Multiple imputation for elliptically symmetric distributions

Pavlo Mozharovskyi [1] & Julie Josse [2] & François Husson [3]

[1] *Centre Henri Lebesgue, IRMAR, Agrocampus Ouest;* pavlo.mozharovskyi@univ-rennes1.fr
[2] *Agrocampus Ouest, INRIA;* julie.josse@agrocampus-ouest.fr
[3] *Agrocampus Ouest, IRMAR;* husson@agrocampus-ouest.fr

Résumé. Nous proposons une méthode d'imputation multiple pour des données issues d'une distribution elliptique. Pour imputer en effectuant un tirage dans la distribution prédictive des données manquantes sachant les données observées, nous exploitons les liens entre distribution elliptique, distance de manahalobis et le concept de profondeur des données (mesure de centralité des données). L'imputation réalisée permet ainsi de préserver la distribution des données. Pour effectuer une imputation multiple au sens de Rubin et refléter l'incertitude associée à la prédiction des données d'une imputation à l'autre, nous utilisons ensuite une approche par boostrap non-parametrique. Les bonnes performances de la méthode sont illustrées via des simulations.

Mots-clés. Imputation multiple, distribution elliptique, imputation stochastique, distance de Mahalanobis, profondeur des données.

Abstract. A method for stochastic and multiple imputation of missing values is proposed for data coming from an elliptically symmetric distribution. For a pair of location and shape estimates, it exploits the Mahalanobis distance and an affine-invariant centrality measure (data depth) for drawing from conditional distributions, and reflects uncertainty by means of the Markov chain Monte Carlo and a bootstrap. As shown by a simulation study, the proposed method imputes close to the data and does not suffer from undercovering.

Keywords. Multiple imputation, elliptical distribution, stochastic imputation, Mahalanobis distance, data depth.

1 Introduction

Many of statistical methods for handling continuous random variables have been developed based on the assumption of normality, and the machinery for imputation of missing data is not an exception. For the multivariate normal distribution with a portion of entries missing (completely) at random (M(C)AR; see Rubin, 1976; Van Buuren, 2012), the point estimate of mean and covariance matrix can be obtained by the EM-algorithm (Dempster

et al., 1977), while inference can be drawn by means of multiple imputation. For the last one, the model uncertainty may be reflected using either bootstrap or Bayesian approach, see Schafer (1997) and, *e.g.*, packages `Amelia` and `norm` for an R-implementation.

In contemporary applications, however, often occur data, whose density shape deviates from the normal one. A natural extension of the multivariate normal model is the elliptically symmetric one, which allows for a broad class of densities but maintains the elliptical geometry of data. Given a vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and a $d \times d$ matrix $\boldsymbol{\Lambda}$, a random vector X is said to be elliptically distributed if

$$X \stackrel{D}{=} \boldsymbol{\mu} + R\boldsymbol{\Lambda}U,$$

for a nonnegative random variable $R \in \mathbb{R}_+$ and a random vector U uniformly distributed on the unit sphere $\mathcal{S}^{d-1}$, or equivalently $X \sim \mathcal{E}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, F_R)$, with $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}'$ and F_R being a cumulative distribution function (c.d.f.) of R. Formally, dimension of U as well as the rank of $\boldsymbol{\Lambda}$ can be smaller than d, and, in general, X may not possess a density (Fang *et al.*, 1990). In the current presentation we restrict to the "nice" case, when F_R is absolutely continuous and $\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma}$ are invertible. Then X possesses density

$$f_X : \mathbb{R}^d \to \mathbb{R}_+ , \boldsymbol{x} \mapsto f_X(\boldsymbol{x}) = \frac{c_{d,f}}{\sqrt{\|\boldsymbol{\Sigma}\|}} f\big(d_{Mah.(\boldsymbol{\mu},\boldsymbol{\Sigma})}(\boldsymbol{x})\big),$$

where $d_{Mah.(\boldsymbol{\mu},\boldsymbol{\Sigma})}(\boldsymbol{x}) = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$ is the Mahalanobis distance (Mahalanobis, 1936) to $\boldsymbol{\mu}$, $f : \mathbb{R}_+ \mapsto \mathbb{R}_{++}$ is the radial density, and $c_{d,f}$ is a constant ensuring that f_X integrates to one.

In what follows, we elaborate on the problem of corruption of a data set consisting of n observations drawn from $\mathcal{E}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, F_R)$ by absent entries appearing due to the M(C)AR mechanism. Denote this by $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$, where for some i-s, entries indexed with $miss(i)$ are missing, and those indexed with $obs(i)$ are observed. First, in Section 2, we present the stochastic imputation scheme and point out its extension to the use of data depth - a statistical measure of centrality w.r.t. a probability measure or a data cloud (Zuo and Serfling, 2000; Mosler, 2013). Then, in Section 3, we suggest a procedure for multiple imputation of elliptical distributions, based on depth distribution and Mahalanobis distance. We end the presentation by a short discussion in Section 4.

