Real donor imputation pools

Nicklas Pettersson[1]

[1]Stockholm University, e-mail: nicklas.pettersson@stat.su.se

Abstract

Real donor matching is associated with hot deck imputation. Auxiliary variables are used to match donee units with missing values to a set of donor units with observed values, and the donee missing values are 'replaced' by copies of the donor values, as to create completely filled in datasets. The matching of donees and donors is complicated by the fact that the observed sample survey data is often both sparse and bounded. The important choice of how many possible donors to choose from involves a trade-off between bias and variance. We transfer concepts from kernel estimators to real donor imputation. In a simulation study we show how bias, variance and the estimated variance of a population behaves, focusing on the size of donor pools.

Keywords: Bayesian Bootstrap, Boundary and nonreponse bias; Multiple imputation

1 Introduction

The 'holes' in an incomplete dataset is always a nuisance, since this precludes most standard statistical methods from being directly applied. Datasets obtained by excluding partially observed units (e.g. due to item nonresponse) give inefficient and usually quite biased results. A better alternative is to impute the missing values. In the extensive book on missing data by Little and Rubin (2002, p.72) it is stated that

"Imputations should generally be:

(a) Conditional on observed variables, to reduce bias due to nonresponse, improve precision, and preserve association between missing and observed variables;
(b) Multivariate, to preserve associations between missing variables;
(c) Draws from predictive distribution rather than means, to provide valid estimates of a wide range of estimands."

The most important factor in imputation is access to auxiliary variables which are predictive of the missing values and the nonresponse propensity. Real donor (hot deck) imputation (Laaksonen, 2000) uses auxiliaries to match a donee unit with missing values to a set (pool) of close (nearest neighbour) donor units with observed values, and then 'replaces' the donee missing values by copies of values of randomly drawn donor units. It is often applied within cells from cross-classified categorical (and sometimes subjectively classified continuous) auxiliares. We only discuss continuous variables with univariate missingness. Point (b) is therefore not relevant here.

Point (c) can be fulfilled by using multiple imputation (Little and Rubin, 2002), which is a method for representing missing data uncertainty. The missing values are then imputed several times, and each imputed dataset is analyzed separately. The final estimates consist of the pooled results.

The size of donor pools becomes important here. Pools with few potential donors give rise to strong correlation between the values imputed for a missing value. In repeated sampling this results in highly variable final estimates, similar to sampling from correlated (e.g. clustered) data. Larger donor pools may instead reduce the quality of matches and increase the bias. The number of potential nearest neighbour donors thus affects the trade-off between bias and variance in imputation, in parallell with pointwise kernel estimators. Features from this area have been applied in imputation to deal with the sparse and bounded data (Aerts et al, 2002; Pettersson, 2012), and to decide the donor pools (Schenker and Taylor, 1995; Marella, Scanu and Conti, 2008). We discuss these issues in the following sections. In simulations we show how different strategies for selecting the number of donors and the features for bias reduction from Pettersson (2012) affects bias, variance, and estimates of variance of a population mean estimate. To yield valid inference, our method is based on the Bayesian Bootstrap (Rubin, 1981).

2 Selecting the donor pool

The choice of bandwidth is important in kernel estimation. Several types of bandwidths exist. A fixed bandwidth corresponds to having imputation donor pools consisting of units with an (auxiliary-based) distance to the donee which is less than a value ϵ. A fixed 'rule-of-thumb' bandwidth based on distributional assumptions is often a good starting point (Silverman, 1986). Fixed bandwidths can be locally adapted by increasing (decreasing) the maximum allowed distance if relatively few (many) donors are close to the donee, i.e. if the density at the donee auxiliary value is low (high). Always using the same number of potential donors in all donor pools corresponds to a nearest neighbour bandwidth, which may find donors that are better matched to the donee in dense regions, and automatically ensures that all donoees obtain at least one potential donor. The trade-off between bias and variance means that gains in precision from increasing the number of donors may result in reduced quality of the matches and increased bias. Different estimators may profit from different strategies of choosing the donor pool size/bandwidth.

3 Bias reduction

A disadvantage of the real donors' methods is that a donee and its pool of donors usually are imperfectly matched. Particularly, this becomes a problem when the donee auxiliary values lie at the boundary (i.e. convex hull) of the donors' auxiliary values, since there may be no or only a few potential donors with observed auxiliary values that lie on one side of the donee auxiliary value. The donor pool is then badly balanced with regard to the donee. If such a pool is used for imputation, the risk is also larger that bias is introduced in the imputed study variable.

