Nittner:

The Additive Model with Missing Values in the Independent Variable - Theory and Simulation

Sonderforschungsbereich 386, Paper 272 (2002)

Online unter: http://epub.ub.uni-muenchen.de/

The Additive Model with Missing Values in the Independent Variable

–

Theory and Simulation

T. Nittner

March 20, 2002

Abstract

After a short introduction of the model, the missing mechanism and the method of inference some imputation procedures are introduced with special focus on the simulation experiment. Within this experiment, the simple additive model $y = f(x) + \epsilon$ is assumed to have missing values in the independent variable according to MCAR. Besides the well–known complete case analysis, mean imputation plus random noise, a single imputation and two ways of nearest neighbor imputation are used. These methods are compared within a simulation experiment based on the average mean square error, variances and biases of $\hat{f}(x)$ at the knots.

KEY WORDS: missing values; complete case analysis; stochastic mean imputation; nearest neighbors; GCV; iteratively reweighted least squares.

1 Introduction

Statistical analysis with incomplete data has been investigated extensively for a variety of models, e.g., for linear regression (e.g. Little (1992)), for logistic regression (e.g. Vach (1994)) or for generalized linear models (e.g. Ibrahim, Lipsitz and Chen (1999)). However, generalized additive models (GAM) affected by missing values hardly are considered. One reason may lie in the short time of twelve years ago where generalized additive models were introduced by Hastie and Tibshirani (1990) in detail. So the paper of Chu and Cheng and their conclusion that *it is not surprising that little attention has been paid to possible nonparametric inference*, Chu and Cheng (1995), p. 86, may be used as motivation for some more research first considering the empirical behavior of some procedures within this context. This paper first considers the necessary theoretical background. Besides the model for the data and the missing mechanism also a short introduction to the inference is given in Section 1.1, 1.2, and 1.3. More of practical interest is the description of the imputation procedures within Section 2. The main part is the simulation experiment with an extensive discussion of the results in Section 3 and Section 4.

1.1 The Model

Given continuous data (x_i, y_i), $i = 1, \ldots, n$, the simplest additive model is given by

$$y = f(x) + \epsilon \,. \tag{1.1}$$

Conforming with the well–known assumption concerning ϵ and X within the linear model also here we have the independence of them denoted by $\mathrm{E}(\epsilon \mid X) = \mathrm{E}(\epsilon)$ with its special case of homoscedastic structure of covariance meaning $\mathrm{E}(\epsilon) = 0$ and $\mathrm{V}(\epsilon \mid X) = \mathrm{V}(\epsilon) = \sigma^2 I_n$. In order to prevent a free constant within the function $f(x)$ (1.1) implies $\mathrm{E}(f(x)) = 0$. The small amount of assumptions of course leads to a large amount of flexibility when fitting the model to the data. This especially is the advantage of generalized additive models (GAM) over other models and the main reason why Hastie and Tibshirani (1990), p. 1, take the opinion that *the data show [...] the appropriate functional form.*

Here, the independent variable X is affected by missing values for $i = n - m + 1, \ldots, n$ according to MCAR [(1.4)] whereas the response vector y is fully observed. This allows to partition the data in 'observed' and 'missing' according to

$$(y, X) = \left(\begin{pmatrix} y_{\mathrm{obs}} \\ y_{\mathrm{mis}} \end{pmatrix}, \begin{pmatrix} X_{\mathrm{obs}} \\ X_{\mathrm{mis}} \end{pmatrix} \right) \,, \tag{1.2}$$

where the indices 'obs' and 'mis' indicate the observed and missing cases based on the observed and missing cases of the independent variable X, respectively. One should note that y_{mis} does not mean that y is affected by missing values; y_{mis} contains the response values belonging to the missing values in X_{mis}.

1.2 Missing Pattern and Missing Mechanism

Within this section two important features are introduced to characterize the situation of an incomplete data set. The first is the missing data pattern concerning the visualization of observed and missing data, the second is the missing data mechanism describing the dependencies between observed and missing data. These two concepts characterize the situation of missing data in a way that may also take into account possible reasons for the missingness. In the context of incomplete data these reasons may affect some methods, their properties or their asymptotical behavior which gives enough reason to take a closer look at these two concepts.

Missing Data Pattern The missing data pattern simply represents the data set variable–by–variable. Each bar represents a variable whereas the length of the bar indicates if there are missing cases for this variable or not. Visualizing the situation where X is incomplete and y is completely observed as denoted in (1.2) leads to Figure 1.1 (Little and Rubin (1987)). Univariate missing data are a special case of the so called monotone pattern of Figure 1.2 where the variables can be ordered in a way that a variable is observed for at least the cases of the

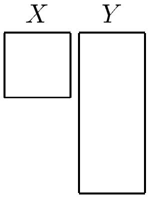

Figure 1.1: Univariate missing data pattern.

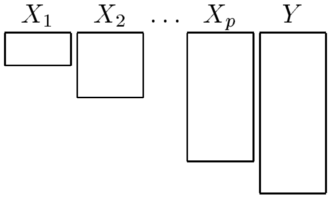

Figure 1.2: Monotone missing data pattern.

previous one. If there is suspect for X missing for large values of y, values could be ordered and a missing data pattern may describe this behavior, too. But obviously, this technique may be swamped with a higher level of dependencies. This defect is one reason why the missing data mechanisms represent a useful tool.

Missing Data Mechanism The main question arising within the statistical inference is whether the missing data mechanism can be ignored. One possibility is to make an assumption that the mechanism is ignorable in the sense described below; the other possibility consists of including the mechanism in the statistical model by including the distribution of an indicator variable indicating if a component is observed or missing. According to Little and Rubin (1987), define the data matrix $Z = (Z_{\text{obs}}, Z_{\text{mis}})$ which represents the data that would occur without missing values. Further define the random variable R indicating the missingness within the data matrix Z according to

$$r_{ij} = \begin{cases} 1 & \text{if} \quad z_{ij} \quad \text{observed} \\ 0 & \text{if} \quad z_{ij} \quad \text{missing} \end{cases} \quad \forall\, i = 1, \dots, n,\, j = 1, \dots, p+1. \tag{1.3}$$

The question whether the missing mechanism can be ignored for the estimation of θ equals the question whether the statistical inference is based on $f(Z_{\text{obs}}, R \mid \theta, \Phi)$—with Φ being an unknown parameter of the missing mechanism and θ being the parameter of the density of $Z_{\text{obs}}, Z_{\text{mis}}$—or on the simpler density $f(Z_{\text{obs}}, \theta)$ ignoring the missing mechanism. Considering the density $f(R \mid Z_{\text{obs}}, Z_{\text{mis}}, \Phi)$ allows classifying the missingness into

1. MCAR (missing completely at random) if

$$f(R \mid Z, \Phi) = f(R \mid \Phi) \quad \forall Z\,, \tag{1.4}$$

2. MAR (missing at random) if

$$f(R \mid Z, \Phi) = f(R \mid Z_{\text{obs}}, \Phi) \quad \forall Z_{\text{mis}}\,, \text{ and} \tag{1.5}$$

