Methodology for multiple imputation for missing data in electronic health record data

International Biometric Conference 2014

Jonathan Bartlett
www.missingdata.org.uk

London School of Hygiene and Tropical Medicine

Acknowledgements

- I am grateful to the UK Medical Research Council for support (MR/K02180X/1).
- Thanks also to Anoop Shah (University College London), who developed one of the random forest imputation techniques [1].
- I have benefited greatly from the excellent book ‘The Elements of Statistical Learning’, by Hastie, Tibshirani and Freidman [2].
- And thank you for the invitation to speak!

Outline

Missing data in electronic health record databases

Trees and random forest for prediction

Trees and random forest for multiple imputation

Simulations

Conclusions

Outline

Missing data in electronic health record databases

Trees and random forest for prediction

Trees and random forest for multiple imputation

Simulations

Conclusions

Electronic health record databases

- Electronic health record databases are increasingly being made available for conducting health research.
- They offer a number of advantages over performing and analysing traditional studies:
  - The obvious one: the data have already been collected (saving time and money)
  - Difficult to study sub-populations or rare outcomes can be examined
  - Large (sometimes very) sample sizes are available
  - They enable assessment of associations and effects in real clinical practice, as opposed to the often less realistic environment of designed studies.
- However, with these come a number of challenges, and a key one is that of missing data.

Multiple imputation for missing data in electronic health record databases

- An obvious approach to consider for tackling missing data in this setting is multiple imputation (MI), which is usually performed using parametric models.
- Ideally we would follow Rubin's original paradigm: the controllers of a database multiply impute missing data, and release the imputed datasets to analysts.
- The problem (in short): if the imputation model is misspecified, analysts may obtain biased estimates, and invalid inferences.
- Moreover, if the imputation model is uncongenial/incompatible with an analyst's model, the analyst may obtain biased estimates [3, 4].
- e.g. the analyst fits a non-linear effect which the imputer assumed was linear.

Nonparametric imputation models

- The obvious solution is to impute using a nonparametric approach.
- e.g. hot-deck imputation / nearest neighbour techniques [5].
- The problem is that we suffer from the curse of dimensionality - as the number of variables increases the nearest neighbours are not very near [2].
- This is particularly acute in electronic health databases where we have a large number of variables.

Imputation using machine learning techniques

- Suggestions have been made (as far back as 1996 [6]) that machine learning methods, such as regression trees, might be used for imputation.
- These techniques relax the strong assumptions of parametric models, and so potentially would be very useful for MI.
- In the last few years some papers have taken up this idea [7, 8, 9, 1].
- These methods are not truly nonparametric - they make certain assumptions, although these (as far as I can see) are often not explicitly stated or even understood yet.
- In the following, I will describe some of the recent proposals for using tree based methods for multiply imputing missing data, and investigate their performance in simulations.

Outline

Missing data in electronic health record databases

Trees and random forest for prediction

Trees and random forest for multiple imputation

Simulations

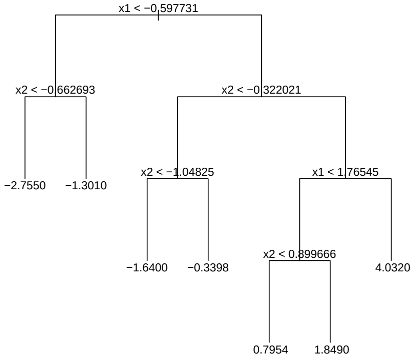
Conclusions

Trees and random forest

- For the moment we leave the issue of missing data aside.
- Our aim is to predict Y using predictors $X_1, .., X_p$.
- I will describe techniques for continuous Y, but the techniques can be adapted for categorical Y.
- I will focus on the random forest technique, proposed by Leo Breiman and Adele Cutler [10].
- Random forest is based on regression/classification trees, which I therefore briefly review.

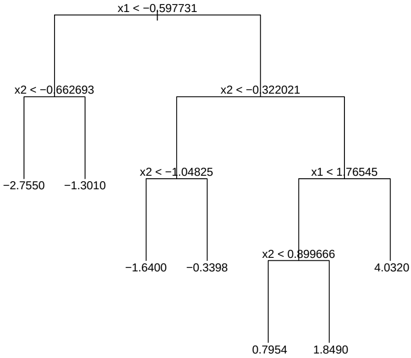
Regression/classification trees

- To start, we consider all predictors X_j, $j = 1, .., p$, and all cut-points s.
- We split the data using the predictor j and cut-point s which reduces the total squared prediction error by the largest amount.
- The data are then divided into two branches, into those with $X_j \leq s$ and $X_j > s$.



