

A numerical study of multiple imputation methods
using nonparametric multivariate outlier identifiers and
depth-based performance criteria with clinical
laboratory data

Xin Dang ¹

University of Mississippi

and

Robert Serfling ²

University of Texas at Dallas

September 2009

¹Department of Mathematics, University of Mississippi, University, MS 38677, USA. Telephone: (662) 915-7409. Fax: (662) 915-2361. Email: xdang@olemiss.edu.

²Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX 75083-0688, USA. Telephone: (972) 883-2361. Fax: (972) 883-6622. Email: serfling@utdallas.edu. Website: www.utdallas.edu/~serfling.

Abstract

It is well known that if a multivariate outlier has one or more missing component values, then *multiple imputation* methods tend to impute non-extreme values and make the outlier become less extreme and less likely to be detected. In this paper, nonparametric depth-based multivariate outlier identifiers are used as criteria in a numerical study comparing several established methods of multiple imputation as well as a new proposed one, nine in all, in a setting of several actual clinical laboratory data sets of different dimension. Two criteria, an “outlier recovery probability” and a “relative accuracy measure”, are developed, based on depth functions. Three outlier identifiers, based on Mahalanobis distance, robust Mahalanobis distance, and generalized PCA, are also included in the study. Consequently, not only the comparison of imputation methods, but also the comparison of outlier detection methods, is accomplished in this study. Our findings show that the performance of a multiple imputation method depends on the choice of depth-based outlier detection criterion, as well as the size and dimension of the data and the fraction of missing components. By taking these features into account, a multiple imputation method for a given data set can be selected more optimally.

AMS 2000 Subject Classification: Primary 62H99, Secondary 62G99.

Key words and phrases: multiple imputation; multivariate; nonparametric; outlier detection; depth functions; missing values.

1 Introduction

Data analysis often becomes difficult in the presence of missing values, and so does outlier detection. In clinical laboratory data, missing values are very common because a patient may miss a follow-up measurement, or because some measurements go unreported due to spoiled samples or instrument malfunction. Past procedures for missing values include *case deletion* (discarding any observation with a missing measurement) and *single imputation* (replacing the missing measurement by a plausible value). Although such methods are simple and easy to implement, they often lead to biased estimates, invalid inferences, loss of power, and underestimated uncertainty (Little and Rubin, 2002). To overcome such drawbacks, a *multiple imputation (MI)* procedure, originally proposed by Rubin (1987), incorporates the degree of imputation uncertainty and also is much less computationally intensive than resampling methods (such as the bootstrap, jackknife), and hence has very broad general applicability.

Outlier detection is of considerable interest in a clinical trial, because it can highlight patients whose laboratory measurements do not follow the same pattern of relationships as the majority of patients. In the multivariate outlier detection context, dealing with missing values is even more difficult. Firstly, marginal checking of each univariate measurement for outlyingness is inadequate. For example, for a patient can be nonoutlying in each coordinate direction and yet still be an outlier in the sense the multivariate measurement lying apart from those of the main body of data. Secondly, outlier detection methods themselves must be robust, for otherwise they will be distorted in the presence of outliers and adversely impact the conclusion by masking and swamping effects. Further, if an outlier has one or more missing component values, then multiple imputation methods will tend to impute non-extreme values and make the outlier become less extreme and less likely to be detected. The first two difficulties arise from the problem of multivariate outlier detection itself. They can be overcome by using robust nonparametric depth-based outlier identifiers (Chen *et al.*, 2009, Dang and Serfling, 2009). Two depth-based outlier identifiers are included in

this study along with three other widely used methods. The third difficulty stems from the presence of missing values and a consequence of the way missing values are handled. In this study, we investigate several leading MI methods as well as a proposed median imputation method in the multivariate outlier detection context. Two criteria based on outlier identifiers are developed to compare performance of MI methods. One, called “outlier recovery probability”, exhibits the degree of agreement between the outliers detected in the original complete data set and those detected in imputed data sets. The other, a “relative accuracy measure”, is a relative metric providing an overall quantity to measure the accuracy of imputation methods. Consequently, not only the comparison of MI methods but also the comparison of outlier detection methods can be carried out in this study, which considerably extends that of Penny and Jolliffe (1999).

In Section 2, we discuss each multiple imputation method and each outlier identification method and address computation issues. Section 3 describes our framework for evaluation of the given multiple imputation methods and introduces two outlier-based criteria to assess performance. Section 4 presents results for four laboratory data sets of differing size and dimension (kindly provided by Dr. Kay Penny). A concluding discussion is provided in Section 5, with practical recommendations.

2 Multiple imputation and outlier detection methods

Following Rubin (1987, 1996) and Little and Rubin (2002), where further details may be found, the multiple imputation (MI) procedure we consider involves the following steps: (i) estimation of $M > 1$ values for each missing value, (ii) creation of M complete data sets differing only in the imputed values, (iii) analysis of each created data set by a complete-data method using standard methods, and finally (iv) combination of the M separate results to produce inferences that account for the uncertainty due to missing data. The MI procedure is attractive from both theoretical and practical standpoints.

In our study, missing values are estimated by each of nine different MI methods, and

each imputed complete data set is examined by five different outlier detection methods. Combining the M sets of results for each combination of imputation method and outlier detection method, we can determine which combinations lead to more accurate results.

2.1 Multiple imputation methods

The study involves several established multiple methods as well as a proposed one, “median imputation”, described as follows:

- **Rand.** Each missing value is replaced by a value randomly selected from the set of all observed sample values for that variable.
- **Mean.** For each variable j , the sample mean and standard deviation s_j^2 are calculated with missing values ignored. A missing value is replaced by this mean plus a random error from the $N(0, s_j^2)$ distribution.
- **Median.** Instead of mean imputation, the median plus an error from $N(0, \text{MAD}_j^2)$ replaces the missing value, where MAD_j is the usual median absolute deviation from the median for the variable j .
- **Reg1.** A missing value is estimated by simple linear regression on the variable most correlated with the variable containing the missing value. If the value of the most correlated variable is also missing, then the second most correlated variable is used for the regression predictor, etc. A residual error is randomly selected from the observed residuals.
- **Reg2.** Missing values of variable j are estimated as with Reg1, except that the residual error term is randomly drawn from $N(0, s_j^2)$, with s_j the standard deviation of the observed residuals.
- **Dreg1.** Missing values are estimated using regression on two variables. We choose the best regression model under the R^2 criterion to predict the missing value. Note that

the regression on the most two highly correlated variables is *not* necessarily the best one. A residual error term is selected as with Reg1.

