Katholieke
Universiteit
Leuven

**Faculty of
Science**

ADVANCED NON PARAMETRIC STATISTICS

A kernel-type sregression estimator for NMAR response variables

Federico Grazi (r1028201)

Academic year 2024–2025

Contents

1 Introduction

The paper by Mojirsheibani and Khudaverdyan [2024] focuses on developing a new methodology for the situation in which the missingness pattern in the response variable is non-ignorable.

In general, missigness patterns may arise in the covariates or in the response variable. The mechanism for which data goes missing may have various structure depending on the data type and its relation with the data entry process.

Little and Rubin [2002] describe the general theory behind missing data as follows. Consider a complete data vector defined as $\boldsymbol{Y} = (\boldsymbol{Y}^{\text{obs}}, \boldsymbol{Y}^{\text{mis}})$ where $\boldsymbol{Y}^{\text{obs}}$ denotes the observed part of $\boldsymbol{Y}$ and $\boldsymbol{Y}^{\text{mis}}$ the missing part. Furthermore, let $\boldsymbol{M}$ be a matrix of missing indicators with density $f(\boldsymbol{M} \mid \boldsymbol{Y}, \boldsymbol{\psi})$ where $\boldsymbol{\psi}$ are some unknown parameters.

Then, whenever the missingness mechanism does not depend on the values of $\boldsymbol{Y}$, the missingness pattern is referred to as Missing Completely At Random (MCAR) such that

$$f(\boldsymbol{M} \mid \boldsymbol{Y}, \boldsymbol{\psi}) = f(\boldsymbol{M} \mid \boldsymbol{\psi}). \tag{1.1}$$

Thus, the mechanism simply does not depend on the data.

Whenever the missingness pattern depens only on the observed vector $\boldsymbol{Y}^{\text{obs}}$ but not on the missing one $\boldsymbol{Y}^{\text{mis}}$, the density of $\boldsymbol{M}$ becomes

$$f(\boldsymbol{M} \mid \boldsymbol{Y}, \boldsymbol{\psi}) = f(\boldsymbol{M} \mid \boldsymbol{Y}^{\text{obs}}, \boldsymbol{\psi}). \tag{1.2}$$

This missingness pattern is referred to at Missing At Random (MAR).

At last, if the missing pattern depends on the missing values $\boldsymbol{Y}^{\text{mis}}$ the Not Missing At Random (NMAR) case arises, for which

$$f(\boldsymbol{M} \mid \boldsymbol{Y}, \boldsymbol{\psi}) = f(\boldsymbol{M} \mid \boldsymbol{Y}^{\text{mis}}, \boldsymbol{\psi}). \tag{1.3}$$

When there are NMAR responses, the observed analysis can result in biased and inefficient parameter estimates, whereas to incorporation additional information from missing responses one needs to assume a parametric model for the missing data mechanism. The importance of a parametric or semiparametric will be discussed later due to identifiability issues.

Another key concept in missing data theory is the concept of ignorability. Rubin [1987] defines an *ignorable* missingness mechanism for the response variable if and only if

$$f(\boldsymbol{Y}^{\text{mis}} \mid \boldsymbol{Y}^{\text{obs}}, \boldsymbol{M}, \boldsymbol{\psi}) = f(\boldsymbol{Y}^{\text{mis}} \mid \boldsymbol{Y}^{\text{obs}}). \tag{1.4}$$

Thus, if the response mechanism is ignorable, the distribution of $\boldsymbol{Y}^{\text{mis}}$ does not depend on the $\boldsymbol{M}$ matrix and statistical analysis of the observed vector under ignorability can be carried out by essentially ignoring the missingness mechanism.

The case of NMAR data is of particular interest due to the fact that studying the underlying missingness structure could lead to a deeper understanding of the problem. This is the case especially for economical or social studies, where the study of the missing response is of particular interest due to nonresponses may arise from causes inherent in the study itself.

A common technique to deal with NMAR respsonse is data imputation introduced by Rubin [1994] whose idea has been broadly discussed and it has been documented by Scheuren [2005]. Nonetheless, imputation is no more valid the the non ignorable case, where the propensity to answer depends not only on observed data but also on unobserved data.

