Kernel imputation with multivariate auxiliaries

Nicklas Pettersson[1]

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

Abstract

We consider a data set with missing observations but known auxiliaries for the sample and develop a real donor imputation. For each unit with missing observations we construct a distribution over a set of possible donors. We want the expectation (or distribution) to be chosen so that the expectation (or distribution) of the imputed values should equal the distribution of the units' true values. This is obtained by letting the expected values of the auxiliaries equal the true value. Several kernel estimation features are introduced to reduce the bias associated with the unbalanced donor sets, due to sparse and bounded data sets. To get the good properties of kernel estimates to carry over, multiple imputation is used. Simulation studies indicate that our method has a good performance compared to competing methods. This is particularly noticable in notoriously difficult situations e.g. when the relationship between the study and auxiliary variables is nonlinear. The displayed simulations are based on two auxiliary variables, but the algorithm is generally formulated for any number of auxiliaries.

Keywords: Boundary bias, Imputation algorithm, Missing data, Real donor imputation, Multiple imputation.

1 Introduction

1.1 Background

Missing data is always a nuisance. The 'holes' in the dataset precludes many simple standard techniques. If only units with no item nonresponse are

considered a complete data set is obtained but the estimate will be less efficient and usually quite biased. A general solution is to fill in the missing values by imputed values which are drawn from the predictive distribution of the missing value (Rubin, 1976). The most important factor for imputation to reduce (nonresponse) bias, improve precision, and preserve associations between variables is using auxiliary variables which are predictive of both the nonresponse propensity (Rosenbaum and Rubin, 1983) and the missing values (Little, 1988).

Real donor imputation (Laaksonen, 2000) is attractive since the imputed values are copies of actually observed (realistic) values on units from a donor pool, but the ever present imbalance between the donor pool and the (donee) unit which is being imputed may result in e.g. biased estimates, especially with sparse or bounded data. The characteristics of donor pools can be improved by adapting features from kernel estimation (Pettersson, 2013), and we believe that this potential has not recieved sufficient attention.

We use simulations to investigate the properties of an algorithm which avoids strong parametric assumptions, appeals to the general properties of kernel smoothers and utilizes known background information. Our goal is to make parameter estimation almost unbiased in combination with multiple imputation and under the assumption of a missing at random response mechanism.

1.2 Notation

Assume that a sample S with n units has been drawn from a population U with N units. Each unit i is characterized by two related variables $\boldsymbol{X_i}$ and Y_i. In our simulations all units' values are assumed to be multivariate iid random variables and thus exchangeable. The goal is to estimate a function $T((\boldsymbol{X_i}, Y_i), i \in U)$, e.g. the mean $T = \sum_{i=1}^{N} Y_i/N$. The vectorvalued (continuous) auxiliary variable $\boldsymbol{X_i}$ is known for all N units in the frame. The study variable Y_i, was meant to be obtained from all the n units in the sample, but there was some nonresponse. This is described by an indicator variable $R_i,\ i \in U$, which indicates whether Y_i will be observed if unit i is selected. When the sample has been drawn R_i is first observed, and Y_i is observed if $R_i = 1$. We will consider $q \geq 2$ auxilary variables, but only a single study variable, so that $r = \sum_{i=1}^{n} R_i$ is the number of observed units. Formally Y_i and R_i may, however, be vectorvalued. Nonresponse may cause

the r responders and $n-r$ nonresponders to differ in distribution of $\boldsymbol{X_i}$ and Y_i.

We suggest an imputation method which can be used for any data set with these properties, but it is mainly intended for a situation when data are missing at random (MAR) (Rubin, 1976), meaning that the response probability is related only to known or observed values. Assuming MAR simplifies the general response mechanism $P_{R|\boldsymbol{X}Y}(r_i|\boldsymbol{x}_i, y_i, \theta_R)$ to $P_{R|\boldsymbol{X}}(r_i|\boldsymbol{x}_i, \theta_R)$. The responders and nonresponders may then be treated as conditionally exchangeable under the full data model

$$f_{\boldsymbol{X}YR}(\boldsymbol{x}_i, y_i, r_i|\theta_{\boldsymbol{X}}, \theta_Y, \theta_R) = f_{\boldsymbol{X}}(\boldsymbol{x}_i|\theta_{\boldsymbol{X}}) f_{Y|\boldsymbol{X}}(y_i|\boldsymbol{x}_i, \theta_Y) P_{R|\boldsymbol{X}}(r_i|\boldsymbol{x}_i, \theta_R) \quad (1)$$

where the parameters $\theta_{\boldsymbol{X}}$, θ_Y and θ_R are the model probability vectors for $\boldsymbol{X}$, Y and R. Using observed responding units as donors when imputing the nonresponding donee units can therefore be justified if MAR holds and the imputation method is derived from the correct model.

In real donor imputation each donee unit i with missing values is assigned a donor pool or subset $S_{i,k}$ of S containing the k_i donor units (from the p currently available donor units, where $p \geq r$) with the closest match to donee i according to some auxiliary-based distance. The specific potential donors in $S_{i,k}$ are determined by a donor pool matrix $H_i = h_i \cdot G_i$, where h_i is a scalar and G_i is a non-negative matrix deciding the shape of the donor pool. To each of the k_i units in the donor pool $S_{i,k}$, we assign a selection probability $\lambda_{i,j}$, creating a probability measure $\lambda_i = (\lambda_{i,1}, ..., \lambda_{i,k_i})$. The missing value y_i on unit i is imputed by randomly drawing a donor j using λ_i, finding its corresponding value y_j, and finally, imputing it.

1.3 Literature review

Our real donor imputation algorithm may be denoted as a hot deck method (Andridge and Little, 2010). In resemblence to Pólya sampling (Feller, 1971), previously imputed donee units are allowed to act as potential donors to forthcoming donees, so an imputed dataset becomes a realization from the Pólya posterior (Ghosh and Meeden, 1997).

Proper consideration of the uncertainty caused by the fact that y_j differs from y_i can be made through explicit formulas, resampling of single imputed datasets, or by multiple imputation (Little and Rubin, 2002). In multiple imputation the missing values are indepently imputed $B \geq 2$ times given

the data. Each imputed dataset is analyzed separately, and the final estimates come from averaging the results. The variability between the imputed datasets should reflect a reasonable level of uncertainty to estimate the variance increase due to the missing data.

The properties of an estimator $\widehat{T}$ of T is related to the donor pools. Imputation approaches to selection of donor pool size (Marella, Scanu and Conti, 2008; Pettersson, 2012; Schenker and Taylor, 1995) parallel bandwidth determination in kernel estimation (Härdle, 1990; Silverman, 1986). Donor pools with few potential donors result in strong dependence between the B values imputed on a missing value, which in repeated sampling can give rise to high variability of $\widehat{T}$. Larger donor pools reduce the imputation variance but may instead increase the bias of $\widehat{T}$.

The number of donors k_i may be determined as those donors in S lying within a range h_i from $\boldsymbol{x}_i$. Approaches to deciding h_i include rules-of-thumb from distributional assumptions (Silverman, 1986), least-squares cross-validation (Scott and Terrell, 1987), plug-in estimates (Chacon and Duong, 2010; Wand and Jones, 1994), and smoothed cross-validation (Duong and Hazelton, 2005; Jones, Marron and Park, 1991). While these approaches are less biased when $\boldsymbol{x}$ is sparse, a nearest neighbour (NN) approach may provide better matching with more dense $\boldsymbol{x}$ and automatically ensures nonzero k_i:s. Given p eligible donors and q auxiliary variables, the ideal NN approach of setting $k_i \approx p^{4/(4+q)}$ is best used when the exact size of k_i is not so important (Silverman, 1986).

Within a donor pool, higher donation probabilities may be assigned to potential donors closer to the donee if λ_i is determined from a kernel (Conti, Marella and Scanu, 2008; Pettersson, 2012; 2013) or some other function (Siddique and Belin, 2008). Given an optimally chosen bandwidth parameter, the Epanechnikov (1969) kernel can be shown to minimize the mean integrated squared error (MISE) (Silverman, 1986) asymptotically with an increasing n in kernel estimation. Canonical kernels (Marron and Nolan, 1989) can be used to neutralize the interplay between the donor pool size and the selection probabilities when donor pool selection approaches are compared. See also (Aerts, Claeskens, Hens and Molenberghs, 2002; Pettersson, 2012; 2013) for determination of donor selection probabilities in kernel imputation.