- **Dreg2.** Missing values are estimated as with Dreg1, but with the residual error term selected as with Reg2.
- **Mreg1.** The missing value is predicted by using regression on all available variables in that particular observation. A residual error is selected as with Reg1 and Dreg1.
- **Mreg2.** The same as Mreg1, but with residual errors selected as with Reg2 and Dreg2.

Note that Rand, Mean and Median only use marginal information without incorporating the relationship with other variables. We call them “marginal imputation methods”. For “regression imputation methods”, if a value of a binary or categorical variable is missing, then logistic or generalized linear regression methods should be used for imputing values for that variable. Also, before applying regression methods, a preliminary study should be conducted to determine necessary and appropriate transformations of variables.

There are many standard routines in the statistical softwares for multiple imputation. Principally there are two major approaches. The first one is based on the joint distribution of all the variables considered. SAS proc mi (MCMC) and Schafer’s NORM (Schafer, 1997, Schafer and Olsen, 1999) assume multivariate normality and an iterative, two-step process is used, in which missing values and sample parameters are alternatively drawn from their distributions. The other approach is based on each conditional density of a variable given all other variables. This is the approach mice of R and Stata’s ice take. (m)ice stands for Multiple Imputation by Chained Equation which generates multiple imputations by Gibbs sampling (Van Buuren *et al.*, 2006 and Van Buuren, 2007). Like SAS mi, (m)ice is an iterative two-step process. The first step is the imputation of a single variable given a set of predictor variables. The second step is the so called “regression switching”, a scheme for cycling through all the variables and executing the first step for each variable with missing values. Repeating these steps a large number of times eventually produces a draw from

each conditional distribution. Our regression methods can be used as a module in (m)ice. Instead of specifying a set of predictor variables for a target variable, we search an “optimal” regression model based on one or two predictor variable(s).

2.2 Outlier detection methods

Depth functions are increasingly being used in building nonparametric outlier identifiers and offer several advantages. First, they are conceptually simple. Compensating for lack of ordering in multi-dimensional space, depth functions provide center-outward orderings of points in \mathbb{R}^d with respect to a given distribution or a data cloud. Hence it is very natural to declare an observation with depth value below some threshold an “outlier”. Secondly, with typical choices of depth functions, the depth-induced outlier identifiers are robust. Dang and Serfling (2009) investigate their robustness properties using masking and swamping breakdown points. Thirdly, depth-based outlier detection is a nonparametric approach, requiring fewer and weaker assumptions than parametric approaches and thus having wider scope of application.

The outlier identifiers in this study consist of two based on the spatial and projection depth functions, and three based on Mahalanobis distance, robust Mahalanobis distance, and generalized PCA, respectively. Indeed, we combine these into the framework of depth functions, using the fact that depth and outlyingness are inverse notions. An outlyingness function $O(x, F)$, $x \in \mathbb{R}^d$, taking values in $(0, 1)$ or $(0, \infty)$, generates a corresponding depth function taking values in $(0, 1)$, defined as either $D(x, F) = 1 - O(x, F)$ or $D(x, F) = 1/(1 + O(x, F))$. In this fashion, along with the spatial depth (SD) and the projection depth (PD), we obtain the Mahalanobis depth (MD), the robust Mahalanobis depth (RMD), and the GPCA depth (GD), respectively.

In this study, without specifying any particular threshold, we apply these to identify the 3 or 6 most outlying observations. Using criteria that we introduce, the performances of the multiple imputation methods may be compared within an outlier detection setting.

2.2.1 Spatial and projection depths

Formally first introduced by Vardi and Zhang (2000), the *spatial depth* of a point x with respect to a distribution F is defined as

$$D_S(x, F) = 1 - \|E_F S(x - X)\|, \quad x \in \mathbb{R}^d,$$

where $S(x) = x/\|x\|$ is the vector sign function ($S(0) = 0$) with Euclidean norm $\|\cdot\|$, and X has distribution F . It relates to the spatial quantile function of Chaudhuri (1996) (see also Serfling, 2002, for discussion). The sample version is

$$D_S(x, F_n) = 1 - \left\| \frac{1}{n} \sum_{i=1}^n S(x - x_i) \right\| = 1 - \left\| \frac{1}{n} \sum_{i=1}^n \frac{x - x_i}{\|x - x_i\|} \right\|, \quad x \in \mathbb{R}^d,$$

where F_n is the empirical distribution function of the data x_1, \dots, x_n . The point of maximal depth is the sample spatial median. The outlier identifier based on spatial depth is robust against the presence of outliers, in the sense that the spatial depth of x does not change if any observation is moved to ∞ along the ray connecting it. That is, the magnitude of any observation has no impact on the the spatial depth of x as long as its direction relative to x doesn't change.

The *projection depth* of x with respect to a distribution F is defined as

$$D_P(x, F) = 1/(1 + O_p(x, F)),$$

in terms of the project outlyingness function

$$O_P(x, F) = \sup_{\|u\|=1} \frac{|u^T x - \mu(F_u)|}{\sigma(F_u)},$$

where μ and σ are location and scale functionals and F_u denotes the distribution of $u^T X$, with X having distribution F . The outlyingness is the worst standardized deviation from the center among all univariate projections. This raises the question of computation. In most cases, approximate computation by stochastic sampling is fast enough and accurate enough for general application. We compute the maximum outlyingness among N random

directions. Choose N large enough to obtain stability of the results. For robustness, μ and σ should be robust estimators, for example, the median and MAD. For more properties and details about projection depth, see Zuo and Serfling (2000), Zuo (2003) and Pan *et al.* (2000).