The approach chosen by Mojirsheibani and Khudaverdyan follows the reasoning of Kim and Yu [2011] for which a Semi-Parametric model was proposed. In the case of Non-Ignorable

NMAR response Variable a parametric exponential tilting model is fitted to the missing part of the data in order to determine the amount of ignorability of the response, then, a semiparametric logistic regression with the tilting parameter is assumed for the response.

The setup of Kim and Yu states that, for a design matrix $\boldsymbol{X} \in \mathbb{R}^d$ and a univariate response variable Y, given a missing indicator $(\Delta \mid \boldsymbol{X}, Y) \sim \text{Bern}(\pi_\varphi)$ with $\pi_\varphi = \pi(\boldsymbol{X}, Y; \varphi)$ a logistic regression model for the missing indicator yields the following model

$$P\big\{\Delta = 1 \mid \boldsymbol{x}, y\big\} = \pi_\varphi(\boldsymbol{x}, y) = \frac{\exp(g(\boldsymbol{x}) + \varphi y)}{1 + \exp(g(\boldsymbol{x}) + \varphi y)} \tag{1.5}$$

for some function $g(\cdot)$ on $\mathbb{R}^d$. This model specification leads to

$$f_0(y \mid \boldsymbol{x}) = f_1(y \mid \boldsymbol{x}) \frac{\exp(\gamma y)}{\mathbb{E}\big[\exp(\gamma Y) \mid \boldsymbol{x}, \Delta = 1\big]} \tag{1.6}$$

where $\gamma = -\varphi$ is the tilting parameter cited before, $f_1(y \mid \boldsymbol{x}) = f(y \mid \boldsymbol{x}, \Delta = 1)$ and similarly for f_0. The importance of the γ parameter can be seen in this equation since it is a measure of the departure from the MAR assumption. Noticeably, this structure resembles a semiparametric Cox proportional hazard model, such that when $\gamma = 0$ one observes $f_0(y \mid \boldsymbol{x}) = f_1(y \mid \boldsymbol{x})$.

The aim of Mojirsheibani and Khudaverdyan is to generalize the setting of Kim and Yu by allowing φ to be any real-valued function $\varphi > 0$ on $\mathbb{R}$ such that

$$P\big\{\Delta = 1 \mid \boldsymbol{x}, y\big\} = \pi_\varphi(\boldsymbol{x}, y) = \Big[1 + \exp(g(\boldsymbol{x})) \cdot \varphi(y)\Big]^{-1} \tag{1.7}$$

where the setup of Kim and Yu is found by setting $\varphi(y) = \exp(\gamma y)$.

It can be seen that the case of ignorability is when $\varphi(y) = 0$, and the missingness does not depend on the observed responses.

Non-identifiability problems arises in the case in which a fully non parametric estimation is also carried out for φ, thus the semiparametric model is chosen. Moreover, Shao and Wang [2016] show that a sufficient condition for the estimation is that

$$\exists \quad \boldsymbol{V} \in \mathbb{R}^q : \boldsymbol{X} = (\boldsymbol{Z}, \boldsymbol{V}) \quad \text{and} \quad (\boldsymbol{V} \mid Y) \perp \Delta \quad \forall \quad 1 \leq q < d$$

.

Nonetheless, to tackle the difficulties that arise from the fact that is not possible to find the instrument $\boldsymbol{V}$ in the case of $d = 1$, Mojirsheibani and Khudaverdyan consider the case where a small follow-up sample of response values from the set of non-respondents is accessed, with the sub-sample that can be negligibly small.

2 Proposed Estimator

To obtain the proposed estimator og $m(\boldsymbol{x}) = \mathbb{Y} \mid \boldsymbol{x}$ we shall first note that the from function π_φ as in 1.7 one can find a usefull representation of the expected value of $1 - \Delta$ that will be usefull later on:

$$\mathbb{E}(1 - \Delta \mid \boldsymbol{x}, Y) = 1 - P(\Delta = 1 \mid \boldsymbol{x}, Y) = 1 - \pi_\varphi(\boldsymbol{x}, Y) = \\ = \frac{\exp(g(\boldsymbol{x})\varphi(Y))}{1 + \exp(g(\boldsymbol{x})\varphi(Y))} = \exp(g(\boldsymbol{x})\varphi(Y))\pi_\varphi(\boldsymbol{x}, Y), \quad (2.1)$$

