Multiple Kernel Imputation

A Locally Balanced Real Donor Method

Nicklas Pettersson

Abstract

We present an algorithm for imputation of incomplete datasets based on Bayesian exchangeability through Pólya sampling. Each (donee) unit with a missing value is imputed multiple times by observed (real) values on units from a donor pool. The donor pools are constructed using auxiliary variables. Several features from kernel estimation are used to counteract unbalances that are due to sparse and bounded data. Three balancing features can be used with only one single continuous auxiliary variable, but an additional fourth feature need, multiple continuous auxiliary variables. They mainly contribute by reducing nonresponse bias. We examine how the donor pool size should be determined, that is the number of potential donors within the pool. External information is shown to be easily incorporated in the imputation algorithm. Our simulation studies show that with a study variable which can be seen as a function of one or two continuous auxiliaries plus residual noise, the method performs as well or almost as well as competing methods when the function is linear, but usually much better when the function is nonlinear.

Key words: Bayesian Bootstrap; Boundary Effects; External Information; Kernel estimation features; Local Balancing; Pólya Sampling

©Nicklas Pettersson
ISBN 978-91-7447-699-6
Printed in Sweden by US-AB, Stockholm 2013
Distributor: Department of Statistics, Stockholm University

To my family

"Well, the way of paradoxes is the way of truth. To test Reality we must see it on the tight-rope. When the Verities become acrobats we can judge them."

Oscar Wilde (1854-1900) in *The picture of Dorian Gray*

List of included papers

I Bias reduction of finite population imputation by kernel methods
To appear in: Statistics in Transition new series

II Real donor imputation pools
Proceedings of the Workshop of the Baltic-Nordic-Ukrainian network on survey statistics, 2012.

III Kernel imputation with multivariate auxiliaries
(Submitted)

IV Informed kernel imputation
(Submitted)

Contents

1 Introduction

Many datasets are incomplete. The reasons for the missingness vary with the setting, e.g. failing measurement devices in an experiment, transmission errors when collecting data from secondary sources, or unit or item nonresponse in a survey due to refusal or skipping of a question in a questionnaire. The issue of handling missing data is thus universal, but in this thesis we typically refer to survey nonresponse, where the general trend has been an increasing amount of missing data over the last decades.

Measures can (and should) be taken to prevent, counteract and learn about the missing data, e.g. by following up nonrespondents in a survey. Despite such efforts one usually faces an incomplete dataset when starting out a statistical analysis. Since almost all statistical methods assume complete rectangular datasets, the question is how to make valid and efficient inferences from an incomplete dataset?

It may be tempting to exclude all units with missing values and simply use the smaller dataset with completely observed units. This is known as a complete cases (or listwise deletion) approach. There will always be a loss of precision with this approach due to the fact that the sample size is smaller than the intended one. If the intended sample size or the number of completely observed units is relatively small this may be of great importance. However, the potential nonresponse bias is usually a more serious problem and is also at the centre of this thesis.

Nonresponse bias can be viewed as a function of the amount of nonresponse and the mechanism(s) which lead to nonresponse. Rubin (1976a) formalized a model for the missing data mechanism with indicators of the missing values viewed as functions of the observed or the unobserved values. The simplest situation is when a missing completely at random (MCAR) mechanism is reasonable to assume, the bias would be zero and a complete case approach would only suffer from the loss of precision. But in practice this approach is often too simple and there is a high risk that the results will be biased. The key to efficient and unbiased estimation with an incomplete dataset lies in utilizing what we know to predict what we do not know, and a good method should be flexible enough to take all relevant and accessible information into account in an adequate way.

Typically we have access to design and register variables measured on all units in the population and study variables measured on the units in the sam-

ple. If the auxiliary variables are associated with the study variable(s) and the (unknown) response probabilities, they may be used to reduce the error in estimation. If they are only associated with the study variable(s) they can only improve the precision. A missing at random (MAR) mechanism states that the response probability is related to the observed (auxiliary) data, and this is generally a more reasonable assumption than MCAR. In Papers I-IV we assume a situation where a MAR assumption is reasonable, so that the distribution of the unknown missing values can be modelled from the known observed values. In Paper IV we also assume access to additional information as support. The third possible mechanism, not missing at random (NMAR), or missing not at random (MNAR), states that the probability of missingness depends on the missing values themselves, and may therefore not be modeled from our observed values but would need additional information to be uncovered, e.g. from a follow-up on nonrespondents. Bias would otherwise be irreducible. In practice a mixture of missingness mechanisms is probably acting simultaneously on the data, and we can only hope for partial reduction of nonresponse bias.

In sample surveys it is common to assume some kind of MAR and compensate for the unit nonresponse through weighting. The design weights, which under the chosen design are used to make the sample representative of the population, are then adjusted. In calibration weighting (Deville and Särndal, 1992), the weights are also adjusted so that resulting estimates comply with known (sub)population quantities. Weighting can counteract bias, but sometimes at the cost of efficiency. Estimates based on weights are also more sensitive to the chosen form for response probability modelling compared to e.g. likelihood methods (Little and Rubin, 2002), though there need not always be a direct trade-off between bias and variance (Little and Vartivarian, 2005).

Item nonresponse is often handled through imputation, where the missing values in the incomplete dataset are replaced by values that are generated under a missing data model. In a report from the EUREDIT project on evaluating methods for data editing and imputation the goal of imputation is well stated as: "Ideally, an imputation procedure should be capable of effectively reproducing the key outputs from a 'complete data' statistical analysis of the data set of interest." (Chambers, 2001, p.15). The properties which are most desirable depend on the goal of inference. The report lists the following properties (from hardest to easiest to achieve):

(1) Predictive Accuracy: The imputation procedure should maximize preservation of true values.

(2) Ranking Accuracy: The imputation procedure should maximize preservation of order in the imputed values.

(3) Distributional Accuracy: The imputation procedure should preserve the distribution of the true data values.

(4) Estimation Accuracy: The imputation procedure should reproduce the lower order moments of the distributions of the true values.

(5) Imputation Plausibility: The imputation procedure should lead to imputed values that are plausible.

In the survey context estimation and distributional accuracy are usually most relevant, while the preservation of true values seldom is, since the statistical statements in principle are never made at the unit level. The goals of imputation can be operationalized as: (I) In order to reduce bias and improve precision the imputations should be conditional on, in principle, all the observed variables. This also enhances the possibilities of preserving the association between missing and observed variables. (II) The imputation prediction model should take contextual and subject matter knowledge about variables being imputed into account. Unless motivated by it, the model should avoid excessive extrapolation beyond the range of the data. (III) If we are interested in preserving the associations between missing variables the imputations should be multivariate. (IV) To be able to provide valid estimates of a wide range of estimands, the missing values should be drawn from their predictive distribution, as to preserve the distribution and enhance variance estimation. (V) Estimation of sample variance should take into account that the imputed values are not the true unobserved values. (Little, 1988; Little and Rubin, 2002).