2.2.2 Mahalanobis depth and robust Mahalanobis depth methods

The Mahalanobis distance $\|x - \mu_F\|_{\Sigma_F} = \sqrt{(x - \mu_F)^T \Sigma_F^{-1} (x - \mu_F)}$ is a widely used method of detecting outliers in multivariate data (see, e.g., Barnett and Lewis, 1994), popular for its affine invariance, mathematical tractability, and computational ease. Using the sample mean \bar{x} and the sample covariance matrix S , the corresponding *sample Mahalanobis depth* is

$$D_M(x, F_n) = \frac{1}{1 + \|x - \bar{x}\|_{S^{-1}}} = \frac{1}{1 + \sqrt{(x - \bar{x})^T S^{-1} (x - \bar{x})}}.$$

Of course, the sample mean and sample covariance matrix are very sensitive to the influence of outlying observations. Hence a robust version (RMD) is obtained using robust location and covariance matrix estimators. In this study, we utilize the package *robust* in R. It first computes the raw minimum covariance determinant (MCD) estimator of Rousseeuw (1984, 1985). Then the MCD estimate is used to assign weights to the observations such that the weighted estimates of location and covariance have good robustness and efficiency properties. For details about the algorithm, see Rousseeuw and Van Driessen (1999).

2.2.3 The Generalized PCA depth

Principle component analysis (PCA) is among the oldest and best known techniques of multivariate analysis. It is used in many different ways and applications. The central idea of PCA is dimensionality reduction by transforming to the first few principal components, which are uncorrelated and retain most of the variation present in the data set. Traditional PCA consists in calculating eigenvalues and eigenvectors of the sample covariance matrix S and hence inherits the sensitivity to outliers of S . One generalization of PCA introduces a metric V on either the observations or the variables. In the context of outlier detection, an

appropriate metric is the inverse of a robust covariance matrix, e.g.,

$$V^{-1} = \frac{\sum_{i=1}^n W(\|x_i - \bar{x}\|_{S^{-1}}^2)(x_i - \bar{x})(x_i - \bar{x})^T}{\sum_{i=1}^n W(\|x_i - \bar{x}\|_{S^{-1}}^2)}$$

with the weight function W a decreasing function. Thus V^{-1} is a weighted covariance matrix with weights based on squared Mahalanobis distance. An observation far from the sample mean is given less weight. If no outlier is present, S and V^{-1} would be equivalent and SV the identity matrix. If there some outliers are present, then the projection on the subspace spanned by q eigenvectors will possibly display the outliers. Thus “generalized PCA” carries out eigen analysis on SV . Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$ be the eigenvalues and $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_d$ the associated eigenvectors of SV . If the first $q \ll d$ principal components are considered, the centered data are projected orthogonally onto the subspace spanned by $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_q$. For more details, see Caussinus and Ruiz (1990, 1993) and Jolliffe (2002).

Here we use Euclidean distance from the two-dimensional GPCA plot, i.e, we take $q = 2$, as an outlyingness measure. That is,

$$O_G(x_i, F_n) = \left(\sum_{j=1}^2 z_{ij}^2 \right)^{1/2},$$

where z_{ij} is the j th component score for the i th observation. Accordingly, the GPCA depth is $D_G(x_i, F_n) = 1/(1 + O_G(x_i, F_n))$.

3 Evaluation procedure and criteria

Here we investigate performance of multiple imputation methods under the assumption that missing values occur completely at random (MCAR) according to some model or general pattern. It is worthwhile to point out that the conclusions are valid and likely same under assumption of missing at random (MAR). Four complete clinical laboratory data sets kindly provided by Dr. Kay Penny are used in the present study: Elect4, Chem12, Pre30, and Haemat7. Elect4 consists of 4 electrolyte measurements, and Chem12 12 chemistry measurements, for each of 230 patients. Pre30 contains 4 electrolyte measurements from Elect4,

12 chemistry measurements from Chem12, and 14 other pre-trial measurements, for each of these patients. Haemat7 consists of 7 haematological measurements for 88 of these patients. These data sets with dimensions 4, 12, 30, and 7 are representative of many situations, and we anticipate that conclusions drawn from our study may be applied quite generally.

3.1 Procedure

We carry out the experiment with the following steps:

1. For each original complete data set, for each outlier detection method the associated depth measure D_i is calculated for each observation x_i , and the 6 most outlying (smallest depth value) observations are identified.
2. A fraction α (10%, 20%, or 30%) of values in the original complete data set are randomly removed to generate missing data. We use suffix .miss1, .miss2, or .miss3 to denote the data after removal of 10, 20 or 30 % of values, respectively. If all components of an observation are missing, the observation will be deleted.
3. The missing values are then imputed by each imputation method.
4. For each complete imputed data set obtained in Step 3, the depth measures D_i are again calculated and the 3 most outlying observations are identified. In this paper, all outlier detection methods are built upon on imputed data. However, for Mean imputed data, mean values used in Mahalanbis depth and Generalized GPCA depth can be computed from the missing data for a better efficiency.
5. Steps 3 and 4 are replicated M times, and the M results from Step 4 are combined, and two performance criteria are computed.
6. Based on these criteria, conclusions are drawn and practical recommendations are developed.

3.2 Criteria

Two criteria have been developed to evaluate the imputation methods. Since a typical result of outlier detection is a list of outliers, our first criterion is based on *outlier recovery*. For each imputation method, there are M sets each listing 3 most outlying observations, where M is the number of multiple imputations carried out. We compare them with the outlier list in the original complete data set. The *Outlier Recovery Probability (ORP)* is calculated by counting the number of observations in the outlier list of the imputed data which also are in the outlier list of the original complete data set, then dividing by the total number $3M$. The outlier recovery probability exhibits the degree of agreement between the outliers detected in an original data set and those detected in an imputed data set.