Since $E[\boldsymbol{x}_j] - \boldsymbol{x}_i \neq \boldsymbol{0}$ real donor pools are always 'individually biased', except when donor pools are formed from categorical auxiliaries where $\boldsymbol{x}_j = \boldsymbol{x}_i \forall j$. However, categorization by introducing subjectively chosen bound-

aries in the data would only mask differences in continuous auxiliaries. The individual bias of donees that lie within the convex hull of their donors-auxiliary values $\{\boldsymbol{x_j}, j \in S_{i,k}\}$ may be eliminated if $\lambda_{i,j}$ is calibrated (Aerts, Claeskens, Hens and Molenberghs, 2002; Pettersson, 2012; 201). Calibration can be made e.g. by viewing the expected imputed value $E[\widehat{y}_i]$ of the study variables as a zero degree polyonomial estimate (Fan and Gijbels, 1996), and then linearize λ_i. Missing values have also been imputed by $E[\widehat{y}_i]$ (Chu and Cheng, 1995). Design weight calibration (Deville and Särndal, 1992) on a pointwise level can also be used to define calibrated selection probabilities.

Special attention is needed when $\boldsymbol{x_i}$ is close to the boundary of the convex hull of $\{\boldsymbol{x_j}, j \in S_{i,k}\}$. The closer a donee is located to the boundary of the convex hull of $\{\boldsymbol{x_j}, j \in S_{i,k}\}$, the more individually biased the imputation becomes due to the area without donors outside the hull. This is known as boundary bias in kernel estimation (Simonoff, 1996) and can not be completely removed through calibration. But bias can be reduced if the donor pool is made oblong along the convex hull of $\{\boldsymbol{x_j}, j \in S_{i,k}\}$. This technique is similar to linear discriminant adaptive nearest neighbour analysis (Hastie and Tibshirani, 1996) although the donor pool is made oblong along the observed boundaries between classes of the study variable, in order to increase the prediction power when finding NN units for classification. At the cost of higher variance, bias can also be reduced by shrinking the total donor pool of boundary donees (Pettersson, 2012; 2013)

1.4 Outline

In Section 2 we adapt the three features from Petterson (2013) to multivariate conditions and include them in an algorithm in Section 3. By setting donor selection probabilities $\lambda_{i,j}$ proportional to the Epanechnikov function, the better matched donors are assigned larger $\lambda_{i,j}$. Next we calibrate the selection probabilities so that $E[\boldsymbol{x}_j|\lambda_i] = \boldsymbol{x}_i$. Thirdly, we change the size and shape of the donor pools to improve the fit for boundary donees. Simulations with these imputation algorithms are then undertaken in Section 4. The MSE of T is seen to be considerably reduced by the proposed features, mainly due to reduction of bias, while variance is less affected. Using the ideal NN donor rate the method show some sensitivity to the number of donors but are generally better than in comparison to other approaches. It performs at least as well as other competing imputation methods. The paper

is concluded in Section 5.

2 Methods

2.1 Our basic real donor multiple imputation method

We have in this description sorted the data, which consists of $q \geq 2$ auxiliary variables and one partially observed study variable, so that the r units with complete data come first followed by the $n - r$ units with missing y-values. At first we assume that no auxiliaries are available. In that case our method will be a version of the finite population Bayesian bootstrap (Lo, 1988) which we describe first in our terminology.

We successively impute the missing values for donee units $i = r + 1, ..., n$. Unit i is imputed using the set $S_{i,p} = S_{i,i-1}$ of all available donors, implying that both the r observed and the $p - r$ previously imputed donee units may be used as potential donors. The p potential donors to donee i are then each assigned donor selection probabilities $\lambda_{i,j} = 1/p$ for $j = 1, ..., p$, where $\lambda_i = (\lambda_{i,1}, ..., \lambda_{i,p})$. When all $n - r$ units are replaced we have a complete data set. We use a multiple imputation technique which means that this is repeated so that we get B complete data sets in the end. For each new data set the $n - r$ units with missing data are randomly permuted.

For the first donee with $i = r + 1$, a donor unit is drawn from the donor pool $S_{r+1,r}$ using probabilities λ_{r+1} where $\lambda_{r+1,j} = 1/r$, $\forall j$. The value of the drawn donor is then copied and imputed on the donee. The potential donor pool $S_{r+2,r+1}$ for donee $i = r + 2$ is then set to be $S_{r+1,r}$ updated by adding the imputed donee unit. Donor selecton probabilities are also updated to $1/(r + 1)$. The procedure of drawing, imputing and updating is then repeated for the remaining non-observed units in S, resulting in a fully imputed dataset $S_{n+1,n}$ from which T can be estimated by some ordinary estimator $\widehat{T}$. This Pólya sampling procedure implies that $S_{n+1,n}$ is a draw from a Dirichlet distribution. By repeating the whole procedure $B \geq 2$ times, we get B different fully imputed datasets. The empirical distribution of the estimates $\widehat{T}_1, ..., \widehat{T}_B$ is then an approximation of the distribution of T given $S_{r+1,r}$.

A point estimate $\widehat{T}$ is given by the multiple imputation combining rules

$$\widehat{T} = B^{-1} \sum_{b=1}^{B} \widehat{T}_b. \tag{2}$$

Its variance can be estimated by

$$\widehat{V}(\widehat{T}) = \widehat{\sigma}^2_{within} + \widehat{\sigma}^2_{between}, \tag{3}$$

where $\widehat{\sigma}^2_{within} = B^{-1} \sum_{b=1}^{B} \widehat{\sigma}^2_b$ is the average of the B estimated variances, and $\widehat{\sigma}^2_{between} = (B+1)B^{-1}(B-1)^{-1} \sum_{b=1}^{B} (\widehat{T}_b - \widehat{T})^2$ is the variance of the estimates. The inflating factor $(B+1)B^{-1}$ is a correction for using a finite B. The approximate large-sample relative efficiency of an estimate based on $B = 20$ compared to $B = \infty$ imputed datasets and 40%(80%) rate of missing information is 99.0%(98.1%) (Rubin, 1987).

In our case we have access to an auxiliary $\boldsymbol{x}$. It is reasonable to alter the process by using its values and only imputing study variable values from units with similar auxiliary varible values. We do this by first defining a donor pool matrix $H_i = h_i \cdot G_i$, where G_i and h_i define the donor pool type and size, and then a distance $D_{H,i,j}$. If e.g. $D_{H,i,j} = (\boldsymbol{x}_j - \boldsymbol{x}_i)H_i^{-1}(\boldsymbol{x}_j - \boldsymbol{x}_i)^t$ and $G_i = \boldsymbol{I}$ then $D_{H,i,j}$ is the Mahalanobis distance between x_j and x_j. A donor pool $S_{i,k}$ is then made up of the $k_i \leq p$ units whose distance $D_{H,i,j} < 1$ for $j = 1, ..., k_i$.

2.2 Probability distributions and individual calibration

With $\boldsymbol{x}$ known we do not only refine the donor pools by using subsets $S_{i,k}$ with close units but also improve the dependence on $\boldsymbol{x}$ within the donor pools by using donor selection probabilities that depend on $\boldsymbol{x}$,

$$\lambda_{i,j} = K(D_{H,i,j}) / \sum_{\upsilon=1}^{k} K(D_{H,i,\upsilon}), \tag{4}$$

where $K()$ is a non-negative kernel function. The closer a donor is to the donee the larger its assigned donor selection probability will be. The expectation of an imputed value

$$E[\widehat{y}_i] = \sum_{j=1}^{k} \lambda_{i,j} y_j \tag{5}$$

is now equivalent to a pointwise kernel smoother.

A uniform function $K^{unif} \propto max\{0, I_{\boldsymbol{x}^t H_i^{-1} \boldsymbol{x} \leq 1}\}$ assigns all units in $S_{i,k}$ equal probabilities of becoming the donor, while an Epanechnikov kernel $K^{epan} \propto max\{0, (1 - \boldsymbol{x}^t H_i^{-1} \boldsymbol{x}) I_{\boldsymbol{x}^t H_i^{-1} \boldsymbol{x} \leq 1}\}$ assign larger probabilities the closer a donor is to the donee. Under some regularity conditions the latter will minimize the MISE of kernel estimators approximately for large p (Härdle, 1990). The difference in scaling between the uniform and Epanechnikov kernels is removed by assigning them the same canonical form (Marron and Nolan, 1989).