Then, by using some simple properties, we can rewrite $m(\boldsymbol{x})$ as

$$m(\boldsymbol{x}) = \mathbb{E}(Y \mid \boldsymbol{x}) = \\ = \mathbb{E}(Y\Delta \mid \boldsymbol{x}) + \frac{\mathbb{E}(Y(1-\Delta) \mid \boldsymbol{x})}{\mathbb{E}(1 - \Delta \mid \boldsymbol{x})}\mathbb{E}(1 - \Delta \mid \boldsymbol{x}).$$

Usign the law of iterated expectation and the result obtained in 2.1 the fraction term further simplifies as

$$\frac{\mathbb{E}(Y(1-\Delta) \mid \boldsymbol{x})}{\mathbb{E}(1 - \Delta \mid \boldsymbol{x})} = \frac{\mathbb{E}_{\boldsymbol{X}}\big[\mathbb{E}(Y(1-\Delta) \mid \boldsymbol{x}, Y) \mid \boldsymbol{X}\big]}{\mathbb{E}_{\boldsymbol{X}}\big[\mathbb{E}(1 - \Delta \mid \boldsymbol{x}, Y) \mid \boldsymbol{X}\big]}$$

$$= \frac{\mathbb{E}_{\boldsymbol{X}}\big[Y(1 - \pi_\varphi(\boldsymbol{x}, Y)) \mid \boldsymbol{X}\big]}{\mathbb{E}_{\boldsymbol{X}}\big[1 - \pi_\varphi(\boldsymbol{x}, Y) \mid \boldsymbol{X}\big]}$$

$$= \frac{\mathbb{E}_{\boldsymbol{X}}\big[Y \exp(g(\boldsymbol{x}\varphi(Y)))\pi_\varphi(\boldsymbol{x}, Y) \mid \boldsymbol{X}\big]}{\mathbb{E}_{\boldsymbol{X}}\big[\exp(g(\boldsymbol{x}\varphi(Y)))\pi_\varphi(\boldsymbol{x}, Y) \mid \boldsymbol{X}\big]}$$

$$= \frac{\mathbb{E}_{\boldsymbol{X}}\big[Y\varphi(Y)\Delta \mid \boldsymbol{X}\big]}{\mathbb{E}_{\boldsymbol{X}}\big[\varphi(Y)\Delta \mid \boldsymbol{X}\big]}$$

To better reduce the notation of the estimator, we define these two quantities for $k = 1, 2$.

$$\Psi_k(\boldsymbol{x}; \varphi) = \mathbb{E}_{\boldsymbol{X}}\big[Y^{2-k}\Delta\varphi(Y) \mid \boldsymbol{x}\big] \quad (2.2)$$

$$\eta_k(\boldsymbol{x}) = \mathbb{E}\big[Y^{2-k}\Delta \mid \boldsymbol{x}\big] \quad (2.3)$$

At last, $m(\boldsymbol{x})$ takes the form of

$$m(\boldsymbol{x}; \varphi) = \mathbb{E}(Y\Delta \mid \boldsymbol{x}) + \frac{\mathbb{E}_{\boldsymbol{X}}\big[Y\varphi(Y)\Delta \mid \boldsymbol{X}\big]}{\mathbb{E}_{\boldsymbol{X}}\big[\varphi(Y)\Delta \mid \boldsymbol{X}\big]}\mathbb{E}(1 - \Delta \mid \boldsymbol{x}) = \quad (2.4)$$

$$= \eta_1(\boldsymbol{x}) + \frac{\Psi_1(\boldsymbol{x}; \varphi)}{\Psi_2(\boldsymbol{x}; \varphi)}\left(1 - \eta_2(\boldsymbol{x})\right) \quad (2.5)$$

Some further notation shall be introduced in order to obtain the actual formula for the non parametric estimator of $m(\boldsymbol{x})$.

