BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 1 2012, pages 112–118
doi:10.1093/bioinformatics/btr597

Data and text mining

Advance Access publication October 28, 2011

MissForest—non-parametric missing value imputation for mixed-type data

Daniel J. Stekhoven[1,2,3,*] and Peter Bühlmann[1,3]

[1]Seminar for Statistics, Department of Mathematics, ETH Zurich, [2]Life Science Zurich PhD Program on Systems Biology of Complex Diseases and [3]Competence Center for Systems Physiology and Metabolic Diseases, Zurich, Switzerland

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Modern data acquisition based on high-throughput technology is often facing the problem of missing data. Algorithms commonly used in the analysis of such large-scale data often depend on a complete set. Missing value imputation offers a solution to this problem. However, the majority of available imputation methods are restricted to one type of variable only: continuous or categorical. For mixed-type data, the different types are usually handled separately. Therefore, these methods ignore possible relations between variable types. We propose a non-parametric method which can cope with different types of variables simultaneously.

Results: We compare several state of the art methods for the imputation of missing values. We propose and evaluate an iterative imputation method (missForest) based on a random forest. By averaging over many unpruned classification or regression trees, random forest intrinsically constitutes a multiple imputation scheme. Using the built-in out-of-bag error estimates of random forest, we are able to estimate the imputation error without the need of a test set. Evaluation is performed on multiple datasets coming from a diverse selection of biological fields with artificially introduced missing values ranging from 10% to 30%. We show that missForest can successfully handle missing values, particularly in datasets including different types of variables. In our comparative study, missForest outperforms other methods of imputation especially in data settings where complex interactions and non-linear relations are suspected. The out-of-bag imputation error estimates of missForest prove to be adequate in all settings. Additionally, missForest exhibits attractive computational efficiency and can cope with high-dimensional data.

Availability: The R package `missForest` is freely available from http://stat.ethz.ch/CRAN/.

Contact: stekhoven@stat.math.ethz.ch; buhlmann@stat.math.ethz.ch

Received on May 3, 2011; revised on September 27, 2011; accepted on October 24, 2011

1 INTRODUCTION

Imputation of missing values is often a crucial step in data analysis. Many established methods of analysis require fully observed datasets without any missing values. However, this is seldom the case in medical and biological research today. The ongoing development of new and enhanced measurement techniques in these fields provides data analysts with challenges prompted not only by high-dimensional multivariate data where the number of variables may greatly exceed the number of observations, but also by mixed data types where continuous and categorical variables are present. In our context, categorical variables can arise as any kind ranging from technical settings in a mass spectrometer to a diagnostic expert opinion on a disease state. Additionally, such datasets often contain complex interactions and non-linear relation structures which are notoriously hard to capture with parametric procedures.

Most prevalent imputation methods, like k nearest neighbours [KNNimpute, Troyanskaya *et al.* (2001)] for continuous data, saturated multinomial model (Schafer, 1997) for categorical data and multivariate imputation by chained equations [MICE, Van Buuren and Oudshoorn (1999)] for mixed data types depend on tuning parameters or specification of a parametric model. The choice of such tuning parameters or models without prior knowledge is difficult and might have a dramatic effect on a method's performance. Excluding MICE, the above methods and the majority of other imputation methods are restricted to one type of variable. Furthermore, all these methods make assumptions about the distribution of the data or subsets of the variables, leading to questionable situations, e.g. assuming normal distributions.

The literature on mixed-type data imputation is rather scarce. Its first appearance was in the developing field of multiple imputation brought up by Rubin (1978). Little and Schluchter (1985) presented an approach based on maximum likelihood estimation combining the multivariate normal model for continuous and the Poisson/multinomial model for categorical data. This idea was later on extended in the book of Little and Rubin (1987). See also Li (1988), Rubin and Schafer (1990) and Schafer (1997). A more refined method to combine different regression models for mixed-type data was proposed by Van Buuren and Oudshoorn (1999) using chained equations. The conditional model in MICE can be specified for the missing data in each incomplete variable. Therefore, no multivariate model covering the entire dataset has to be specified. However, it is assumed that such a full multivariate distribution exists and missing values are sampled from conditional distributions based on this full distribution (for more details see Section 3). Another similar method using variable-wise conditional distributions was proposed by Raghunathan *et al.* (2001) called sequential regression multivariate imputation. Unlike in MICE, the predictors must not be incomplete. The method is focussed on survey data and therefore includes strategies to incorporate restrictions on

*To whom correspondence should be addressed.

subsamples of individuals and logical bounds based on domain knowledge about the variables, e.g. only women can have a number of pregnancies recorded.