We denote donees close to (or at) the boundary of the convex hull of $\{\boldsymbol{x_j}, j \in S_{i,k}\}$ as boundary donees, and other donees as interior donees. We first consider interior donees. Special features for boundary donees are discussed in Section 2.4.

Let $\widehat{\boldsymbol{x}}_i$ denote the value $\boldsymbol{x}_j$ of the randomly selected donor j to donee i. In a pool $S_{i,k}$ the $\boldsymbol{x}_j$ values of the donors usually scattered unevenly around the donee $\boldsymbol{x}_i$. The expectation of $\widehat{\boldsymbol{x}}_i$

$$E[\widehat{\boldsymbol{x}}_i] = \sum_{j=1}^{k} \lambda_{i,j} \boldsymbol{x}_j \tag{6}$$

will usually differ from the donee value $\boldsymbol{x}_i$, so $\widehat{\boldsymbol{x}}_i$ will have an individual bias

$$B[\widehat{\boldsymbol{x}}_i] = E[\widehat{\boldsymbol{x}}_i] - \boldsymbol{x}_i = \sum_{j=1}^{k} \lambda_{i,j} \boldsymbol{x}_j - \boldsymbol{x}_i. \tag{7}$$

Similarly we define the individual bias of $\widehat{y}_i$ as

$$B[\widehat{y}_i] = E[\widehat{y}_i] - y_i = \sum_{j=1}^{k} \lambda_{i,j} y_j - y_i. \tag{8}$$

The individual properties of (7) and (8) are often directly related to overall properties, where e.g. the bias of a sample total is the sum of the imputed values individual biases. Asymptotic results (Silverman, 1986) give that the bias (8) tends to zero as $p \to \infty$ if $\frac{k}{p} \to 0$, that the variance tends to zero if $k = O(p)$, and that the MSE tends to zero if $k = O(p^{4/(4+q)})$.

The variables $\boldsymbol{x}$ and y are often locally linearily related. By making (7) equal to zero through calibration we hope to reduce (8) and thus the total

bias. The calibrated $\lambda^*_{i,j}$ may for each i be found through minimization of a Lagrange function (Pettersson, 2013, p. 8)

$$min \sum_{j=1}^{k} L(\lambda^*_{i,j} - \lambda_{i,j}) + \mathbf{\Lambda_1}(\sum_{j=1}^{k} \lambda^*_{i,j}(\boldsymbol{x}_j - \boldsymbol{x}_i)) + \Lambda_2(\sum_{j=1}^{k} \lambda^*_{i,j} - 1) \qquad (9)$$

with respect to $\mathbf{\Lambda_1}$,Λ_2 and $\lambda^*_{i,j}, j = 1, ...k$, where $L()$ is a distance function and $\mathbf{\Lambda_1}$ and Λ_2 are Lagrange multipliers. As long as there are possible donors at all sides of $\boldsymbol{x}_i$, i.e. $\boldsymbol{x}_i$ belongs to the interior of the convex hull of $\{\boldsymbol{x_i}, i \in S_{i,k}\}$, it is possible to obtain donor selection probabilities whích fully eliminates the individual bias of a donee. Features to counteract the bias for boundary donees are discussed in Section 2.4.

2.3 Deciding donor pools

The properties of final estimates are strongly influenced by the choice of pool size and shape. We mimic several suggested approaches in kernel estimation for simultaneously selecting h_i and k_i, and apply them to imputation. The choice is always a trade-off since a small h_i (or k_i) leads to small bias while a large h_i (or k_i) leads to small variance.

In four approaches we first determine the pool size h_i, from which k_i is indirectly determined. In the first approach we use a simple rule-of-thumb based on distributional assumptions (Scott, 1992, p. 152); secondly we use least-squares cross-validation (Duong and Hazelton, 2005, p. 489); thirdly we use a plug-in approach (Duong and Hazelton, 2003, p. 24); and finally smoothed cross-validation (Duong and Hazelton, 2005, p. 489). To ensure at least one donor in each pool we use the restriction that $k_i \geq q + 1, \forall i$.

We also use a nearest neighbour approach where we first decide the number of donors $k_i \approx p^{q/(q+2)}$, from which the pool size is indirectly determined. This approach may find donors that are better matched to the donee in dense regions and easily ensures that no donee get zero donors.

2.4 Boundary adjustment of donor pools

We start by denoting a donee i as a boundary donee if $\boldsymbol{x}_i$ lies outside or close to the boundary of the convex hull of $\{\boldsymbol{x_j}, j \in S_{i,k}\}$, and other donees as

interior donees, see Figure 1a for an example. Generally it is easier to find a relatively balanced donor pool for interior donees since the donee vector $\boldsymbol{x}_i$ is (closely) surrounded by $\boldsymbol{x}_j$ of potential donors.

Since the potential donors $\boldsymbol{x}_j$ tend to lie only on one side of boundary donees $\boldsymbol{x}_i$ the expected individual bias (7) is nonzero. Expected bias of boundary donees diminishes when $\boldsymbol{x}_i$ approaches the boundary, except for when NN is used, since the number of donors then is independent of the location of $\boldsymbol{x}_i$. It may still be possible to make the bias zero unless the donee has no potential donors at all on one side.

Two measures can be taken to, at least, mitigate (7) in this case. The first one is to remove the furthest (and thus most bias-threatening) potential donors and increase the selection probabilities of the closest ones by shrinking h_i (or k_i). Only the closest donors will then remain in the donor pool, so (7) is expected to approach zero although with higher variance. Secondly, when $\boldsymbol{x}$ has at least two dimensions it is possible to change G_i so that the donor pool becomes oblong along the boundary of $\boldsymbol{x}$. Donors along the boundary may then be substituted for donors that were orthogonally distant from the boundary in $S_{i,k}$, see Figure 1. Their donor probabilities $\lambda_{i,j}$ are also altered accordingly.

3 Kernel imputation algorithm

The algorithm is divided into three parts; each is described in Sections 3.1-3.3 respectively. The algorithm is described for one donee and one imputation round. The initation section (A1) consists of initial donor selection and boundary definition. In the calibration section (A2) an attempt to eliminate the individual bias in $\widehat{\mathbf{x}}_i$ is made through restricted linearization of the donor selection probabilities. The boundary section (A3) aims at reducing the individual bias in $\widehat{\mathbf{x}}_i$ of boundary units by changing H_i. It is always ensured that $k_i \geq q + 1$.

3.1 (A1) Initiation

Suppose we are going to impute y_i for a certain i. For notational convenience we first standardize the data set so that mean and variance is equal to $\overline{\boldsymbol{x}} = \mathbf{0}$ and $\widehat{\boldsymbol{\Sigma}}_i = \boldsymbol{I}$. We then calculate distances $D_{H,i,j}$ for all $j = 1, ..., p$ units, and find h_i and k_i. In the NN approach the donor pool $S_{i,k}$ border is interpolated

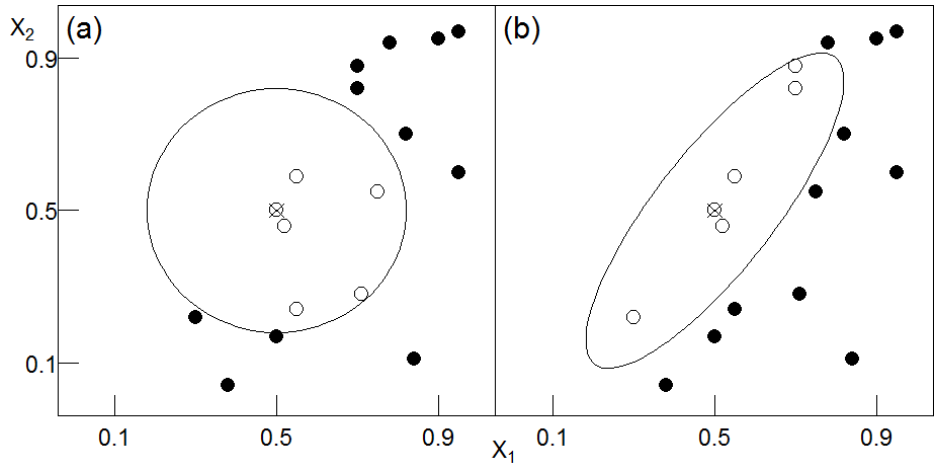

Figure 1: Donor pool of boundary unit decided by $(\boldsymbol{x}_j - \boldsymbol{x}_i)G_i^{-1}(\boldsymbol{x}_j - \boldsymbol{x}_i)^t < h$. *(a) Original donor pool with* $G = \boldsymbol{I}$*; (b) reoriented donor pool with* $G \neq \boldsymbol{I}$. *Donees are represented by (⊗), potential donors by (○), and non-potential donors by (●).*

between the k_ith and (k_i+1)th potential donor. If $G_i = \boldsymbol{I}$ the donor set $S_{i,k}$ is identified as the k_i complete units for which $D_{H,i,j} \leq 1$. The k_i donors are assigned nonzero donor selection probabilities $\lambda_{i,j}$ as in (4).

