Nonparametric imputation based on data depth

Pavlo Mozharovskyi[123], Julie Josse[34], François Husson[23]

[1] Centre Henri Lebesgue, Rennes, France
[2] Institut de Recherche Mathématique de Rennes, France
[3] Agrocampus Ouest, Rennes, France
[4] Institut National de Recherche en Informatique et en Automatique, Orsay, France

E-mail for correspondence: `pavlo.mozharovskyi@univ-rennes1.fr`

Abstract: A method for single imputation of missing values is presented. It consists in iterative maximization of data depth of each observation with missing values, and can be used with any properly defined depth. The method is robust, distribution-free, and applicable to general elliptically symmetric densities. Its particular case has direct connection to the well know treatments for multivariate normal model.

Keywords: Missing data; Data depth; Single imputation; Elliptical symmetry.

1 Introduction

The problem of missing values exists since the earliest attempts of exploiting data as a source of knowledge as it lies intrinsically in the process of obtaining, recording, and preparation of the data itself. The most naïve treatment consists in dropping rows or columns, depending on the view on the data, but by deleting the entire row (column) present data is removed as well. And if a data set contains one or a few missing values in a large portion of rows, substantial part of data can be missed by this list-wise deletion. To exploit all the information present in the data set, a statistical method may be adapted to missing values, but this requires developing such a one for each estimator and inference of interest. A more universal way is to impute missing data first, and then apply the statistical method of interest to the completed data set (Little and Rubin, 2002). Lastly, the multiple imputation has gained a lot of attention: for a data set containing missing values, a number of completed data sets is generated reflecting uncertainty of the imputation process, which enables not only estimating the value of interest but also drawing an inference on it (Van Buuren, 2012). Nevertheless, single imputation, *i.e.* just meaningfully replacing missing values, is still paid attention in the statistical literature. This can be appropriate when one needs just to complete a single data set, when no inference is required, when the applied statistical method is computationally too demanding for multiple data sets, or when a few values are missing only but one seeks an alternative to the list-wise deletion.

2 Proposal

One of the existing approaches to single imputation is to replace a missing value by its conditional mean, based on a specific joint model. Being of highest importance, multivariate normal distribution and perturbing this mechanisms have gained a lot of attention in the imputation literature. In the present work, we propose a single imputation method able to properly work for a broader class of elliptically symmetric distributions — a natural generalization of the multivariate normal model. The suggested technique is based on the notion of statistical centrality measure — data depth, and is generic in it. Before presenting the approach in Section 2.2, we refer to the notion of data depth in Section 2.1.

2.1 Data depth

Consider a point $\boldsymbol{x}_0 \in \mathbb{R}^d$ and a random sample $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ in $\mathbb{R}^d$. A statistical data depth is a function $D(\boldsymbol{x}_0|\boldsymbol{X}) : \mathbb{R}^d \to [0, 1]$ that describes how deep, or central the observation $\boldsymbol{x}_0$ is located w.r.t. $\boldsymbol{X}$. To be a well behaving depth, $D(\cdot|\cdot)$ should satisfy elementary postulates: be affine invariant, vanishing at infinity, non-increasing on any ray from the deepest point ($\arg\max_{\boldsymbol{x}_0 \in \mathbb{R}^d} D(\boldsymbol{x}_0|\boldsymbol{X})$) or even quasi-concave, and upper semi-continuous; see Mosler (2013) for a recent survey. $D(\boldsymbol{x}_0|\boldsymbol{X})$ provides a multivariate center-outward ordering, *i.e.* points closer to the center should have higher depth, and those more outlying smaller one. During the last decades, a number of notions of statistical depth function differing in properties and areas of application have been developed. For shortness and demonstrative reasons we proceed with the historically first Tukey depth below.

The Tukey (or halfspace, also location) depth (Tukey, 1975) of $\boldsymbol{x}_0$ w.r.t. $\boldsymbol{X}$ is defined as the smallest portion of $\boldsymbol{X}$ that can be contained in a closed halfspace with $\boldsymbol{x}_0$ on its boundary

$$D(\boldsymbol{x}_0|\boldsymbol{X}) = \frac{1}{n} \min_{\boldsymbol{r} \in S^{d-1}} \#\{i|\boldsymbol{x}_i'\boldsymbol{r} \geq \boldsymbol{x}_0'\boldsymbol{r}, i = 1, ..., n\}. \qquad (1)$$

2.2 Iterative approach

Given (a complete) $\boldsymbol{X}$, let $\boldsymbol{x} \in \mathbb{R}^d$ be an observation with missing coordinates, and index its existing entries by $\boldsymbol{x}_{obs}$ and missing with $\boldsymbol{x}_{miss}$. Denoting $D_\alpha(\boldsymbol{X})$ an α-upper-level set of $D(\cdot|\boldsymbol{X})$ (or depth-trimmed region), and denoting interior by *int*, let

$$\alpha^* = \inf_{\alpha\in(0;1)} \bigl\{\alpha \,|\, int D_\alpha(\boldsymbol{X}) \cap \{\boldsymbol{y} \,|\, \boldsymbol{y} \in \mathbb{R}^d\,,\, \boldsymbol{y}_{obs} = \boldsymbol{x}_{obs}\} = \emptyset\bigr\} \qquad (2)$$