3. MNAR (missing not at random)

$$f(R \mid Z, \Phi) = f(R \mid Z_{\text{obs}}, Z_{\text{mis}}, \Phi)\,. \tag{1.6}$$

Following Little and Rubin (1987), the missing data mechanism can be ignored in the context of likelihood inference when the distribution of the missing mechanism is independent of the missing values [(1.5)]: Compute the density of the actual observed data obtained by integrating Z_{mis} out of the density

$$f(Z_{\text{obs}}, R \mid \theta, \Phi) = \int f(Z_{\text{obs}}, Z_{\text{mis}} \mid \theta) f(R \mid Z_{\text{obs}}, Z_{\text{mis}}, \Phi) dZ_{\text{mis}} \tag{1.7}$$

which by the help of (1.5) leads to

$$\begin{aligned} f(Z_{\text{obs}}, R, \theta, \Phi) &= f(R \mid Z_{\text{obs}}, \Phi) \int f(Z_{\text{obs}}, Z_{\text{mis}}, \theta) dZ_{\text{mis}} \\ &= f(R \mid Z_{\text{obs}}, \Phi) f(Z_{\text{obs}} \mid \theta) \,. \end{aligned} \tag{1.8}$$

If the parameters θ and Φ concerning the density of the data Z and the missing mechanism, respectively, are distinct in the sense that each parameter contains no information about the other (see for example Schafer (1997)) then the likelihood–based inferences for θ based on $f(Z_{\text{obs}}, R \mid \theta, \Phi)$ and for θ based on $f(Z_{\text{obs}} \mid \theta)$ are the same.

1.3 Inference

The well–known trade–off between wiggliness of the estimated curve and closeness to the data motivates the minimization of the target function

$$\sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \int f''(t)^2 dt \,. \tag{1.9}$$

The value of λ controls the trade–off, f' and f'' have to be continuous, f'' has to be quadratically integrable. For the two extreme cases $\lambda \to \infty$ and $\lambda \to 0$ the estimated curve equals a straight line and an interpolating spline, respectively.

Following Wood (2000) the problem of estimating the parameters $\beta^{(k+1)}$ of the nonlinear function f with $\mathrm{E}(f(Y_i)) = f(\beta)$ by Fisher–Scoring is equivalent to iteratively solving the weighted penalized least squares problem

$$\min \lambda \mid\mid W^{\frac{1}{2}} (z^{(k)} - X\beta) \mid\mid^2 + \sum_i \theta_i \beta' S_i \beta \tag{1.10}$$

with the iteratively least squares (IRLS) where the least squares problem at each iterate is replaced by a penalized one. S_i is a non–negative definite matrix of coefficients defining the ith penalty which is associated with the smoothing parameter θ_i, W is a diagonal matrix of weights, and λ is the overall smoothing parameter. From a practical point of view in the case of an estimable parameter θ_i it is more interesting to consider generalized cross validation (GCV). The problem here is to minimize the GCV–Scores

$$V = \frac{\mid\mid W^{\frac{1}{2}} (y - A(\lambda, \theta_i) y) \mid\mid^2 / n}{[1 - \mathrm{tr}(A(\lambda, \theta_i))/n]^2} \tag{1.11}$$

with respect to θ_i/λ. Combining the two procedures solves problem (1.9) and could be written in two steps.

```
1.  estimate $\mu$ and the variances $V_i$ for each $y_i$
    with the help of $\beta^{(k)}$; compute

    (i)   the diagonal weight matrix $W$ with
          $W_{ii} = (g'(\mu_i)^2 V_i)^{-1}$
    (ii)  the vector
          $z = X\beta + \Gamma(y - \mu)$ ,
          of pseudo–data with the diagonal matrix
          $\Gamma_{ii} = (g'(\mu_i))^{-1}$

2.  Compute $\theta_i$ by minimizing
          $\frac{||W^{\frac{1}{2}}(z-X\beta)||^2}{(\mathrm{tr}(I-A))^2}$
    where $\beta$ is the solution of minimizing
          $|| W^{\frac{1}{2}}(z - X\beta) ||^2 + \sum \theta_j \beta' S_j \beta$
    with respect to $\beta$, $A$ denoting the hat matrix
          $A = X(X'WX + \sum \theta_j \beta' S_j \beta)^{-1} X'W$ .
```

Table 1.1: Iteratively Reweighted Least Squares with GCV.

For a more detailed description of additive models and estimation concepts consider Hastie and Tibshirani (1990), Fahrmeir and Tutz (2001) or Wood (2000). The settings of the simulation experiment—also concerning parameters for estimating the model—in detail are described in Section 3.

2 Imputation Methods

Generally one has to distinguish between methods for analyzing the data set as it is and procedures for the imputation of missing data. It should be noted that usually the completed data set is analyzed as if there weren't any missing data however there exist also methods assigning weights to the imputed data which are different from 0 (complete case analysis) and 1 (treat the data set as completely observed), see for example Toutenburg, Fieger and Srivastava (1999). In the following a short introduction to the complete case analysis is given.

2.1 Complete Case Analysis

The complete case analysis (CCA) simply discards all cases containing at least one missing value. Based on the partitioning according to (1.2) the analysis is restricted to the estimation of

$$y_{\mathrm{obs}} = f(X_{\mathrm{obs}}) + \epsilon_{\mathrm{obs}} \, . \tag{2.1}$$

An apparent problem is the large waste of information. Estimates may also be biased if there are stratified data. According to Schafer (1997) the percentage of missing values the CCA is thought to be suitable is about 5%.

The estimates of the CCA are unbiased if the missingness does not depend on y, i.e., if $f(R \mid y, X) = f(R \mid X)$ holds. Then

$$f(y \mid R, X) = \frac{f(y, R \mid X)}{f(R \mid X)} = \frac{f(R \mid y, X) f(y \mid X)}{f(R \mid X)} = f(y \mid X)\,. \tag{2.2}$$

Equation (2.2) means that the conditional density of y given R and $X = (X_{\text{obs}}, X_{\text{mis}})$ is independent of the value of R, i.e., the conditional expectation of $f(y \mid x)$ is the same for $R = 0$ and $R = 1$ yielding unbiased estimates for an analysis based on the complete cases if the missingness does not depend on y. Further, the estimates of the complete case analysis are consistent even under MNAR.

2.2 Mean Imputation

The unconditional mean imputation, also known as *zero order regression* (ZOR) has been first described in Wilks (1932). It is of nonnegligable interest for users doing analysis with popular software where this method often is implemented. A missing value x_{ij} is imputed by

$$\hat{x}_{ij} = \bar{x}_j = \frac{1}{n - m_j} \sum_{i \notin \Phi_j} x_{ij}\,, \tag{2.3}$$

where $\Phi_j = \{i : x_{ij} \text{ missing}\}$ denotes the indices of the missing values and m_j denotes the number of missing values for X_j. If X_j is non–continuous mode and median are suited alternatives.