If $\overline{\boldsymbol{x}}_j = \boldsymbol{x}_i$ so that the donee has no individual bias, A1 is terminated; see Appendix A. Otherwise a boundary matrix Q_i, to be defined below, is used to test whether unit i lies at (or close to) the boundary of the donor pool, which means that none (or few) of the $\boldsymbol{x}_j$ are located on one side of $\boldsymbol{x}_i$. Let the boundary vector

$$Q_i^{(vector)} = \frac{\overrightarrow{\overline{\boldsymbol{x}}_j \boldsymbol{x}_i}}{\left|\overrightarrow{\overline{\boldsymbol{x}}_j \boldsymbol{x}_i}\right|} = \frac{1}{k_i} \sum_{j=1}^{k_i} \frac{(\boldsymbol{x}_j - \boldsymbol{x}_i)}{\sqrt{D_{H,i,j}}} \tag{10}$$

be the normed normal from the mean $\overline{\boldsymbol{x}}_j$ of the k_i donors in $S_{i,k}$, and let $Q_i^{(plane)}$ be a matrix spanning the $(q-1)$ dimensional normed boundary plane (which is a single vector if $q = 2$) orthogonal to $Q_i^{(vector)}$, so that $Q_i = \left[Q_i^{(vector)} \quad Q_i^{(plane)}\right]$, see Figure 2.

Also let $Q_{i,j}^{(vector)} = \overrightarrow{\boldsymbol{x}_j \boldsymbol{x}_i}$ be vectors from donee i to the k_i potential donors. To be able to determine whether the donors lie above or below the

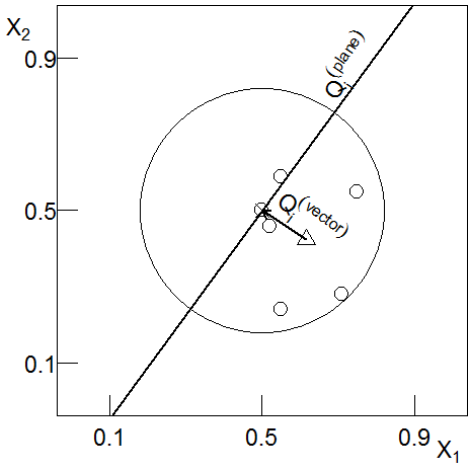

Figure 2: Initial detection of boundary donees. The donee is represented by (⊗), potential donors by (○), and the mean of potential donors by (△).

boundary plane, we first decide η as the minimum acceptable proportion, out of the k_i potential donors in the donor pool, that may lie either above or below the boundary plane. We then let

$$a_j = \begin{cases} 1/k_i & \text{if } Q_i^{(vector)} \cdot Q_{i,j}^{(vector)} > 0 \text{ (above)} \\ 0 & \text{if } Q_i^{(vector)} \cdot Q_{i,j}^{(vector)} = 0 \text{ (on)} \\ -1/k_i & \text{if } Q_i^{(vector)} \cdot Q_{i,j}^{(vector)} < 0 \text{ (below)} \end{cases} \quad \text{for } j = 1, ..., k_i,$$

where $Q_i^{(vector)} \cdot Q_{i,j}^{(vector)}$ are scalar products. Donee i is then definied to be a boundary unit if $\left|\sum_{j=1}^{k_i} a_j\right| > |2\eta - 1|$.

3.2 (A2) Calibration of selection probabilities

By calibrating the donor selection probabilities the individual bias (7) can often be almost eliminated. In order to prevent individual selection probabilities to become very large we introduce a constraint $\lambda_{i,j} \leq max(\lambda^{max}, 1/k_i)$ where $\lambda^{max} \in (0, 1]$. Setting $\lambda^{max} = 1$ means no restriction while the strongest possible restriction $\lambda^{max} \leq 1/k_i$ results in equal selection prob-

abilities for all k_i donors. When possible, the following steps calibrate $\lambda_{i,j} \forall j$, and otherwise leave them unchanged.

1. Set $\lambda_i^{sum} = 1$ and $\lambda_i^{max} = max(\lambda^{max}, \frac{1}{k_i})$.

2. Minimize $L(\lambda_{i,j}^* - \lambda_{i,j})$ subject to $\{\sum_{j=1}^{k} \lambda_{i,j}^* (\mathbf{x}_j - \mathbf{x}_i)\} = 0$ and $(\sum_{j=1}^{k} \lambda_{i,j}{}^*) = \lambda_i^{sum}$, where $L()$ is a distance function. With the Euclidean distance $L()^2$ the calibrated selection probabilities are given by
$\boldsymbol{\lambda}_i^* = -[\boldsymbol{\lambda}_i [W^t]^{-1} W^{-1} + [0\ 0\ 1][W^{-1}]^t]/2$, where $W = \begin{bmatrix} \boldsymbol{x}_1 - \boldsymbol{x}_i & 1 \\ \cdots & \\ \boldsymbol{x}_{k_i} - \boldsymbol{x}_i & 1 \end{bmatrix}$.

3. If the solution is singular, set $\lambda_{i,j}^* = \lambda_{i,j}$ and go to step 7.

4. If any $\lambda_{i,j}^* < 0$, set those $\lambda_{i,j}^* = 0$ and temporarily remove unit(s) j from the calculations. Then return to step 2.

5. If $max(\lambda_{i,j}^*) > \lambda_i^{(max)}$, set that $\lambda_{i,j}^* = \lambda_i^{(max)}$ and temporarily remove unit j from the calculations. Then set $\lambda_i^{(sum)} = \lambda_i^{(sum)} - \lambda_i^{(max)}$ and return to step 2.

6. Restore the units that were temporarily removed from the calculations in steps 4 and 5.

7. Use $\lambda_{i,j}^*$ as donor selection probabilities.

3.3 (A3) Bias reduction through pool shape and width

To reduce individual bias (7) of boundary donees the donor pool $S_{i,k}$ the pool is reoriented (by changing the matrix G_i) or shrunk (by shrinking the scalar h_i). The maximum number of times (c) the two features may be applied on a donee i is equal to the number of auxiliary variables q; see also Appendix A.

The total shrinkage after q processings is determined by a constant $\alpha \in [0, 1]$, where $\alpha = 0$ is the largest possible and $\alpha = 1$ is no shrinkage. The possible shrinkage is curtailed since each new donor pool is restricted to contain at least $q + 1$ potential donors.

The degree of reorientation is determined by a constant $\beta \in [0, \infty]$, where $\beta = 0$ is no reorientation, and $\beta = \infty$ is full reorientation such that the norm of $Q_i^{(vector)}$ becomes zero.

1. Set $Q_i = \left[Q_i^{(vector)} p^{-\beta} \quad Q_i^{(plane)} p^{\beta/(q-1)} \right]$ to change the boundary matrix according to the location of the donors.

2. Set $G_i = (\left(Q_i^T Q_i G_i^{-1} \right))^{-1}$ to adjust the pool shape through the boundary matrix.

3. In the nearest neighbour approach, to adjust the width, set $k_i = k_i p^{\alpha(q-c)/q}$ and then determine h_i. Else, set $h_i = h_i p^{\alpha(c-q)/q^2}$.

4. Use $H_i = h_i \cdot G_i$ as in A1 to find the new set of k_i donors.

4 Simulation study

In the simulations we illustrate the kernel imputation algorithm with its different features using a bivariate auxiliary variable. Simulations with a trivariate auxiliary gave similar results and are not shown here. We study different ways of finding the donor pools, and comparisons are made to competing imputation methods. Similar simulations with a univariate auxiliary variable may be found in Pettersson (2012; 2013). In line with those studies we hypothesize that including the features will help remove bias, and also improve variance estimation. The effect on the variance itself is less obvious. Since the NN approach puts relatively more focus on locally adapting the donor pools, it should be more able to reduce bias, while the other approaches may be relatively more shifted towards variance reduction with uniform auxiliaries, although the differences among them might not be very pronounced.