Let $\mathbb{D}_n = \{(\boldsymbol{X}_1, Y_1, \Delta_1), \ldots, (\boldsymbol{X}_n, Y_n, \Delta_n)\}$ represente a sample of size n, then, the kernel estimator $\hat{m}(\boldsymbol{x})$ for $m(\boldsymbol{x})$ can be obtained starting from

$$\hat{\Psi}_k(\boldsymbol{x}; \varphi) = \frac{\sum_i Y^{2-k} \Delta_i \varphi(Y_i) \mathcal{K}\left(\frac{\boldsymbol{x}-\boldsymbol{X_i}}{h}\right)}{\sum_i \mathcal{K}\left(\frac{\boldsymbol{x}-\boldsymbol{X_i}}{h}\right)} \tag{2.6}$$

$$\hat{\eta}_k(\boldsymbol{x}) = \frac{\sum_i Y^{2-k} \Delta_i \mathcal{K}\left(\frac{\boldsymbol{x}-\boldsymbol{X_i}}{h}\right)}{\sum_i \mathcal{K}\left(\frac{\boldsymbol{x}-\boldsymbol{X_i}}{h}\right)} \tag{2.7}$$

for $\mathcal{K} : \mathbb{R}^d \to \mathbb{R}^+$ a kernel with bandwidth $h > 0$.

Given this notation, the only element left to estimate to obtain $\hat{m}(\boldsymbol{x})$ is a parametric estimate of $\varphi(Y)$. Mojirsheibani and Khudaverdyan, following the idea of Kim and Yu, divide the sample into a training and validation set, such that we have $\mathbb{D}_n = \{\mathbb{D}_m, \mathbb{D}_\ell\}$ with corresponding index sets $\mathcal{I}_m = \left\{i \in \{1, \ldots, n\} : (\boldsymbol{X}_i, Y_i, \Delta_i) \in \mathbb{D}_m\right\}$ and $\mathcal{I}_\ell = \left\{i \in \{1, \ldots, n\} : (\boldsymbol{X}_i, Y_i, \Delta_i) \in \mathbb{D}_\ell\right\}$. Then, from the validation set $\mathbb{D}_\ell$ a sub-sample of non-respondants is sampled with probability of success p_n

$$\delta_i \sim \text{Bern}(p_n), \quad i = 1, \ldots, \ell, \quad \text{wtih} \quad p_n \to 0 \quad \text{as} \quad n \to \infty,$$

where a non-respondant is included in the follow-up if and only if $(1 - \Delta_i)\delta_i = 1$, $i \in \mathcal{I}_\ell$.

Moreover, let us define $\hat{m}_m(\boldsymbol{x}; \varphi)$ to be the estimator constructed based solely on the training set $\mathbb{D}_m$, hence only for the units such that $i \in \mathcal{I}_m$. Then, the estimator for φ can be obtained by minimizing the empirical L_2 error of the estimator $\hat{m}_m(\boldsymbol{x}; \boldsymbol{\varphi})$ evaluated on the validation set $\mathbb{D}_\ell$:

$$\hat{\varphi}_n = \arg\min_{\varphi \in \mathcal{F}_\varepsilon} \frac{1}{\ell} \left[\sum_{i \in \mathcal{I}_\ell} \Delta_i \left(\hat{m}_m(\boldsymbol{X_i}; \boldsymbol{\varphi}) - Y_i \right)^2 + \sum_{i \in \mathcal{I}_\ell} (1 - \Delta_i) \frac{\delta_i}{p_n} \left(\hat{m}_m(\boldsymbol{X_i}; \boldsymbol{\varphi}) - Y_i \right)^2 \right]$$

where $\mathcal{F}$ is a given class of functions $\varphi : [-L, L] \to (0, B]$ for some finite B and positive L and $\mathcal{F}_\varepsilon = \{\varphi_1, \ldots, \varphi_{N(\varepsilon)}\} \subset \mathcal{F}$ is an ε-cover of $\mathcal{F}$ such that the cardinality of the smaller ε-cover of $\mathcal{F}$ is denoted as $\mathcal{N}_\varepsilon$. Moreover, if $\mathcal{N}_\varepsilon < \infty \, \forall \, \varepsilon > 0$, then the family $\mathcal{F}$ is said to be *totally bounded*.

The empirical L_2 error can be seen as a weighted sum of the observed points in the validation set that have assigned a mass of $1/\ell$ and the resampled points from the validation set that have a weight of $1/\ell p_n$. Thus more importance is given to predicting the re-sampled points.

Finally, the proposed estimator is found to be

$$\hat{m}(\boldsymbol{x}; \hat{\varphi}_n) = \hat{m}_m(\boldsymbol{x}; \varphi)\Big|_{\varphi = \hat{\varphi}_n}. \tag{2.8}$$

2.1 Remarks on the Estimator

First of all, altough the estimation process requires new observation to be observed, the actual estimator is evaluted only using the observed respondants from $\mathbb{D}_m$. The observation sampled from the validation set are used only to obtain an estimate for the function φ.