Not all methods in the plethora of imputation methods meet all these requirements. As with other prediction methods they can be classified into parametric, semiparametric or nonparametric ones. A correct parametric method is naturally most efficient, whereas semi- and nonparametric methods with fewer and weaker assumptions become more robust when the assumptions fail, especially when sample sizes are large. The actually imputed values can be classified as model-donor values, which are derived from a

(behavioural) model and thus are values that are non-observable and even impossible in a real life world; or real-donor values, which are derived from a set of observed values on respondents and thus are natural possible values (Laaksonen, 2000). Model-donor values may be preferred if the observed values do not cover all potential values exhaustively, e.g. if there are no respondents within an area, or if the share of respondents is low. Parametric methods often employ model-donors (Tanner and Wong, 1984; Schafer, 1997) but also real-donors (Rubin, 1987; Little, 1988; Heitjan and Little, 1991; Laaksonen, 2000). Nonparametric methods often employ real-donors (Sande, 1983; Andridge and Little, 2010) but can also be a weighted mixture of real-donors (Kim and Fuller, 2004), which thus is best described as a model-donor approach.

The nonparametric real-donor hot (and cold) deck imputation methods originate from the time of punch card machines. Nowadays these methods typically refer to when the conditional predictive model for a unit with missing value (the donee or recipient) is obtained through matching to a donor pool of nearest neighbour units with observed values (the donors). Hot deck imputation refers to methods where the donors are found within the same sample as the donees, as in Papers I-IV, and cold deck imputation to methods where the donors are found from outside of the sample, as in Paper IV. In Paper IV we use a common donor approach where several values are imputed simultaneously from the same unit, as a means to try to preserve the associations between the imputed variables.

Exact matching (Fechner, 1966 [1860]) on continuous auxiliaries is impossible, and though sometimes practical, categorization of such variables would introduce false boundaries in the data. A variety of metrics have been used in the survey context (Rubin, 1976b; Little and Rubin, 2002). Two common approaches are propensity score matching (Rosenbaum and Rubin 1983) and matching based on the Mahalanobis distance (Cochran and Rubin, 1973; Rubin 1979). They can also be combined (Rosenbaum and Rubin 1985; Rubin, 2001). Given auxiliaries with ellipsoidal distributions the metrics possess the desirable property of reducing bias in all linear combinations of the auxiliaries (Rubin and Thomas, 1992). We use Mahalanobis matching in Papers I-IV.

An estimate of a study variable mean from a dataset imputed by a reasonable conditional parametric (Buck, 1960) or nonparametric (Cheng, 1994) real-donor imputation method giving good mean predictions of the missing

values should be approximately unbiased. But by treating the imputed values as actual responses the precision would be overstated for two reasons. First the sampling variability would be underestimated, and secondly the uncertainty about the imputation model would not be propagated. The latter is especially difficult to handle in the single imputation approaches to variance estimation (Särndal, 1992; Little and Rubin, 2002; Brick, Kalton and Kim, 2004). The estimate could be improved upon by using a random (real- or model-donor) draw instead of a conditional mean, but this would generally not fully suffice to propagate the uncertainty. However, a general simulation approach which approximately could accomplish this, while providing unbiased estimates, is multiple imputation (Rubin, 1976a; 1978; 1987; Herzog and Rubin, 1985; Little and Rubin, 2002). Multiple imputation has a Bayesian justification but usually also shows good frequentist properties (Wang and Robins, 1998).

The idea of multiple imputation is to set up a posterior predictive model for the missing values, and then replace each value by multiple draws from this posterior. If each missing value is replaced $B > 1$ times, standard statistical procedures can be implemented separately on each of the resulting B datasets. Using a combination rule on the B estimates would give a final estimate with the proper precision (Rubin, 1987). The process of combining results from different imputed data sets is essentially the same, regardless of the type of statistical analysis. This can provide valid inference even in complicated cases (Herzog and Rubin, 1985).

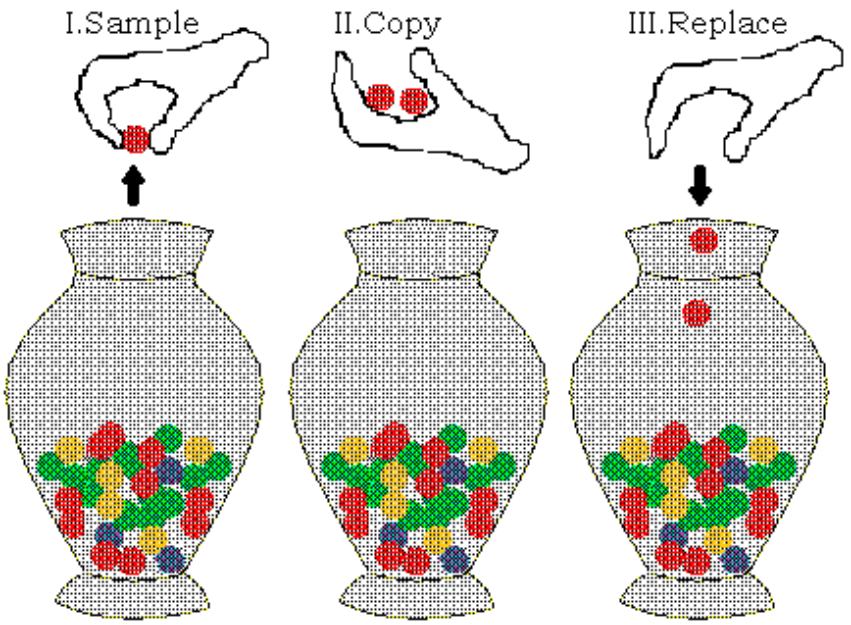

Figure 1: Pólya urn sampling

The base for the imputation methods proposed in this thesis is a nonparametric multiple real-donor approach, originating from (Eggenberger-)Pólya urns (Eggenberger and Pólya, 1923; Feller, 1968). The Pólya urn in Figure 1 is commonly used to illustrate (elementary) sampling procedures. It contains balls of different colours, which are sampled by randomly drawing, copying, and replacing balls in the urn. The sampling probabilities are thus given by the composition of the balls of the different colours at the time of each random drawing. If infinitely many balls are sampled, the composition of the sampled balls would equal a draw from a Dirichlet distribution with parameters corresponding to the composition of the balls in the initial urn. Likewise, if a finite number of balls are sampled, a draw from a Dirichlet-Multinomial distribution gives the number of balls of each colour.

By treating the units in a completely observed sample as the balls in an initial Pólya urn, and then perform Pólya urn sampling up to the size of a finite population (Lo, 1988), a bootstrap draw from a 'Pólya posterior' distribution (Ghosh and Meeden, 1997) is generated. For large populations,

such draws are approximately finitely exchangeable (Diaconis and Freedman, 1980), and has the 'Bayesian bootstrap' (Rubin, 1981) as the limiting case.