A second performance criterion is based on *relative accuracy*, comparing depth values in an imputed data set to depths in the original data set. Let D_i be a depth measure of the i th observation in the original data set, and for each imputation method let $D_{i(m)}$ represent the corresponding measure for observation i in the m th imputed data set, $m = 1, \dots, M$. In order to compare across imputations, we standardize $D_{i(m)}$ by $D_{i(m)}^*$, where

$$D_{i(m)}^* = D_{i(m)} \frac{\sum_{j=1}^n D_j}{\sum_{j=1}^n D_{j(m)}}. \quad (1)$$

Then we define the relative accuracy of observation i as

$$\text{Racc}_i = \frac{1}{M} \sum_{m=1}^M \frac{|D_{i(m)}^* - D_i|}{D_i}. \quad (2)$$

The average of Racc_i over the observations gives an overall measure, the *Relative Accuracy Measure (RAM)*

$$\text{RAM} = \frac{1}{n} \sum_{i=1}^n \text{Racc}_i,$$

for the given depth measure. The RAM provides the average absolute value of the change in depth between the original data and the imputed data. A small RAM implies a good imputation method. The RAM also may be considered as the weighted average of absolute deviation of depth between the original data and the imputed data, with weight proportional

to the reciprocal of depth. An observation with small depth in the original data has a large weight on the RAM. Hence the RAM is more sensitive to change of depth for outliers than for “normal” observations.

In our study, the number of replications M is equal to 5, which is recommended by Rubin (1987) and suffices to produce valid inference due to the missing components. Our experiments also confirm this choice for M if the positions of the missing values are also given.

4 Results

We present and discuss results separately for the data sets in order of increasing dimension, with Tables 1-4 containing the numerical results, respectively. We note that the relative accuracy measures are not comparable across different outlier detection methods, i.e., the numbers along each column in the bottom parts of Tables 1-4 are not comparable. This is because the depth values D under different methods have differing ranges and distributions, and it is not suitable here to perform normalization or standardization operations.

In Section 5 we summarize all the findings taken together and develop some useful general recommendations.

4.1 Results for Elect4

In the original complete Elect4 data set, observations 11 and 69 are most outlying. Both have low values on the first measurement and high values on the fourth, which is inconsistent with the positive relationship between those two measurements in the majority of the data. All 5 outlier identifiers successfully identify these two cases.

The study results based on this data set are presented in Table 1, which we discuss these separately for each choice of α .

With randomly generated missing data, Elect4.miss1 has 10% of values missing, affecting 72 observations out of 230. Of the affected observations, 11 have 2 measurements missing

Table 1: Results for the Elect4 data set. Bold-face indicates best in each row.

	α	Mean	Med	Rand	Reg1	Reg2	Dreg1	Dreg2	Mreg1	Mreg2	
OUTLIER RECOVERY PROBABILITY	0.1	MD	0.67	0.60	0.53	0.87	0.80	0.87	0.80	0.80	0.87
		RMD	0.73	0.67	0.60	0.93	0.93	0.93	0.93	0.80	1.00
		GD	0.87	0.53	0.60	0.93	0.87	0.80	0.87	0.80	0.93
		SD	0.93	0.73	0.60	1.00	1.00	1.00	0.93	0.93	1.00
		PD	0.80	0.53	0.53	0.93	0.80	0.80	0.93	0.80	0.93
	0.2	MD	0.27	0.20	0.13	0.40	0.40	0.40	0.40	0.33	0.40
		RMD	0.40	0.20	0.07	0.67	0.47	0.47	0.53	0.60	0.47
		GD	0.73	0.27	0.13	0.80	0.73	0.60	0.87	0.87	0.47
		SD	0.73	0.27	0.33	0.60	0.67	0.87	0.80	0.73	0.80
		PD	0.53	0.20	0.07	0.47	0.47	0.33	0.67	0.63	0.67
	0.3	MD	0.47	0.07	0.07	0.60	0.67	0.60	0.67	0.53	0.60
		RMD	0.80	0.33	0.40	0.53	0.87	0.80	0.80	0.53	0.67
		GD	0.60	0.13	0.33	0.47	0.53	0.47	0.67	0.53	0.53
		SD	0.47	0.27	0.40	0.53	0.73	0.87	0.80	0.60	0.60
		PD	0.67	0.07	0.07	0.53	0.60	0.60	0.60	0.67	0.73
RELATIVE ACCURACY MEASURE	0.1	MD	0.075	0.105	0.079	0.065	0.069	0.070	0.074	0.062	0.058
		RMD	0.009	0.013	0.009	0.008	0.008	0.008	0.010	0.007	0.007
		GD	0.222	0.371	0.189	0.198	0.205	0.251	0.206	0.239	0.134
		SD	0.119	0.158	0.121	0.098	0.104	0.101	0.118	0.088	0.092
		PD	0.087	0.115	0.089	0.078	0.082	0.080	0.087	0.078	0.075
	0.2	MD	0.109	0.147	0.111	0.116	0.116	0.109	0.105	0.106	0.107
		RMD	0.013	0.019	0.014	0.014	0.013	0.013	0.013	0.012	0.011
		GD	0.383	0.439	0.338	0.350	0.334	0.371	0.310	0.382	0.320
		SD	0.201	0.284	0.208	0.214	0.204	0.191	0.183	0.190	0.189
		PD	0.124	0.161	0.123	0.131	0.125	0.123	0.127	0.117	0.113
	0.3	MD	0.151	0.190	0.153	0.152	0.148	0.145	0.141	0.146	0.142
		RMD	0.019	0.026	0.019	0.019	0.018	0.018	0.018	0.017	0.017
		GD	0.377	0.497	0.415	0.369	0.387	0.391	0.354	0.387	0.372
		SD	0.271	0.363	0.277	0.269	0.267	0.256	0.270	0.258	0.253
		PD	0.166	0.205	0.166	0.170	0.163	0.164	0.166	0.160	0.154

and 1 has 3 variables missing. Observation 11 has 1 value missing, while observation 69 is complete. In terms of the *outlier recovery probability criterion*, regression imputation methods are found to be superior to marginal imputations. *Mreg2* consistently performs best, with highest outlier recovery probabilities, under each outlier detection method. *SD* consistently has the highest recovery rate, for each imputation method. And *RMD* performs better than *MD*. By the *relative accuracy measure criterion*, similar results are observed, for comparison of the imputation methods.