An important disadvantage of the ZOR is the underestimation of variance. Therefore, small confidence intervals distort corresponding tests. Modifying the imputed value in terms of an additive random error may be a way to improve the ZOR; let us denote this procedure by ZOR+, the zero order regression plus random noise, a kind of stochastic mean imputation.

2.3 Single Imputation

A further method is called single imputation and was is for example described in Little and Rubin (1987). In comparison to the mean imputation the single imputation should provide substitutes representing more variation than the ZOR+ did because the variation of the distribution of the complete cases is somewhat larger than the variance of the additive random error within the ZOR+. As already indicated the single imputation (sI) could be based on the distribution of the complete cases, i.e., impute a random number out of the distribution characterized by its estimated parameters. This distribution sometimes is known. Otherwise one may consider conditional distributions based on the complete cases and an auxiliary model used for predicting the missing values which however here is not of interest because of having just the simple model $y = f(x) + \epsilon$ with one covariate.

Example 1 *Assume a linear model $y = X\beta + \epsilon$ where $X = (\mathbf{1}, X_2, X_3)$. X_3 is supposed to be binary and partially incomplete. Compute an auxiliary model, for example a logistic regression for the complete cases with X_3 being the response vector, X_1, X_2 and may be y representing the independent variables. The resulting estimates are used to compute conditional probabilities π_i via the logit link using the values of the observed variables X_1, X_2, y for the missing indices. The π_i could be considered as parameters of row–wise binomial distributions which motivate the following imputation steps for $i = 1, \ldots, m$*

1. *Draw a random number z_i from a continuous uniform distribution over the interval [0;1]*
2. *Impute*

$$x_i = \begin{cases} \text{‘1’} & \text{if} \quad z_i \le \pi_i \\ \text{‘0’} & \text{if} \quad z_i > \pi_i \end{cases} \quad . \tag{2.4}$$

2.4 Nearest Neighbor Imputation

Within this section two kinds of nearest neighbor imputations are introduced. The first one is the ‘classical’ nearest neighbor imputation (NN1) the second one is a modified version (NN2).

2.4.1 Nearest Neighbor Imputation—Version I (NN1)

The nearest neighbor imputation has a long history but according to Chen and Shao (2001) is still not fully investigated although it is used in many surveys. Referring to the data structure (1.2) with m missing values for the row indices $i = n - m + 1, \ldots, n$ visualized by

$$\underbrace{x_1, \ldots, x_{n-m}}_{\text{observed}}, \underbrace{x_{n-m+1}, \ldots, x_n}_{\text{missing}} \quad \text{and} \tag{2.5}$$

$$\underbrace{y_1, \ldots, y_{n-m}, y_{n-m+1}, \ldots, y_n}_{\text{observed}} \quad , \tag{2.6}$$

a missing value $x_j, j = n - m + 1, \ldots, n$, is imputed by choosing that value x_i, $1 \le i \le n - m$, which is the nearest neighbor of j. In this context the distance determining the nearest neighborhood is measured in y–values such that i satisfies

$$\mid y_i - y_j \mid \quad = \quad \min_{1 \le l \le n-m} \mid y_l - y_j \mid \ . \tag{2.7}$$

If the solution is not unique the mean of the corresponding x–values is imputed.

The nearest neighbor imputation is a hot deck imputation procedure which yields values unlikely to be nonsensical. Population means and quantiles are asymptotically unbiased and consistent (see Chen and Shao (2000)). Since it is a nonparamteric method it is expected to be somewhat more robust against model violations. Chen and Shao (2001) give a detailed overview over several possibilities for adjusting the procedure in order to get asymptotically unbiased and consistent variance estimates. Additive models, however, here and yet haven't been investigated prior to the current paper.

2.4.2 Nearest Neighbor Imputation—Version II (NN2)

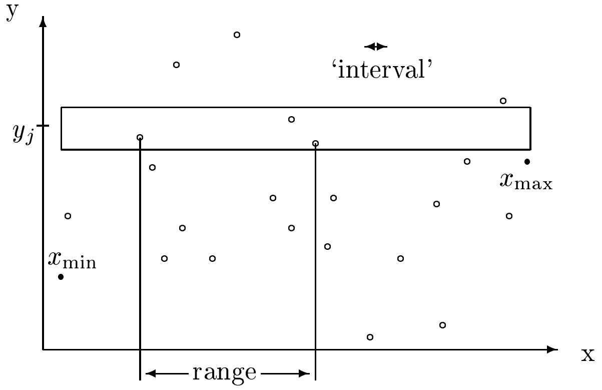

Figure 2.1: Fixed neighborhood based on $k = 3$.

Because of investigating a smooth function relating x and y, e.g., a cubic function, also the nearest neighbor could lead to substitutes being far away from the 'true' value. This is the reason why a modified version of the classical NNI has been implemented based on some plausibility. Let x_j again denote a missing value and y_j be the corresponding response value. Consider a neighborhood of y_j based on a fixed number of neighbors k. The main idea is to control the range of this fixed neighborhood in comparison to a percentage of the length of the data interval ('interval'). Figure 2.1 illustrates a data situation with three nearest neighbors, an artificial range between the three nearest neighbors and the reference value 'interval' with a 5%–data rate.

The neighborhood is defined according to (2.7) whose solution for $k = 3$ is a (3×1)–vector containing the ordered values $x_{[s]}$ for $s = 1, 2, 3$ satisfying (2.7). The range for $k = 3$ defined by $x_{[3]} - x_{[1]}$ is the first relevant value for the procedure of the NN2; 'interval' is defined as a fixed percentage—here 5%—of $x_{\max} - x_{\min}$ and should be a reference value for the range of the neighborhood. A short introduction to the NN2 is given in Table 2.1.

It must be said that this procedure is characterized by decisions based on a kind of plausibility. Following Table 2.1 step–by–step clarifies this. If condition (1) is true one would give the probability that the missing value is within $[x_{[1]}; x_{[3]}]$ a large value and therefore may impute as described. If not, as it is the case in the following steps, one would try to quantify the difference between the range and 'interval' as in (2). Further, one may be induced to compare this value with a random number based on the same range of values. The smaller z, the more likely it is to be expected that the range is nearer to the value of 'interval' which justifies to impute a value representing the center of a quantum

```
(1)  IF ‘range’ $\leq$ ‘interval’

         impute a random number out of a continuous uniform distribution
         over $[x_{[1]}; x_{[3]}]$

(2)  ELSE

         compute $z =$ ‘range’ $-$ ‘interval’     (note: $z_{\max} = 0.95 \cdot (x_{\max} - x_{\min})$)

         draw a random number $u$ out of a continuous uniform distribution
         over $(0; z_{\max}]$

     (3)  IF $u > z$

              impute $1/3\ (x_{[1]} + x_{[2]} + x_{[3]})$

     (4)  ELSE

              compute the empirical distribution $N(\bar{X}; S^2)$ and the probabilities
              $P(X \leq x_{[1]}), P(X > x_{[3]})$ and $P(x_{[1]} < X \leq x_{[3]})$ and order them

              impute a value satisfying the condition of the maximum probability
              and satisfying $x_{\min} \leq X \leq x_{\max}$
```