The Pólya posterior is a noninformative unique stepwise Bayes approach (Hsuan, 1979; Ghosh and Meeden, 1997). Imputation of a sample, or population, can then be achieved by treating a response set as a sample, and the intended sample as the population. Pólya urn sampling is then carried out until the sample is fully imputed. Alternatively, if the imputation model is compatible with the sampling model, the full population may be imputed instead of the intended sample. If the process of imputing the data is repeated $B > 1$ times, the generated data distribution is an approximation to the multiple imputation posterior distribution.

The proposed multiple kernel imputation method in this thesis should be able to fulfil the above criteria on imputation methods. It sympathizes strongly with the quotations that "the data will be allowed to speak for themselves" (Silverman, 1986, p.1) to achieve robustness, and that "any method that is consistent for a wide class of functions must ultimately be 'local'" (Scott and Wand, 1991, p.204). These statements surely become more relevant the more data there is at hand, though our belief is that their relevance also may hold for moderate sample sizes, with sparse and bounded data. This belief is based on the fact that the method is aided by features from kernel estimation.

2 Kernel estimation

In estimation we often assume that the relationship between e.g. a study variable Y and an auxiliary variable X can be described as

$$Y = f(X) + e, \tag{1}$$

where the functional form, including the degree of parameterization, of f is controlled by the analyst. While a parametric model is restrictive, in the sense that a finite set of parameters is used, a nonparametric model instead is very flexible, allowing essentially an infinite number of parameters. Semi-parametric models lie in between. But the flexibility comes at a price, namely that we need a sample large enough. While parametric estimates typically converge at a rate $n^{-1/2}$, convergence of semi- and nonparametric estimates is slower. On the other hand, with a larger sample it is more probable to observe the full data dispersion. Results can then be derived in finer details using semi- and nonparametric models, but approximations or smoothing are typically needed. More detailed descriptions of semi- and nonparametric models can be found in Hastie and Tibshirani, (1990), Ruppert, Wand, and Carroll (2003), Härdle, Müller, Sperlich, and Werwatz (2004).

When the object of study (Y) is the density at a point X, a nonparametric estimate can be obtained by taking a weighted average using the X_i points that are close to X, where the weight function is a (typically symmetric and unimodal) kernel function $K(\cdot)$. A bandwidth parameter $H = g \cdot h$, consisting of the orientation (g) and size (h), determines the size of $K(\cdot)$ and thus which units are used in the estimate $\widehat{f}(X) = \sum_{i=1}^{n} K\left(X_i - X, H\right)$ (Rosenblatt, (1956); Parzen, 1962). The kernel smoother (Nadaraya, 1964; Watson, 1964) is of particular interest to this thesis, where $f(X)$ is estimated by

$$\widehat{f}(X) = \sum_{i=1}^{n} Y_i \frac{K\left(X_i - X, H\right)}{\sum_{i=1}^{n} K\left(X_i - X, H\right)}. \tag{2}$$

For a particular class of this estimator the bandwidth H is indirectly determined by the number of k nearest neighbour (kNN) units (Loftsgaarden and Quesenberry, 1965; Mack and Rosenblatt, 1979; Silverman, 1986), rather than the units that are located within a chosen bandwidth.

A key issue is choosing the bandwidth size (Jones, Marron and Sheather, 1996). Too small a neighborhood produces a highly variable undersmoothed

estimate of f, while too large a neighbourhood produces a biased over-smoothed estimate. Generally an optimal bandwidth should decrease with the dimensionality of X and increase with the sample size. It may further depend on the smoothness of the underlying distribution f, in that it should be decreased (increased) at a high (low) density of f. It may further depend on the type of kernel function $K(\cdot)$.

Automatic techniques determining the bandwidth usually minimize some mean squared error function of $\widehat{f}(X)$. They include rules-of-thumb (Silverman, 1986), least-squares cross-validation (Bowman, 1984; Scott and Terrell, 1987; Sain, Baggerly, and Scott, 1994), plug-in estimates (Sheather and Jones 1991; Wand and Jones, 1994; Chacon and Duong, 2010), smoothed cross-validation (Jones, Marron and Park, 1991; Duong and Hazelton, 2005). It may be advisable to compare several bandwidths (Scott, 1992; Marron and Chung, 2001). If the density is low (high), the bandwidth can be locally adapted by increasing (decreasing) it. In comparison to automatic methods, the local neighbourhood is never empty with a k-nearest-neighbour (kNN) approach. Given p eligible units and q auxiliary variables, the ideal kNN approach of setting $k \propto p^{4/(4+q)}$ (Silverman 1986, p.99) is best used when the exact size of k is not so important. A general recommendation when $q = 1$ is to use $k \approx \sqrt{p}$ (Silverman, 1986, p.19).

There are some early references to kernel-based model-donor methods for density estimation from incomplete data (Titterington and Mill, 1983) and kernel density estimation for imputing values (Titterington and Sedransk, 1989) using a MCAR mechanism. Early references to kernel-based imputation under MAR include imputation of missing values by local mean estimates (Cheng, 1994) and by local constant polynomial regression estimates (Chu and Cheng, 1995). However, the papers included in this thesis concern real-donor imputation, where we would interpret Y_i in Equation (2) as the potential donor values, and $\frac{K(X_i-X,H)}{\sum_{i=1}^{n} K(X_i-X,H)}$ as the donor selection probabilities. The latter are denoted by λ_i. Examples of real-donor imputation of missing values using semi- and non-parametric techniques includes (Rubin and Schenker, 1986; Heitjan and Little, 1991; Aerts, Claeskens, Hens, and Molenberghs, 2002; Marella, Scanu, Conti, 2008). Asymptotic results for real donor nearest neighbour imputation when estimating a population mean are found in Chen and Shao (1997).

In parallel with kernel smoothers, the number of donors in imputation regulates the trade-off between bias and variance (Schenker and Taylor, 1995),

so that fewer potential donors translate into a tighter bandwidth. 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. Donor pools with few potential donors give rise to strong correlation between the values imputed for a missing value in multiple imputation. This may lead to both higer variances and to biased variance estimators that underestimate the true variance, as with repeated sampling from correlated (e.g. clustered) data. Larger donor pools may instead reduce the quality of matches and increase the bias.

Using sample data means that the support of the investigated data is usually finite and bounded. While prediction with parametric models works directly, nonparametric models give rise to a bias at the boundary of the convex hull of the data (Silverman, 1986, p.29-32; Wand and Jones, 1995, p.46-47; Simonoff, 1996, p.49-50). The bias is expected to be worse the closer we are to the boundary end points, and can be reproduced into the final estimates. The rate of convergence may also be slower.

Several methods have been used to reduce boundary effects, e.g. boundary kernel methods (Gasser and Müller, 1979; Müller, 1991), transformation methods to enforce more smoothing at the boundary (Marron and Ruppert, 1994), and reflection techniques (Karunamuni and Alberts, 2005). Not to violate the definition of probability, for real-donor imputation one has to restrict the choice of methods to those where $K(\cdot)$, which is directly proportional to the selection probabilities λ, is non-negative.