The data set `Elect4.miss2` with 20% missing values affects 128 observations. Among them, 90 have 1 variable missing, and 31 including observations 11 and 69 have 2 missing components. Observation 11 misses the first and fourth measurements, while observation 69 is missing the first and the second variables. This presents difficulties for detecting both outliers in the missing data, since imputation methods tend to impute non-extreme values and make outliers become less extreme and less likely to be detected. This results in low outlier recovery rates. Most of them are lower than 60%, which indicates that both outliers remaining undetected in the imputed data. However, the regression imputation methods have better chance to recover true outliers than marginal imputations. *Median* and *Rand* perform extremely poorly. *Dreg2* and *Mreg2* are stronger than others according to both the outlier recovery probability and the relative accuracy measure criteria. *SD* and *GD* have higher outlier recovery rates than other methods.

There are only 59 complete observations in the data set `Elect4.miss3` with approximately 30% of values missing. Here 78 observations have 1 measurement missing, 64 are missing 2 components, and 18 are missing 3. Observation 27 which misses all 4 components makes regression methods infeasible, although the marginal imputation methods may still apply. In order to obtain a fair and convenient comparison, the results for all imputations are based on all observations except 27. Based on both criteria, *Dreg2* performs the best for *MD* and *GD*, while *Mreg2* is better than others for *PD*. For *RMD*, *Reg2* has a higher recovery rate than others but is not necessary the best according to the relative accuracy measure. For

SD, the two criteria prefer different imputation methods, one for *Dreg1*, the other *Mreg2*.

4.2 Results for Haemat7

For the original complete data, the four robust outlier identifiers all find that observations 84, 82 and 85 are most outlying, but *MD* fails to do so and claims incorrectly that observations 30 and 64 are outliers. This is caused by the poor masking and swamping performance of the outlier detector based on *MD*. Observations 84, 82 and 85 are very outlying in the 7th measurement, with raw values 669, 636 and 669 respectively. They pull the sample mean in that direction and inflate the variation in that component. The sample mean and standard deviation of values for that measurement are 275 and 105 respectively, comparing with median 252 and MAD 70. This results in masking of observation 82, which is supposed to be detected, and in swamping of observations 30 and 64, incorrectly identifying them as outliers by *MD*.

There are 48 affected observations in Haemat7.miss1 generated by randomly removing approximately 10% of values. Here 11 observations have 2 components missing and 3 have 3 values missing. Observations 84 and 85 have one missing measurement each, while observations 82, 30 and 64 are complete. The two criteria ORP and RAM agree on best performance of *Mreg2* in *GD* and of *Dreg2* in *PD*. On the other hand, *Mreg1* seems strong for *RMD* and *SD*, with the best RAM and the second best ORP. *Rand* appears to be the least accurate.

The data set Haemat7.miss2 with roughly 20% values randomly missing affects 77 observations. Among them, 23 including observation 30 miss 1 variable, 29 including observation 64 have 2 components missing, 11 including outlier 85 contain 3 missing measurements, and 3 including outlier 84 miss 4 values. Outlier 82 contains no missing value. *Reg1* is preferred by both criteria for all outlier detection methods except *MD*. *RMD* has consistently higher outlier recovery rate than other outlier methods, no matter the imputation method.

The missing data Haemat7.miss3 contains 30% missing values, with only 9 complete observations. Observations 84, 82 and 30 have 3 missing variables each, and observation

Table 2: Results for the Haemat7 data set. Boldface indicates best in each row.

	α	Mean	Med	Rand	Reg1	Reg2	Dreg1	Dreg2	Mreg1	Mreg2	
OUTLIER RECOVERY PROBABILITY	0.1	MD	0.40	0.60	0.53	0.73	0.80	0.93	0.87	0.60	0.33
		RMD	1.00	1.00	0.80	0.87	0.93	0.93	1.00	0.93	1.00
		GD	0.93	1.00	0.67	0.87	1.00	0.93	0.93	0.87	1.00
		SD	0.93	0.73	0.53	1.00	0.80	0.87	0.87	0.93	0.73
		PD	0.40	0.20	0.13	0.80	0.93	0.93	1.00	0.93	0.93
	0.2	MD	0.53	0.47	0.53	0.60	0.40	0.47	0.40	0.67	0.60
		RMD	0.93	1.00	0.73	0.93	1.00	0.73	0.87	0.80	0.93
		GD	0.87	0.93	0.67	1.00	0.93	0.73	0.80	0.73	0.80
		SD	0.60	0.67	0.47	0.87	0.67	0.67	0.80	0.73	0.73
		PD	0.33	0.33	0.27	0.67	0.67	0.53	0.47	0.50	0.60
	0.3	MD	0.33	0.33	0.20	0.33	0.40	0.20	0.33	0.40	0.33
		RMD	0.93	1.00	0.73	0.80	0.93	0.87	1.00	0.33	0.33
		GD	0.93	1.00	0.67	0.67	0.93	0.87	1.00	0.13	0.27
		SD	0.53	0.67	0.53	0.47	0.67	0.67	0.53	0.13	0.20
		PD	0.47	0.33	0.27	0.53	0.53	0.73	0.40	0.27	0.13
RELATIVE ACCURACY MEASURE	0.1	MD	0.102	0.124	0.139	0.098	0.094	0.102	0.094	0.095	0.105
		RMD	0.186	0.194	0.217	0.197	0.237	0.193	0.206	0.140	0.213
		GD	0.538	0.310	0.542	0.497	0.461	0.468	0.374	0.339	0.277
		SD	0.141	0.150	0.155	0.108	0.135	0.129	0.124	0.105	0.116
		PD	0.144	0.201	0.189	0.133	0.204	0.174	0.130	0.184	0.171
	0.2	MD	0.145	0.174	0.147	0.138	0.133	0.137	0.141	0.131	0.124
		RMD	0.276	0.252	0.224	0.197	0.286	0.270	0.208	0.244	0.253
		GD	0.611	0.434	0.647	0.479	0.650	0.774	0.565	0.524	0.594
		SD	0.277	0.260	0.285	0.231	0.299	0.323	0.250	0.232	0.271
		PD	0.167	0.219	0.206	0.148	0.222	0.192	0.155	0.207	0.204
	0.3	MD	0.175	0.197	0.183	0.165	0.180	0.176	0.175	0.174	0.181
		RMD	0.355	0.345	0.349	0.404	0.393	0.412	0.403	0.401	0.447
		GD	0.813	0.875	0.956	0.900	0.954	1.074	0.797	0.919	0.938
		SD	0.616	0.737	0.559	0.671	0.691	0.749	0.690	0.657	0.647
		PD	0.230	0.290	0.241	0.215	0.258	0.235	0.215	0.240	0.236