Pettersson (2012) employs three methods to reduce this bias. First, since the closest donors provide a better match to the donee, they are given higher selection probabilities than more distant donors. Due to the optimality properties in estimation the donor selection probabilities are decided by an Epanechnikov function (Silverman, 1986). Secondly, the selection probabilities are calibrated so that the expected imputed auxiliary value equals the auxiliary value of the donee. The third method not only reduces the selection probabilities of the furthest donors in the pool but also completely removes these furthest donors (which match the donee least and thus contribute most to the bias), and only keeps the best matches, which gain larger selection probabilities. The bias will be reduced, but donor pool variance is expected to increase.

4 Simulation

We used the setup in Pettersson (2012) with a population of $N = 1600$ units, from which $G = 1000$ samples of size $n = 400$ were drawn, and with each study variable imputed $B = 20$ times using the auxiliary variable from which it was generated. Since bandwidth behaviour may depend on the underlying distribution we used three auxiliaries; $X_{Uniform} \sim U(\pi/6, 2\pi)$; $X_{Normal} \sim N(13\pi/12, 11\pi/48)$; and $X_{Gamma} \sim Z + \pi/6$, where $Z \sim Gamma(1, 1/2)$. All auxiliaries had approximately a range of $(\pi/2, 2\pi)$, although X_{Gamma} had an outlier at 6.28. We chose a logistic missingness mechanism: $logit(Pr(R = 1|X) = -1 + \beta_z \sum_{i=1}^{5}(X - (2i\pi - 2)/4)^2$, where the β_z was adjusted to give on average 25% missingness irrespective of the auxiliary ($z = Uniform, Normal, Gamma$).

Imputation methods rely on the relation between study and auxiliary variables, so we generated a linear $Y_X = X_z + e_{X_z}$, a nonlinear $Y_{cosX} = cos(4X_z) + e_{cos(4X_z)}$, and a mixed $Y_{X+cosX} = X_z + cos(4X_z) + e_{X_z+cos(4X_z)}$ study variable. The error terms e_t were generated from $N(0, Var(t))$. The probability of nonresponse on the study variables induced by the missingness mechanism is thus highest (lowest) as $cos(4X_z) = 1(0)$. Means and covariances are found in Table 1. We abbreviate z to u, n and g for Uniform, Normal and Gamma.

The number of potential donors k was determined in three ways. The first method (kNN) initially used $k = 2, ..., 30$ potential donors, and gradually increased the number as more values were imputed. Secondly, we used a rule-of-thumb method (fix) where the donor pool consisted of units with distance less than $\epsilon \propto s_{X_z} m^{-1/5}$ from the donee, where m is the number of potential donors and s_{X_z} is the standard deviaton of varible X_z. Thirdly, we used a locally adapted version ($adap$) of fix, where ϵ was increased (decreased) if the density at the donee auxiliary value was low (high) (see Silverman

1986, p.101). We also used versions with all the bias reduction features from section 3 applied on kNN_b, fix_b and $adap_b$.

We compute $Bias = \frac{1}{G}\sum_{g=1}^{G}(\widehat{\overline{Y}}_g - \overline{Y})$, where $\widehat{\overline{Y}}_g = \sum_{b=1}^{B}\widehat{\overline{Y}}_{b,g}$ is the overall estimated mean in the B imputed datasets and $\overline{Y}$ is the true mean, $Var = \frac{1}{G}\sum_{g=1}^{G}(\widehat{\overline{Y}}_g - \frac{1}{G}\sum_{g=1}^{G}\widehat{\overline{Y}}_g)^2$, and relative error of estimated variance $\frac{Var - \widehat{Var}}{Var}$, where $\widehat{Var} = \frac{1}{G}\sum_{g=1}^{G}(s_{Y_g} + \frac{B+1}{B(B-1)}\sum_{b=1}^{B}(\widehat{\overline{Y}}_g - \widehat{\overline{Y}}_{b,g})^2)$ is the average estimated variance.