As stated before, a complete non-parametric estimation of the model suffers from identifiability issues, thus a semiparametric specification of the probability of missing is used. The

specification of Kim and Yu is the mostly used, but the estimator in 2.8 can be obtained with any semiparametric specification of $\varphi(Y;\gamma)$ with γ the unkown parameter.

The estimation procedure simplifies to

$$\hat{\gamma}_n = \arg\min_{\gamma\in\Upsilon} \frac{1}{\ell}\left[\sum_{i\in\mathcal{I}_\ell} \Delta_i \left(\hat{m}_m(\boldsymbol{X_i};\boldsymbol{\gamma}) - Y_i\right)^2 + \sum_{i\in\mathcal{I}_\ell}(1-\Delta_i)\frac{\delta_i}{p_n}\left(\hat{m}_m(\boldsymbol{X_i};\boldsymbol{\gamma}) - Y_i\right)^2\right]$$

where Υ is the possible parameter space such that the conditions specified for $\varphi(Y;\gamma)$ are still satisfied.

A rather counter-intuitive fact from the paper is the function φ. The class of function $\mathcal{F}$ is defined such that it yields positive and bounded values by $B > 0$. This would mean that the exponential used by Kim and Yu should not be a valid function to be in the class of the estimator, nonetheless it is still proposed by the authors and studied in the simulation study. For sake of reproducibility of the paper, the proposed simulation study in this report will also use the exponential.

Another proposed function in the paper is $\pi_\gamma(y) = (0.1 + (\gamma y)^2)^{-1}$ which can be seen to be bounded by $B = 10$ and satisfies the function property.

One last remark is about convergence rate. If we assume that the cardinality of the ε-cover $\mathcal{F}_\varepsilon$ is such that $\log \mid \mathcal{F}_\varepsilon \mid = \mathcal{O}(n)$, then it can be seen that

$$\mathbb{E}\left[\hat{m}(\boldsymbol{X},\hat{\varphi})_n) - m(\boldsymbol{X})\right] = \mathcal{O}\left(\sqrt{\frac{\log n}{nh^d p_n^2}}\right) \tag{2.9}$$

whereas if p_n is fixed to a constant value $c \in [0,1]$ the rate of convergence can be lowered to $\mathcal{O}(\log(n)/nh^d)$. Nonetheless, this can be seen to be a slower than the estimator without missing data, which is $\mathcal{O}(1/nh^d + h^2)$.

3 Simulation Study

A simulation study was replicated as in the original paper. The setup for the simulation was the following

$$X \sim \mathcal{U}(-10, 10)$$

$$Y = 2.6 + \sin^2(0.1X + 2)\cos^2(0.1X + 3) + (0.1X)^3$$

$$\varphi(Y; \gamma) = \exp(\gamma Y)$$

$$g(X) = 0.5 + 0.25X$$

$$\pi_\gamma(X, Y) = \big(1 + \exp g(X)\varphi(Y; \gamma)\big)^{-1}$$

$$\Delta \sim \text{Bern}(\pi_\gamma)$$

The chosen value for gamma was $\gamma = 0.98$ and resulted in a percentage of missing values of approximately 30%.

Using this setup, a sample of $n = 1001$ was sampled. As Figure 1 illustrates,the pattern of missing data has a clear dependance with both variables due to how π_γ wass constructed. Thus the missing data are NMAR and the mechanism of missingness is also non-ignorable.