82 contains 1 measurement missing. *Dreg2* and *Median* perfectly recover the outliers when *RMD* and *GD* are used. *Mreg1* and *Mreg2* perform poorly in terms of recovery rate as well as relative accuracy. This is mainly because this 7-dimensional data set Haemat7.miss3 has relatively small sample size $n = 88$ but a high fraction of missing values. The multiple regression prediction will be very poor when the number of available observations is limited. An even more serious problem may occur when the number of available observations is less than the number of parameters in multiple regression. In that case, *Mreg1* and *Mreg2* become inapplicable.

4.3 Results for Chem12

In the original complete Chem12 data set, every outlier method concludes observation 46 as outlying. Its 6th measurement is extremely large, 6 standard deviations away from the sample mean. Also, observations 7 and 163 are fairly outlying according to *RMD*, *GD* and *SD*. Observation 110 is regarded as an outlier by *MD*, *GD* and *PD*.

Chem12.miss1 with 10% missing values has 169 affected observations. Of these, 96 contain 1 missing measurement each, 51 miss 2 variables each, and the remaining observations have 3 or 4 values missing. The outliers 46 and 163 contain one missing value each, while observation 110 is complete. *MD* and *PD* have higher outlier recovery rate than others for most of imputation methods, especially improving upon the others for marginal imputations. *Mreg2* performs strongly with *MD* and *GD*, while *Mreg1* is the most accurate in *SD* and *Dreg2* is the best in *PD* for both criteria.

Chem12.miss2 affects about 94 observations. Outlier 46 misses the 6th and 8th measurements, while outliers 7 and 110 have 1 variable missing each. Multiple regression methods are less accurate, with very low recovery rates. The reason, as mentioned before, is that the number of available observations is small relative to the number of parameters in the regression equation. Other regression methods perform well, however. In particular, *Dreg2* is the best for *PD*, and *Reg1* the best for *RMD*, under both criteria.

Table 3: Results for the Chem12 data set. Boldface indicates best in each row.

	α	Mean	Med	Rand	Reg1	Reg2	Dreg1	Dreg2	Mreg1	Mreg2	
OUTLIER RECOVERY PROBABILITY	0.1	MD	0.93	1.00	0.80	0.80	1.00	1.00	1.00	0.93	1.00
		RMD	0.60	0.20	0.27	0.80	0.73	0.67	0.80	0.87	0.87
		GD	0.53	0.27	0.33	0.73	0.60	0.80	0.80	0.73	0.80
		SD	0.53	0.27	0.40	0.87	0.80	0.80	0.80	0.93	0.87
		PD	0.93	1.00	0.80	0.93	1.00	0.80	1.00	0.83	0.87
	0.2	MD	1.00	0.87	0.67	0.87	1.00	1.00	0.93	0.13	0.00
		RMD	0.67	0.20	0.27	0.93	0.80	0.87	0.87	0.33	0.33
		GD	0.60	0.33	0.47	0.87	0.87	0.73	0.93	0.27	0.27
		SD	0.80	0.27	0.33	1.00	0.73	0.93	0.87	0.33	0.33
		PD	0.87	0.87	0.67	0.87	1.00	1.00	1.00	0.33	0.33
	0.3	MD	0.67	0.67	0.40	0.67	0.73	0.60	0.73	-	-
		RMD	0.73	0.07	0.20	0.93	0.53	0.80	0.60	-	-
		GD	0.67	0.47	0.40	0.80	0.80	0.67	0.80	-	-
		SD	0.60	0.13	0.13	0.93	0.67	0.80	0.67	-	-
		PD	0.67	0.67	0.40	0.67	0.73	0.60	0.80	-	-
RELATIVE ACCURACY MEASURE	0.1	MD	0.074	0.072	0.098	0.072	0.073	0.074	0.074	0.073	0.067
		RMD	0.097	0.120	0.100	0.095	0.084	0.095	0.100	0.093	0.085
		GD	0.552	0.481	0.539	0.503	0.524	0.546	0.413	0.431	0.354
		SD	0.124	0.171	0.131	0.124	0.118	0.121	0.126	0.112	0.113
		PD	0.111	0.126	0.153	0.107	0.196	0.158	0.107	0.199	0.185
	0.2	MD	0.107	0.136	0.117	0.104	0.125	0.117	0.105	0.117	0.112
		RMD	0.131	0.156	0.134	0.129	0.135	0.144	0.133	0.136	0.132
		GD	0.548	0.470	0.531	0.519	0.524	0.436	0.537	0.451	0.498
		SD	0.185	0.237	0.202	0.187	0.203	0.208	0.188	0.199	0.200
		PD	0.166	0.176	0.209	0.158	0.242	0.213	0.156	0.236	0.230
	0.3	MD	0.138	0.165	0.309	0.144	0.489	0.316	0.143	-	-
		RMD	0.197	0.211	0.359	0.184	0.516	0.355	0.192	-	-
		GD	0.661	0.734	0.845	0.671	0.810	0.795	0.757	-	-
		SD	0.306	0.354	0.453	0.299	0.584	0.445	0.307	-	-
		PD	0.211	0.212	0.365	0.204	0.526	0.372	0.194	-	-

Chem12.miss3 with about 30% of values missing only contains 2 complete observations. This makes it impossible to implement the multiple regression methods *Mreg1* and *Mreg2*, since there are too few available observations to estimate the regression coefficients. (This disadvantage of multiple regression methods also occurs below for the 30-dimensional data set Pre30, even with only 10% missing.) *Dreg2* and *Reg1* appear to be the best, while *Mean* also provides good results.