Table 1: Means and (co)variances of simulated data

	Mean			Variance			Covariance with X_z		
z	u	n	g	u	n	g	u	n	g
Y_{X_z}	1.08	3.47	3.42	3.04	3.19	5.65	0.32	0.19	0.51
Y_{cX_z}	-0.44	0.04	0.01	0.62	0.76	0.72	0.51	-0.04	0.48
$Y_{X_z+cX_z}$	0.64	3.51	3.44	6.08	5.95	8.74	2.89	0.17	3.08
X_z	1.03	3.42	3.38	0.27	0.51	2.81	0.27	0.51	2.81

5 Results

We present the results in Figures 1-3. The number of initial donors for kNN and kNN_b is plotted against bias, variance and the relative error in estimated variance. We also add horizontal lines for fix, fix_b, $adap$ and $adap_b$. Due to the shrinkage feature, the initial number of donors is expected to be larger then the final number of donors for the bias corrected methods.

Except for the least complex data $Y_{X_{uniform}}$ and $Y_{X_{normal}}$ with small initial k, bias is always smaller for kNN_b compared to kNN. Bias tends to increase as k increases for both methods, but kNN_b at a lower rate. kNN_b also has lower bias than the fixed and adaptive versions with a few exceptions for Y_{X_z} when they are comparable. Bias corrected versions fix_b and $adap_b$ always give lower bias than their noncorrected counterparts fix and $adap$, except for $Y_{X_{normal}}$ with $adap$.

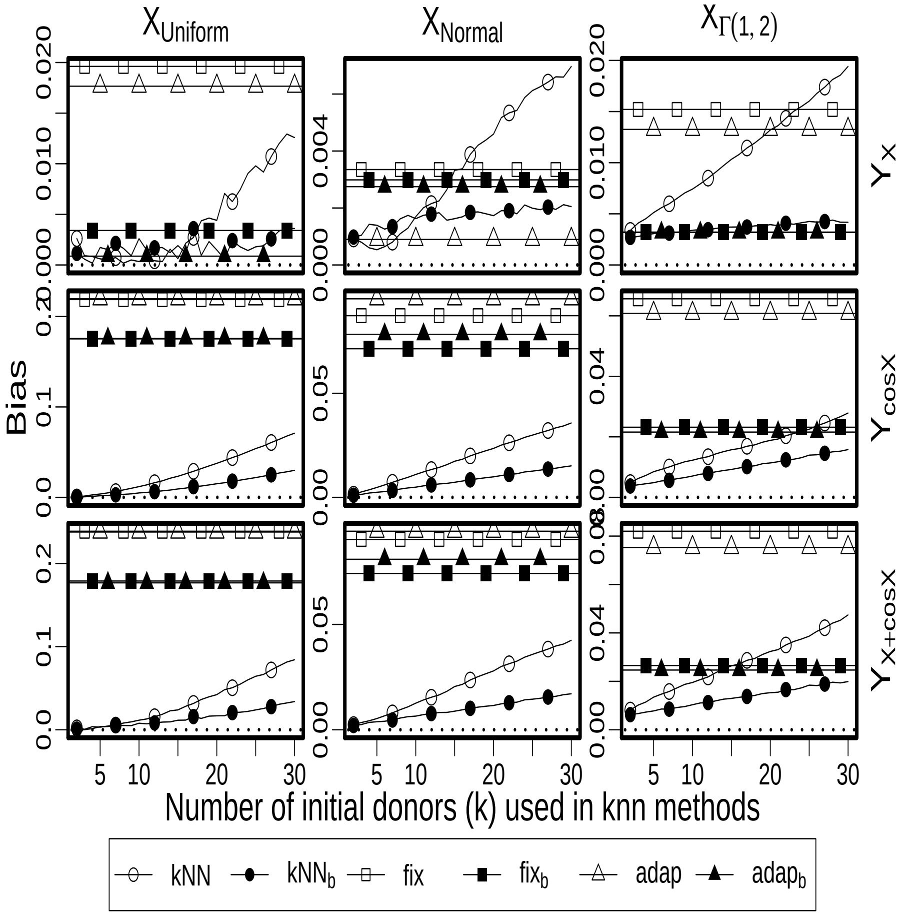

Figure 1: Bias of estimates from simulations

For small k, variance always falls as k is increased but is slightly higher for kNN_b compared to kNN. For larger k, the variance continues to fall or flatten out, except for kNN where it sometimes increases, especially with auxiliary X_{Gamma}. Both fixed and adaptive methods generally have lowest variance, but nearest neighbour methods usually approached them as k was increased, and for $Y_{cosX_{uniform}}$ the variance for kNN_b was always low. For $Y_{X_{uniform}}$, methods without bias correction (which on average also used most donors) had variance not far from complete data.