2 Stochastic imputation

The task of stochastic (or improper) imputation is to reflect uncertainty due to the distribution. Thus, in the case of multivariate normality, one can use the EM point estimates to draw $\boldsymbol{x}_{miss}$ from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ conditioned on $\boldsymbol{x}_{obs}$. For elliptical symmetry the task is more complicated. First, an algorithm estimating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ with missing data able to work for any elliptical distribution is required, and second, the density to draw from is generally unknown. To overcome the first difficulty, we design a Markov chain Monte Carlo

(MCMC) allowing to use estimators on complete data. Again, we first stick to the "nice" assumption and additionally admit moment estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to easily derive their conditional counterparts. The second problem seems to be more involved, we consider it in detail right below.

As our aim is imputation, *i.e.* rather to make a draw close to the data than to estimate the density, we exploit the idea of data depth when employing the following strategy. For an absolutely continuous elliptically symmetric distribution, any data depth satisfying corresponding postulates from Mosler (2013) or Zuo and Serfling (2000) possesses the characterization property: it is a monotone function of the radial density, which in order is expected to be a monotone function of the Mahalanobis distance. For the Mahalanobis depth, which is just a monotone transformation of the Mahalanobis distance, this relationship is the most intuitive. In what follows we denote a generic depth function w.r.t. $\boldsymbol{X}$ by $D_{\boldsymbol{X}}(\cdot)$.

Different to normal case, the shape of conditional distribution, and of the c.d.f. of depth as well, will differ from the unconditional one. The corresponding transformation will be mainly defined by Mahalanobis distances of the points and of the conditional mean. Let (for shortness) $\boldsymbol{\mu}^* \in \mathbb{R}^d$ denote the conditional mean for a point with missing entries. Further let $f_{D_{\boldsymbol{X}}(X)}$ denote the density of the depth for a random vector $X \in \mathbb{R}^d$ w.r.t. a sample $\boldsymbol{X}$. The depth should then be drawn as a quantile Q uniformly on $[0, F_{\boldsymbol{\mu}^*}\big(D_{\boldsymbol{X}}(\boldsymbol{\mu}^*)\big)]$, with (we omit mean and covariance for shortness)

$$
\begin{aligned}
F_{\boldsymbol{\mu}^*}(x) \;=\; & \int_0^x f_{D_{\boldsymbol{X}}(X)}(y) \frac{\Big(\sqrt{d^2_{Mah.}(y) - d^2_{Mah.}(\boldsymbol{\mu}^*)}\Big)^{\#miss-1}}{d^{d-1}_{Mah.}(y)} \times \\
& \times \; \frac{d_{Mah.}(y)}{\sqrt{d^2_{Mah.}(y) - d^2_{Mah.}(\boldsymbol{\mu}^*)}} dy,
\end{aligned}
\tag{1}
$$

and projected back on the support by $D = F^{-1}_{\boldsymbol{\mu}^*}(Q)$. The aim of this transformation is to normalize the volume. Any constant normalization factor can be omitted here as $F_{\boldsymbol{\mu}^*}(\cdot)$ is used exceptionally for drawing. The square root in the formula could be avoided, but this way Mahalanobis distance is exploited as a function of depth for joint distribution only, which can be obtained from the data without further transformations. For instance, when using the Mahalanobis depth, one can substitute directly in the equation (1) $d_{Mah.}(y)$ by $\sqrt{1/y - 1}$.

Now we should find the point on the region of depth D (this point will lie in its intersection with the hyperplane of missing values). We draw U^* uniformly on $\mathcal{S}^{\#miss-1}$, set, and transform it by conditional scatter matrix obtaining $U \in \mathbb{R}^d$ having $U_{miss} = \boldsymbol{\Lambda}^* U^*$ (with $\boldsymbol{\Sigma}_{miss,miss} - \boldsymbol{\Sigma}_{miss,obs}\boldsymbol{\Sigma}^{-1}_{obs,obs}\boldsymbol{\Sigma}_{obs,miss} = \boldsymbol{\Lambda}^*(\boldsymbol{\Lambda}^*)'$) and $U_{obs} = \mathbf{0}$. Such U is distributed uniformly on the conditional density contour. Then $\boldsymbol{x}$ having missing coordinates is imputed as $\boldsymbol{x} = \boldsymbol{\mu}^* + \alpha U$, where α is a positive scalar obtained as the solution of the quadratic equation $(\boldsymbol{\mu}^* + \alpha U - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^* + \alpha U - \boldsymbol{\mu}) = d^2_{Mah.}(D)$.