Regression/classification trees

- Within each branch, we then repeat the process iteratively until each terminal node is less than or equal to a given size (e.g. 5).
- The predicted value of Y for a particular combination of predictor values is taken as the mean of the corresponding terminal node.



Strengths and limitations of trees

- Trees are capable of automatically capturing complex relationships between variables (e.g. interactions and non-linearities).
- They should therefore have relatively little bias in terms of predicting $E(Y|X_1,..,X_p)$.
- However they are noisy / highly variable.
- To improve stability, random forest performs 'bagging':
  - take k bootstrap samples of the data
  - grow a tree on each bootstrapped dataset
  - for input values $x_1,..,x_p$, average predictions from the the k trees to form prediction
- This averaging process reduces the variance of predictions, without affecting bias.

De-correlating predictions from bootstrapped trees

- Random forest also modifies the tree growing process in order to attempt to reduce correlation between predictions from bootstrapped trees.
- At each node, rather than considering all predictors for splitting, a random $m \leq p$ predictors are considered.
- m is chosen in some way. For continous Y, the default is $p/3$.

Statistical properties

- Random forest doesn't start with an explicit statistical model or even explicit assumptions, and so deducing its statistical properties is difficult.
- A number of papers have derived results for algorithms which are modifications of random forest in order to tackle the problem [11, 12].
- The bootstrapping and covariate selection reduce variance, although the covariate selection may induce bias [2].
- Despite the lack of exact formal results, Hastie *et al* state that it often performs remarkably well [2].

Outline

Missing data in electronic health record databases

Trees and random forest for prediction

Trees and random forest for multiple imputation

Simulations

Conclusions

Using trees for imputation

- Hastie *et al* first proposed that classification/regression trees may be useful for imputing missing data [2].
- Specifically, suppose we want to impute missing values in Y using $X_1, .., X_p$ (which for now we assume are fully observed).
- Let y^{obs} and y^{mis} denote the observed and missing values in Y, and let x^{obs} denote the predictor values corresponding to y^{obs}.
  1. grow a tree for predicting Y from $X_1, .., X_p$, using y^{obs}, x^{obs}
  2. for a subject who is missing Y, find their terminal node based on their values of $X_1, .., X_p$
  3. impute the missing Y using a random sample from the observed Ys in the given terminal node

MICE using trees for imputation

- Burgette and Reiter then proposed that this be embedded within the chained equations (MICE) / full conditional specification (FCS) technique [7].
- This enables missing data in multiple variables to be imputed.
- Burgette and Reiter used the Bayesian bootstrap within the terminal node before sampling.
- However, their approach does not appear to incorporate uncertainty about the node which a given set of predictor values leads to.

Simulation results from Burgette and Reiter [7]

- Burgette and Reiter performed a simulation study with non-linearities and interactions, and missingness in multiple variables.
- Regression tree imputation was less biased than standard MICE (ignoring non-linearities and interactions), but CI coverage was stated as being poor (coverage results were not given).

Random forest for multiple imputation

- Recently, Doove *et al* proposed using random forest for multiple imputation, again within the MICE framework [9].
- To impute Y using fully observed $X_1, .., X_p$:
  1. apply random forest to (y^{obs}, x^{obs}), using k bootstraps
  2. for a given subject with missing Y with predictor values $x_1, .., x_p$, take the observed values of Y in the terminal nodes of all k trees
  3. randomly sample one observed value of Y from these as the imputation of the missing Y
- Again this can be embedded into MICE, and repeated to create multiple imputations.
- The approach is included in van Buuren's MICE package in R.

Simulation results from Doove *et al* [9]

- Doove *et al* performed simulations with missing values in Y and a number of fully observed predictors.
- With Y having expectation a quadratic function of predictors, random forest was less biased than predictive mean matching imputation.
- However, for some scenarios/parameters, random forest had large biases, and CIs had coverage below nominal level in general.
- Qualitatively similar results were found with a model where predictors interacted in their effects on Y.
- They suggested that biases may be due to the fact that tree based methods may struggle to recreate smooth, linear associations between variables.

Allowing for uncertainty in the (implicit) model parameters

- For given (y^{obs}, x^{obs}) and observed predictor values $(x_1, .., x_p)$, as $k \rightarrow \infty$, Doove *et al* 's procedure draws from y^{obs} with particular (fixed) probabilties.
- This means that (I believe) uncertainty in the (implicit) model parameters is not being propagated.
- In simulations and data analysis to follow, I therefore also consider a slightly modified version, where Doove *et al* 's random forest procedure is applied to a bootstrap sample $(y^{obs,bs}, x^{obs,bs})$, rather than to (y^{obs}, x^{obs}).