Our motivation is to introduce a method of imputation which can handle any type of input data and makes as few as possible assumptions about structural aspects of the data. Random forest [RF, Breiman (2001)] is able to deal with mixed-type data and as a non-parametric method it allows for interactive and non-linear (regression) effects. We address the missing data problem using an iterative imputation scheme by training an RF on observed values in a first step, followed by predicting the missing values and then proceeding iteratively. Mazumder *et al.* (2010) use a similar approach for the matrix completion problem using a soft-thresholded SVD iteratively replacing the missing values. We choose RF because it can handle mixed-type data and is known to perform very well under barren conditions like high dimensions, complex interactions and non-linear data structures. Due to its accuracy and robustness, RF is well suited for the use in applied research often harbouring such conditions. Furthermore, the RF algorithm allows for estimating out-of-bag (OOB) error rates without the need for a test set. For further details, see Breiman (2001).

Here we compare our method with k-nearest neighbour imputation [KNNimpute, Troyanskaya *et al.* (2001)] and the Missingness Pattern Alternating Lasso (MissPALasso) algorithm by Städler and Bühlmann (2010) on datasets having continuous variables only. For the cases of categorical and mixed type of variables, we compare our method with the MICE algorithm by Van Buuren and Oudshoorn (1999) and a dummy variable encoded KNNimpute. Comparisons are performed on several datasets coming from different fields of life sciences and using different proportions of missing values.

We show that our approach is competitive to or outperforms the compared methods on the used datasets irrespectively of the variable type composition, the data dimensionality, the source of the data or the amount of missing values. In some cases, the decrease of imputation error is up to 50%. This performance is typically reached within only a few iterations which makes our method also computationally attractive. The OOB imputation error estimates give a very good approximation of the true imputation error having on average a proportional deviation of no more than 10–15%. Furthermore, our approach needs no tuning parameter, and hence is easy to use and needs no prior knowledge about the data.

2 APPROACH

We assume $\mathbf{X}=(\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_p)$ to be a $n\times p$-dimensional data matrix. We propose using an RF to impute the missing values due to its earlier mentioned advantages as a regression method. The RF algorithm has a built-in routine to handle missing values by weighting the frequency of the observed values in a variable with the RF proximities after being trained on the initially mean imputed dataset (Breiman, 2001). However, this approach requires a complete response variable for training the forest.

Instead, we directly predict the missing values using an RF trained on the observed parts of the dataset. For an arbitrary variable $\mathbf{X}_s$ including missing values at entries $\mathbf{i}_{\text{mis}}^{(s)}\subseteq\{1,\ldots,n\}$ we can separate the dataset into four parts:

(1) The observed values of variable $\mathbf{X}_s$, denoted by $\mathbf{y}_{\text{obs}}^{(s)}$;

(2) the missing values of variable $\mathbf{X}_s$, denoted by $\mathbf{y}_{\text{mis}}^{(s)}$;

(3) the variables other than $\mathbf{X}_s$ with observations $\mathbf{i}_{\text{obs}}^{(s)}=\{1,\ldots,n\}\setminus\mathbf{i}_{\text{mis}}^{(s)}$ denoted by $\mathbf{x}_{\text{obs}}^{(s)}$; and

(4) the variables other than $\mathbf{X}_s$ with observations $\mathbf{i}_{\text{mis}}^{(s)}$ denoted by $\mathbf{x}_{\text{mis}}^{(s)}$.

Note that $\mathbf{x}_{\text{obs}}^{(s)}$ is typically not completely observed since the index $\mathbf{i}_{\text{obs}}^{(s)}$ corresponds to the observed values of the variable $\mathbf{X}_s$. Likewise, $\mathbf{x}_{\text{mis}}^{(s)}$ is typically not completely missing.

To begin, make an initial guess for the missing values in $\mathbf{X}$ using mean imputation or another imputation method. Then, sort the variables $\mathbf{X}_s, s=1,\ldots,p$ according to the amount of missing values starting with the lowest amount. For each variable $\mathbf{X}_s$, the missing values are imputed by first fitting an RF with response $\mathbf{y}_{\text{obs}}^{(s)}$ and predictors $\mathbf{x}_{\text{obs}}^{(s)}$; then, predicting the missing values $\mathbf{y}_{\text{mis}}^{(s)}$ by applying the trained RF to $\mathbf{x}_{\text{mis}}^{(s)}$. The imputation procedure is repeated until a stopping criterion is met. The pseudo Algorithm 1 gives a representation of the missForest method.

Algorithm 1 Impute missing values with RF.