With real-donor imputation, one implication of having sparse and bounded data is that the donor pool will be unbalanced to the donee, in the sense that the observable donee X will differ from the expected donor auxiliary variables X_i when weighted by their selection probabilities.

The bias-variance trade-off has implications for the choice of kernel function $K(\cdot)$. Three common univariate kernel functions are shown in Figure 2, and their functional forms and efficency relative to the Epanechnikov kernel are given in Table 1.

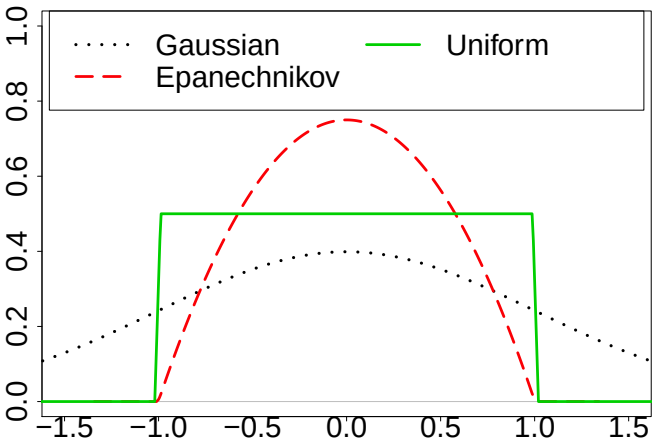

Figure 2: Three common univariate kernel functions.

Table 1: Three common univariate kernel functions and their efficiencies. (Simonoff, 1996, p.44)

	Epanechnikov	Gaussian	Uniform
Kernel function	$\frac{3}{4}(1-x^2)I_{\lvert x\rvert<1}$	$\frac{e^{-x^2/2}}{\sqrt{2\pi}}$	$\frac{1}{2}I_{\lvert x\rvert<1}$
Univariate efficiency	1	1.051	1.076

In an imputation context, the choice of kernel function directly determines the donor selection probabilities. The described Pólya urn sampling means that uniform selection probabilities are applied to all potential donors. By conditioning on X, only those with closest match are given uniform selection probabilities.

As an example of imputation from a donor pool we use a donor pool of juvenile offenders from the population used in Paper IV. The variable *family*

type is missing for one donee. It can take the value 'living in a split family' (white ball) or 'living with both parents' (black ball). The number of school credits, X1, is an auxiliary variable, which is a completely observed. It is standardised and in Figure 3 the donee is assumed to have the value 0.

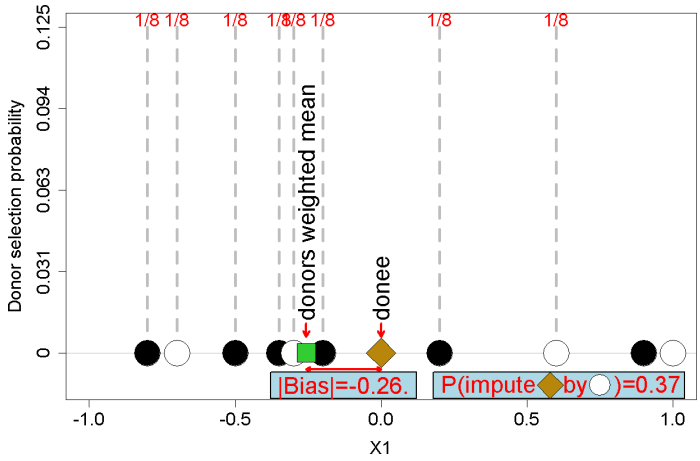

Figure 3: One auxiliary and uniform selection probabilities.

In Figure 3 the potential donors are given uniform selection probabilities. Since the donors' weighted mean of X1 differs from the donee value on X1, this donor pool is unbalanced.

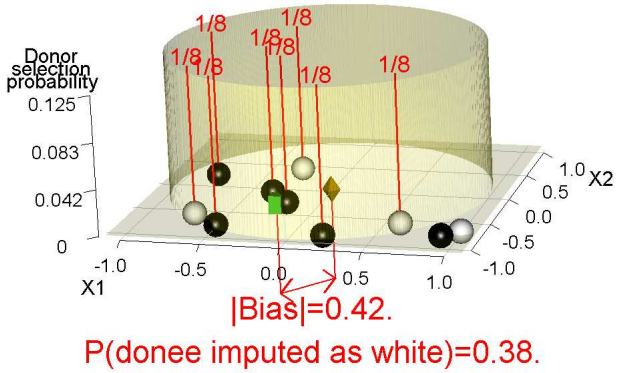

Figure 4: Two auxiliaries and uniform selection probabilities.

With an additional (standardized) auxiliary variable X2, representing the number of prosecutions, we have the situation in Figure 4. As before, the donor pool is unbalanced.

With exact conditioning on categorical variables, as in adjustment cells imputation, all units within the donor pool (including the donee) are exchangeable. It is therefore motivated to use uniform donor selection probabilities under such a model. But in our example in Figures 3 and 4, the auxiliaries are enumerable but viewed as continuous. The closest donors should thus provide a better match to the donee.

The donors' different distances to the donee can be reflected by assigning higher selection probabilities to closer donors than to more distant ones. This would be fulfilled by using a Gaussian kernel function (Conti, Marella and Scanu, 2008), meaning that all possible values are given a nonzero donor selection probability. This might be reasonable in some situations, but as pointed out, we believe that estimates should be local.

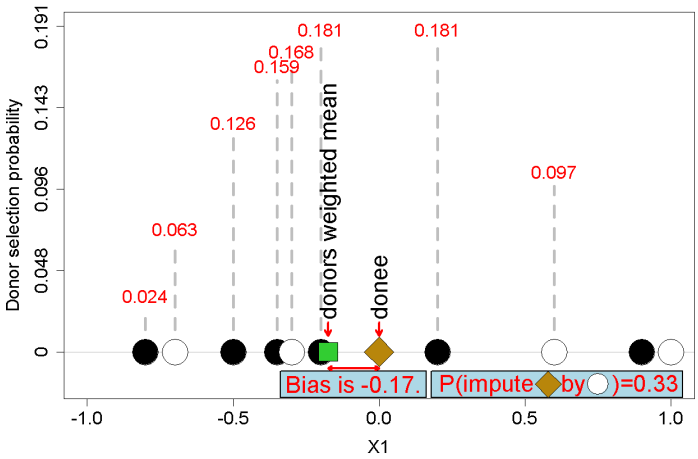

Figure 5: One auxiliary and Epanechnikov selection probabilities.

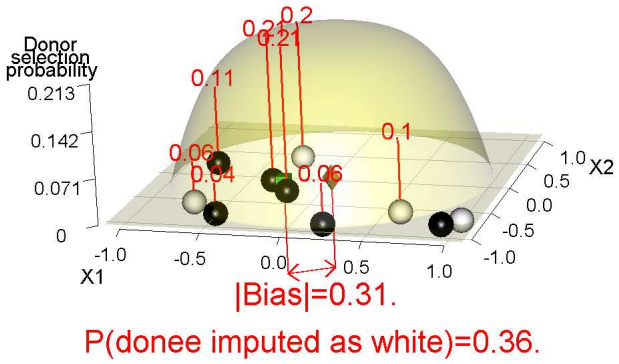

Figure 6: Two auxiliaries and Epanechnikov selection probabilities.