Table 2.1: NN2—a modified version of the nearest neighbor imputation, $k = 3$.

according to (3). ‘The smaller’ is tried to be evaluated by the random number u. If (4) becomes true, the empirical distribution—which type is supposed to be known—is estimated based on its empirical parameters. The probabilities dividing it with respect to $x_{[1]}, x_{[2]}$ and $x_{[3]}$ have to be computed and ordered. Impute a value according to the property belonging to the largest probability. Within the simulation experiments about 2–3% of the data satisfied condition (1) for $\sigma = 1.0$ and about 20–22% for $\sigma = 0.1$. The mean according to (3) was imputed in 58–76% and data fulfilling (4) in about 11–37%. We see that a larger deviation of the errors, i.e., the variance of the data in y–direction leads to a little amount of cases satisfying (1).

3 A Simulation Experiment

This section gives a short introduction to the simulation experiment including the data, the model, its different settings and some technical details. R programming language (see Venables and Smith (2001)) was used to implement the imputation methods within an additive model estimated by mgcv, a statistical feature of R for GCV.

3.1 Model and Data

As within Section 1 we assume a model $y = f(x) + \epsilon$ with pairs (x_i, y_i) being continuous whereas X is assumed to be affected by missing values according to MCAR and the response y assumed to be completely observed. X follows a normal distribution with mean ν and varying variance δ^2. With the help of

$$f(x) = a + bx + cx^2 + dx^3 \text{ and } \epsilon \sim N(\mu, \sigma^2) \tag{3.1}$$

the corresponding response vector could be generated. Depending on σ^2 the true function $f(x)$ is overlaid with some noise illustrated in Figure 3.1. The missing values—here according to MCAR—are simply created by drawing a random number out of the row indices $i = 1, \ldots, n$ with the restriction not to draw the upper or lower bound of the data interval $[inta, intb]$. Using partition (1.2) enables us to create complete case vectors which are simply filled up with substitutes subject to the implemented imputation procedures.

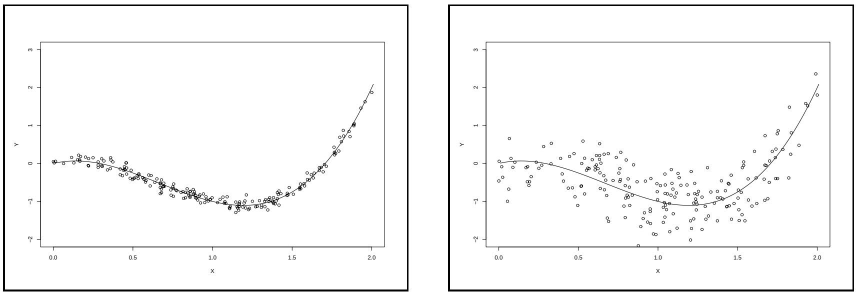

Figure 3.1: $f(x)$, $\sigma = 0.1$ (left hand side), $\sigma = 1.0$ (right hand side).

Because of considering four imputation methods (ZOR+, sI, NN1, NN2) overall six estimators have to be compared. Inference follows Section 1.3. Ten knots were chosen for fitting the smooth term whose regression spline basis is based on cubic Hermite polynomials. The knots are evenly placed throughout the (ordered) covariate values, i.e., for 10 knots and $n = 500$ the knots may be weighted averages of observations because of no whole–numbered ranks which would for example have been the case for $n = 505$. So the different procedures lead to models estimated from possibly different knot location but are compared based on the fixed location of the 'true' model. Apart from the estimation of the smoothing parameter λ it should be noted that the selection of the estimated degrees of freedom is an integral part of model fitting.

3.2 Settings and Criteria

name	meaning	value
n	sample size	500
`knots`	number of knots for the cubic spline	10
`replications`	number of replications	1000
k	number of the 'nearest neighbors'	3
m_p	missing percentage	0.1 (0.3)
`interval`	length of the interval	2.0

Table 3.1: Program variables.

For illustrating the settings of the experiment a few tables are given. Table 3.1 contains program variables, their meaning and values. Values in paranthesis indicate alternative settings. Further parameter values concerning the data and the nearest neighbor imputation are shown in Table 3.2 and Table 3.3.

parameters of $f(x)$				$x_i \in$ [inta; intb]		$x_i \sim N(\nu;\delta^2)$		$\epsilon \sim N(0;\sigma^2)$
a	b	c	d	inta	intb	ν	δ	σ
0.0	1.0	-4.0	2.0	0.0	2.0	1.0	0.3 (0.7)	0.1 (1.0)

Table 3.2: Parameter for generating the data.

Number of 'nearest neighbors'	interval criterion
k	interval
3	0.05 * (intb - inta)

Table 3.3: Parameter for the 'nearest neighbor imputation'.

An important note concerns the generation of the data. Because of having an interval for X its empirical distribution cannot exactly follow the postulated normal distribution. This is the reason why the empirical distribution of X is a truncated normal distribution depending on the variance δ^2. Integrating a truncated normal distribution with $\mu = 0$ from -1 to 1 shows that the maximal standard deviation is about 0.5. That is the reason why here and in the following δ is assumed to be 0.5 instead of 0.7. Apart from some empirical values controlling for the quality of the experiment the different models were compared based on the sample mean squared error (see (4.1)), the variance and the bias of the estimates $\hat{y}$. The values $\hat{y}$ for each imputation method were computed by predicting $\tilde{f}(\tilde{x}_j)$ with $\tilde{x}_j$ being the fixed knots of the 'true' model and $\tilde{f}$ being the estimated function based on the corresponding imputation procedure. The variances at the knots are the variances between the replications, the biases are the differences between $\tilde{f}(\tilde{x}_j)$ and $f(x_j)$. The number of knots has to satisfy

$$\#(\text{knots}) < n - m$$

with m denoting the number of missing values. This condition is necessary else the complete case analysis can't be compared with the imputation procedures. Following Wood (2001) in an R newsletter, *the choice of the number of knots is not very critical, but should be somewhat larger than the estimated degrees of freedom plus 1*; this condition was taken into account by controlling it during the experiment. Altogether we ran eight simulation experiments which are shown in Table 3.4. Models 5–8 for example correspond to 1–4 with $\delta = 0.7(0.5)$.

model	m_p	σ	δ
1	0.1	0.1	0.3
2	0.3	0.1	0.3
3	0.3	1.0	0.3
4	0.1	1.0	0.3
5	0.1	0.1	0.5
6	0.3	0.1	0.5
7	0.3	1.0	0.5
8	0.1	1.0	0.5

Table 3.4: Different models for n=500, 10 knots, 1000 replications.