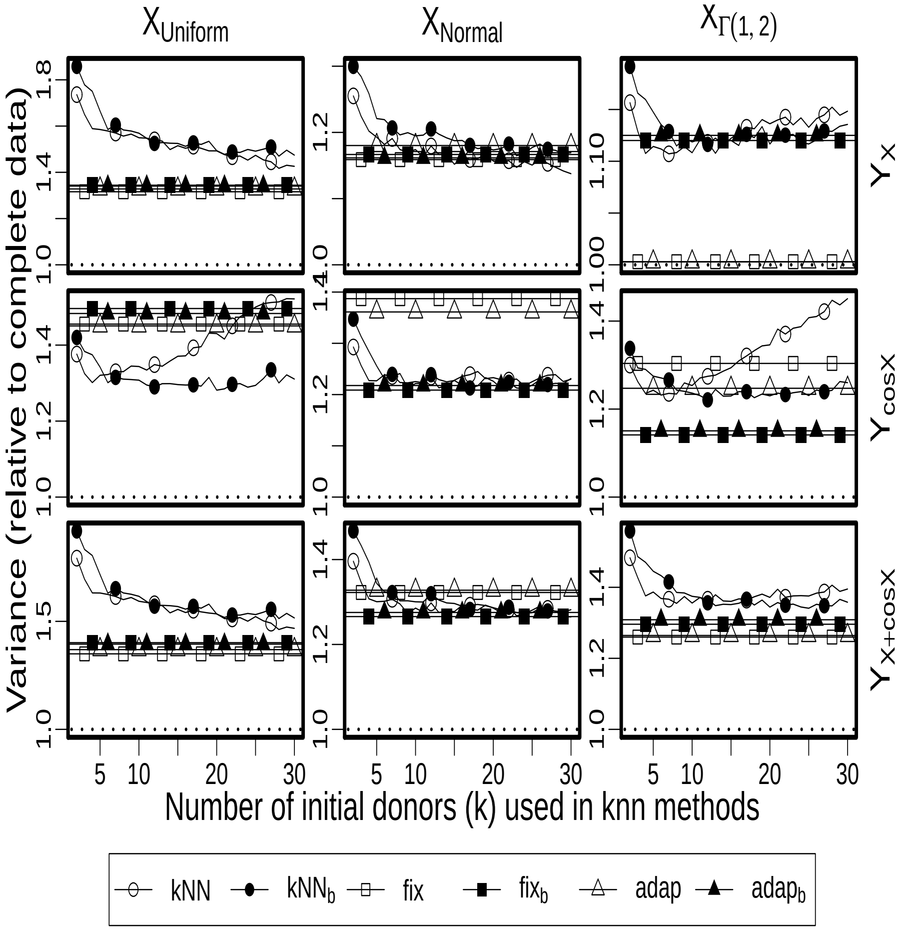

Figure 2: Variance of estimates from simulations

6 Conclusions

Multiple real donor imputation has the advantage of requiring few model assumptions. Moreover, the imputed values are actually observed values. But some difficulties with continuous auxiliaries arise that need to be dealt with. Since boundary donee units with missing values can only be matched to donors on one side, donor pools will be biased. Relative sparseness of donors also worsen the probability of forming good predictive donor pools. The size of donor pools is important since it involves a trade-off between bias and variance and affects the ability of estimating variances. This was clearly seen here where the fixed/adaptive methods, which generally had