Alternative random forest imputation

- Independently of Doove *et al* , Shah *et al* proposed using random forest for imputation [1].
- For continuous Y, Shah *et al* use a somewhat different approach:
  1. take a bootstrap sample $(y^{obs,bs}, x^{obs,bs})$ from (y^{obs}, x^{obs})
  2. standard random forest is applied to $(y^{obs,bs}, x^{obs,bs})$, giving $\hat{E}(Y|X_1, .., X_p)$
  3. missing Y values are imputed by taking a normal draw, centred on $\hat{E}(Y|X_1, .., X_p)$ and residual variance equal to the 'out of bag' mean square error
- This is implemented in the R package CALIBERrfimpute.

Shah *et al* 's random forest imputation approach

- In simulations, Shah *et al* found that their random forest imputation implementation gave estimates with little bias and good CI coverage.
- A drawback of the approach however is the assumption of conditional normality and constant variance.
- The ‘out of bag’ error is also not residual variance – it is residual variance plus bias [13].
- For both random forest imp. approaches, an open question is how best to choose of the number of trees k, the size of terminal nodes in trees (default is 5), and m (number of variables to consider at each split).

Outline

Missing data in electronic health record databases

Trees and random forest for prediction

Trees and random forest for multiple imputation

Simulations

Conclusions

Simulation study

- Here I report a series of simulations to investigate the performance of random forest imputation.
- In each, 1,000 simulations were performed, on $n = 1,000$ subjects.
- One or more predictors $X_1, .., X_p$ are used, and these are fully observed.
- Y is generated as normal conditional on $X_1, .., X_p$, and values are made missing.
- Missing Y values are imputed 5 times, and a correctly specified analysis model for $Y|X_1, .., X_p$ (or a subset of predictors) fitted to each.
- Rubin's rules are used to combine estimates from the 5 imputed datasets.

Imputation methods

1. Imputation using correctly specified normal imputation model.
2. Imputation using incorrectly specified normal imputation model, with default assumptions of linearity and no interactions.
3. Predictive mean matching, using the same default imputation model as 2.
4. Random forest imputation proposed by Doove *et al* (‘RF-Doove’).
5. Random forest imputation proposed by Doove *et al* with additional bootstrap (‘RF-Doove2’).
6. Imputation assuming conditional normality, with mean and variance from random forest (‘RF-Shah’).

Scenario 1 setup

- $X \sim N(0,1)$
- $Y = \beta_0 + \beta_1 X + \epsilon$, $\epsilon \sim N(0,1)$
- $\beta_0 = 0$, $\beta_1 = 1$
- Y MCAR, with $P(R = 1) = 0.5$

Scenario 1 results

Results shown for $\beta_1 = 1$

Imp. method	Mean	Emp. SD	Mean SE	CI Cov
Norm correct	1.00	0.047	0.046	95.1
PMM	0.99	0.047	0.045	93.6
RF-Doove	1.00	0.047	0.038	90.0
RF-Doove2	1.00	0.049	0.049	94.2
RF-Shah	1.00	0.048	0.045	93.2

- All methods are unbiased.
- The estimated SE from RF-Doove is too small.
- RF results were also good with $n = 100$ and an MAR missingness mechanism (not shown).

Scenario 2 setup

- $X \sim N(0,1)$
- $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$, $\epsilon \sim N(0,1)$
- $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 1$
- Y MCAR, with $P(R=1) = 0.5$, or Y MAR with $P(R=1|X) = \text{expit}(X)$.

Scenario 2 (MCAR, $n = 1,000$)

Results shown for $\beta_2 = 1$

Imp. method	Mean	Emp. SD	Mean SE	CI Cov
Norm correct	1.00	0.035	0.033	94.5
Norm wrong	0.50	0.060	0.046	0
PMM	0.50	0.060	0.046	0
RF-Doove	0.97	0.045	0.032	79.8
RF-Doove2	0.97	0.045	0.039	86.2
RF-Shah	0.97	0.045	0.038	84.3

- Default normal imputation is (as expected) badly biased, as is PMM.
- RF methods now show slight bias. Est SEs are too small, and coverage is below nominal, but not too badly.