Require: $\mathbf{X}$ an $n\times p$ matrix, stopping criterion γ

1. Make initial guess for missing values;
2. $\mathbf{k}\leftarrow$ vector of sorted indices of columns in $\mathbf{X}$ w.r.t. increasing amount of missing values;
3. **while** not γ **do**
4. $\quad\mathbf{X}_{\text{old}}^{\text{imp}}\leftarrow$ store previously imputed matrix;
5. $\quad$**for** s in $\mathbf{k}$ **do**
6. $\quad\quad$Fit a random forest: $\mathbf{y}_{\text{obs}}^{(s)}\sim\mathbf{x}_{\text{obs}}^{(s)}$;
7. $\quad\quad$Predict $\mathbf{y}_{\text{mis}}^{(s)}$ using $\mathbf{x}_{\text{mis}}^{(s)}$;
8. $\quad\quad\mathbf{X}_{\text{new}}^{\text{imp}}\leftarrow$ update imputed matrix, using predicted $\mathbf{y}_{\text{mis}}^{(s)}$;
9. $\quad$**end for**
10. $\quad$update γ.
11. **end while**
12. **return** the imputed matrix $\mathbf{X}^{\text{imp}}$

The stopping criterion γ is met as soon as the difference between the newly imputed data matrix and the previous one increases for the first time with respect to both variable types, if present. Here, the difference for the set of continuous variables $\mathbf{N}$ is defined as

$$\Delta_N=\frac{\sum_{j\in\mathbf{N}}(\mathbf{X}_{\text{new}}^{\text{imp}}-\mathbf{X}_{\text{old}}^{\text{imp}})^2}{\sum_{j\in\mathbf{N}}(\mathbf{X}_{\text{new}}^{\text{imp}})^2},$$

and for the set of categorical variables $\mathbf{F}$ as

$$\Delta_F=\frac{\sum_{j\in\mathbf{F}}\sum_{i=1}^{n}\mathbf{I}_{\mathbf{X}_{\text{new}}^{\text{imp}}\neq\mathbf{X}_{\text{old}}^{\text{imp}}}}{\#\text{NA}},$$

where #NA is the number of missing values in the categorical variables.

After imputing the missing values, the performance is assessed using the normalized root mean squared error [NRMSE, Oba *et al.* (2003)] for the continuous variables which is defined by

$$\text{NRMSE}=\sqrt{\frac{\text{mean}\left((\mathbf{X}^{\text{true}}-\mathbf{X}^{\text{imp}})^2\right)}{\text{var}\left(\mathbf{X}^{\text{true}}\right)}},$$

where $\mathbf{X}^{\text{true}}$ is the complete data matrix and $\mathbf{X}^{\text{imp}}$ the imputed data matrix. We use mean and var as short notation for empirical mean and variance computed over the continuous missing values only. For categorical variables, we use the proportion of falsely classified entries (PFC) over the categorical missing values, Δ_F. In both cases, good performance leads to a value close to 0 and bad performance to a value around 1.

When an RF is fit to the observed part of a variable, we also get an OOB error estimate for that variable. After the stopping criterion γ was met, we average over the set of variables of the same type to approximate the true imputation errors. We assess the performance of this estimation by comparing the absolute difference between true imputation error and OOB imputation error estimate in all simulation runs.

3 METHODS

We compare missForest with four methods on 10 different datasets where we distinguish among situations with continuous variables only, categorical variables only and mixed variable types.

The most well-known method for imputation of continuous datasets especially in the field of gene expression analysis is the KNNimpute algorithm by Troyanskaya *et al.* (2001). A missing value variable $\mathbf{X}_j$ is imputed by finding its k nearest observed variables and taking a weighted mean of these k variables for imputation. Thereby, the weights depend on the distance of the variable $\mathbf{X}_j$. The distance itself is usually chosen to be the Euclidean distance.

When using KNNimpute the choice of the tuning parameter k can have a large effect on the performance of the imputation. However, this parameter is not known beforehand. Since our method includes no such parameter, we implement a cross-validation (Algorithm 2) to obtain a suitable k.

Algorithm 2 Cross-validation KNN imputation.

Require: $\mathbf{X}$ an $n \times p$ matrix, number of validation sets l, range of suitable number of nearest neighbours $\mathbf{K}$

1. $\mathbf{X}^{\text{CV}} \leftarrow$ initial imputation using mean imputation;
2. **for** t in $1, \ldots, l$ **do**
3. $\quad \mathbf{X}^{\text{CV}}_{\text{mis},t} \leftarrow$ artificially introduce missing values to $\mathbf{X}^{\text{CV}}$;
4. $\quad$ **for** k in $\mathbf{K}$ **do**
5. $\quad\quad \mathbf{X}^{\text{CV}}_{\text{KNN},t} \leftarrow$ KNN imputation of $\mathbf{X}^{\text{CV}}_{\text{mis},t}$ using k nearest neighbours;
6. $\quad\quad \varepsilon_{k,t} \leftarrow$ error of KNN imputation for k and t;
7. $\quad$ **end for**
8. **end for**
9. $k_{\text{best}} \leftarrow \underset{k}{\text{argmin}} \frac{1}{l} \sum_{t=1}^{l} \varepsilon_{k,t}$;
10. $\mathbf{X}^{\text{imp}} \leftarrow$ KNN imputation of $\mathbf{X}$ using k_{best} nearest neighbours.

In the original paper of Troyanskaya *et al.* (2001), the data were not standardized before applying the KNNimpute algorithm. This constitutes no issue in the case of gene expression data, because such data generally consist of variables on similar scales. However, we are applying the KNNimpute algorithm to datasets with varying scales in the variables. To avoid variance-based weighting of the variables, we scale them to a unit SD. We also centre the variables at zero. After imputation, the data are retransformed such that the error is computed on the original scales. This last step is performed because missForest does not need any transformation of the data and we want to compare the performance of the methods on the original scales of the data.

