{R}^d \mid D(\boldsymbol{x}|X) = a\} = 0, \beta = 0$ almost surely and thus $\boldsymbol{z}$ is imputed with $\boldsymbol{z}^{\prime} = \boldsymbol{R}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y} - \boldsymbol{\mu})$. Q

Proof of Theorem 2:
(The challenge here is that the resulting distribution is not elliptical.)

For $X \sim E_d(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X, F_R)$ allow the transform $X \mapsto Z = \boldsymbol{R}\boldsymbol{\Sigma}^{-1/2}(X - \boldsymbol{\mu})$, with $\boldsymbol{R}$ being a rotation operator such that w.l.o.g. $\boldsymbol{x} \mapsto \boldsymbol{z}$, such that $miss(\boldsymbol{z}) = 1$. (Z has spherical density contours and missing values are in the first coordinate only.)

Let $Z^{\prime} = (0, (Z^{\prime\prime})^{\top})^{\top}$ with $Z^{\prime\prime} \sim E_{d-1}(\boldsymbol{0}, \boldsymbol{I}, F_R)$, where $\boldsymbol{I}$ is the diagonal matrix. Consider a random vector $U \sim (1-p)Z + pZ^{\prime}$ which is a mixture of d- and $(d-1)$-dimensional spherical distributions. $Z^{\prime}$ corresponds to the imputed missing values—let us now show that this is true. Due to the fact that $D_{n,a}(\boldsymbol{U}) \xrightarrow[n\to\infty]{a.s.} D_a(U)$, in what follows we focus on the population version only. Missing values constitute one-dimensional affine subspaces parallel to the first coordinate. Thus, due to the affine invariance property (P1 in Definition 2), $D_a(U) \cap \{\boldsymbol{u} \in \mathbb{R}^d \mid \boldsymbol{u}_1 \geq 0\} = D_a(U) \cap \{\boldsymbol{u} \in \mathbb{R}^d \mid \boldsymbol{u}_1 \leq 0\} \times (-1, 0, \ldots, 0)^{\top}$. To see this, it suffices to note that the symmetric reflection of U w.r.t. the linear space normal to $(1, 0, \ldots, 0)^{\top}$ equals U. Now, for $\lambda \in \mathbb{R}$ let $\boldsymbol{u} = (\lambda, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_d)^{\top}$ be this one-dimensional affine subspace of missingness for a point. In (3), ave $D_a(U) \cap \boldsymbol{u} =$ ave $D_a(U) \cap \{\boldsymbol{v} \in \mathbb{R}^d \mid \boldsymbol{v}_1 > 0\} \cap \boldsymbol{u} \cup D_a(U) \cap \{\boldsymbol{v} \in \mathbb{R}^d \mid \boldsymbol{v}_1 < 0\} \cap \boldsymbol{u} \cup D_a(U) \cap \{\boldsymbol{v} \in \mathbb{R}^d \mid \boldsymbol{v}_1 = 0\} \cap \boldsymbol{u} = D_a(U) \cap \{\boldsymbol{v} \in \mathbb{R}^d \mid \boldsymbol{v}_1 = 0\} \cap \boldsymbol{u} = (0, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_d) = \boldsymbol{R}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y} - \boldsymbol{\mu})$ (with the obvious correspondence between $\boldsymbol{u}$ and $\boldsymbol{y}$). Q

Proof of Corollary 1:
(P1)–(P5) are obviously satisfied for the Tukey, zonoid and Mahalanobis depths. In Theorem 1, $D_{n,a}(\boldsymbol{X}) \xrightarrow[n\to\infty]{a.s.} D_a(X)$ is clearly satisfied for the Mahalanobis depth, following Corollary 3.11 by Mosler (2002) for the zonoid depth, and by Theorem 4.2 in Zuo and Serfling (2000*b*) for the Tukey depth. In Theorem 2, the same logic holds for the Mahalanobis and zonoid depths, but not for the Tukey depth as Z is not elliptical. Using techniques similar those in the proof of Theorem 3.4 in Zuo and Serfling (2000*b*), one can show that $P\{\boldsymbol{x} \in \mathbb{R}^d \mid D(\boldsymbol{x}|Z) = a\} = 0$, from which, together with the vanishing at infinity property (P4) and $\sup_{\boldsymbol{x}\in\mathbb{R}^d} |D_n(\boldsymbol{x}|\boldsymbol{Z}) - D(\boldsymbol{x}|Z)| \xrightarrow[n\to\infty]{a.s.} 0$ (see Donoho and Gasko, 1992), it follows that $D_{n,a}(\boldsymbol{Z}) \xrightarrow[n\to\infty]{a.s.} D_a(Z)$. Q