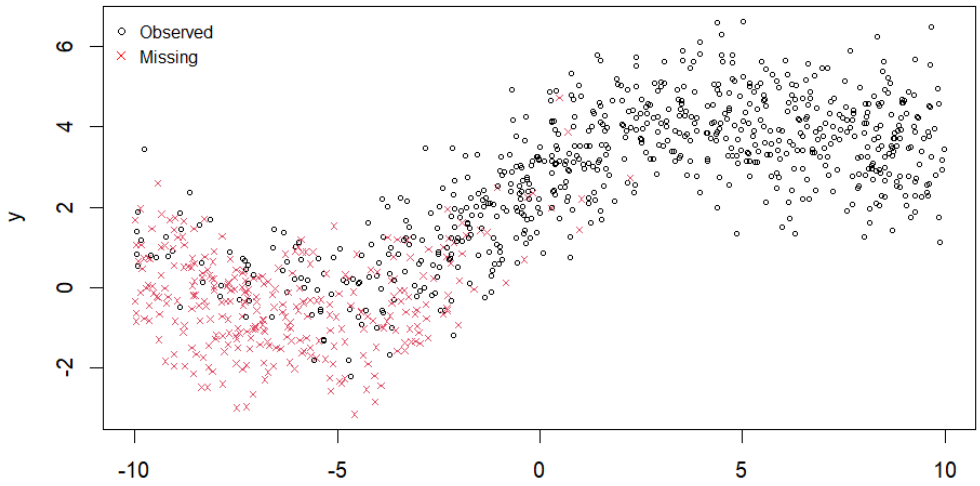


Figure 1: Simulated Data

During this simulation four estimators will be compared:

1. Mojirsheibani and Khudaverdyan (MOJ) as in 2.8;
2. Nadaraya-Watson on the observed data (NW);
3. Nadaraya-Watson with follow-up sub-sample (NW+);
4. Nadaraya-Watson estimator using the values of Kim and Yu (KY).

The Nadaraya-Watson estimator was implemented using the `np` package developed by Hayfield and Racine [2008]. All bandwith are obtained using the cross-validation method described by Racine and Li [2004] and implemented in R with the function `npregbw`.

The fourth estimator is built upon the proposed solution of Kim and Yu to estimate the expected value of Y_i with non-ignorable missing data. Their proposed estimator for μ_Y takes the form of

$$\hat{\mu}_Y = n^{-1} \sum_i \big(\Delta_i Y_i + (1 - \Delta_i)\hat{m}_0(x_i; \hat{\gamma})\big) \tag{3.1}$$

where

$$\hat{m}_0(x; \hat{\gamma}) = \sum_i \frac{\Delta_i \mathcal{K}\big(\frac{x - X_i}{h}\big) \exp\{\hat{\gamma} Y_i\}}{\sum_j \Delta_j \mathcal{K}\big(\frac{x - X_j}{h}\big) \exp\{\hat{\gamma} Y_j\}} Y_i, \tag{3.2}$$

and $\hat{\gamma}$ is found as the solution of

$$\sum_i (1 - \Delta_i)\delta_i\big(Y_i - \hat{m}_0(x_i; \gamma)\big) = 0. \tag{3.3}$$

Thus, the fourth estimator will be a simple Nadaraya-Watson based on the couple $(X_1, \psi_1), \ldots, (X_n, \psi_n)$ where $\psi_i = \Delta_i Y_i + (1 - \Delta_i)\hat{m}_0(x_i; \hat{\gamma})$.

3.1 Finding γ

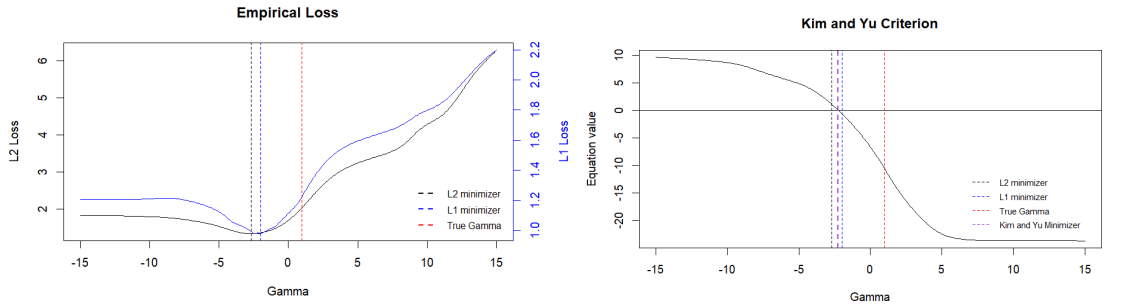
Using the empirical L_2 and the L_1 error we can find a value of γ consistent in the L_2 and L_1 sense. Thus the values obtained with the empirical loss will lead to a consistent estimate not necessarily in the point-wise sense.

The value chosen for p_n used for the follow-up sub-sample was based on the following formula

$$p_n = \sqrt{\frac{\log(n)^{0.25}}{n\lambda^{1-\alpha}}} = 0.042 \tag{3.4}$$

with $\lambda = 0.95$ and $\alpha = 0.01$

Figure 2a show that the empirical loss are convex near the optimal values with both losses finding a unique minimizer. The ranged explored for the values of γ was $[-15, 15]$.