An Epanechnikov (1969) function would give higher selection probability the closer a potential donor is to the donee, and at the same time only allow local potential donors. The risk of imputing (extreme) outliers in the interior of the data is thereby avoided. This function possesses optimal properties in terms of being able to minimize mean squared error (Silverman, 1986). Using Epanechnikov selection probabilities without changing the number of donors in our example, the probability mass is shifted from the boundary to the center and the donor pool is expected to be less unbalanced, see Figures 5 and 6.

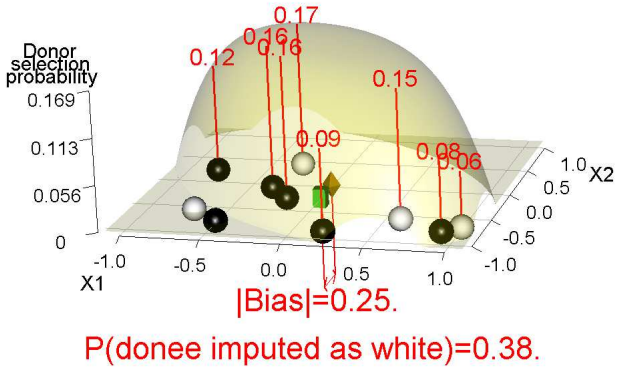

Figure 7: Two auxiliaries, Epanechnikov selection probabilities and reorientation of the donor pool

The large area in the upper right corner of the spherical donor pool in Figure 6 without any potential donors suggests that the donee is located at the boundary of the data. In this case it may be possible to reduce the bias by shifting the donor selection probabilities away from this area, as in Figure 7 where the donor pool is reoriented along the boundary.

A third kernel feature is inspired by Rice (1984), who used a linear combination of two kernel estimators with different bandwidths to reduce bias at the boundary. This may thus be viewed as tightening the bandwidth such

that the donor pool is shrunk at the boundary. In Figures 8 and 9 this is achieved by allowing fewer potential donors, compared to the normal case with the donee in the interior of the convex hull of the data.

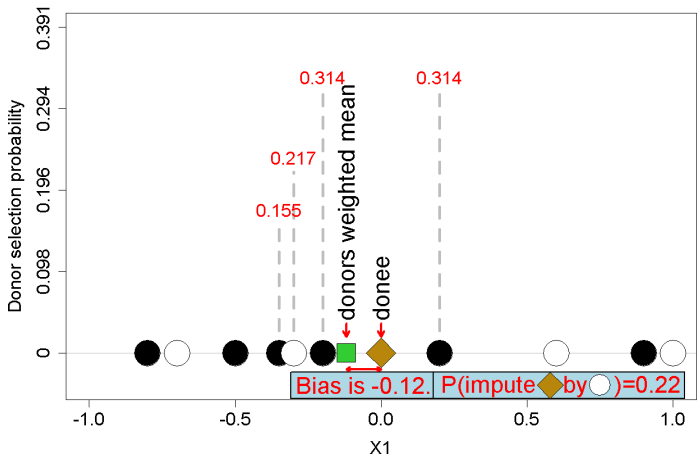

Figure 8: One auxiliary, Epanechnikov selection probabilities and shrinkage of the donor pool.

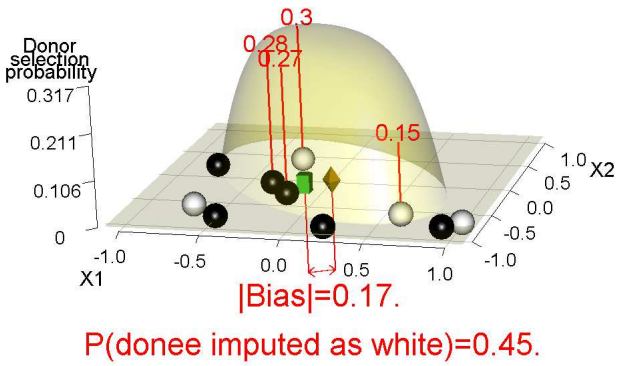

Figure 9: Two auxiliaries, Epanechnikov selection probabilities and reorientation and shrinkage of the donor pool.

In comparison to Figures 5 and 7 several effects are seen here. Not only are the most distant donors, which match the donee the least and thus contribute most bias, removed. The selection probabilities of the most distant donors in this new pool are also reduced, while those with best matches gain larger selection probabilities. As expected, the bias is also reduced. However, in a repeated sampling sense the donor pool variance is also increased (Schenker and Taylor, 1996).

Given enough data and a reasonable degree of smoothing, the estimate in Equation (2) should be able to capture the essence of $f(X)$. But even for a relationship which globally appears to be far from linear, if studied more locally around a point X_i, modelling $f(X_i)$ as linear will often be a good approximation. By imagining ourselves as looking through a magnifying glass, the much smaller proportion of data that we will be able to spot should for the sake of efficiency make such a conclusion easier to embrace.

Actually, Equation (2) may be interpreted as the estimate of the parameter α_j which minimizes the local constant polynomial fit (Fan and Gijbels,

1996)

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{J} \alpha_j (X_i - X)^j \right)^2 K\left(X_i - X, H\right). \tag{3}$$

for $J = 0$. If our inspection supports an approximately local linear relationship, it can be obtained by solving Equation (3) for $J = 1$. In the imputation context the result translates to having the donee value X being perfectly balanced to the donor pool estimate $f(X)$. Local polynomials are also known to resolve the issue of boundary bias (Simonoff, 1996) and possess good asymptotic properties (Cheng, Fan and Marron, 1993).

However, the resulting selection probabilites may become very unevenly distributed or even negative. This may occur with donees that are located at the boundary of the range of the potential donors. Linearized selection probabilities that are constrained to be non-negative can e.g. be found by calibration weighting (Särndal, 2007), normally applied on global design weights, or asymptotic equivalents to kernel weights (Aerts, Claeskens, Hens and Molenberghs, 2002) obtained by the b-bootstrap (Hall and Presnell, 1999).

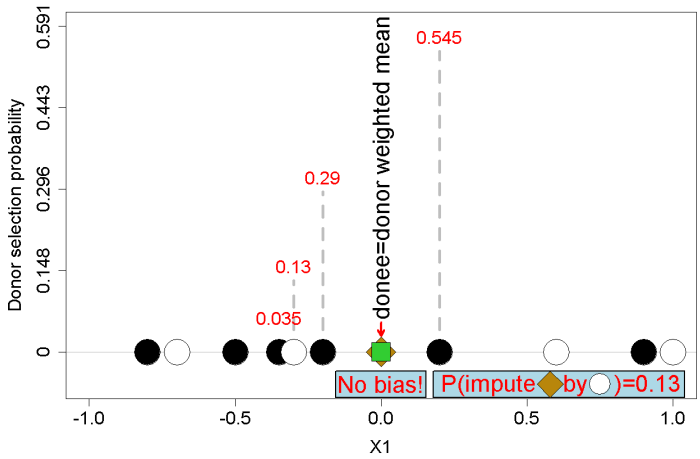

Figure 10: One auxiliary, Lagrange-adjusted Epanechnikov selection probabilities and shrinkage of the donor pool.