Proof of Proposition 1: w.l.o.g. we restrict ourselves to the case $i = 1$. Let $\boldsymbol{Z}$ be $\boldsymbol{X}$ transformed in such a way that it is an $n \times d$ matrix with $\boldsymbol{\mu}_{\boldsymbol{Z}} = \boldsymbol{0}$ and $\boldsymbol{z}_{1,miss(1)} = \boldsymbol{\Sigma}_{\boldsymbol{Z}\, miss(1),obs(1)}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{Z}\, obs(1),obs(1)}\boldsymbol{z}_{1,obs(1)}$. Denote the argument $\boldsymbol{a} = (0, \ldots, 0, \boldsymbol{y}^{\top})^{\top} \in \mathbb{R}^d$. Replacing $\boldsymbol{z}_1$ with $\boldsymbol{z}_1 + \boldsymbol{a}$ and subtracting the column-wise average $\frac{\boldsymbol{a}}{n}$ from each row gives the covariance matrix estimate:

$$
\begin{aligned}
n\boldsymbol{\Sigma}_{\boldsymbol{Z}}(\boldsymbol{y}) &= \boldsymbol{Z}^{\top}\boldsymbol{Z} - \boldsymbol{z}_1\boldsymbol{z}_1^{\top} + (\boldsymbol{z}_1 + \boldsymbol{a})(\boldsymbol{z}_1 + \boldsymbol{a})^{\top} - \frac{1}{n}\boldsymbol{a}\boldsymbol{a}^{\top} \\
&= \boldsymbol{Z}^{\top}\boldsymbol{Z} + 2\boldsymbol{z}_1\boldsymbol{a}^{\top} + \frac{n-1}{n}\boldsymbol{a}\boldsymbol{a}^{\top} .
\end{aligned}
$$

Since $\boldsymbol{z}_1^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{a} = 0$ due to Mahalanobis orthogonality, by simple algebra for the determinant, one obtains:

$$
\begin{aligned}
n^d |\boldsymbol{\Sigma}_{\boldsymbol{Z}}(\boldsymbol{y})| &= \left| \boldsymbol{Z}^\top \boldsymbol{Z} + \sqrt{2}\boldsymbol{z}_1 \boldsymbol{a}^\top \sqrt{2} + \sqrt{\frac{n-1}{n}} \boldsymbol{a}\boldsymbol{a}^\top \sqrt{\frac{n-1}{n}} \right| \\
&= \left| \boldsymbol{Z}^\top \boldsymbol{Z} + \sqrt{2}\boldsymbol{z}_1 \boldsymbol{a}^\top \sqrt{2} \cdot \left(1 + \sqrt{\frac{n}{n-1}} \boldsymbol{a}^\top \left(\boldsymbol{Z}^\top \boldsymbol{Z} + \sqrt{2}\boldsymbol{z}_1 \boldsymbol{a}^\top \sqrt{2} \right)^{-1} \boldsymbol{a} \sqrt{\frac{n-1}{n}} \right) \right| \\
&= |\boldsymbol{Z}^\top \boldsymbol{Z}| \cdot \left(1 + \sqrt{2}\boldsymbol{a}^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{z}_1 \sqrt{2} \right) \cdot \left(1 + \sqrt{\frac{n-1}{n}} \boldsymbol{a}^\top \left(\boldsymbol{Z}^\top \boldsymbol{Z} \right)^{-1} \boldsymbol{a} \sqrt{\frac{n-1}{n}} - \right. \\
&\quad \left. - \frac{\sqrt{\frac{n-1}{n}} \boldsymbol{a}^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{z}_1 \sqrt{2} \cdot \sqrt{2}\boldsymbol{a}^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{a} \sqrt{\frac{n-1}{n}}}{\sqrt{2}\boldsymbol{a}^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{z}_1 \sqrt{2}} \right) \\
&= |\boldsymbol{Z}^\top \boldsymbol{Z}| \cdot \left(1 + \frac{n-1}{n} \boldsymbol{a}^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{a} \right).
\end{aligned}
$$