Scenario 2 (MCAR, with small ($n = 100$) sample size)

Results shown for $\beta_2 = 1$

Imp. method	Mean	Emp. SD	Mean SE	CI Cov
Norm correct	1.00	0.127	0.122	96.2
Norm wrong	0.47	0.173	0.151	15.8
PMM	0.47	0.173	0.151	14.0
RF-Doove	0.84	0.179	0.114	70.7
RF-Doove2	0.83	0.182	0.132	75.6
RF-Shah	0.84	0.176	0.133	80.5

- RF methods now show larger bias.
- Without a parametric model, sparsity of data is a problem for RF, as for other hot-deck approaches.

Scenario 2 (MAR, $n = 1,000$)

Results shown for $\beta_2 = 1$

Imp. method	Mean	Emp. SD	Mean SE	CI Cov
Norm correct	1.00	0.040	0.040	95.0
Norm wrong	0.35	0.052	0.044	0
PMM	0.35	0.052	0.044	0
RF-Doove	0.88	0.084	0.036	32.0
RF-Doove2	0.86	0.086	0.064	57.2
RF-Shah	0.87	0.083	0.063	58.0

- RF methods have larger bias under MAR.
- Under MAR, less data in some regions of x-space, so again we have sparsity.

Scenario 3 setup

- $X \sim N(0,1)$, $Z|X \sim N(0.5X, 1)$
- $Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \epsilon$, $\epsilon \sim N(0,1)$
- $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = -1$, $\beta_3 = 1$
- Y MCAR, with $P(R = 1) = 0.5$

Scenario 3 (MCAR, $n = 1,000$)

Results shown for $\beta_3 = 1$

Imp. method	Mean	Emp. SD	Mean SE	CI Cov
Norm correct	1.00	0.038	0.038	95.0
Norm wrong	0.49	0.058	0.049	0
PMM	0.54	0.061	0.048	0
RF-Doove	0.91	0.050	0.045	57.7
RF-Doove2	0.89	0.051	0.051	56.4
RF-Shah	0.90	0.049	0.044	47.9

- Again, RF outperforms mis-specified parametric and PMM imp. models.
- Here only ($p/3 = 2/3$) one predictor is chosen at random for consideration at each split.
- Increasing to $m = 2$ reduces bias of RF somewhat further.

Scenario 4 setup

- $X_1, ., .X_{50} \sim N(0,1)$, $Corr(X_j, X_{j'}) = 0$ for $j \neq j'$
- $Y = \beta X_1 + \epsilon$, $\epsilon \sim N(0,1)$
- $\beta = 1$
- Y MCAR, with $P(R = 1) = 0.5$
- The analysis model is regression of Y on X_1.

Scenario 4 (MCAR, $n = 1,000$)

Results shown for $\beta = 1$

Imp. method	Mean	Emp. SD	Mean SE	CI Cov
Norm correct	1.00	0.048	0.048	94.0
PMM	0.99	0.049	0.046	93.2
RF-Doove	0.86	0.048	0.053	33.7
RF-Doove2	0.85	0.050	0.058	34.0
RF-Shah	0.85	0.047	0.053	30.2

- RF is now biased. This is likely due to the fact that at each split, there is only a $1/3$ probability of the only X which is important being considered.
- To alleviate this, we can set $m = p$, so that all variables are considered at each split...

Scenario 4 (MCAR, $n = 1,000$)

Results shown for $\beta = 1$

Imp. method	Mean	Emp. SD	Mean SE	CI Cov
RF-Doove ($m = p$)	0.96	0.048	0.043	83.1
RF-Doove2 ($m = p$)	0.95	0.048	0.049	85.1
RF-Shah ($m = p$)	0.96	0.050	0.046	86.3

- Choosing $m = p$ here reduces bias considerably, with no cost in increased variance.

Scenario 5 setup

- $X_1, ., .X_{50} \sim N(0,1)$, $Corr(X_j, X_{j'}) = 0$ for $j \neq j'$
- $Y = \sum_{j=1}^{50} \beta X_j + \epsilon$, $\epsilon \sim N(0,1)$
- $\beta = 1/\sqrt{50} = 0.141$
- Y MCAR, with $P(R = 1) = 0.5$
- The analysis model is regression of Y on X_1.

Scenario 5 (MCAR, $n = 1,000$)

Results shown for $\beta = 0.141$

Imp. method	Mean	Emp. SD	Mean SE	CI Cov
Norm correct	0.139	0.059	0.058	94.5
PMM	0.138	0.058	0.057	94.8
RF-Doove	0.081	0.042	0.056	88.3
RF-Doove ($m = p$)	0.082	0.044	0.056	86.6
RF-Doove2	0.082	0.043	0.057	89.3
RF-Doove2 ($m = p$)	0.083	0.044	0.057	88.3
RF-Shah	0.084	0.044	0.052	82.8
RF-Shah ($m = p$)	0.086	0.045	0.053	85.8

- RF shows downward bias.
- Choosing $m = p$ makes little difference, now that all predictors are important.