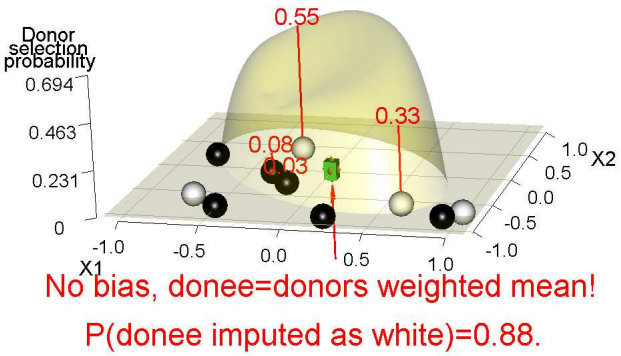

Figure 11: Two auxiliaries, Lagrange-adjusted Epanechnikov selection probabilities and reorientation and shrinkage of the donor pool.

In our example, the selection probabilities are derived from the restricted Lagrange function which is used in the four papers of this thesis. The restrictions on the selection probabilities are that they may not be negative and that they may have a maximum value.

In Figures 10 and 11 bias is now totally eliminated. This would not have been possible had the donee been lying at the utmost border of the convex hull of the data. Here seems to be the limit to what real donor imputation methods can achieve, since the estimate of $f(X)$ will always be biased away from the boundary towards the interior of the data.

3 Simulations

A summary of the simulations in the included papers is given in Table 2.

Table 2: Simulation setups of the included articles. In Paper IV, 'Con.' cover both continuous and enumerable variables, and 'Cat.' are binary variables. CC stands for complete cases, SI and MI for single and multiple imputation.

	Paper I	Paper II	Paper III	Paper IV
Estimators	mean	mean	mean	mean, regr. coef.
Pop.origin	generated	generated	generated	SOU(1971)
Pop.size	N=1600	N=1600	N=1600	N=3650
Sample	SRS, n= 100,400,900	SRS, n=400	SRS, n=400	STSRS, n_h=100, h=4
Imp. var. (Con+Cat)	separately (2+0)	separately (3+0)	separately (3+0)	simultaneously (1+0 or 2+1)
Aux. var. (Con+Cat)	(1+0)	(1+0)	(2+0)	(2+2 or 1+1)
Nonresponse	item=unit	item=unit	item=unit	item and unit
Kernel features	E, L, S combined	U, ELS	E, L, R, S combined	E, L, R, S combined
#donors $k \propto$	$\sqrt{n_r}$	$n_r^{4/(4+q)}$	$n_r^{4/(4+q)}$	$n_r^{4/(4+q)}$
Additional features	nonsampled imputed	canonical kernel	decide pool in 5 ways	restrictions, external units
Comparison methods	CC, 10 SI, 13 MI	fixed, adaptive	CC, 5 MI	CC

Since the incorporation of several kernel features in an imputation method is relatively novel, the simulations in all four papers have estimations of a population mean in focus. Population means are often the main target, or are almost always at least one of the targets, of estimation in the survey

context. The same argument also holds, though slightly more weakly, for the extension to regression coefficients in Paper IV.

In order to have full control of the data generating processes, all data in Papers I-III are generated under known models, but then a real dataset is used in Paper IV. For the same reason of control, the missing data mechanisms are also generated under known models in all four papers.

The sample size is chosen to represent a significant proportion (25 %) of the population size, except in Paper I where different sample sizes are examined specifically, and in Paper IV, where the overall sampling rate is 11 % and the sample is stratified into four equal size strata.

In Papers I-III with generated datasets, several study variables are imputed separately. Since the auxiliary variables are completely observed while there is only one variable with missing values, this may be viewed either as item or unit nonresponse. In Paper IV we distinguish between item and unit nonresponse, and impute both jointly. In addition to imputation of the missing units, in Paper I we compare to a 'mass imputation' approach and also impute all nonsampled values.

The study variables are always derived as functions of the auxiliary variables plus a residual noise from a normal distribution. Both parts contribute a significant portion to the total variance. All the generated auxiliaries are uniformly distributed, except in Paper II, where we also use one auxiliary each from a normal and a gamma distribution. The uniform auxiliaries provides simple and basic distributions with clearly defined and indicated boundaries of the data. The possible conclusions are of course constrained, but the simulations with the non-uniform auxiliaries in Paper II does not contradict that the results would be generally viable, even though the sometimes ambigous results in Paper IV may indicate that this needs further investigation.

An issue we do not discuss in detail is the choice of distance measure. Because of the way that data is generated in Paper I-III with only one or two auxiliary variables, this should not have any significant effect on these results. Possibly, the impact could be larger in Paper IV.

One study variable is always chosen such that a parametric linear regression imputation model would be suitable, while the second (in Papers I-III) and third (in Papers II and III) study variables are constructed so that a parametric linear regression imputation model would not be tenable. In all cases, the missing data mechanism is designed such that a mean estimate based on the complete case is biased.

The study variables could in principle be at any measurement level, though for simplicity in the generated datasets in Papers I-III they are only continuous. In Paper IV the enumerable variables are treated as continuous, and no kernel features are applied on the binary auxiliaries. One of the binary variables are also used for estimation of the regression coefficients.

The kernel features are applied to continuous variables. As described in Section 2, the origin of all the four features are from kernel estimation. We summarize their univariate effect (from Papers I, II and IV) on bias in comparison to the basic case with uniform (U) selection probabilities as follows, when estimating a mean using a nearest neighbour approach:

- Epanechnikov selection probabilities (E) effectively reduced bias in principle in all studied situations.
- Lagrange calibration of selection probabilities (L) effectively reduced bias in principle in all studied situations.
- Reorientation of the donor pool for boundary donees (R) slightly reduced bias except with a large donor pool in Paper II.
- Shrinkage of the donor pool for boundary donees (S) slightly reduced bias except with a small donor pool in Paper II.

The four features generally contributed in error reduction, mainly through an additative reduction of bias. But a trade-off is also seen between bias and variance. These results are all in line with general results from kernel estimation. More details on the interactions between the features are given in Paper I (with one auxiliary and three features) and in Paper III (with two auxiliaries and four features), where all possible combinations of the features are presented.