Thus $|\boldsymbol{\Sigma}_Z(\boldsymbol{y})|$ is a quadratic function of $\boldsymbol{y}$, which is clearly minimized in $\boldsymbol{y} = (0, \ldots, 0)^\top$. Q

Proof of Theorem 3: The first point can be checked by elementary algebra. The second point follows from the coordinate-wise application of Proposition 1. For the third point, it suffices to prove the single-output regression case. The regularized PCA algorithm will converge if

$$
\boldsymbol{y}_{id} = \sum_{s=1}^{d} u_{is} \sqrt{\lambda_s} v_{ds} = \sum_{s=1}^{d} u_{is} \left(\sqrt{\lambda_s} - \frac{\sigma^2}{\sqrt{\lambda_s}} \right) v_{ds}
$$

for any $\sigma^2 \leq \lambda_d$. W.l.o.g. we prove that

$$
\boldsymbol{y}_d = \boldsymbol{\Sigma}_{d(1,\ldots,d-1)} \boldsymbol{\Sigma}^{-1}_{(1,\ldots,d-1)(1,\ldots,d-1)} \boldsymbol{y}_{(1,\ldots,d-1)} \iff \sum_{i=1}^{d} \frac{u_i v_{di}}{\sqrt{\lambda_i}} = 0,
$$

denoting $\boldsymbol{\Sigma}(\boldsymbol{Y})$ simply $\boldsymbol{\Sigma}$ for the centered $\boldsymbol{Y}$, and an arbitrary point $\boldsymbol{y}$. Using matrix algebra,

$$
\begin{aligned}
\boldsymbol{y}_d &= \boldsymbol{\Sigma}_{d(1,\ldots,d-1)} \boldsymbol{\Sigma}^{-1}_{(1,\ldots,d-1)(1,\ldots,d-1)} \boldsymbol{y}_{(1,\ldots,d-1)} = -(\boldsymbol{\Sigma}^{-1})_{dd}^{-1} (\boldsymbol{\Sigma}^{-1})_{d(1,\ldots,d-1)} \boldsymbol{y}_{(1,\ldots,d-1)}, \\
\sum_{i=1}^{d} u_i \sqrt{\lambda_i} v_{di} &= -\left(\sum_{i=1}^{d} \frac{v_{di}^2}{\lambda_i} \right)^{-1} \left(\sum_{i=1}^{d} \frac{v_{di} v_{1i}}{\lambda_i}, \sum_{i=1}^{d} \frac{v_{di} v_{2i}}{\lambda_i}, \ldots, \sum_{i=1}^{d} \frac{v_{di} v_{(d-1)i}}{\lambda_i} \right) \times \\
&\quad \times \left(\sum_{i=1}^{d} u_i \sqrt{\lambda_i} v_{1i}, \sum_{i=1}^{d} u_i \sqrt{\lambda_i} v_{2i}, \ldots, \sum_{i=1}^{d} u_i \sqrt{\lambda_i} v_{(d-1)i} \right)^\top.
\end{aligned}
$$

After reordering the terms, one obtains

$$
\sum_{i=1}^{d} u_i \sqrt{\lambda_i} \sum_{j=1}^{d} \frac{v_{dj}}{\lambda_j} \sum_{k=1}^{d} v_{ki} v_{kj} = 0.
$$

Due to the orthogonality of $\boldsymbol{V}$, $d^2 - d$ terms from the two outer sum signs are zero. Gathering non-zero terms, i.e., those with $i = j$ only, we have that

$$
\sum_{i=1}^{d} u_i \sqrt{\lambda_i} \frac{v_{di}}{\lambda_i} = \sum_{i=1}^{d} \frac{u_i v_{di}}{\sqrt{\lambda_i}} = 0.
$$