Now we are ready to formulate the MCMC algorithm (Algorithm 1) for improper imputation, which we start with single imputation (`impute.single` can be any single imputation method) to shorten the burn-in period. A step of Algorithm 1 is demonstrated in Figure 1.

Algorithm 1 Improper imputation

1: **function** IMPUTE.ELL.IMPROPER($\boldsymbol{X}$, num.burnin)
2: $\quad\boldsymbol{Y} \leftarrow$IMPUTE.SINGLE($\boldsymbol{X}$) ▷ Start MCMC with single imputation
3: $\quad$**for** $k = 1 : (num.burnin + 1)$ **do**
4: $\quad\quad\boldsymbol{\mu} \leftarrow \hat{\boldsymbol{\mu}}(\boldsymbol{Y})$
5: $\quad\quad\boldsymbol{\Sigma} \leftarrow \hat{\boldsymbol{\Sigma}}(\boldsymbol{Y})$
6: $\quad\quad$Estimate $f_{D_{\boldsymbol{Y}}(\boldsymbol{Y})}$
7: $\quad\quad$**for** $i = 1 : n$ **do**
8: $\quad\quad\quad$**if** $miss(i) \neq \emptyset$ **then**
9: $\quad\quad\quad\quad\boldsymbol{\mu}^*_{miss(i)} \leftarrow \boldsymbol{\mu}_{miss(i)}$ ▷ Calculate conditional mean
10: $\quad\quad\quad\quad\quad+ \boldsymbol{\Sigma}_{miss(i),obs(i)}\boldsymbol{\Sigma}^{-1}_{obs(i),obs(i)}(\boldsymbol{y}_{i,obs(i)} - \boldsymbol{\mu}_{obs(i)})$
11: $\quad\quad\quad\quad\boldsymbol{\mu}^*_{obs(i)} \leftarrow \boldsymbol{y}_{i,obs(i)}$
12: $\quad\quad\quad\quad$Calculate $F_{\boldsymbol{\mu}^*}(\cdot)$
13: $\quad\quad\quad\quad Q \leftarrow Unif\big([0, F_{\boldsymbol{\mu}^*}\big(D(\boldsymbol{\mu}^*)\big)]\big)$ ▷ Draw depth
14: $\quad\quad\quad\quad D \leftarrow F^{-1}_{\boldsymbol{\mu}^*}(Q)$
15: $\quad\quad\quad\quad U^* \leftarrow Unif(\mathcal{S}^{\#miss(i)-1})$ ▷ Draw random direction
16: $\quad\quad\quad\quad U_{miss(i)} \leftarrow U^*\boldsymbol{\Lambda}^*$
17: $\quad\quad\quad\quad U_{obs(i)} \leftarrow 0$
18: $\quad\quad\quad\quad\alpha \leftarrow$ positive solution of ▷ Intersection with contour
19: $\quad\quad\quad\quad\quad d^2_{Mah.}(D) = (\boldsymbol{\mu}^* + \alpha U - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^* + \alpha U - \boldsymbol{\mu})$.
20: $\quad\quad\quad\quad\boldsymbol{y}_{i,miss(i)} \leftarrow \boldsymbol{\mu}^*_{miss(i)} + \alpha U_{miss(i)}$ ▷ Impute missing entries
21: $\quad$**return** $\boldsymbol{Y}$

3 Multiple imputation

To reflect the uncertainty of model parameters as well, one can make use of multiple imputation. For an elliptically symmetric distribution, the natural parameters are the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. As no distributional assumptions are made, we resort to bootstrap, which leads to a slight modification of the Algorithm 1. First, a bootstrap sequence of indices is generated $(b_1, ..., b_n)$, each b_i , $i = 1, ..., n$ drawn as $Unif(\{1, ..., n\})$. Then, in each iteration, for estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (lines 4–5 of Algorithm 1) $\boldsymbol{Y^{(b)}} = \{\boldsymbol{y}_{b_1}, ..., \boldsymbol{y}_{b_n}\}$ should be used instead of $\boldsymbol{Y}$.