The number of potential donors plays a crucial rule. In Paper I with only one auxiliary variable, we essentially ignore this and rely on an automatic rule-of-thumb which only takes account of the number of eligible potential donors. In Paper II-IV we use a more general rule which accounts for the number of eligible donors and the number of auxiliary variables. In order to exploratively study the effect of using different donor pool sizes, we also compare a range of these. The four automatized ways of deciding the donor pool size in Paper III might have been more competitive to the nearest neighbour approach had the auxiliaries been non-uniform, as in Paper II. In addition,

in Paper II we compare with fixed and adaptive approaches of selecting the donor pool. Canonical kernels where used to neutralise the difference in variance due to choice of kernel.

The criterion to make use of all available information is discussed in more detail in Paper IV. Specifically, the simulations are used to show the effect of restricting the imputations to comply with known quantities, and the effect of using external units in a cold deck manner.

The comparison methods in Paper I are selected as those readily available in the *Packages* section on *cran.r-project.org* in the R program (R Development Core Team, 2011). This does of course not give a complete coverage of all possible imputation methods, but should at least cover a large share of those that are readily available to users of statistics. In Paper III we are more restrictive and mainly select a smaller set of comparison methods along those that performed well in Paper I.

4 Concluding remarks and future research

The kernel features behave largely in accordance with general results from kernel estimation. This is also true for the strong dependence on the donor pool size. Even though some features do not contribute in all situations, the combination of features generally contributes more to bias reduction compared to when used separately. Though the kernel features seemed to ease up the reliance on donor pool size, we stress that it is always important to explore this effect, and to make statements conditional on the chosen degree of smoothing.

The simulations show that our imputation method performs almost as well as competing methods when the study variable is a linear function of the auxiliaries, and better when the study variable is a nonlinear function of the auxiliaries. Our method seems to benefit a lot from increasing the sample size. Since we generally used a rather modest sample size of 400 units with on average 50% response rate, this seems beneficial for applications to larger datasets.

Various topics would be interesting to pursue in future research. Two possible tracks, or presumably a compromise between them, is to develop the handling om non-continuous auxiliary variables in line with the proposed methods, or to incorporate already existing methods. A more practical long-term goal is to set up a R package to make the methods available to a wider public. It would then be beneficial to develop the handling of deciding the degree of smoothing, which has not been the main focus of this thesis. Also, we would like to increase the degree of automatization for selection algorithms and methods. Another area of great interest, which has not been discussed to a great extent in this thesis, is the choice of distance function. It may have a large impact when there is access to several auxiliaries of different types, or if we would be utilizing e.g. the response propensity, or samples based on unequal probability designs.

5 Acknowledgements

So, these are the final words of my thesis. I feel a certain ambivalence towards the notion of dichotomizing something into being final or not. Although the classification ability that we as humans possess, which provides us with the ability of rapid decision making, and has contributed to our (global) survival as a species, short-sighted rationality and judgments from first impressions will certainly not always provide a fruitful long run path on a (local) individual level. Working with this thesis has been an important part of my own ongoing path in life, which occupied me and gave me several new insights, rather than something that is final. Hopefully, my efforts will also be of some help to others.

However, in daily life we do need to make decisions, and we will then make use of statistics, be it consciously or not. Finding myself in hesitant situations, I tend to ask myself what's the worst that may come out of a seemingly rational but challenging choice. Usually I come up with the answer that, within reasonable limits, the outcome of the hard choice is seldom threatening enough for me to refrain from it. I hope that I will always keep an open mind, maintain my playfulness, believe that there is something good in everyone, and that we always can learn more about the world we live in.

We sometimes say that we are standing on the shoulders of giants. When it comes to my supervisor Daniel, I hope that I at least have managed to climb up on some of your toes. Thank you for always being patient with me and my habits, including the frequent jumps from one tussock to another, with a never ending thirst to discover new areas, and in an attempt to capture the big picture, though the picture always seemed to expand in numerous directions.

There are so many more people that I would like to thank, including the lady who this morning helped my holding my cup of coffee for a while on my way to the university, and if you are not mentioned here, you are not forgotten. In particular, I want to express my gratitude towards my co-supervisor Dan for all the pragmatic and joyful views on the state of affairs, and especially for all the last minute commenting on my thesis. Further, my gratitude goes to Håkan for providing me with good views on various issues and always allowing me additional computing power for my simulations. A big hug to all my other present and former colleagues at the departmen of Statistics (AA, BB, BS, CKK, EFF, GG, GR, HL, HN, HR, JF, JO, JSS,

LL, LW, MC, MH, MV, PC, PP, RC, RH, TvR) and especially my fellow doctoral students (AT, BM, BW, CH, CT, FL, JM, KS, MM, MQ, OS, PS, SN, TS, YL). I also would like to thank FS at the department of Criminology for giving me the opportunity to apply my methods on their data.

Not to be overlooked are new and old friends, and others who supported and helped me in various ways, and who managed to brighten my life (AE, DE, GS, IL, JJ, JL, JW, MB, MD, MÖ, LW).

Closest to me are my family, Mum (Ann-Christine) and Dad (Kjell) - your infinite love and support of me obviously see no boundaries, Krister - for always being such a happy and helpful man, Mattias, I hope that I still succeed in being such a good role model for you as you are to me, Evelina - for always trusting your own judgement and believing in me, Jenny - no words can describe what you have meant and will always mean to me, I love you!, Vile - my first son who was born during my time as a doctoral student and who made it the best time of my life, Lovar - my son to soon be born, I will really try to be a good father to you!

On a windy and sunny afternoon!
Nicklas Pettersson,
Frescati
24th of April, 2013

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(1), 40-64.

Bowman, A. (1984). An alternative method of cross-validation for the smoothing of density estimates. Biometrika 71 353360.

Brick, J.M., Kalton, G., and Kim, J.K. (2004). Variance Estimation with Hot Deck Imputation using a model. Survey Methodology, 30, 57-66.

Buck, S.F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. Journal of the Royal Statistical Society, Series B 22, 302306.

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

Chambers, R., (2001). Evaluation Criteria for Statistical Editing and Imputation. National Statistics Methodological Series, United Kingdom, 28, 1-41.

Chen, J.H., and Shao, J., (2000). Nearest neighbor imputation for survey data. Journal of Official Statistics, vol.16, 113-131.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. Journal of the American Statistical Association, 89, 81-87.

Cheng, M.Y., Fan, J. and Marron, J.S. (1993). Minimax efficiency of local polynomial fit estimators at boundaries. Mimeo Series 2098, Inst. Statist., Univ. North Carolina, Chapel Hill.

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

Cochran, W.G. and Rubin, D.B. (1973). Controlling Bias in Observational Studies: A Review. Sankhya, Ser. A 35: 417-446.

Conti, P. L., Marella, D. and Scanu, M. (2008). Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators. Computational statistics and data analysis, 53, 354-365.

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

Diaconis and Freedman (1980). "Finite exchangeable sequences". Annals of Probability 8 (4): 745–76

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

Eggenberger, F. and Polya, G., (1923). Über die Statistik verketter Vorgänge, Zeitschrift für angewandte Mathematik und Mechanik, 3:279-289.

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, Chapman and Hall, London.