Another approach for continuous data, especially in the case of high-dimensional normal data matrices, is presented by Städler and Bühlmann (2010) using an EM-type algorithm. In their Missingness Pattern Alternating Imputation and l_1-penalty (MissPALasso) algorithm, the missing variables are regressed on the observed ones using the lasso penalty by Tibshirani (1996). In the following E step, the obtained regression coefficients are used to partially update the latent distribution. The MissPALasso has also a tuning parameter λ for the penalty. As with KNNimpute, we use cross-validation to tune λ (cf. Algorithm 2). When applying MissPALasso, the data are standardized as regularization with a single λ requires the different regressions to be on the same scale.

In the comparative experiments with categorical or mixed-type variables, we use the MICE algorithm by Van Buuren and Oudshoorn (1999) based on the multivariate multiple imputation scheme of Schafer (1997). In contrast to the latter, the conditional distribution for the missing data in each incomplete variable is specified in MICE, a feature called fully conditional specification by Van Buuren (2007). However, the existence of a multivariate distribution from which the conditional distribution can be easily derived is assumed. Furthermore, iterative Gibbs sampling from the conditional distributions can generate draws from the multivariate distribution. We want to point out that MICE in its default setup is not mainly intended for simple missing value imputation. Using the multiple imputation scheme, MICE allows for assessing the uncertainty of the imputed values. It includes features to pool multiple imputations, choose individual sampling procedures and allows for passive imputation controlling the sync of transformed variables. In our experiments, we used MICE with either linear regression with normal errors or mean imputation for continuous variables, logistic regression for binary variables and polytomous logistic regression for categorical variables with more than two categories.

For comparison across different types of variables, we apply the KNNimpute algorithm with dummy coding for the categorical variables. This is done by coding a categorical variable $\mathbf{X}_j$ into m dichotomous variables $\tilde{\mathbf{X}}_{j,m} \in \{-1, 1\}$. Application of the KNNimpute algorithm for categorical data can be summarized as:

(1) Code all categorical variables into $\{-1, 1\}$-dummy variables;
(2) standardize all variables to mean 0 and SD 1;
(3) apply the cross-validated KNNimpute method from Algorithm 2;
(4) retransform the imputed data matrix to the original scales;
(5) code the dummy variables back to categorical variables; and
(6) computed the imputation error.

For each experiment, we perform 50 independent simulations where 10, 20 or 30% of the values are removed completely at random. Each method is then applied and the NRMSE, the PFC or both are computed (Section 2). We perform a paired Wilcoxon test of the error rates of the compared methods versus the error rates of missForest. In addition, the OOB error estimates of missForest is recorded in each simulation.

4 RESULTS

4.1 Continuous variables only

First, we focus on continuous data. We investigate the following four publicly available datasets:

- Isoprenoid gene network in *Arabidopsis thaliana*: this gene network includes $P = 39$ genes each with $n = 118$ gene expression profiles corresponding to different experimental conditions. For more details on this dataset, see Wille *et al.* (2004).
- Voice measures in Parkinson's patients: the data described by Little *et al.* (2008) contains a range of biomedical voice measurements from 31 individuals, 23 with Parkinson's disease (PD). There are $P = 22$ particular voice measurements and $n = 195$ voice recordings from these individuals. The dataset

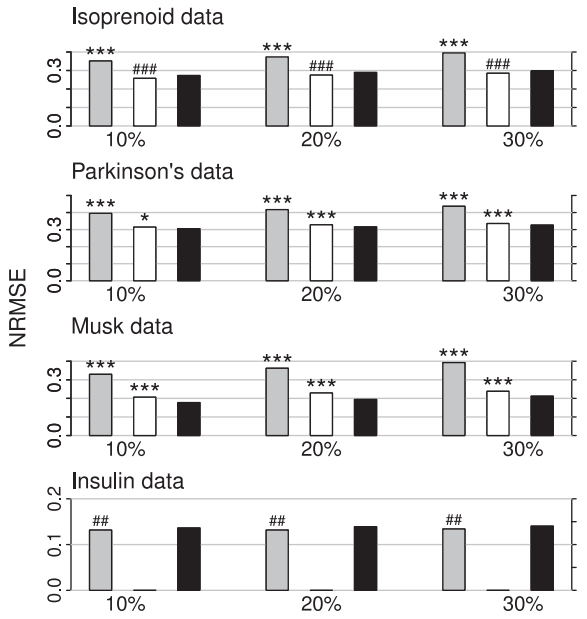

Fig. 1. Continuous data. Average NRMSE for KNNimpute (grey), MissPALasso (white) and missForest (black) on four different datasets and three different amounts of missing values, i.e. 10, 20 and 30%. Standard errors are in the order of magnitude of 10^{-4}. Significance levels for the paired Wilcoxon tests in favour of missForest are encoded as '*' <0.05, '**' <0.01 and '***' <0.001. If the average error of the compared method is smaller than that of missForest, the significance level is encoded by a hash (#) instead of an asterisk. In the lowermost dataset, results for MissPALasso are missing due to the implementations limited capability with regard to high dimensions.

also contains a response variable giving the health status. Dealing only with continuous variables, the response was removed from the data. We will return to this later on.