4.1 Setup of simulation study

A population of N=1600 units is constructed. Two auxiliary variables are generated from independent uniform distributions, $x_t \sim U(0,1), t = 1, 2$. Three study variables $y_u = f_u(x_1, x_2) + e_u, u = 1, 2, 3$, are generated from the auxiliaries, where $f_1 = \Phi^{-1}(x_1) + \Phi^{-1}(x_2)$, $f_2 = (\Phi^{-1}(x_1))^2 + (\Phi^{-1}(x_2))^2$

and $f_3 = sin(\pi x_1) + sin(2\pi x_2)$, and Φ is the standard normal distribution. The expected values of these random variables are 0, 2, $2/\pi$ for u=1, 2 and 3 and the variances 2, 4 resp 1-$4/\pi^2$. We finally add random noise terms, generated from independent normal distributions $e_u \sim N(0, Var[f_u])$. The population is described in Table 1.

Table 1: Means, variances and correlations in the simulated population.

Variable	x_1	x_2	y_1	y_2	y_3
Mean	0.50	0.49	0.03	2.14	0.67
Variance	0.001	0.001	4.143	8.111	1.073
Correlation to x_1	1	0.012	0.482	-0.013	-0.019
Correlation to x_2	0.012	1	0.516	-0.016	-0.490

We draw 1000 independent samples of size $n = 400$. To avoid very few or many missing observations we independently create exactly 50 % nonresponse on y_u in each sample with Poisson sampling. For each unit we draw a value $z \sim U(0, 1)$, and then let the 200 units in each sample for which $(z(x_1 + 2x_2)/3)^{1/4}$ was largest become nonresponders.

The missing data is imputed $B = 20$ times. We compare the methods described in Section 2.3 for selecting donor pool size. The nearest neighbours are found by Mahalanobis distances. We also compare choices of donor pool size. These are fixed to be k=1, 2, ..., 30 as p=200. As more donees are imputed, the number of potential donors increase. The size of the donor pool is then set to $[Ap^{4/(4+q)}]$, where [ ] denotes the integer part and A is determined so that this holds also for p=200. Four other methods suggested in the literature are also included for comparison. In the rule-of-thumb (RT), least-squares cross-validation (CV), plug-in (PI), and smoothed cross-validation (SC) approaches all 400 sampled units are used when the h is estimated. To account for the fact that the number of potential donors varies, h is adjusted by the factor $(p/400)^{1/(4+q)}$.

We use all 16 combinations of the four features in our proposed algorithm to impute the missing data. Each method is denoted by the components it

contains. The basic method (U) uses uniform selection probabilities and has no features included. In the initiation section A1 Epanechnikov selection probabilities (E) can be used instead of the uniform probabilities. The selection probabilities may be calibrated by solving the Lagrange equation (L) in A2 using the Euclidean distance function. In A3 the matrix G of donee units at the boundary (as detected in A1 or A2) may be modified so that the donor pool is reoriented (R), and h may also be shrunk (S). For boundary donees we fix the parameters for shrinkage and reorientation at $\alpha = .97$ and $\beta = .03$. This may be a conservative choice but should be sufficient to illustrate the effect of features R and S.

Comparisons are made to what an analysis made on the complete data set without any missing values would have given (CD) and to what an analysis based only on the complete cases would have given if they were assumed to be the full sample (CC). We impute a single nearest neighbour ($1NN$) donor, which would be the deterministic outcome if our method was used with the donees always obtaining the values of the closest donor units. Comparison is also made to imputation methods which draw from Bayesian predictive distributions. First we use multiple linear regression imputation from the R-package MICE (Buuren van and Groothuis-Oudshoorn, 2011), both with random model donors (REG^{mod}) (Rubin, 1987, p167) and predicitive mean matching (Rubin, 1987, p168) real donors (REG^{real}). These linear models should fit imputation of y_1 well. Imputation is also made with random model donors (SPL^{mod}) and predictive mean matching real donors (SPL^{real}) by regressing the study variables on restricted cubic spline transformations of the auxiliaries on replacement samples using the R-package Hmisc (Harrell, 2010). The R-package sbgcop (Hoff, 2010) is used for imputation when semiparametrically estimating a Gaussian copula (COP^{mod}) with the univariate marginal distributions treated as nuisance parameters, and draws made from the posterior distribution of the correlation matrix based on a scaled inverse-Wishart prior distribution and an extended rank likelihood.

For all the compared estimation techniques we calculate the same estimates and measures of goodness of fit. First, we estimate the population means $\overline{y}$ of the three study variables. For each of them we calculate root mean squared error RMSE= $\sqrt{1000^{-1}\sum_{g=1}^{1000}(\widehat{\overline{y}}_g - \overline{y})^2}$, where $\widehat{\overline{y}}_g = \sum_{b=1}^{B}\widehat{\overline{y}}_{b,g}$ is the overall estimated mean in the B imputed datasets; BIAS= $1000^{-1}\sum_{g=1}^{1000}(\widehat{\overline{y}}_g - \overline{y})$; variance VAR= $1000^{-1}\sum_{g=1}^{1000}(\widehat{\overline{y}}_g - 1000^{-1}\sum_{g=1}^{1000}\widehat{\overline{y}}_g)^2$;

standard error SE=$\sqrt{VAR}$; and relative error of estimated variance REEV= $(\widehat{VAR} - VAR)/VAR$, where $\widehat{VAR} = 1000^{-1} \sum_{g=1}^{1000} (\widehat{\sigma}_{y_g} + (B+1)B^{-1}(B-1)^{-1} \sum_{b=1}^{B} (\widehat{\overline{y}}_g - \widehat{\overline{y}}_{b,g})^2)$ is the average estimated variance from equation (3). In our comparisons we divide bias and standard error by the root mean squared error of the complete data RMSE(CD), and multiply all figures by 100.

4.2 Bias

4.2.1 Comparison of features and donor pool approaches

The estimated biases are given in Table 2. For many donor pool approaches the four features (E, L, R and S) improve estimation, so that the feature combination ELRS is among the least biased. For $\overline{y}_1$ and $\overline{y}_3$ PI, CV and SC show rather small bias and are relatively unaffected by the features. L seems to have the strongest, E the second strongest, and S the weakest effect. Effects are generally additative. One exception occurs when estimating $\overline{y}_2$ where L is better used without E, except for NN. Adding feature L with RT also causes bias to become positive instead of negative.

The bias reductive effect from including each of the features is relatively evident for NN except when a small k is chosen, see Figure 3. Estimates of $\overline{y}_1$ are least biased when $k \approx 4$ irrespective of the included features, and unbiased estimation is not attainable. By including feature L when estimating $\overline{y}_2$ and $\overline{y}_3$ almost unbiased estimation is enabled if $k \geq 13$ (depending on which other features that are included).

4.2.2 Comparison to other methods

In the simple monotone case with y_1 the model donor linear regression REG^{mod} with random errors estimates $\overline{y}_1$ without bias, see Table 3. The real donor regression and the two spline regression methods attain a slightly lower bias than NN and $1NN$, while COP^{mod} only manages to remove about half of the bias compared to the case CC without any imputation. When estimating $\overline{y}_2$ the comparison methods show similar or larger bias compared to CC, while NN^{ELRS} is almost unbiased. The real donor methods show smallest bias when estimating $\overline{y}_3$.

Table 2: Average bias with features E, L, R or S. RT=Rule-of-thumb; LS=Least-squares cross-validation; PI=Plug-in; SC=Smoothed cross-validation; NN=Nearest neighbour with $k = 15$. 1000 simulations and n=400. Bold values are not significantly different from zero.