4 Results

Before considering statistical results a few remarks. The time which took an experiment to run was between 3 and 13 hours, especially depending on the missing percentage and δ. Comparing theoretical and estimated values of ϵ, σ and δ showed precision from 10^{-2} up to 10^{-5}—satisfactory results. Considering the empirical distributions of ϵ and X showed worse results with increasing variances. Classifying 'worse' is based on the attempt to rate the concerning distribution plots with respect to unimodality, smoothness and symmetry. Concerning the smoothing parameters it can be stated that its values tend to increase with increasing variances of X and ϵ and with an increasing missing percentage.

4.1 The sample mean squared error (SMSE)

The sample mean square error follows

$$\widehat{\text{SMSE}}(\hat{y}, y) = \sum_{j=1}^{\text{knots}} \hat{\text{V}}(\hat{y}_j) + [\hat{\text{B}}(\hat{y}_j, y_j)]^2 \,, \tag{4.1}$$

the well–known formula here based on the values at the knots. Variance and bias at the knots are computed as described on page 11.

The single imputation always had the largest SMSE. Ordering the SMSEs for each of the eight experiments and summing up the ranks lead to the ranking of Table 4.1.

TRUE	CC	NN2	ZOR+	NN1	sI
10	16	28	32.5	33.5	48

Table 4.1: Sum of ranks for all procedures summed up over all experiments.

This first impression has to be analyzed further, especially by considering the two components of the SMSE. But first let's take a look at further properties regarding the SMSE.

Analyzing the maximum of the SMSE for the experiments showed that it increases with

- an increasing percentage of missing values,
- an increasing variance of ϵ, and
- a decreasing variance of X.

Except for m_p the minimum also tended to increase according to these conditions. The exact values of the SMSE for the models 1–8 are listed in Table 4.2.

	1	2	3	4	5	6	7	8
TRUE	0.096	0.094	1.519	1.495	0.017	0.015	0.304	0.326
CC	0.089	0.078	1.691	1.537	0.018	0.017	0.401	0.355
ZOR+	0.123	0.153	1.671	1.548	0.077	0.251	0.555	0.369
sI	0.247	0.776	2.685	1.826	0.131	0.567	0.889	0.466
NN1	0.128	0.203	1.854	1.591	0.041	0.074	0.443	0.369
NN2	0.102	0.122	1.743	1.537	0.035	0.082	0.542	0.377

Table 4.2: $\widehat{\text{SMSE}}$ for Models 1–8.

Figure 4.1 shows the SMSE for $\sigma = 0.1$, Figure 4.2 for $\sigma = 1.0$ both with $\delta = 0.3$ on the left hand side and $\delta = 0.5$ on the right hand side—with 10% (grey bars) and 30% (white bars) missing percentage. The properties concerning maximum and minimum also can be seen here. Comparing left– and right hand side we can see with one exception (ZOR+ for $\sigma = 0.1$ and 30% missing percentage) that the SMSE increases with a decreasing variance of X. One could justify this fact with the larger amount of values at the margins of the interval which may lead to more precise estimates. Concerning σ which means comparing upper and lower graphics tells us that the SMSEs increase with an increasing variance of the errors. This increase is weaker for $\delta = 0.5$. Last but not least we can state that for the larger percentage of missing values all imputation procedures show larger sample mean square errors which could be seen by comparing grey and white bars.

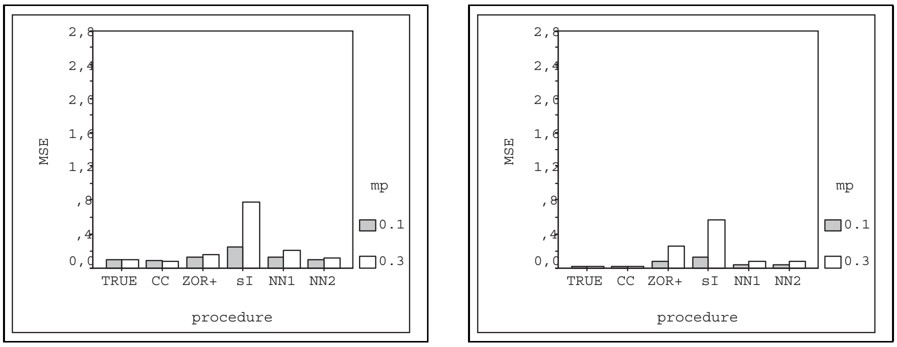

Figure 4.1: $\widehat{\text{SMSE}}$ of all procedures, $\sigma = 0.1$, $\delta = 0.3$ (left) and $\delta = 0.5$ (right).

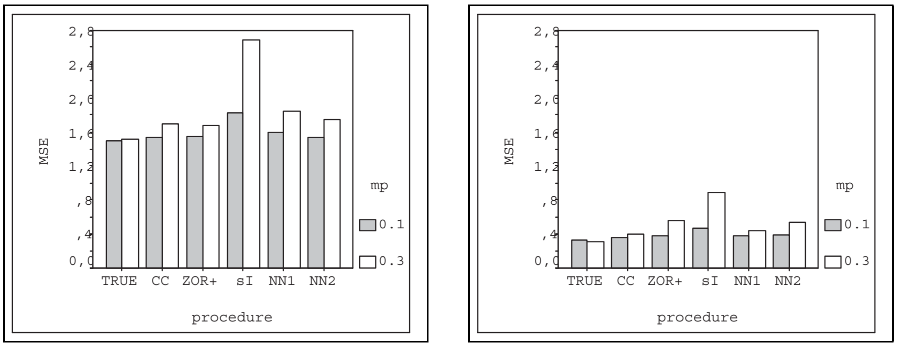

Figure 4.2: $\widehat{\text{SMSE}}$ of all procedures, $\sigma = 1.0$, $\delta = 0.3$ (left) and $\delta = 0.5$ (right).

Altogether it can be said that it tends to exist obvious cohesion between the SMSE and the interesting parameters. However, these are just trends especially because of having only two levels for σ, δ and m_p.

At last we want to consider the SMSE for each procedure. The **complete case analysis** shows the smallest SMSE in comparison to the imputation procedures (except for $\sigma = 1.0, \delta = 0.3$ compared to the ZOR+) and differs just slightly from the 'true' model. Most obvious are the largest values of the **single imputation**. Both **nearest neighbor imputations** seem to be more adequate than the alternative imputation procedures because of their minor SMSE. The 'classical' version tends to have a smaller SMSE for data being more homogeneous which can deduced from the results for $\delta = 0.5$ and an experiment with X being uniformly distributed. The **zero order regression** obviously tends to be less appropriate for a larger variance of X where its SMSE is not smaller than these of the nearest neighbor imputations (except for Model 8). This could be justified as well—the smaller the variance of a distribution the more representative the mean of the distribution. In order to be able to differ the methods at all, variance and bias are considered in the next section. Within these sections there's special focus on the changes depending on σ, δ and m_p. The procedures are compared within the Conclusion.