- Shapes of musk molecules: this dataset describes 92 molecules of which 47 are musks and 45 are non-musks. For each molecule $P=166$ features describe its conformation, but since a molecule can have many conformations due to rotating bonds, there are $n=476$ different low-energy conformations in the set. The classification into musk and non-musk molecules is removed.
- Insulin gene expression: this high-dimensional dataset originates from an analysis by Wu *et al.* (2007) of *vastus lateralis* muscle biopsies from three different types of patients following insulin treatment. The three types are insulin-sensitive, insulin-resistant and diabetic patients. The analysis involves $P=12'626$ genes whose expression levels were measured from $n=110$ muscle biopsies. Due to computation time we only perform 10 simulations instead of 50.

Results are given in Figure 1. We can see that missForest performs well, sometimes reducing the average NRMSE by up to 25% with respect to KNNimpute. In case of the musk molecules data, the reduction is even >50%. The MissPALasso performs slightly better than missForest on the gene expression data. However, there are no results for the MissPALasso in case of the Insulin dataset, because the high dimension makes computation not feasible.

For continuous data, the missForest algorithm typically reaches the stopping criterion quite fast needing about five iterations. The imputation takes ∼10 times as long as performing the cross-validated KNNimpute where $\{1,\ldots,15\}$ is the set of possible numbers of neighbours. For the Insulin dataset, an imputation takes on average 2 h on a customary available desktop computer.

4.2 Categorical variables only

We also consider datasets with only categorical variables. Here, we use the MICE algorithm described in Section 3 instead of the MissPALasso. We use a dummy implementation of the KNNimpute algorithm to deal with categorical variables (Section 3). We apply the methods to the following datasets:

- Cardiac single photon emission computed tomography (SPECT) images: Kurgan *et al.* (2001) discuss this processed dataset summarizing over 3000 2D SPECT images from $n=267$ patients in $P=22$ binary feature patterns.
- Promoter gene sequences in *Escherichia coli*: the dataset contains sequences found by Harley and Reynolds (1987) for promoters and sequences found by Towell *et al.* (1990) for non-promoters totalling $n=106$. For each candidate, a sequence of 57 bp was recorded. Each variable can take one of four DNA nucleotides, i.e. adenine, thymine, guanine or cytosine. Another variable distinguishes between promoter and non-promoter instances.
- Lymphography domain data: the observations were obtained from patients suffering from cancer in the lymphatic of the immune system. For each of the $n=148$ lymphoma, $P=19$ different properties were recorded mainly in a nominal fashion. There are nine binary variables. The rest of the variables have three or more levels.

In Figure 2, we can see that missForest is always imputing the missing values better than the compared methods. In some cases, namely for the SPECT data, the decrease of PFC compared with MICE is up to 60%. However, for the other datasets the decrease is less pronounced ranging around 10–20%, but there still is a decrease. The amount of missing values on the other hand seems to have only a minor influence on the performance of all methods. Except for MICE on the SPECT data, error rates remain almost constant increasing only by 1–2%. We pointed out earlier that MICE is not primarily tailored for imputation performance, but offers additional possibilities of assessing uncertainty of the imputed values due to the multiple imputation scheme. Anyhow, the results using the cross-validated KNNimpute (Algorithm 2) on the dummy-coded categorical variables is surprising. The imputation for missForest needs on average five times as long as a cross-validated imputation using KNNimpute.

4.3 Mixed-type variables

In the following, we investigate four datasets where the first one has already been introduced, i.e. *musk molecules* data including the categorical response yielding the classification. The other datasets are as follows:

- Proteomics biomarkers for Gaucher's disease: Gaucher's disease is a rare inherited enzyme deficiency. In this dataset, Smit *et al.* (2007) present protein arrays for biomarkers

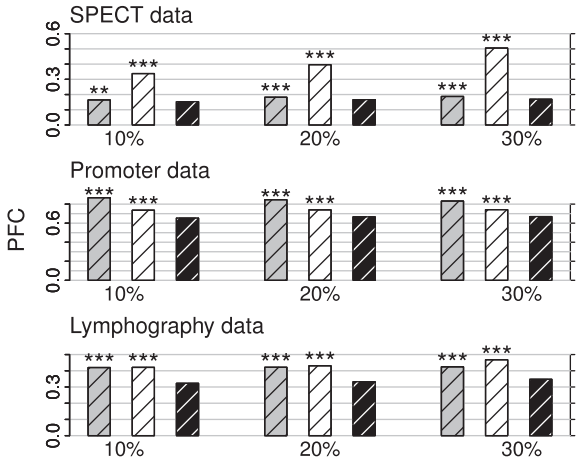

Fig. 2. Categorical data. Average PFC for cross-validated KNNimpute (grey), MICE (white) and missForest (black) on three different datasets and three different amounts of missing values, i.e. 10, 20 and 30%. Standard errors are in the order of magnitude of 10^{-4}. Significance levels for the paired Wilcoxon tests in favour of missForest are encoded as '*' <0.05, '**' <0.01 and '***' <0.001.