	$\overline{y}_1$					$\overline{y}_2$					$\overline{y}_3$				
	NN	*RT*	*PI*	*CV*	*SC*	*NN*	*RT*	*PI*	*CV*	*SC*	*NN*	*RT*	*PI*	*CV*	*SC*
U	42	39	19	20	20	-181	-145	-99	-97	-104	51	37	17	18	19
R	40	37	20	22	20	-163	-122	-94	-93	-98	41	26	17	19	17
S	40	38	19	21	21	-171	-145	-99	-98	-104	47	38	18	19	20
RS	38	36	20	22	20	-153	-117	-92	-92	-96	38	25	17	19	16
E	37	39	21	21	19	-157	-138	-90	-93	-94	43	34	18	20	17
ER	37	35	18	22	18	-149	-115	-89	-94	-88	38	18	15	18	14
ES	39	37	17	19	17	-169	-146	-94	-101	-98	48	32	14	18	15
ERS	35	37	22	23	19	-137	-107	-83	-88	-86	34	21	18	21	15
L	33	28	22	21	22	-16	52	-69	-84	-58	15	**-6**	17	18	15
LR	32	27	23	23	21	**7**	76	-60	-78	-47	**7**	-14	16	19	12
LS	35	28	22	21	22	-14	52	-70	-85	-59	17	**-6**	17	18	16
LRS	30	26	23	23	22	**5**	76	-60	-79	-47	**6**	-15	17	19	12
EL	30	24	19	19	19	-12	71	-78	-96	-62	13	-18	16	18	13
ELR	30	22	20	23	20	**9**	97	-70	-85	-47	**7**	-26	15	19	11
ELS	32	24	18	20	19	-10	73	-78	-94	-62	15	-18	15	18	13
ELRS	28	22	20	24	19	**9**	97	-71	-86	-50	**6**	-26	15	19	10

4.2.3 Conclusions

Naturally, the model donor linear regression imputation method is good at providing unbiased estimation when y_1 with its monotonic relation to the auxiliaries is imputed, but most other imputation methods show acceptable levels of bias. The situation is opposite with y_2 where all imputation methods result in larger biased estimates of $\overline{y}_2$, except for *NN* which enables almost unbiased estimation if a sufficiently large k and (at least) feature L is included in the algorithm. Generally all four of our features are able to improve

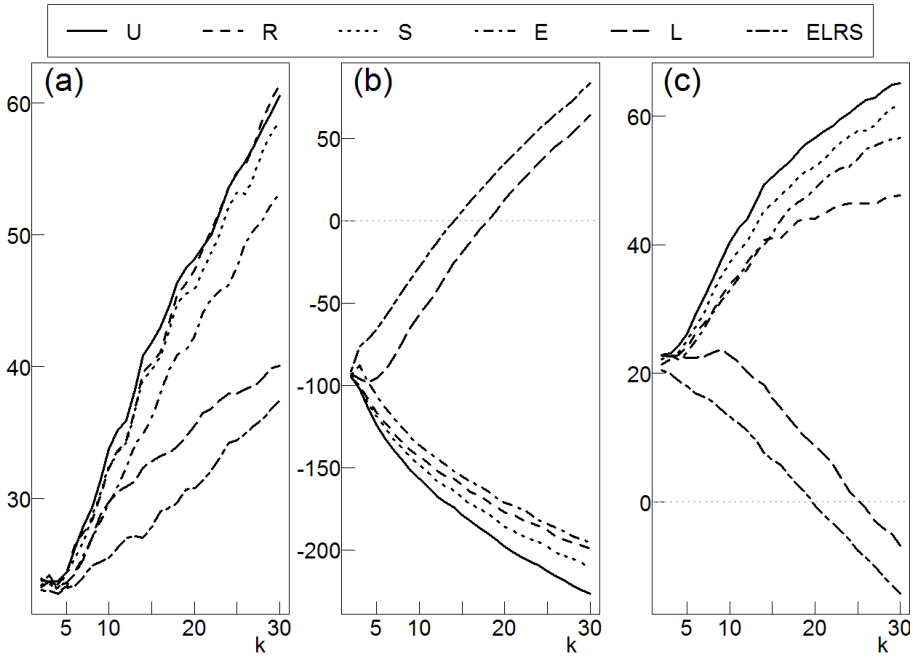

Figure 3: Average bias of $\widehat{\overline{y}}_1$ (a); $\widehat{\overline{y}}_2$ (b); $\widehat{\overline{y}}_3$ (c); for nearest neighbour (NN) with k donors and feature combinations E, L , R, S and ELRS. 1000 simulations and n=400.

Table 3: Average bias for complete data (CD), complete cases (CC), nearest neighbour (1NN), model and real donor linear (REG^{mod} and REG^{real}) and splines (SPL^{mod} and SPL^{real}) regression, Gaussian copula (COP^{mod}), and our method (NN_{ELRS}) with $k = 15$ and all features. 1000 simulations and n=400. Bold values are not significantly different from zero.

	CD	CC	$1NN$	REG^{mod}	REG^{real}	SPL^{mod}	SPL^{real}	COP^{mod}	NN^{ELRS}
$\overline{y}_1$	**-3**	236	27	**3**	18	13	22	128	28
$\overline{y}_2$	-9	-84	-79	-131	293	-126	-83	-123	**9**
$\overline{y}_3$	**-4**	-118	24	35	8	48	20	-84	**6**

estimation. Among the donor pool approaches CV is least and RT is most affected by the features. The effect on NN is most obvious if k is relatively

large. L is the most important and E seems to be the second most important feature in bias reduction. Both S and R contribute to bias reduction. Since the parameters α and β were chosen conservatively their contributions might be increased.

4.3 Standard error

4.3.1 Comparison of features

In Table 4 the standard errors of the methods are given. They are relatively unaffected by the type of donor pool approach and which features are included. For estimates of $\overline{y}_2$ with *NN* and *RT* standard error tend to increase slightly when feature L is included.

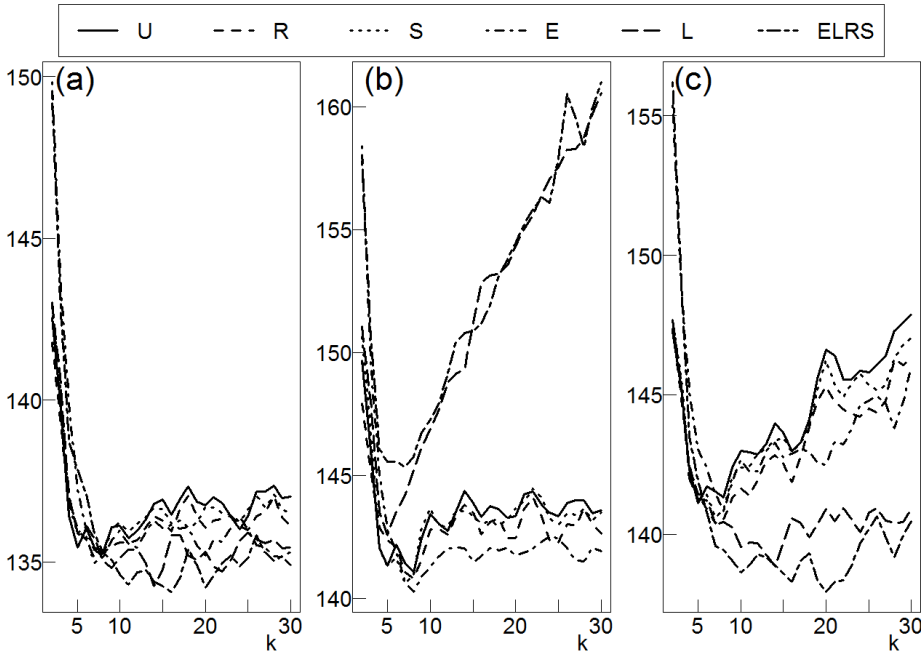

Figure 4: Standard error of $\widehat{\overline{y}}_1$ (a), $\widehat{\overline{y}}_2$ (b), and $\widehat{\overline{y}}_3$ (c), for nearest neighbour (NN) with k donors and feature combinations E, L, R, S and ELRS. 1000 simulations and n=400.

Table 4: Standard error with features E, L, R or S. RT=Rule-of-thumb; LS=Least-squares cross-validation; PI=Plug-in; SC=Smoothed cross-validation; NN=Nearest neighbour with $k = 15$. 1000 simulations and n=400. Any two values differing by at least 4 units are (almost) always signficantly different in a pairwise test not adjusted for multiplicity.