be the depth of the region with the smallest depth not touching the missing affine subspace of $\boldsymbol{x}$, or exactly touching it when $D(\cdot|\cdot)$ is continuous. We impute $\boldsymbol{x}$ by

$$\boldsymbol{x} = ave\Bigl(\operatorname*{arg\,min}_{\boldsymbol{y}\in\mathbb{R}^d\,,\,\boldsymbol{y}_{obs}=\boldsymbol{x}_{obs}} \{\|\boldsymbol{y}-\boldsymbol{z}\| \,|\, \boldsymbol{z}\in\mathbb{R}^d\,,\, \boldsymbol{z}\in D_{\alpha^*}(\boldsymbol{X}), \|\}\Bigr), \qquad (3)$$

with *ave* being the averaging operator. In this way, discrete depth functions as well those vanishing immediately beyond the convex hull of data (as, *e.g.*, the Tukey depth) are accounted for, also computationally; see Figure 1 for a data set from `http://stat.ethz.ch/Teaching/Datasets/`. On the other hand, as noted above, if $D(\cdot|\cdot)$ is continuous, one can explicitly write

$$\boldsymbol{x} = \operatorname*{arg\,max}_{\boldsymbol{y}\in\mathbb{R}^d\,,\,\boldsymbol{y}_{obs}=\boldsymbol{x}_{obs}} D(\boldsymbol{y}|\boldsymbol{X}), \qquad (4)$$

i.e. (instead of taking conditional mean) a point of the highest depth conditioned on $\boldsymbol{x}_{obs}$ and on $\boldsymbol{X}$ is taken.

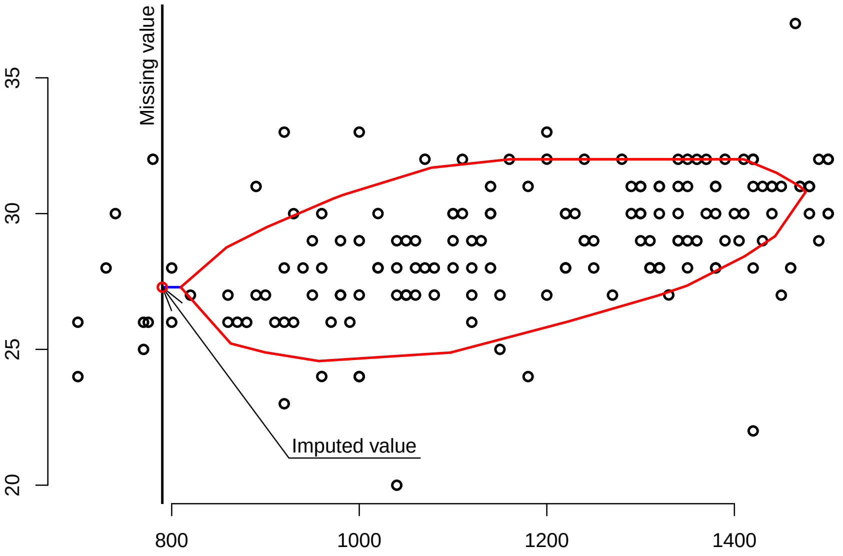

FIGURE 1. An imputation for the Babies data set using Tukey depth.

Given a data set with missing entries, we first fill not available data with starting values (say coordinate wise mean of the existing ones). Then to each observation with initially missing entries, (3, respectively 4) is applied, in this way updating all the missing entries. The process is iterated till convergence.

3 Discussion

The proposed method is general and generic, and can be coupled with any measure of centrality essentially defining its properties. Thus when employed with Mahalanobis (1936) depth, it imputes by iterated multiple-output regression, which coincides exactly with single imputation by iterated regression. Additionally, it can be shown that it yields exactly the same solution as imputation by the regularized PCA (Josse and Husson, 2012) when assuming rank equal to $d-1$ and any admissible variance of noise. Indeed, after convergence, each missing entry lies in the hyperplane of regressing on other coordinates; if for some $\boldsymbol{x}$ $\#miss > 1$, then on the intersection of several such regression hyperplanes, *i.e.* in general in a multiple output regression affine subspace of dimension $\#obs$.
With Tukey or projection depth, it yields a distribution-free imputation scheme, fitting missing value close to the data geometry. It does not exploit any estimates of location or scatter, avoiding problems with, *e.g.*, mean and covariance matrix, in a natural way. The approach is robust both in sense of outliers and heavy-tailed distributions, and for the class of continuous elliptically symmetric distributions imputed points converge to the points of the highest conditional density.

References

Josse, J. and Husson, F. (2012). Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique*, **153**, 1 – 21.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data.* Hoboken: John Wiley & Sons.

Mosler, K. (2013) Depth statistics. In: *Robustness and Complex Data Structures, Festschrift in Honour of Ursula Gather*, Springer, Berlin, 17 – 34.

Tukey, J.W. (1975). Mathematics and the picturing of data. In: *Proceeding of the International Congress of Mathematicians (Volume 2)*, Canadian Mathematical Congress, Vancouver, 523 – 531.

Van Buuren, S. (2012). *Flexible Imputation of Missing Data.* Boca Raton: Chapman & Hall/CRC.