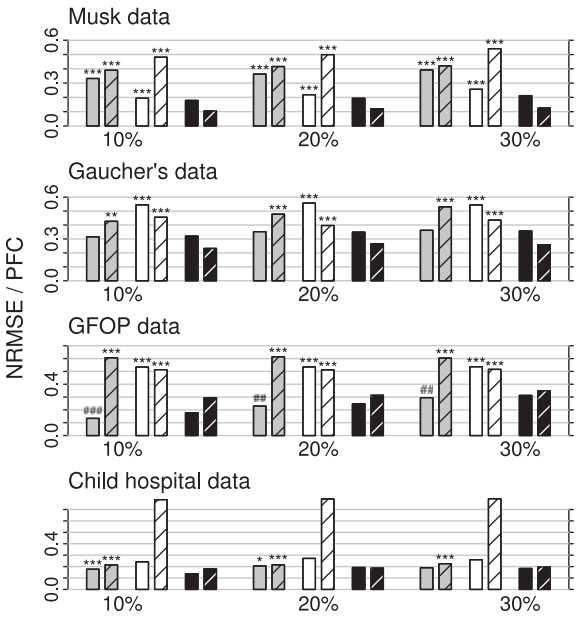

Fig. 3. Mixed-type data. Average NRMSE (left bar) and PFC (right bar, shaded) for KNNimpute (grey), MICE (white) and missForest (black) on four different datasets and three different amounts of missing values, i.e. 10, 20 and 30%. Standard errors are in the order of magnitude of 10^{-3}. Significance levels for the paired Wilcoxon tests in favour of missForest are encoded as '*' <0.05, '**' <0.01 and '***' <0.001. If the average error of the compared method is smaller than that of missForest, the significance level is encoded by a hash (#) instead of an asterisk. Note that, due to ill-distribution and near dependence in the Child hospital data, the results for MICE have to be treated with caution (Section 4.3).

($P = 590$) from blood serum samples ($n = 40$). The binary response distinguishes between disease status.

- Gene finding over prediction (GFOP) peptide search: this dataset comprises mass-spectrometric measurements of $n = 595$ peptides from two shotgun proteomics experiments on the nematode *Caenorhabditis elegans*. The collection of $P = 18$ biological, technical and analytical variables had the aim of novel peptide detection in a search on an extended database using established gene prediction methods.
- Children's Hospital data: this dataset is the product of a systematic long-term review of children with congenital heart defects after open heart surgery. Next to defect- and surgery-related variables, also long-term psychological adjustment and health-related quality of life was assessed. After removing observations with missing values, the dataset consists of $n = 55$ patients and $P = 124$ variables of which 48 are continuous and 76 are categorical. For further details see Latal *et al.* (2009).

The results of this comparison are given in Figure 3. We can see that missForest performs better than the other two methods, again reducing imputation error in many cases by >50%. For the GFOP data, KNNimpute has a slightly smaller NRMSE than missForest but makes twice as much error on the categorical variables. Generally, with respect to the amount of missing values the NRMSE tends to have a greater variability than the PFC which remains largely the same.

The imputation results for MICE on the Children's Hospital data have to be treated cautiously. Since this dataset contains ill-distributed and nearly dependent variables, e.g. binary variables with very few observations in one category, the missingness pattern has a direct influence on the operability of the MICE implementation in the statistical software R. The imputation error illustrated in Figure 3 was computed from 50 successful simulations by randomly generating missingness patterns, which did not include only complete cases or no complete cases at all within the categories of the variables. Therefore, the actual numbers of simulations were >50 for all three missing value amounts. Furthermore, nearly dependent variables were removed after each introduction of missing values. This leads to an average of seven removed variables in each simulation. Due to this *ad hoc* manipulation for making the MICE implementation work, we do not report significance statements for the imputation error.

4.4 Estimating imputation error

In each experiment, we get for each simulation run an OOB estimate for the imputation error. In Figure 4 the differences of true imputation error, $\mathrm{err}_{\mathrm{true}}$, and OOB error estimates, $\widehat{\mathrm{err}}_{\mathrm{OOB}}$, are illustrated for the continuous and the categorical datasets. Also, the mean of the true imputation error and the OOB error estimate over all simulations is depicted.