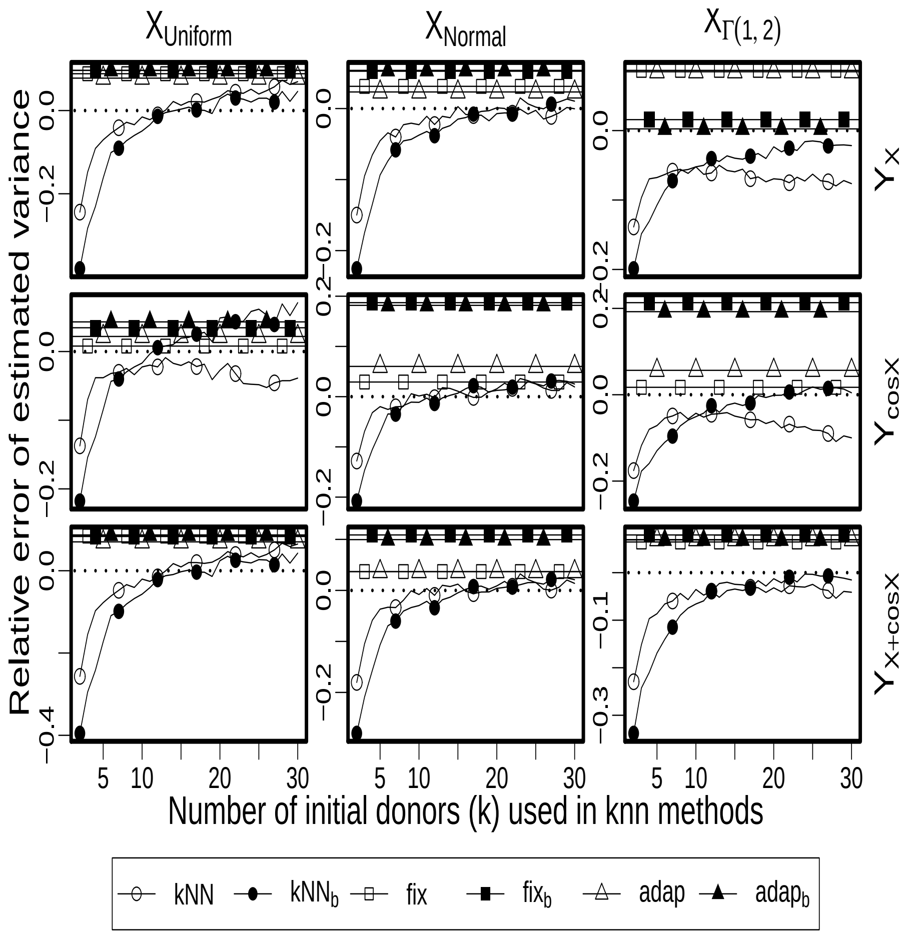

Figure 3: Relative error of estimated variance of estimates from simulations

large donor pools, also had larger bias but smaller variance. Without any bias reduction applied there is certainly a risk of increased bias (and variance) associated with increasing donor pool sizes. Increasing the number of donors for boundary donees naturally aggrevates the already insufficient matching. Too few donors is, on the other hand, associated with high variance and too-low variance estimates.

The bias reduction techniques adress the boundary bias and matching problems by adapting the donor pools. Given sufficiently many initial donors, it can make bias of the nearest neighbour method less dependent on the exact number of donors, and also improve bias of fixed/adaptive methods. We only study one fixed (and adaptive) rule-of-thumb method in our simulation, and blind use of it obviously involved a risk of large bias. Compared to a reasonably large nearest neighbour it only had lower MSE when the study variable was a linear function of a uniform auxiliary. This seems to be associated with its generally larger donor pools giving rise to larger bias. Simulations with several other fixed methods (not presented here) generally also gave larger bias but smaller variance then nearest neighbour methods. The effects from local adaption of the fixed method seemed relatively small here and need further investigation.

References

Aerts, M. Claeskens, G. Hens, N. Molenberghs G., (2002). Local multiple imputation. *Biometrika*, 89(2), 375-388.

Laaksonen, S., (2000). Regression-based nearest neighbour hot decking, *Computational Statistics*, 15(1), 65-71.

Little, R.J.A. Rubin, D.B., (2002). *Statistical Analysis with Missing Data.* New York: Wiley.

Marella, D. Scanu, M. Conti, P.L., (2008). On the matching noise of some nonparametric imputation. *Statistics and Probability Letters*, 78, 1593-1600.

Pettersson, N., (2012) Bias reduction of finite population imputation by kernel methods. *To appear in Statistics in Transitions.*

Rubin, D.B., (1981). The Bayesian bootstrap, *Annals of Statistics*, 9, 130-134.

Schenker, N. Taylor, J.M.G., (1996), Partially Parametric Techniques for Multiple Imputation, *Computational Statistics and Data Analysis*, 22, 425-446.

Silverman, B.W., (1986). *Density estimation for statistics and data analysis.* Chapman and Hall, London.