4.2 The variance

Analyzing the sum over the knots showed exactly the same trend as was observed for the SMSE, i.e., the sum of the variance with one exception increases with

- an increasing percentage of missing values,
- an increasing variance of ϵ, and
- a decreasing variance of X.

This increase was most obvious for a change of the error variance σ^2, it increased least depending on the missing percentage. An exact analysis of the behavior of the variance depending on procedure, knots, missing percentage, σ^2 and δ^2 is extensive and forces us to just show a couple of graphics.

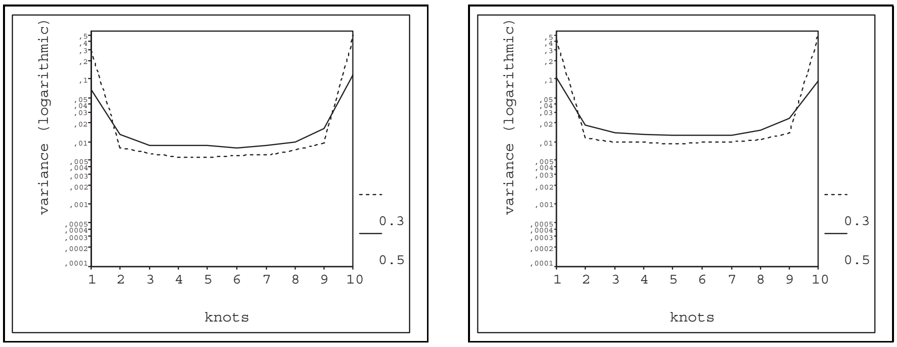

Figure 4.3: $\hat{\mathrm{V}}(\hat{y})$ (logarithmic scale) at all knots, $\sigma = 1.0, m_p = 0.3, \delta = 0.3$ (dashed line) and $\delta = 0.5$ (continuous line); sI (left) and NN1 (right).

The influence of an increasing **standard deviation of X** is somewhat differentiated. As already mentioned the sum of the variance over all knots decreased with enlarging δ. A more exact inspection yielded an increasing variance at the inner knots and a decreasing variance at the outer ones as can be seen in Figure 4.3 for the single imputation and the NN1 for example. For a better illustration the ordinate was changed to logarithmic scale. The decrease of the variance at the outer knots could be explained by the larger amount of values within this area which may give more stable estimates.

Figure 4.4 shows the variances of the zero order regression and the single imputation depending on the **error variance** $\boldsymbol{\sigma^2}$, for example, at the knots for δ and m_p fixed. The dashed line pictures the values for $\sigma = 0.1$, the continuous one for $\sigma = 1.0$. We see that at each knot the variance of the two imputation procedures is larger for $\sigma = 1.0$ than for 0.1. This is the case for the 'true' model and the other imputation procedures, too. The larger the variance, the larger the 'distance' of the scatterplot from the 'true' curve which is expected to result in an increase of the variance.

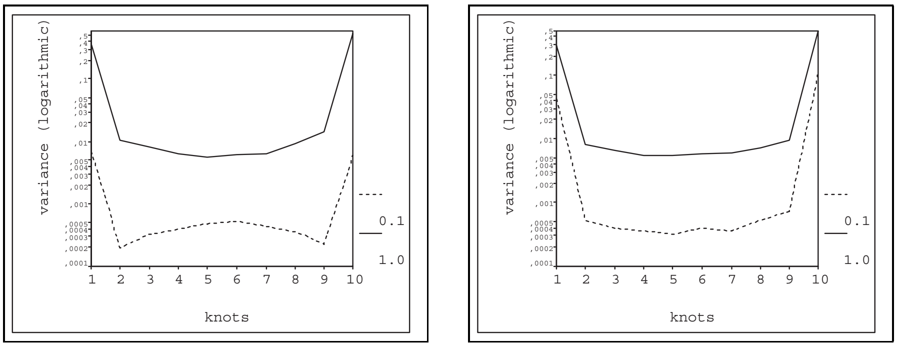

Figure 4.4: $\hat{\mathrm{V}}(\hat{y})$ (logarithmic scale) at all knots, $\delta = 0.3, m_p = 0.3, \sigma = 0.1$ (dashed line) and $\sigma = 1.0$ (continuous line); ZOR+ (left) and sI (right).

A less precise situation can be observed analyzing the dependence on the **missing percentage $\mathbf{m_p}$**. But as one might see from Figure 4.5 there is a trend to an increase of the variances with raising m_p from 10 to 30% which can here be seen for the complete case analysis. The zero order regression for example doesn't show a unique trend yet.

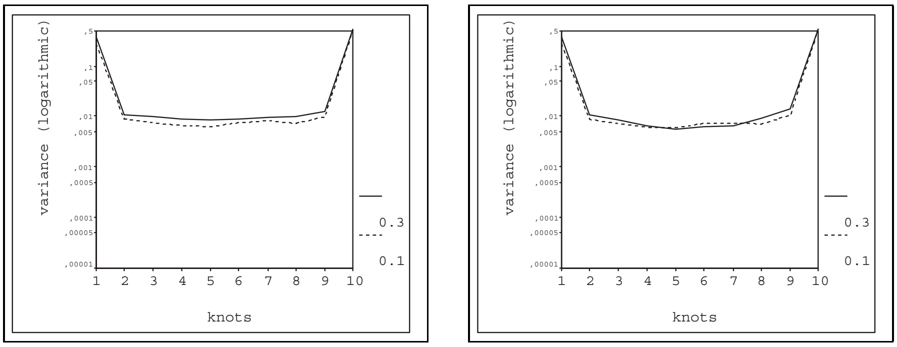

Figure 4.5: $\hat{\mathrm{V}}(\hat{y})$ (logarithmic scale) at all knots, $\sigma = 1.0, \delta = 0.3, m_p = 0.1$ (dashed line) and $m_p = 0.3$ (continuous line); CCA (left) and ZOR+ (right).

However, we are especially interested in analyzing differences among the procedures themselves coming up in the next paragraph.

Whereas the **zero order regression** had minimum variance at the outer knots for $\sigma = 0.1$ an increase lead to larger variances of the ZOR+. The variance of the ZOR+ therefore is suspected to raise more relative to the alternatives in this context. The **single imputation** seemed to have largest variances for all knots but the increase of σ improved its situation to the alternatives and even tended to the smallest variances independent of the knots. Within the imputation methods the **nearest neighbor imputation NN1** tended to have minimum variances for a smaller σ. The increase of σ totally changed the situation between the two nearest neighbor imputation procedures, the **nearest neighbor imputation NN2** increased its variance less than the NN1 and nearly has smaller values at all knots. The **complete case analysis** tended to show the same behavior like the imputation methods, especially like the NN1. For the smaller σ the CCA had small variances and for $\sigma = 1.0$ the variance increased but still has good values. Essentially there weren't large differences between the methods for a change in δ or m_p. The procedures and some differences are analyzed within the Conclusion.

Based on this short summary the imputation procedures behave somewhat different in comparison to the analysis of the SMSE which yet motivates some guesswork about the biases.