(a) Empirical Losses **(b)** Values of Equation 3.3 for different γ

Figure 2

As it can be seen from then behaviour of the losses, the minimizer are indeed not point-wise consistent with the true γ.

Given the simulation study, the empirical L_2 error had weigths of $1/\ell = 0.0033$ and $1/\ell p_n = 0.078$, thus the resampled values were almost 24 times more important as those of the validation set. This highly emphasize how the proposed estimator uses the sub-sample to obtain a regression estimate that better fits the missing data.

The Kim and Yu criteria also obtains a similar value for γ, as Figure 2b shows.

With the estimated γ the missing probabilities can be estimated using $[1+\varphi(y;\hat{\gamma})]^{-1}$. Figure 3 displays the the estimated missing probabilities against the actual probabilities obtained using both $g(x)$ and $\varphi(y)$.

The probabilities exhibit similar behaviour, with the estimated one being inside the range of possible values. This difference can be explained by the fact that the estimator does not consider $g(x)$ in the estimation process, and it focuses on the mechanism that is responsable for the non-ignorability of the data, hence, its relation with the Y.

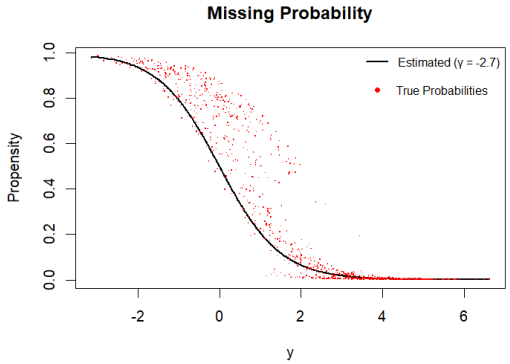


Figure 3: Missing probabilities

3.2 Estimators

The dataset $\mathbb{D}_m$ was created by using 70% of the units in $\mathbb{D}_n$ and 6 units where resampled from the non-respondants in $\mathbb{D}_\ell$

Figure 4 illustrates the behaviour of the four estimators compared. For MOJ estimator the optimal empirical L_2 γ was employed.

All four estimators exhibit similar behaviour where there is no to few missing values, while where the most part of the missing values lie the four estimator split. Both MOJ and KY exhibit similar behaviour, with the estimated $\exp\{\hat{gamma}\}$ driving the function downwards with respect to the NW estimator.

interestingly enough, adding the follow-up sub-sample does not significantely improve the NW estimator, and the the MOJ and KY estimators with only the training set $\mathbb{D}_m$.

To better capture how the different estimator fit the missing data process,95% CI using bootstrap were computed for all four estimates using the function `boot` in the package `boot` of R.

Figure 5 displays the four estimators and the corresponding bootstrap confidence interval. In every plot also the NW and MOJ are also plotted with lighter stroke to facilitate comparison with the other estimators.

The most important result that can be seen from Panel 5a and 5b is that the Mojirsheibani and Khudaverdyan not only fit better the missing gap, but it also achieve smaller confidence interval where all the data is observed; thus it also improves the general fit of the data.