We can see that for the Isoprenoid and Musk datasets, the OOB estimates are very accurate only differing from the true imputation error by a few percents. In the case of Parkinson's dataset, the OOB estimates exhibit a lot more variability than in all other datasets. However, on average the estimation is comparably good. For the categorical datasets, the estimation accuracy behaves similarly over all scenarios. The OOB estimates tend to underestimate the imputation error with increasing amount of missing values. Apparently, the absolute size of the imputation error seems to play a minor role in the accuracy of the OOB estimates, which can be seen nicely when comparing the SPECT and the Promoter data.

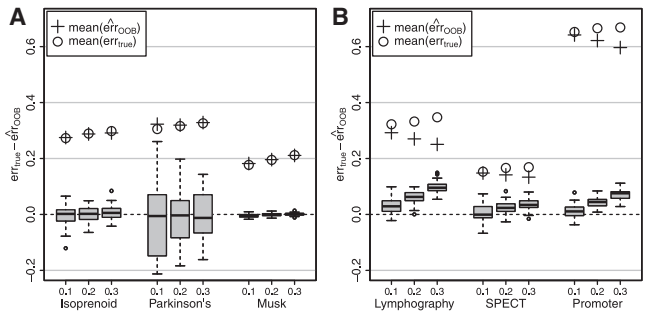

Fig. 4. Difference of true imputation error $\mathrm{err}_{\mathrm{true}}$ and OOB imputation error estimate $\widehat{\mathrm{err}}_{\mathrm{OOB}}$ for the continuous datasets (**A**) and the categorical datasets (**B**) and three different amounts of missing values, i.e. 0.1, 0.2 and 0.3. In each case, the average $\mathrm{err}_{\mathrm{true}}$ (circle) and the average $\widehat{\mathrm{err}}_{\mathrm{OOB}}$ (plus) over all simulations is given.

Table 1. Average runtimes (in seconds) for imputing the analysed datasets

Dataset	n	P	KNN	MissPALasso	MICE	missForest
Isoprenoid	118	39	0.8	170	–	5.8
Parkinson's	195	22	0.7	120	–	6.1
Musk (cont.)	476	166	13	1400	–	250
Insulin	110	12626	1800	NA	–	6200
SPECT	267	22	1.3	–	37	5.5
Promoter	106	57	14	–	4400	38
Lymphography	148	19	1.1	–	93	7.0
Musk (mixed)	476	167	27	–	2800	500
Gaucher's	40	590	1.3	–	130	29
GFOP	595	18	2.7	–	1400	40
Children	55	124	2.7	–	4000	110

Runtimes are averaged over the amount of missing values since this has a negligible effect on computing time. NA, not available.

4.5 Computational efficiency

We assess the computational cost of missForest by comparing the runtimes of imputation on the previous datasets. Table 1 shows the runtimes in seconds of all methods on the analysed datasets. We can see that KNNimpute is by far the fastest method. However, missForest runs considerably faster than MICE and the MissPALasso. In addition, applying missForest did not require antecedent standardization of the data, laborious dummy coding of categorical variables nor implementation of CV choices for tuning parameters.

There are two possible ways to speed up computation. The first one is to reduce the number of trees grown in each forest. In all comparative studies, the number of trees was set to 100 which offers high precision but increased runtime. In Table 2, we can see that changing the number of trees in the forest has a stagnating influence on imputation error, but a strong influence on computation time which is approximately linear in the number of trees.

The second one is to reduce the number of variables randomly selected at each node (m_{try}) to set up the split. Table 2 shows that increasing m_{try} has limited effect on imputation error, but computation time is strongly increased. Note that for $m_{\mathrm{try}} = 1$ we no longer have an RF, since there is no more choice between variables to split on. This leads to a much higher imputation error, especially for the cases with low numbers of bootstrapped trees. We use for all experiments $\lfloor\sqrt{p}\rfloor$ as default value, e.g. in the GFOP data this equals 4.

Table 2. Average imputation error (NRMSE/PFC in percent) and runtime (in seconds) with different numbers of trees (n_{tree}) grown in each forest and variables tried (m_{try}) at each node of the trees

m_{try}	n_{tree}				
	10	50	100	250	500
1	36.8/35.5 2.5 s	27.4/32.3 3.2 s	20.4/31.3 3.9 s	17.2/30.0 5.8 s	16.0/30.8 9.2 s
2	34.9/31.8 6.9 s	24.8/29.2 11.8 s	18.3/28.8 15.0 s	16.0/28.6 25.2 s	15.5/29.1 39.3 s
4	34.9/31.3 16.5 s	24.4/28.9 25.1 s	17.9/28.2 35.0 s	15.4/28.2 49.0 s	15.8/28.7 83.3 s
8	34.7/31.4 39.2 s	24.3/28.9 57.4 s	18.1/27.8 84.4 s	15.2/27.8 130.2 s	15.7/28.6 190.8 s
16	34.6/30.9 68.7 s	24.3/28.7 99.7 s	18.1/28.0 172.2 s	15.4/27.8 237.6 s	15.6/28.5 400.7 s

Here, we consider the GFOP dataset with artificially introduced 10% of missing values. For each comparison, 50 simulation runs were performed using always the same missing value matrix for all number of trees/randomly selected variables for a single simulation.