4.4 Results for Pre30

This 30-dimensional data set of size 220 contains 4 variables from Elect4 and 12 measurements from Chem12. The outliers 46, 7, 163, and 110 in Chem12 are also detected as outliers in Pre30 by most methods.

For pre30.miss1 with 10% missing data, only 10 observations have all 30 measurements. Half of these observations have 2 or 3 components missing. Outlier 46 contains 4 missing variables, but its most extreme measurement still remains. Outliers 7 and 163 have 3 measurements missing. *Reg1* and *Dreg2* tend to be more accurate than others by the relative accurate measure, but *Reg2* has higher outlier recovery probability. *PD* is consistently better than others by *ORP*. while *GD* appears to be inferior since we only consider the projections to the space spanning the first two principle components. The dimension is reduced too much, from 30 to 2, and as a result we lose too much information. We found that the generated PCA depth through projection onto the 4-dimensional eigenvector subspace improves results. However, the result is even worse for *GD* based on projection onto the 6-dimensional space than that in the 2-dimensional projection. This raises a practical issue when applying the GPCA depth: how to choose the lower dimension, especially with high-dimensional data.

There is no complete observation in pre30.miss2 generated by random removal of about 20% values. The *Reg1* and *Mean* imputations result in better RAMs. *PD* yields the strongest results in terms of outlier recovery probability. Using *PD*, each imputation method except *Rand* perfectly recovers the outliers of the original complete data set.

Table 4: Results for the Pre30 data set. Boldface indicates best in each row.

	α	Mean	Med	Rand	Reg1	Reg2	Dreg1	Dreg2	
OUTLIER RECOVERY PROBABILITY	0.1	MD	0.80	0.93	0.93	0.87	1.00	0.80	0.80
		RMD	0.67	0.40	0.40	0.80	0.80	0.60	0.53
		GD	0.33	0.33	0.40	0.67	0.53	0.47	0.40
		SD	0.80	0.40	0.47	0.80	0.80	0.67	0.67
		PD	1.00	0.93	0.73	0.87	1.00	1.00	1.00
	0.2	MD	0.53	0.67	0.60	0.40	0.53	0.43	0.37
		RMD	0.47	0.13	0.20	0.53	0.60	0.47	0.47
		GD	0.07	0.07	0.07	0.33	0.33	0.33	0.27
		SD	0.60	0.07	0.40	0.60	0.73	0.60	0.60
		PD	1.00	1.00	0.80	1.00	1.00	1.00	1.00
	0.3	MD	0.53	0.67	0.47	0.53	0.67	0.37	0.43
		RMD	0.47	0.07	0.07	0.40	0.20	0.33	0.40
		GD	0.33	0.13	0.07	0.33	0.27	0.33	0.33
		SD	0.60	0.07	0.13	0.53	0.33	0.40	0.53
		PD	0.80	0.67	0.53	0.73	0.80	0.73	0.87
RELATIVE ACCURACY MEASURE	0.1	MD	0.080	0.102	0.266	0.076	0.454	0.262	0.079
		RMD	0.111	0.118	0.287	0.105	0.464	0.289	0.099
		GD	0.518	0.607	0.696	0.482	0.713	0.682	0.505
		SD	0.150	0.176	0.321	0.146	0.489	0.320	0.138
		PD	0.111	0.125	0.292	0.103	0.467	0.287	0.108
	0.2	MD	0.103	0.118	0.282	0.100	0.465	0.280	0.103
		RMD	0.138	0.164	0.312	0.144	0.490	0.314	0.144
		GD	0.636	0.605	0.624	0.565	0.849	0.636	0.606
		SD	0.210	0.258	0.364	0.218	0.533	0.369	0.216
		PD	0.128	0.151	0.303	0.121	0.483	0.305	0.126
	0.3	MD	0.116	0.126	0.295	0.113	0.471	0.294	0.116
		RMD	0.178	0.200	0.346	0.181	0.516	0.338	0.175
		GD	0.614	0.632	0.670	0.626	0.791	0.690	0.702
		SD	0.285	0.332	0.431	0.293	0.587	0.427	0.288
		PD	0.174	0.186	0.345	0.167	0.508	0.339	0.166

With increase to 30% of values missing in pre30.miss3, the outlier recovery rate decreases dramatically. *Dreg2* and *Mean* yield better results than others. *PD* still yields the best performance.

5 Summary and practical recommendations

Combining the results for all four data sets, we summarize our findings and provide some practical recommendations.

5.1 Multiple imputation methods

The performance of imputation methods depends on the choice of outlier detection method, the size and dimension of the data, the fraction of data missing, and also the position of missing values in the data.

In general, *marginal imputations are outperformed by regression imputation methods*, as expected. *Mean*, *Median*, and *Rand* are less accurate than regression methods, since they perform only componentwise, ignoring relations with other variables. We should avoid using marginal imputations, except when the missing data has low ratio of size to dimension combined with high fraction of values missing. In the latter case, *Mean* or *Median* should be considered along with *simple regression*.

For low dimensional data, the general recommendation is *Mreg2*, which is most likely to impute values which lead to the detection of the true outliers in the data.

For moderate dimensional data with relatively large sample size and low missing fraction, *Mreg2* is still a good choice. Otherwise, *Dreg2* is superior to the others.