4.3 The bias

The analysis of the bias may after all show where differences between the methods concerning their SMSE come from, especially when the SMSE can't be followed straight from the variance. Like SMSE and variance also the sum of the squared biases increased with an increasing percentage of missing values, an increasing error variance, and a decreasing variance of X.

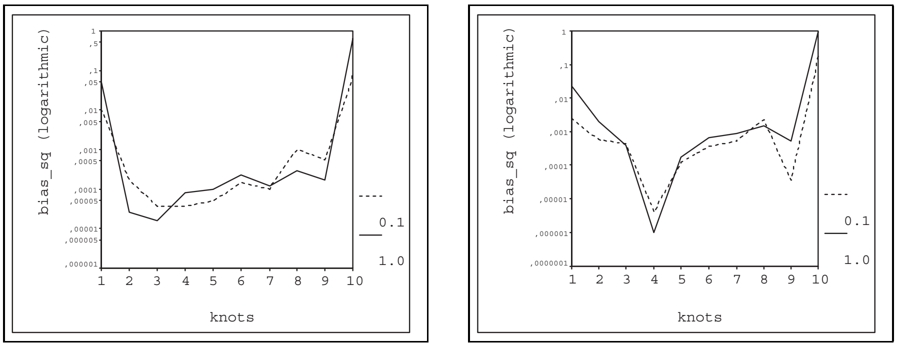

Figure 4.6: $\hat{\mathrm{B}}(\hat{y}, y)$ (logarithmic scale) at all knots, $m_p = 0.1, \delta = 0.3, \sigma = 0.1$ (dashed line) and $\sigma = 1.0$ (continuous line); NN2 (left) and sI (right).

Less obvious as within the analysis of the variance was the change of the bias depending on the **error variance** $\boldsymbol{\sigma^2}$. An increase could be noticed with some exceptions for the inner knots mainly for the imputation methods. Figure 4.6 especially shows a similar behavior of the bias depending on the knots when the variance changes from 0.1 to 1.0.

A clear decrease of the bias can be observed with an increasing **variance** $\mathrm{V}(X)$ at the outer knots, an increase at the inner knots as is shown by Figure 4.7 where the bias is plotted for the nearest neighbor imputation NN2 and the zero order regression. The complete case analysis and the nearest neighbor imputation NN1 totally tended to decrease their bias.

With some exception at outer knots the biases increase with the **missing percentage** raising from 10 up to 30%. Differences between the procedures are hard to identify because of a diffuse behavior at the knots as one might see in Figure 4.8—again a logarithmic scale was chosen—where the single imputation shows an explicit trend the nearest neighbor imputation NN2 however differs between inner and outer knots. The complete case analysis has the smallest differences when raising the missing percentage and even tends to smaller biases with $m_p = 0.3$.

Before analyzing the components together the behavior of the methods is shortly summarized. The **zero order regression** tends to have a smaller bias than the **single imputation** for the outer knots but for the inner ones there's reason to state that the single imputation shows less deviation—independent of the settings. Version 2 of the **nearest neighbor imputation** has smaller biases than

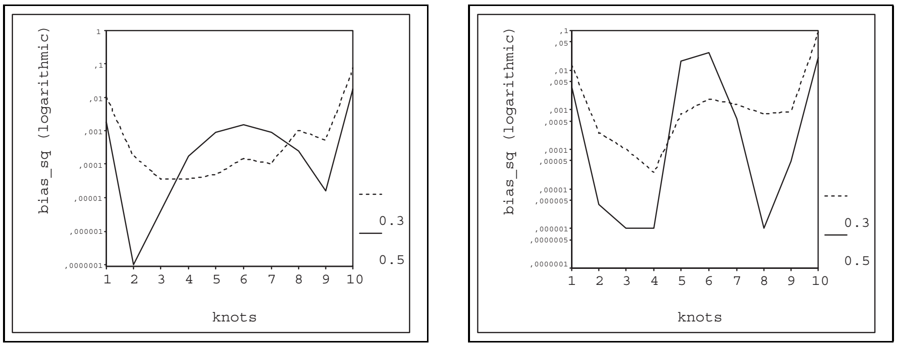

Figure 4.7: $\hat{\mathrm{B}}(\hat{y}, y)$ (logarithmic scale) at all knots, $m_p = 0.1, \sigma = 0.1, \delta = 0.3$, (dashed line) and $\delta = 0.5$ (continuous line); NN2 (left) and ZOR+ (right).

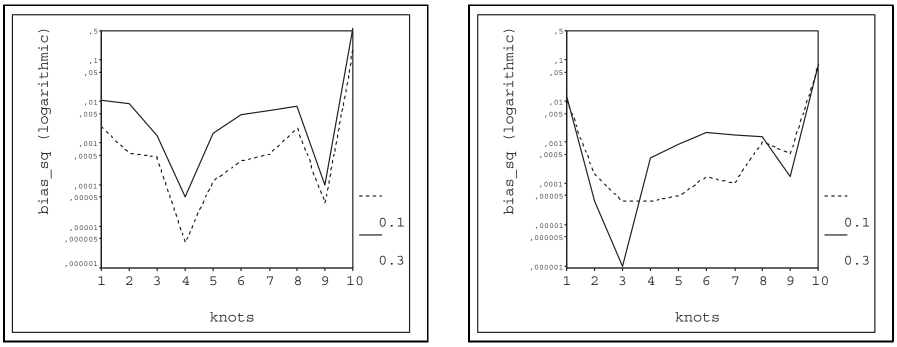

Figure 4.8: $\hat{\mathrm{B}}(\hat{y}, y)$ (logarithmic scale) at all knots, $\delta = 0.3, \sigma = 0.1, m_p = 0.1$, (dashed line) and $m_p = 0.3$ (continuous line); sI (left) and NN2 (right).

version 1 at the outer knots especially for a smaller σ, NN1 has less biased estimates for the inner knots. Altogether it seems that at the inner knots the **zero order regression** has maximal biases, both **nearest neighbor imputations** and the **complete case analysis** have minimal biases.

4.4 Variance and bias and their effect on the SMSE

The share of variance and bias in the SMSE Because of having the same weights in specifying the SMSE, an increase of the proportion of the bias corresponds to a decrease of the proportion of the variance. To condense this item it is just mentioned that the bias proportion tends to decrease with increasing δ and σ and a decreasing missing percentage (especially for the imputation procedures).

The share of the outer knots on $\sum \hat{\mathrm{B}}^2$ and $\sum \hat{\mathrm{V}}$ The **zero order regression** clearly has the smallest proportions of the outer knots concerning $\sum \hat{\mathrm{B}}^2$. One may follow that both procedures tend to have larger bias for the inner knots. For $\sigma = 0.1$ especially the **ZOR+** and the **NN2** have minimal proportions for

$\sum \hat{\mathrm{V}}$ which may denote more stable estimates at the outer knots.