Scenario 6 setup

- $X_1, ., .X_{50} \sim N(0, 1)$, first 25 have mutual correlation 0.5, second 25 have mutual correlation 0.25, but the two sets are independent.
- $Y = \sum_{j=1}^{25} X_j + \epsilon$, $\epsilon \sim N(0, 325)$ (so that $R^2 = 0.5$)
- Y MCAR, with $P(R = 1) = 0.5$
- The analysis model is regression of Y on X_1, which has true coefficient $\beta = 13$.

Scenario 6 (MCAR, $n = 1,000$, **100 simulations**)

Results shown for $\beta = 13$

Imp. method	Mean	Emp. SD	Mean SE	CI Cov
Norm correct	13.0	0.78	0.98	98
PMM	12.9	0.79	0.92	96
RF-Doove	11.8	0.69	0.87	75
RF-Doove ($m = p$)	11.8	0.70	0.86	76
RF-Doove2	11.9	0.65	0.90	83
RF-Doove2 ($m = p$)	11.9	0.66	0.92	88
RF-Shah	12.1	0.62	0.84	91
RF-Shah ($m = p$)	12.2	0.70	0.86	90

- RF shows a downward bias, although proportionately smaller here.
- Again choosing $m = p$ makes little difference to results here.

Outline

Missing data in electronic health record databases

Trees and random forest for prediction

Trees and random forest for multiple imputation

Simulations

Conclusions

Conclusions

- Imputation based on random forest shows promise, and in particular in the context of missing data in electronic health databases may be useful.
- Simulation evidence suggest it may be able to automatically allow for interactions and non-linearities.
- If imputed datasets are to be released to many researchers, this would be very useful.
- Limited simulation results not shown also suggest RF may be useful when $p \approx n$, where standard parametric imputation results in highly variable estimates.

Conclusions

- However, we have also seen in some simple setups that it can lead to biased estimates.
- Small sample sizes, and non-MCAR missingness in particular seem to lead to bias, since RF cannot extrapolate in the same way as a smooth parametric model can.
- Simulation results show the default choice of $m = p/3$ for continuous variables can lead to bias, suggesting that using $m = p$, where feasible, may be preferable.
- Moreover, further research is clearly needed to better understand RF's statistical properties, and consequently its properties when used for multiple imputation.

References I

[1] Anoop D Shah, Jonathan W Bartlett, James Carpenter, Owen Nicholas, and Harry Hemingway.

Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study.

American Journal of Epidemiology, 179(6):764–774, 2014.

[2] Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani.

The Elements of Statistical Learning.

Number 1. Springer, 2nd edition, 2009.

[3] X L Meng.

Multiple-imputation inferences with uncongenial sources of input (with discussion).

Statistical Science, 10:538–573, 1994.

References II

[4] J W Bartlett, S R Seaman, I R White, and J R Carpenter.

Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model.

Statistical Methods in Medical Research, epub:epub, 2014.

[5] Rebecca R Andridge and Roderick JA Little.

A review of hot deck imputation for survey non-response.

International Statistical Review, 78(1):40–64, 2010.

[6] Nathaniel Schenker and Jeremy MG Taylor.

Partially parametric techniques for multiple imputation.

Computational Statistics & Data Analysis, 22(4):425–446, 1996.

[7] Lane F Burgette and Jerome P Reiter.

Multiple imputation for missing data via sequential regression trees.

American Journal of Epidemiology, 172:1070–1076, 2010.

References III

[8] Daniel J Stekhoven and Peter Bühlmann.

Missforestnon-parametric missing value imputation for mixed-type data.

Bioinformatics, 28(1):112–118, 2012.

[9] LL Doove, Stef Van Buuren, and Elise Dusseldorp.

Recursive partitioning for missing data imputation in the presence of interaction effects.

Computational Statistics & Data Analysis, 72:92–104, 2014.

[10] Leo Breiman.

Random forests.

Machine learning, 45(1):5–32, 2001.

[11] Yi Lin and Yongho Jeon.

Random forests and adaptive nearest neighbors.

Journal of the American Statistical Association, 101(474):578–590, 2006.

References IV

[12] Gérard Biau.
Analysis of a random forests model.
The Journal of Machine Learning Research, 13(1):1063–1095, 2012.

[13] Guillermo Mendez and Sharon Lohr.
Estimating residual variance in random forest regression.
Computational Statistics & Data Analysis, 55(11):2937–2950, 2011.