5 CONCLUSION

Our new algorithm, missForest, allows for missing value imputation on basically any kind of data. In particular, it can handle multivariate data consisting of continuous and categorical variables simultaneously. MissForest has no need for tuning parameters nor does it require assumptions about distributional aspects of the data. We show on several real datasets coming from different biological and medical fields that missForest outperforms established imputation methods like k-nearest neighbours imputation or multivariate imputation using chained equations. Using our OOB imputation error estimates, missForest offers a way to assess the quality of an imputation without the need of setting aside test data nor performing laborious cross-validations. For subsequent analysis, these error estimates represent a mean of informal reliability check for each variable. The full potential of missForest is deployed when the data include complex interactions or non-linear relations between variables of unequal scales and different type. Furthermore, missForest can be applied to high-dimensional datasets where the number of variables may greatly exceed the number of observations to a large extent and still provides excellent imputation results.

ACKNOWLEDGEMENTS

Except for the Isoprenoid, the Lymphography, the Children's Hospital and the GFOP data all other datasets were obtained from the UCI machine learning repository (Frank and Asuncion, 2010). The GFOP dataset was obtained from the Institute of Molecular Systems Biology, Zurich, Switzerland. Thanks to L. Reiter for

providing the data. The Lymphography dataset was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Slovenia. Thanks to M. Zwitter and M. Soklic for providing the data. The Children's Hospital dataset was obtained from the Child Development Center at the University Children's Hospital, Zürich, Switzerland. Thanks to B. Latal and I. Beck for providing the data. Finally, we thank two anonymous referees for their constructive comments.

Funding: The work was partly financed with a grant of the Swiss SystemsX.ch Initiative to the project LiverX of the Competence Center for Systems Physiology and Metabolic Diseases. The LiverX project was evaluated by the Swiss National Science Foundation.

Conflict of interest: none declared.

REFERENCES

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Frank,A. and Asuncion,A. (2010) UCI machine learning repository. Available at http://archive.ics.uci.edu/ml.

Harley,C.B. and Reynolds,R.P. (1987) Analysis of e. coli promoter sequences. *Nucleic Acids Res.*, **15**, 2343–2361.

Kurgan,,L. *et al.* (2001) Knowledge discovery approach to automated cardiac SPECT diagnosis. *Artif. Intell. Med.*, **23**, 149–169.

Latal,B. *et al.* (2009) Psychological adjustment and quality of life in children and adolescents following open-heart surgery for congenital heart disease: a systematic review. *BMC Pediatr.*, **9**, 6.

Li,K. (1988) Imputation using Markov chains. *J. Stat. Comput. Simul.*, **30**, 57–79.

Little,M. *et al.* (2008) Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *Nature Precedings*. Available at http://hdl.handle.net/10101/npre.2008.2298.1.

Little,R. and Rubin,D. (1987) *Statistical Analysis with Missing Data*. Wiley, New York.

Little,R. and Schluchter,M. (1985) Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, **72**, 497–512.

Mazumder,R. *et al.* (2010) Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.*, **11**, 2287–2322.

Oba,S. *et al.* (2003) A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, **19**, 2088–2096.

Raghunathan,T. *et al.* (2001) A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv. Methodol.*, **27**, 85–96.

Rubin,D. (1978) Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, American Statistical Association, pp. 20–34.

Rubin,D. and Schafer,J. (1990) Efficiently creating multiple imputations for incomplete multivariate normal data. In *Proceedings of the Statistical Computing Section of the American Statistical Association*, American Statistical Association, pp. 83–88.

Schafer,J. (1997) *Analysis of Incomplete Multivariate Data*. Chapman & Hall, UK.

Smit,S. *et al.* (2007) Assessing the statistical validity of proteomics based biomarkers. *Anal. Chim. Acta*, **592**, 210–217.

Städler,N. and Bühlmann,P. (2010) Pattern alternating maximization algorithm for high-dimensional missing data. *Arxiv preprint arXiv:1005.0366*.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* , **58**, 267–288.

Towell,G. *et al.* (1990) Refinement of approximate domain theories by knowledge-based neural networks. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, AAAI press, pp. 861–866.

Troyanskaya,O. *et al.* (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Van Buuren,S. (2007) Multiple imputation of discrete and continuous data by fully conditional specification. *Stat. Methods Med. Res.*, **16**, 219–242.

Van Buuren,S. and Oudshoorn,K. (1999) *Flexible Multivariate Imputation by MICE*. TNO Prevention Center, Leiden, The Netherlands.

Wille,A. *et al.* (2004) Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biol.*, **5**, R92.

Wu,X. *et al.* (2007) The effect of insulin on expression of genes and biochemical pathways in human skeletal muscle. *Endocrine*, **31**, 5–17.