The ranks of $\sum \hat{\mathrm{B}}^2$ and $\sum \hat{\mathrm{V}}$ For an increasing percentage of missing values the ranks of the procedures show little change for $\sum \hat{\mathrm{B}}^2$, analogously for the sum of the variances. For an increasing variance of X the two **nearest neighbor imputations** keep their rank relative to each other whereas the **zero order regression** increases its rank of bias and decreases its rank of variance. An increase of the error variance σ^2 leads to a lower rank of the **single imputation** and the **NN2** for the sum of variances and higher ranks of the **NN1** concerning $\sum \hat{\mathrm{B}}^2$ and $\sum \hat{\mathrm{V}}$. The **zero order regression** again shows a kind of trade–off, here improving the rank situation for the bias and downgrading it for the variance.

Uniformly distributed X The main result for an additional experiment with X being uniformly distributed was a reduced influence of the outer knots on variance and bias for the **zero order regression**. Solely 3% of the bias and 15% of the variance of the ZOR+ are explained through the outer knots.

5 Conclusion

Some of the following results, especially those comparing the methods could be seen in Figure 5.1 and Figure 5.2. Both compare two groups of methods—one containing the single imputation with the ZOR+, the other comparing the complete case analysis with the nearest neighbor imputations—with respect to bias (Figure 5.1) and variances (Figure 5.2) at all knots for varying missing percentages; here δ was chosen to 0.5 and σ to 1.0.

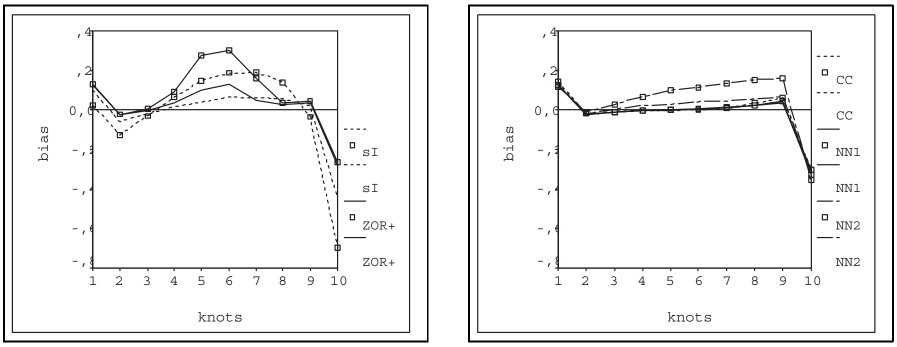

Figure 5.1: Estimated bias for all knots, $\sigma = 1.0, \delta = 0.5$; ZOR+ and sI (dashed line) left hand side, NN1, NN2 (semi–dashed line) and CCA (dashed line) right hand side; 30% missing percentage marked by squares.

The **complete case analysis** tends to minimum variance and bias very often, resulting in a least SMSE and to the ranking in Table 4.1. Large variances and small biases often occur in statistical analysis, then called ‘trade–off’. This trade–off clearly has been observed for the **zero order regression**. It tends to

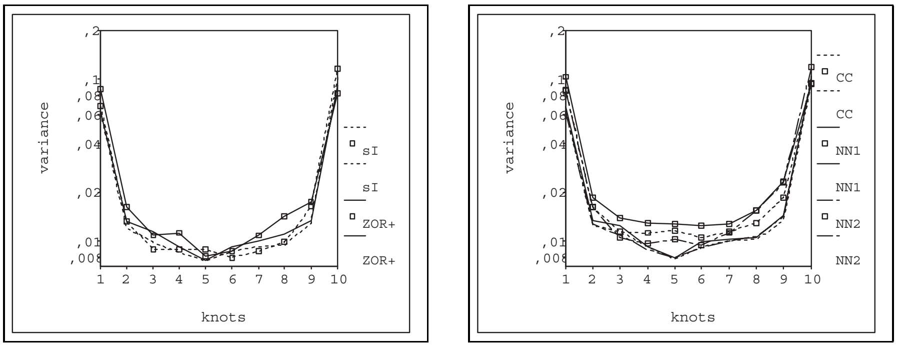

Figure 5.2: Estimated variance for all knots (logarithmic scale), $\sigma = 1.0, \delta = 0.5$; ZOR+ and sI (dashed line) left hand side, NN1, NN2 (semi–dashed line) and CCA (dashed line) right hand side; 30% missing percentage marked by squares.

have a large bias and little variance for inner knots. Except for some knots (first, fourth, ninth) the **single imputation** has strong biased estimates. Comparing these two methods gives some advantage for the single imputation in the middle where bias and variances are smaller and a slight superiority for the ZOR+ at the outer knots which may however result from an underestimated variance especially for a smaller error variance. This contradicts the analysis of the SMSE but we saw that the large SMSE of the single imputation might be based on strong biased estimates at the outer knots where additionally the ZOR+ shows small variances and biases. The two **nearest neighbor imputations** show values near to the complete case analysis. Whereas the NN1 has a smaller bias in the center of the interval the NN2 shows less bias at the outer knots. The NN2 also shows less deviation nearly at all knots.

Altogether one might prefer the classical method of a complete case analysis—which is however not an alternative for all practical problems—and the nearest neighbor imputation. Firstly, the NNI is a nonparametric method and is supposed to lead to more or less sensible substitutes and secondly, it provides good results near to the properties of the complete case analysis. Additionally, the second version of the nearest neighbor imputation is a flexible procedure because of the possible variability in its parameters.

In order to stick to a scientific character it should be a goal to study missing data within additive and generalized models on a more theoretical basis. These results may have given a first impression of some behavior.

References

Chen, J., and Shao, J. (2000). Biases and variances of survey estimators based on nearest neighbor imputation, *Technical report.*

Chen, J., and Shao, J. (2001). Jackknife variance estimation for nearest–neighbor imputation, *Journal of the American Statistical Association*

96(453): 260–269.

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

Fahrmeir, L., and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2 edn, Springer–Verlag, New York.

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

Ibrahim, J. G., Lipsitz, S. R., and Chen, M.-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable, *Journal of the Royal Statistical Society, Series B* **61**(1): 173–190.

Little, R. J. A. (1992). Regression with missing X's: A review, *Journal of the American Statistical Association* **87**(420): 1227–1237.

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

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

Toutenburg, H., Fieger, A., and Srivastava, V. K. (1999). Weighted modified first order regression procedures for estimation in linear models with missing X-observations, *Statistical Papers* **40**: 351–361.

Vach, W. (1994). *Logistic Regression with Missing Values and Covariates*, Vol. 86 of *Lecture Notes in Statistics*, Springer–Verlag, Berlin.

Venables, W., and Smith, D. (2001). *An Introduction to R.*

Wilks, S. S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples, *Annals of Mathematical Statistics* **3**: 163–195.

Wood, S. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties, *Journal of the Royal Statistical Society, Series B* **62**(2): 413–428.

Wood, S. (2001). *mgcv: GAMs and Generalized Ridge Regression for R.*