	$\overline{y}_1$					$\overline{y}_2$					$\overline{y}_3$				
	NN	*RT*	*PI*	*CV*	*SC*	*NN*	*RT*	*PI*	*CV*	*SC*	*NN*	*RT*	*PI*	*CV*	*SC*
U	137	134	138	138	136	143	136	141	142	139	143	139	141	142	139
R	136	134	137	137	136	143	137	140	141	140	142	138	140	140	139
S	136	134	138	138	135	143	135	141	141	139	143	139	141	141	139
RS	136	135	138	137	136	143	137	142	141	140	142	138	140	141	139
E	136	135	138	140	137	141	136	142	145	141	142	139	140	142	140
ER	136	130	139	143	135	141	137	146	151	142	142	136	143	146	141
ES	136	131	139	141	135	142	138	144	150	142	143	138	142	144	140
ERS	135	135	138	139	137	141	136	142	145	142	142	138	140	142	140
L	136	140	138	139	138	153	150	143	143	142	141	136	139	140	139
LR	136	140	139	138	138	154	152	144	143	143	140	137	140	140	139
LS	136	140	138	138	138	153	151	143	143	142	140	137	140	141	139
LRS	136	140	139	138	138	153	152	145	142	143	140	136	141	140	139
EL	134	135	140	143	138	150	150	147	152	145	139	135	141	146	140
ELR	134	135	139	143	138	151	150	148	151	146	138	135	143	146	141
ELS	134	135	139	141	137	150	150	146	151	145	139	135	142	144	139
ELRS	134	135	140	143	138	151	150	149	151	146	138	135	143	146	140

For *NN* a small k always leads to a relatively large standard error irrespective of the features in Figure 4. The features additatively reduce variance except for L with $\overline{y}_2$, so that ELRS is preferred for $\overline{y}_1$ and $\overline{y}_3$, and ERS for $\overline{y}_2$ (not shown in Figure). For $\overline{y}_1$ the differences are relatively small between the features and are relatively independent of the size of k. The situation is similar for $\overline{y}_2$ except for L. For $\overline{y}_3$ standard error tends to increase as k increases, except when L is included.

4.3.2 Comparison to other methods

The standard error of estimates of $\overline{y}_1$ and $\overline{y}_3$ in Table 5 is improved for all methods, except for *1NN*, when compared to complete cases *CC*. For $\overline{y}_2$ the methods are comparable to *CC*, except for *1NN* which is slightly higher, and REG^{real} which is much higher.

Table 5: Standard error for complete data (CD), complete cases (CC), nearest neighbour (1NN), model and real donor linear (REG^{mod} and REG^{real}) and splines (SPL^{mod} and SPL^{real}) regression, Gaussian copula (COP^{mod}), and our method (NN^{ELRS}) with $k = 15$ and all features. 1000 simulations and n=400. Two values differing by at least 5 units are signficantly different.

	CD	*CC*	*1NN*	REG^{mod}	REG^{real}	SPL^{mod}	SPL^{real}	COP^{mod}	NN^{ELRS}
$\overline{y}_1$	100	152	155	125	133	127	129	129	134
$\overline{y}_2$	100	152	162	150	379	153	144	148	151
$\overline{y}_3$	100	156	158	147	146	148	141	149	138

4.3.3 Conclusions

For standard error it is not so important which type of donor pool approach or which features that are included in kernel imputation, except for *NN* where a very small k is bad. The effect on standard error when including different features may be different depending on the type of data, see Figure 4. Two of the real donor methods (*NN* and REG^{real}) may give higher standard errors. This is reasonable for *NN* since it is a deterministic method which neglects the imputation variance. The failure of REG^{real} when estimating $\overline{y}_2$ is probably a combination of a similar previously noted problem (Buuren van and Groothuis-Oudshoorn, 2011, p19) with a badly suited linear imputation model.

Table 6: Relative error of estimated variance with features E, L, R or S. RT=Rule-of-thumb; LS=Least-squares cross-validation; PI=Plug-in; SC=Smoothed cross-validation; NN=Nearest neighbour with $k = 15$. 1000 simulations and n=400. Bold values are not significantly different from zero.

	$\overline{y}_1$					$\overline{y}_2$					$\overline{y}_3$				
	NN	*RT*	*PI*	*CV*	*SC*	*NN*	*RT*	*PI*	*CV*	*SC*	*NN*	*RT*	*PI*	*CV*	*SC*
U	-14	-9	-23	-27	-20	-10	4	-17	-24	-14	-6	**2**	-16	-21	-12
R	-13	-9	-22	-26	-19	-9	3	-16	-21	-12	-4	4	-14	-18	-10
S	-14	-9	-23	-27	-19	-11	4	-18	-24	-13	-8	**2**	-16	-20	-12
RS	-13	-10	-23	-26	-19	-9	3	-17	-21	-13	-5	**2**	-15	-19	-11
E	-14	-11	-27	-34	-24	-9	**2**	-25	-33	-20	-8	**2**	-19	-27	-17
ER	-14	**-2**	-29	-38	-21	-8	4	-29	-40	-19	-6	6	-23	-33	-16
ES	-14	-4	-29	-37	-22	-10	**2**	-28	-40	-20	-8	**2**	-23	-32	-16
ERS	-13	-10	-27	-33	-23	-9	4	-23	-31	-20	-6	3	-19	-26	-16
L	-11	-14	-25	-29	-23	-14	-9	-20	-26	-17	-5	3	-16	-21	-13
LR	-10	-14	-25	-27	-22	-14	-10	-20	-23	-16	-3	3	-15	-18	-12
LS	-11	-15	-24	-28	-22	-14	-10	-20	-25	-16	-4	3	-16	-21	-13
LRS	-10	-14	-26	-28	-23	-13	-10	-21	-23	-17	-3	3	-17	-19	-13
EL	-10	-7	-32	-40	-24	-11	-5	-32	-43	-24	-4	5	-24	-34	-16
ELR	-8	-7	-29	-38	-24	-10	-4	-31	-40	-23	**-1**	6	-24	-32	-17
ELS	-9	-6	-30	-38	-23	-11	-4	-31	-41	-23	-3	5	-23	-32	-15
ELRS	-9	-7	-31	-39	-24	-12	-4	-33	-40	-23	**-2**	5	-25	-33	-17

4.4 Relative error of estimated variance

4.4.1 Comparison of features

NN and *RT* have smaller REEV and the effects from adding the features are relatively small, and almost estimates $\overline{y}_3$ without bias, see Table 6. The REEV for *PI*, *CV* and *SC* is higher and is generally worsened (or unchanged) when any of the features are included.

In Figure 5 REEV is considerably increased if k is increased, with a

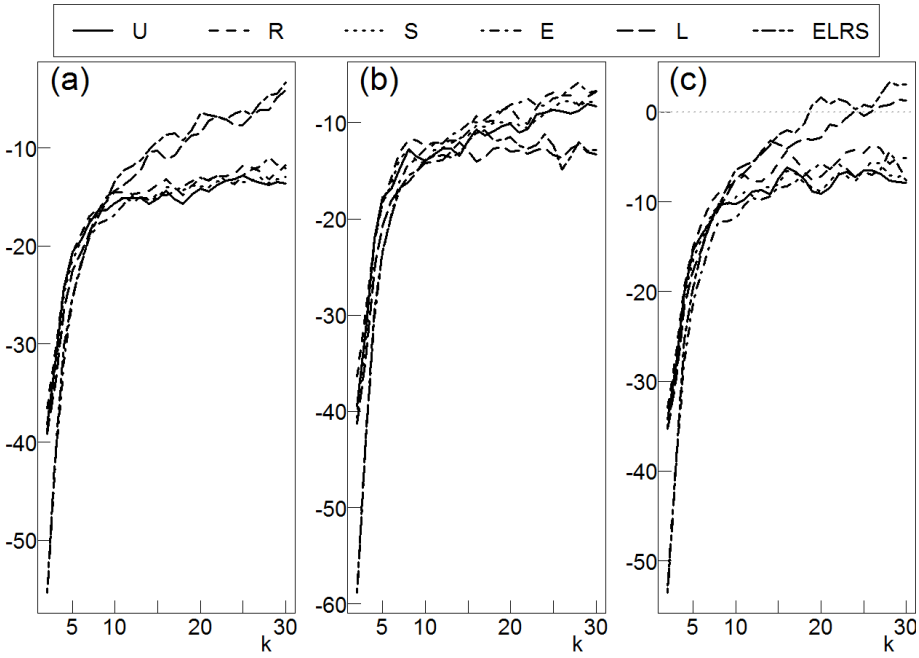

Figure 5: Relative error of estimated variance of $\widehat{\overline{y}}_1$ (a); $\widehat{\overline{y}}_2$ (b); $\widehat{\overline{y}}_3$ (c); for nearest neighbour (NN) with k donors and feature combinations E, L, R, S and ELRS. 1000 simulations and n=400.

strong leverage for including L when $\overline{y}_1$ or $\overline{y}_3$ is estimated. REEV is always underestimated except for when $\overline{y}_3$ is estimated with a large k and feature L is included.

4.4.2 Comparison to other methods

In Table 7 the model donor splines regression method SPL^{mod} almost shows an error-free estimation of variance for all estimates. Both CC and REG^{mod} are comparable to NN^{ELRS} with a large k, while *1NN* severly underestimates variance.