We apply multiple imputation to a model $\boldsymbol{\beta}'(1, \boldsymbol{x}')' + \epsilon$, with $\boldsymbol{\beta} = (0.5, 1, 3)'$, $\boldsymbol{x} \sim$

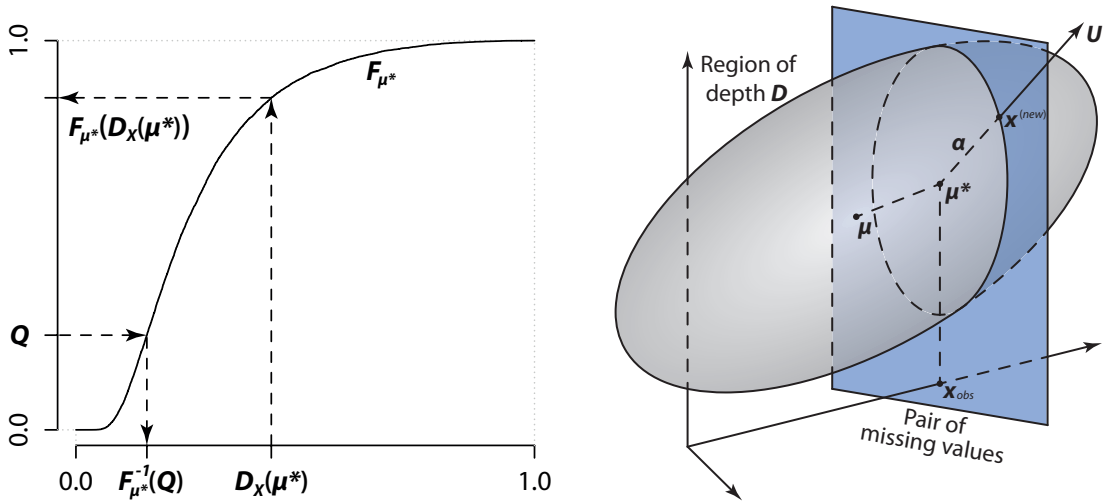

Figure 1: Illustration of Algorithm 1. Drawing depth $D = F^{-1}_{\mu^*}(Q)$ via the depth c.d.f. F_{μ^*} (left) and locating the corresponding imputed point $\boldsymbol{x}^{(new)}$ (right).

$N\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$, and $\epsilon \sim N(0, 0.25)$, generating $m = 5$ and $m = 20$ multiply-imputed samples. This yields a three dimensional multivariate normal distributions of the vector $(\boldsymbol{x}', y)'$. We include R-packages `Amelia` and `mice` in the comparison. The results are presented in Table 1. Columns "med", "cov", and "confi" indicate respectively median of the estimated regression coefficients, coverage by the 95% confidence interval calculated according to the Rubins' rules, and median width of the confidence interval, over 1000 runs.

Table 1: Simulation results for Model 2

	β_0			β_1			β_2		
	med	cov	confi	med	cov	confi	med	cov	confi
	$m = 5$								
Amelia	0.5	0.946	0.536	1.005	0.939	0.438	2.999	0.94	0.226
mice	0.525	0.984	1.464	1.063	0.975	1.476	2.92	0.976	0.88
Ell	0.513	0.974	0.719	0.989	0.957	0.589	3	0.961	0.295
	$m = 20$								
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
Ell	0.504	0.971	0.613	0.989	0.979	0.519	3.003	0.97	0.26

One can see that the proposed method has a slightly higher coverage due to wider confidence intervals, which, on the other hand, is never below 95%. Additional study

(not presented here) confirms these to be reasoned by the higher between sample variation, which can be explained by the semi-parametric nature of the model. Another not presented simulation shows very precise reconstruction of quantiles for distributions with heavier tails, *e.g.* Student-t with 3 and 5 degrees of freedom. It is interesting to notice, that for Model 2, where the correlation between second and third dimensions is very high (≈ 0.988), `mice` with normal imputation turns out to be biased.

4 Discussion

The proposed method in a unified way performs stochastic and multiple imputation for the natural generalization of the multivariate normal model – elliptically symmetric distributions. The considered (joint) model is of semi-parametric nature. While its parametric part (location and shape estimates) is dealt with by the transform based on Mahalanobis distance, the non-parametric one (radial density) is accounted for by a c.d.f. of a centrality measure. By that, it is able to impute close to the underlying radial density, and thus reflects quantiles and further distribution-dependent statistics. When using statistical data depth or an outlier-persistent estimator for the covariance matrix, the technique can be used for robust imputation. As the simulation study shows, the proposed method never suffers from under-covering, and thus, though somewhat conservative, may be preferred in practice.

References

[1] 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*, 39, 1–38.

[2] Fang, K.-T., Kotz, S., and Ng, K. W. (1990), *Symmetric Multivariate and Related Distributions*, Chapman & Hall, London.

[3] Mahalanobis, P. (1936), On the generalized distance in statistics, *Proceedings of the National Academy India*, 12, 49–55.

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

[5] Rubin, D. B. (1976), Inference and missing data, *Biometrika*, 63, 581–592.

[6] Schafer, J. (1997), *Analysis of Incomplete Multivariate Data*, Chapman & Hall/CRC, Boca Raton.

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

[8] Zuo, Y. J. and Serfling, R. (2000), General notions of statistical depth function, *The Annals of Statistics*, 28, 461–482.