For high dimensional data with high fraction of values missing, *Reg1* is also strong. Furthermore, except for simple regression methods, regression imputation methods with random errors from the normal model seem to be better than those using observed errors, for most outlier identifiers. In other words, *Mreg2* is more accurate than *Mreg1*, and *Dreg2* performs better than *Dreg1*.

5.2 Outlier detection methods

Among outlier detection methods, MD is not suitable for the goal of finding outliers, since it is not robust, suffering from masking and swamping effects. In the Haemat7 data, we encounter this drawback of MD . Although RMD overcomes the influence of outliers, it has the limitation of elliptical symmetry. GD performs well in data with low or moderate dimensions. For high dimensional data, since there is no general guidance for choosing the right size of dimension reduction especially in the outlier detection context, GD may be limited and perform poorly. Both of the nonparametric depths, spatial and projection, yield strong results. PD has many good properties such as high breakdown point and affine invariance, but in practice its computation is approximate one and more expensive than others. SD is consistently good over all data sets and imputation methods. Overall, the spatial depth SD is a good measure both for detecting outliers and for comparing the multiple imputation methods. In practice, of course, a so-called *transformation-retransformation* version of SD is used, which achieves full affine invariance. Essentially, this is accomplished by standardization using a suitable robust covariance matrix. See Chakraborty, Chaudhuri, and Oja (1998), Chakraborty (2001), and Serfling (2009) for details.

5.3 Practical recommendations

Since the performance of a multiple imputation method depends on the choice of outlier detection method, it is preferred to choose the imputation and depth measure together.

If SD is used, then, for low dimensional data, $Mreg1$, $Dreg1$ or $Reg1$ are better than others, while for high dimensional data, $Reg1$ or $Mean$ are recommended.

If PD is used, then we should choose $Mreg2$ for multiple imputation in low dimensional data sets, and $Dreg2$ or $Reg1$ in a high dimensional data set.

If GD is used, then the best imputation methods are $Mreg2$ for a low fraction of missing values and $Dreg2$ for a high fraction.

If RMD is used, then it is advisable to choose $Mreg2$ and $Dreg2$ for low fraction missing,

and *Reg1* and *Mean* for high dimensional data sets with high fraction missing.

Finally, from an overall standpoint, the more favorable multiple imputation methods are *Reg1*, *Dreg2*, and *Mreg2* for *low dimensional* data sets. *Reg1* also gives promising results for *high dimensional* data sets, in our extended study.

Our study and that of Penny and Jolliffe (1999) agree that *Dreg2* performs very well. However, while their study concludes that multiple regression imputation methods are not very reliable at all, our more extensive study using depth-based performance criteria and several robust outlier identifiers finds that *Mreg2* does perform the best for low dimensional data. In general, we should choose the imputation method in companion with the outlier method, the size and dimension of data, and the fraction of values missing.

Acknowledgments

The authors highly value an anonymous reviewer's insightful comments and helpful suggestions, which led to significant improvements. We also thank Dr. Kay Penny for providing the data sets used here. Support under National Science Foundation Grants DMS-0103698, CCF-0430366, and DMS-0805786 is gratefully acknowledged.

References

- [1] Caussinus, H. and Ruiz, A. (1990). Interesting projections of multidimensional data by means of generalized principle component analysis. *Compstat* **90** 121–126.
- [2] Caussinus, H. and Ruiz, A. (1993). Projection pursuit and generalized principal component analyses. In *New Directions in Statistical Data Analysis and Robustness*, Morgenthaler *et al.* (Eds.), Birkhäuser Verlag, Berlin, 35–46.
- [3] Chakraborty, B. (2001). On affine equivariant multivariate quantiles. *Annals of the Institute of Statistical Mathematics* **53** 380–403.

- [4] Chakraborty, B., Chaudhuri, P., and Oja, H. (1998). Operating transformation and retransformation on spatial median and angle test. *Statistica Sinica* **8** 767–784.
- [5] Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association* **91** 862–872.
- [6] Chen, Y., Dang, X., Peng, H. and Bart, H. (2009). Outlier detection with the kernelized spatial depth function. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(2) 288–305.
- [7] Dang, X. and Serfling, R. (2009). Nonparametric multivariate outlier identifiers, and robustness properties. *Journal of Statistical Planning and Inference*, to appear.
- [8] Jolliffe, I. (2002). *Principal Component Analysis (2nd edition)*. Springer-verlag, New York.
- [9] Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data (2nd edition)*. Wiley, New York.
- [10] Pan, J., Fung, W. and Fang, K. (2000). Multiple outlier detection in multivariate data using projection pursuit techniques. *Journal of Statistical Planning and Inference* **83** 153–167.
- [11] Penny, K. and Jolliffe, I. (1999). Multivariate outlier detection applied to multiply imputed laboratory data. *Statistics in Medicine* **18** 1879–1895.
- [12] Rousseeuw, P. J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41** 212–223.
- [13] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.
- [14] Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91** 473–489.

- [15] Schafer, J (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.
- [16] Schafer, J. and Olsen, M. (1998). Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behavioral Research* **33** 545–571.
- [17] Serfling, R. (2002). A depth function and a scale curve based on spatial quantiles. In *Statistical Data Analysis Based on the L_1 -Norm and Related Methods* (Y. Dodge, ed.), 25–38. Birkhäuser.
- [18] Serfling, R. (2009). Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardization. Preprint.
- [19] Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* **16** 219–242.
- [20] Van Buuren, S., Brand, J., Groothuis-Oudshoorn, C. and Rubin, D. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* **76**(12) 1049–1064.
- [21] Vardi, Y. and Zhang, C. (2000). The multivariate L_1 -median and associated data depth. *Proceedings of the National Academy of Sciences* **97** 1423–1436.
- [22] Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Annals of Statistics* **28** 461–482.
- [23] Zuo, Y. (2003). Projection-based depth functions and associated medians. *Annals of Statistics* **31** 1460–1490.