4.4.3 Conclusions

With few exceptions the variance is underestimated with any method, except for SPL^{mod}. For kernel imputation methods, REEV shows the same depen-

Table 7: Relative error of estimated variance for complete data (CD), complete cases (CC), nearest neighbour (1NN), model and real donor linear (REG^{mod} and REG^{real}) and splines (SPL^{mod} and SPL^{real}) regression, Gaussian copula (COP^{mod}), and our method (NN^{ELRS}) with $k = 15$ and all features. 1000 simulations and n=400. Bold values are not significantly different from zero.

	CD	*CC*	*1NN*	REG^{mod}	REG^{real}	SPL^{mod}	SPL^{real}	COP^{mod}	NN^{ELRS}
$\overline{y}_1$	-7	-13	-63	-7	-10	**2**	-26	-15	-9
$\overline{y}_2$	**-1**	-6	-66	-10	-13	**1**	-21	-16	-12
$\overline{y}_3$	**1**	-3	-59	-10	-23	**1**	-24	-18	**-2**

dence on donor pool size as the standard error did, where low (e.g. *NN* with large k) or high (e.g. *NN* with small k) gave rise to low or high REEV. This is also seen in the failure of the deterministic *1NN*, which does not account for the uncertainty about the imputed values.

5 Discussion

As in previous studies (Pettersson, 2012; 2013), due to its different features, the presented kernel imputation algorithm allows us to approach the goal of almost unbiasedness when estimating a population mean from a data set with missing values. This is in line with how similar features operate in kernel estimation; see e.g. Simonoff (1996). As in kernel estimation, it is not surprising that the donor pool size was strongly related to bias. Without any bias reduction features applied increasing donor pool sizes were associated with increased bias. For boundary donees an increased number of donors could worsen the already insufficient matching. When unbalanced, small donor pools might also result in bias because the donors could be badly located. Having few donors might also cause high variance and too low variance estimates.

The two methods with largest donor pools, *NN* and *RT*, initially also had largest bias but improved if the four features were added. While *RT* tended to overcompensate, *NN* was able to provide almost unbiased estimation given a large enough k. The other methods *PI*, *CV* and *SC* for selecting the donor pool was much less dependent on the exact feature setup. The small

differences between them may be due to the uniformly distributed auxiliary variables. Other type of auxiliaries may give other results; see e.g. Pettersson (2012).

Among the features, the local linearization from balancing with Lagrange calibration (L) enabled almost unbiased estimation when the relation between the auxiliaries and the study variable was nonlinear. Given the canonical fixed kernel variance, using Epanechnikov (E) instead of uniform (U) selection probablities meant that the donor pools became larger and created more possibilities for better balancing the donor pools. Reorientation (R) and shrinkage (S) gave good but small contributions. Their relative contributions could possibly be larger with higher dimensional auxiliaries with more severe boundary problems.

Standard errors were less affected by the donor pool size. The trade off to bias was not evident, which is rather surprising. However, as in other studies, the reduction of bias does lead to an improvement of variance estimation (Pettersson, 2012).

In the displayed simulations bias contributed largely to MSE in the *CC* case without imputation. The compared imputation methods were successful in removing most of the bias, except for the variable y_2 and for COP^{mod}. With nonlinearities in the data our algorithm was most effective in removing bias, especially for $\overline{y}_2$ where the other methods failed and the parametric methods even increased the bias as seen in Table 3.

Our algorithm is therefore very promising since it can handle nonlinearities very well and is only slightly less efficient when parametric imputation methods would do better. The simulations are relatively limited in scope, though other simulations not reported here show similar results. But in order to test the method further it is highly recommended that future studies investigate its behaviour with other nonresponse mechanisms, variable assocations, sample functions, and more auxiliaries.

6 References

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

Andridge, R.R. and Little, R.J.A. (2010). A review of hot deck imputation for survey non-response. International Statistical Review, 78, 40-64.

Chacon, J.E. and Duong, T. (2010). Multivariate plug-in bandwidth selection with unconstrained pilot matrices. Test, 19, 375-398.

Chu, C.K. and Cheng, P.E. (1995). Nonparametric regression estimation with missing data. Journal of Statistical Planning and Inference, 48, 85-99.

Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, 87, 376-382.

Duong, T. and Hazelton, M.L. (2003). Plug-in bandwidth matrices for bivariate kernel density estimation. Nonparametric Statistics. 15(1), 1730.

Duong, T. and Hazelton, M.L. (2005). Cross-validation bandwidth matrices for multivariate kernel density estimation. Scandinavian Journal of Statistics. 32, 485-506.

Epanechnikov, V. (1969). Nonparametric estimation of a multivariate probability density. Theory of Probability & Its Applications. 14, 153-158.

Fan, J. and Gijbels, I. (1996). Local Polynomial Modelling and its Applications. London: Chapman & Hall.

Feller, W. (1971). An introduction to probability theory and its applications (Second edition). Vol. 2. New York: Wiley.

Ghosh, M. and Meeden, G. (1997). Bayesian methods for finite population sampling. London: Chapman & Hall.

Harrell, F.E. (2010). Package 'Hmisc'. Available at: URL=http://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf (Accessed November 2012)

Hastie, T. and Tibshirani, R. (1996). Discriminant adaptive nearest neighbor classification. IEEE transaction on Pattern Analysis and Machine Intelligence, 18, 607615.

Hoff, P. (2010). Package 'sbgcop'. Available at: URL=http://cran.r-project.org/web/packages/sbgcop/sbgcop.pdf (Accessed November 2012)

Härdle, W. (1990). Applied nonparametric regression. New York: Cambridge university press.

Jones, M.C., Marron J.S. and Park B.U. (1991). A simple root n bandwidth selector. Annals of statistics, 19, 1919-1932.

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

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. International Statistical Review, 54, 139-157.

Little, R.J.A. and Rubin, D.B. (2002). Statistical analysis with missing data. New York: Wiley.

Lo, A.Y. (1988). A bayesian bootstrap for a finite population. The annals of statistics, 16, 1684-1695.

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

Marron, J.S. and Nolan D. (1989). Canonical kernels for density estimation. Statistics and Probability Letters, 7, 195-199.

Pettersson, N. (2012). Real donor imputation pools. Proceedings of the Workshop of Baltic-Nordic-Ukrainian network on survey statistics, 162-168.

Pettersson, N. (2013). Bias reduction of finite population imputation by kernel methods. To appear in Statistics in Transitions new series.

Rosenbaum, P.R. and Rubin, D.B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. The American Statistician, 39, 3338.

Rubin, D.B. (1976). Inference and missing data. Biometrika, 63, 581592.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Sain, S.R. and Scott, D.W. (1996). On locally adaptive density estimation. Journal of the American Statistical Association, 91, 15251534.

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

Scott, D.W. (1992). Multivariate Density Estimation: Theory Practice and Visualization. New York: Wiley.

Scott, D.W. and Terrell, G.R. (1987). Biased and unbiased cross-validation in density estmation. Journal of the American Statistical Association, 82, 1131-1146.

Siddique, J. and Belin, T.R. (2008). Using an Approximate Bayesian Bootstrap to multiply impute nonignorable missing data. Computational Statistics & Data Analysis, 53, 405-415.

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

Simonoff, J.S. (1996). Smoothing methods in statistics. New York: Springer.

Buuren van, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45, 1-67.

Wand, M.P. and Jones, M.C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. Journal of the American Statistical Association, 88, 520-529.

7 Acknowledgements

I would like to express my gratitude to my supervisor Professor Daniel Thorburn for his useful critiques of this research work.

A Appendix - The full algorithm

Start from the first donee, $i = r + 1$:

1. Set the processing indicator to $c = 0$ and apply A1. If A1 is stopped due to zero boundary bias (i.e. $\overline{\boldsymbol{x}}_j = \boldsymbol{x}_i$) go to step 5.
2. If donee unit i is detected as a boundary unit in A1, set $c = 1$ and go to step 4.
3. Apply A2. If the solution is non-singular (i.e. the last step in A2 is reached) or if $c = q$, go to step 5.
4. Apply A3. Then set $c = c + 1$ and go to step 3.
5. Use selection probabilities $\lambda_{i,j}$ to draw a unit j out of the k_i donors.
6. Replace the missing value y_i by a copy of $y_{i,j}$ from the drawn donor j.
7. If $i = n$, end the algorithm. Otherwise set $i = i + 1$, and return to step 1.