Fechner, Gustav Theodor. 1966 [1860]. Elements of psychophysics, Vol 1. New York: Rinehart and Winston. Translated by Helmut E. Adler and edited by D.H. Howes and E.G. Boring.

Feller, W. (1968). An introduction to probability theory and its applications. Vol. 1. Wiley, New York.

Gasser, T., and Müller, H-G., (1979). Kernel estimation of regression functions. Smoothing Techniques for Curve Estimation (eds. T. Gasser and M. Rosenblatt). Springer-Verlag, Heidelberg. p.23-68.

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

Hall, P. and Presnell, B. (1999). Intentionally biased bootstrap methods. Journal of the Royal Statistical Society series B, 61, 143-58.

Hastie, T.J., Tibshirani, R.J. (1990). Generalized Additive Models. Chapman and Hall, London.

Heitjan, D. and Little, R. (1991) Multiple imputation for the fatal accident reporting system. Applied Statistics, 40, 1329.

Herzog, T.N., and Rubin, D.B. (1983). Using multiple imputations to handle nonresponse in sample surveys. In Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliography, New York: Academic Press, 209-245.

Hsuan, F.C., (1979). A Stepwise Bayesian Procedure. The Annals of Statistics, Vol. 7, No. 4, 860-868.

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). Nonparametric and Semiparametric Models. Springer Series in Statistics, Springer, New York.

Jones, M.C., Marron J.S. and Park B.U. (1991). A simple root n bandwidth

selector. Annals of statistics, 19, 1919-1932.

Jones, M.C., Marron, J.S. and Sheather, S.J. (1996). A brief survey of bandwidth selection for density estimation. Journal of the American Statistical Association. 91, 401407.

Karunamuni, R.J., Alberts, T., (2005). A generalized reflection method of boundary correction in kernel density estimation. The Canadian Journal of Statistics 33 (4), 497-509.

Kim, J.K., and Fuller, W.A. (2004). Fractional hot deck imputation, Biometrika, 91, 559-578.

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

Little, R. (1988). Missing-Data Adjustments in Large Surveys. Journal of Business Economics Statistics, vol 6, No. 3, pp. 287-296 + comments

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

Little, R.J.A. and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? Surv. Method., 31, 161168.

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

Loftsgaarden D. O. and Quesenberry C. P. (1965). A Nonparametric Estimate of a Multivariate Density Function. Ann. Math. Statist. Volume 36, Number 3, 1049-1051.

Mack, Y.P. and Rosenblatt, M., (1979). Multivariate k-nearest neighbor density estimates. Multivariate Anal, 9, 1-15.

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

Marron, J.S. and Chung, S.S. (2001). Presentation of smoothers: the family approach, Computational Statistics, 16, 195-207.

Marron, J., Ruppert, D., (1994). Transformations to reduce boundary bias in kernel density estimation. Journal of the Royal Statistical Society Series B 56, 653-671.

Müller, H-G., (1991). Smooth optimum kernel estimators near endpoints. Biometrika, 78, 521-530.

Nadaraya, E. A. (1964). On estimation regression. Theory of Probability and Its Applications, 9, 141-142.

Parzen E. (1962). On estimation of a probability density function and mode,

Ann. Math. Statist., vol. 33, pp. 1065-1076

R Development Core Team (2011). R: A language and environment for statistical computing. Vienna, Austria. ISBN 3-900051-00-3, URL http://www.R-project.org.: R Foundation for Statistical Computing.

Rice, J., (1984). Boundary modification for kernel regression. Communications in statstistcs- Theory and methods, 13(7), 893-900.

Rosenbaum, P.R., and Rubin, D. B., (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika, 70, 41-55.

Rosenbaum, P.R., and Rubin, D. B., (1983). The bias due to incomplete matching. Biometrics, 41, 103-116.

Rosenblatt M. (1956). Remarks on some nonparametric estimates of a density function, Ann. Math. Statist., vol 27, pp. 832-837

Rubin, D.B., (1976a). Inference and missing data (with discussion), Biometrika 63, 581-592.

Rubin, D. B. (1976b). "Multivariate Matching Methods That are Equal Percent Bias Reducing, I: Some Examples." Biometrics 32 (1): 109-120.

Rubin, D.B. (1978). Multiple imputations in sample surveys, Proc. Survey Res. Meth. Sec., Am. Statist. Assoc. 20-34.

Rubin, D.B. (1979). "Using Multivariate Sampling and Regression Adjustment to Control Bias in Observational Studies." Journal of the American Statistical Association 74: 318-328.

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

Rubin, D.B. (1987). Multiple imputation for nonresponse in surveys. John Wiley and Sons, Hoboken/New York.

Rubin, D.B. (2001). Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation. Health Services and Outcomes Research Methodology 2 (1): 169-188.

Rubin, D.B. and Schenker, N. (1986) Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. Journal of the American Statistical Association, 81, 366374.

Rubin, D.B. and Thomas. N. (1992). "Affinely Invariant Matching Methods with Ellipsoidal Distributions." Annals of Statistics 20 (2): 1079-1093.

Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). Semiparametric Regression. Cambridge: Cambridge University Press.

Sain, S.R., Baggerly, K.A., and Scott, D.W. (1994). Cross-Validation of

Multivariate Densities. Journal of the American Statistical Association, 89, 807817.

Sande, I.G. (1983). Hot-deck imputation procedures. Incomplete Data in Sample Surveys,3, 339-349.

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. Chapman and Hall, London.

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

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

Scott, D.W. and Terrell, G.R. (1987). Biased and unbiased cross-validation in density estimation. J. Amer. Statist. Assoc. 82 11311146.

Scott, D. W., Wand. M. P., (1991). Feasibility of Multivariate Density. J. Amer. Statist. Assoc.

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

Simonoff (1996). Smoothing methods in statistics. Springer-Verlag, New York.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. J. Roy. Statist. Soc. Ser. B 53 683690.

SOU, (1971). 1956 års klientelundersökning rörande ungdomsbrottslingar. Unga lagöverträdare. 1, Undersökningsmetodik. Brottsdebut och återfall. (in Swedish). Stockholm.

Särndal, C.-E. (1992). Methods for Estimating the Precision of Survey Estimates when Imputation Has Been Used. Survey Methodology, Vol. 18, No 2, pp. 241-252

Särndal, C-E., (2007). The calibration approach in survey theory and practice. Survey Methodology, Vol. 33, pp. 99119

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association, 82, 528-550.

Titterington, D.M. and G.M. Mill (1983), Kernel-based density estimates from incomplete data, J. Roy. Statist. Sot. Ser. B 45, 258-266.

Titterington, D. and Sedransk, J. (1989). Imputation of missing values using density estimation. Statistics and Probability Letters, 8, 411–418.

Wand, M.P. and Jones, M.C. (1995). Kernel Smoothing. Chapman and Hall, London.

Wang, N., and Robins, J.M., (1998). Large-sample theory for parametric multiple imputation procedures. Biometrika (to appear).

Watson, G.S., (1964). Smooth regression analysis. Sankhya Series A, 26(4), 359-372.