$i=1$ i $i=1$ i

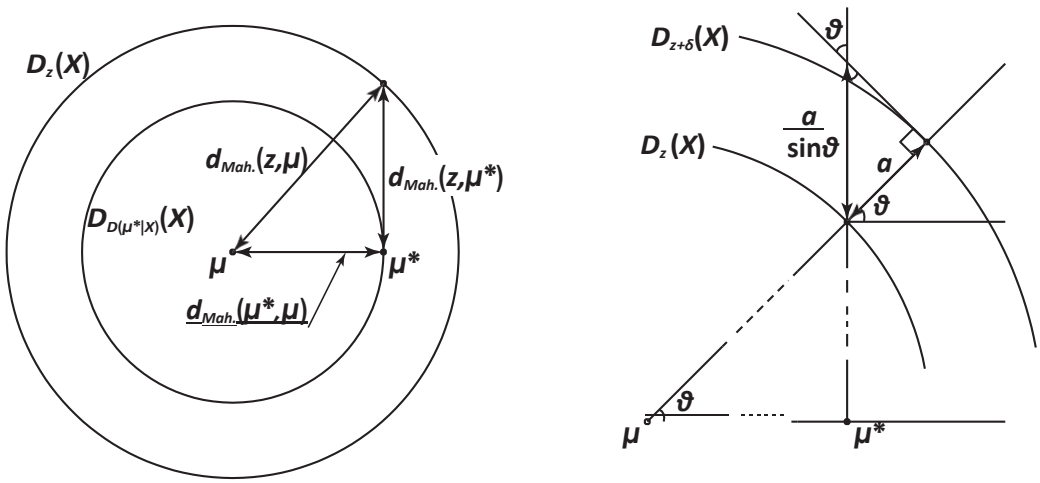

Figure 11: Illustration of the derivation of (4).

Q

Derivation of (4): The integrated quantity is the conditional depth density that can be obtained from the joint one by the volume transformation (denoting $d_M(z, \boldsymbol{\mu})$ the Mahalanobis distance between a point of depth z and $\boldsymbol{\mu}$):

$$f_{D((X|X_{obs}=\boldsymbol{x}_{obs})|X)}(z) = f_{D(X|X)}(z) \cdot C \cdot T_{down}(d_M(z,\boldsymbol{\mu})) \cdot T_{up}(d_M(z,\boldsymbol{\mu}^*)) \times$$
$$\times\, T_{angle}(d_M(z,\boldsymbol{\mu}), d_M(z,\boldsymbol{\mu}^*))$$

Any constant C is ignored as it is unimportant when drawing. The three terms below correspond to descaling the density to dimension one (downscaling), re-scaling it to the dimension of the missing values (upscaling), and the linear transformation from dimension d to dimension $|miss|$ =number of missing coordinates of a point (angle transformation):

$$T_{down}(d_M(z,\boldsymbol{\mu})) = d_M^{1-d}(z,\boldsymbol{\mu}) = \frac{1}{d_M^{d-1}(z,\boldsymbol{\mu})}.$$

$$T_{up}(d_M(z,\boldsymbol{\mu}^*)) = d_M^{|miss(\boldsymbol{x})|-1}(z,\boldsymbol{\mu}^*)$$
$$= \left(\sqrt{d_M^2(z,\boldsymbol{\mu}) - d_M^2(D(\boldsymbol{\mu}^*|X),\boldsymbol{\mu})}\right)^{|miss(\boldsymbol{x})|-1}.$$

$$T_{angle}(d_M(z,\boldsymbol{\mu}), d_M(z,\boldsymbol{\mu}^*)) = \frac{1}{\sin\theta} = \frac{1}{\frac{d_M(z,\boldsymbol{\mu}^*)}{d_M(z,\boldsymbol{\mu})}}$$
$$= \frac{d_M(z,\boldsymbol{\mu})}{\sqrt{d_M^2(z,\boldsymbol{\mu}) - d_M^2(D(\boldsymbol{\mu}^*|X),\boldsymbol{\mu})}}.$$

T_{down} and T_{up} are illustrated in Figure 11 (left); for T_{angle} see Figure 11 (right). Setting $d_M(z, \boldsymbol{\mu}) = d_M(z)$ to shorten notation gives (4).