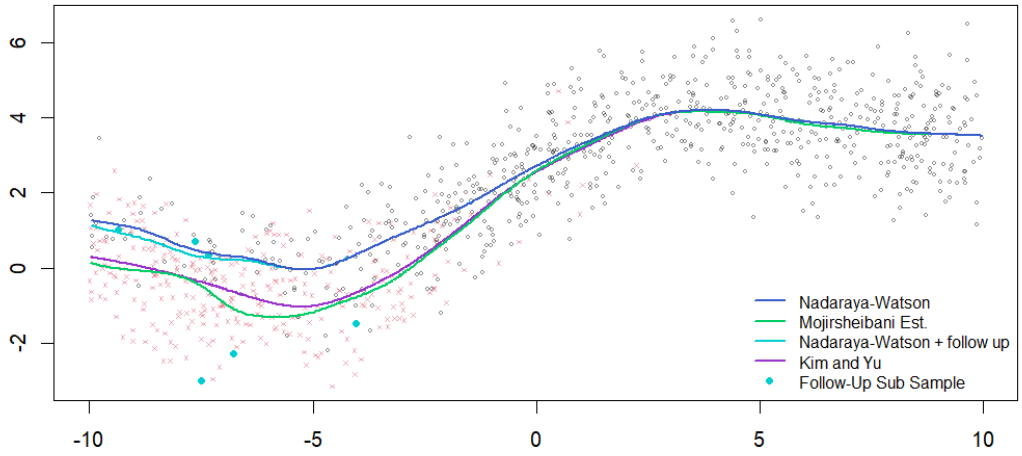
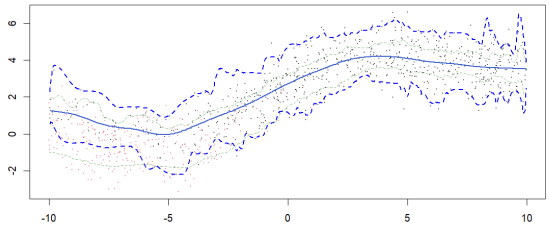
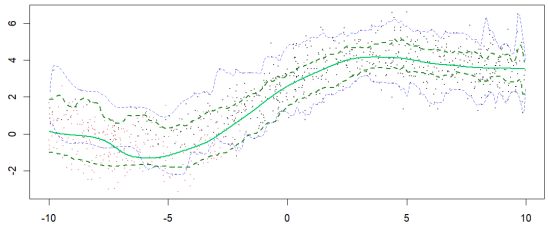


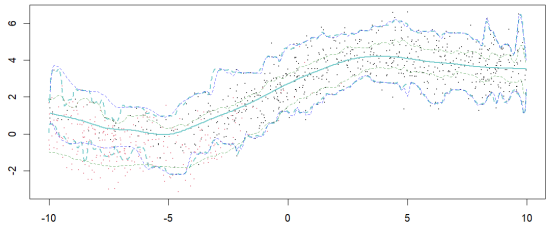
Figure 4: The four estimators plotted with the follow-up sub-sample



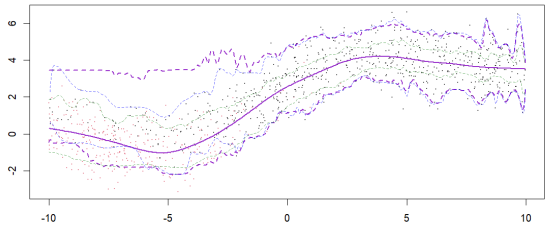
(a) Nadaraya-Watson Estimator



(b) Mojirsheibani Estimator



(c) Nadaraya-Watson with Follow-up



(d) Kim and Yu Estimator

Figure 5: 95% Boostrap Confidence Interval results

The estimator based on the Kim and Yu paper with calculated ψ_i values achieves a lower bound which is very similar to the Mojirsheibani and Khudaverdyan one, but it fails to achieve a similar upper bound. Moreover, due to the fact that when $\Delta_i = 1$ we are fitting a simple Nadaraya-Watson estimator to the observed values, thus the effect of the tilting parameter γ is observed only where the data is missing, while the MOJ estimator fits the data with less variation all throughout the values.

Another comparison between the four estimator can be done using many replications of this setup and obtaining the distribution of the MSE for the four estimators. The MSE was obtained by taking estimator built with the samples from $\mathbb{D}_m$ and evaluating it over all the data points X_i in $\mathbb{D}_n$, thus also those in $\mathbb{D}_\ell$ and the X_i with $\Delta_i = 0$.

From each sample size and noise variance combination 100 datasets were obtained. TFor each dataset the processes described above were computed and the MSE error was obtain by averaging the MSE for each dataset, calculated as

$$n^{-1} \sum_{i \in \mathcal{I}_n} \left(\hat{m}_m(X_i) - Y_i \right)^2$$

Thus, we refer to a *population*-level MSE as it compares how the estimator performe also on the missing values. The result of the distribution of each MSE across the various combinations in Figure 6 clearly suggest that the estimator proposed by Mojirsheibani and Khudaverdyan outperforms the other three estimators in each scenarios.

Moreover, the L_2 consistency is suggested from the densities as the MSE (which is a measure of the L_2 type error) obtained by the MOJ estimator always has small variance. On the converse, the Kim and Yu values do not manage to achieve such small variance in every scenario: nonetheless, with a very small sample size, the two estimator seem to yield the same result, but as the sample size increase also the KIM estimator starts to be less consistent

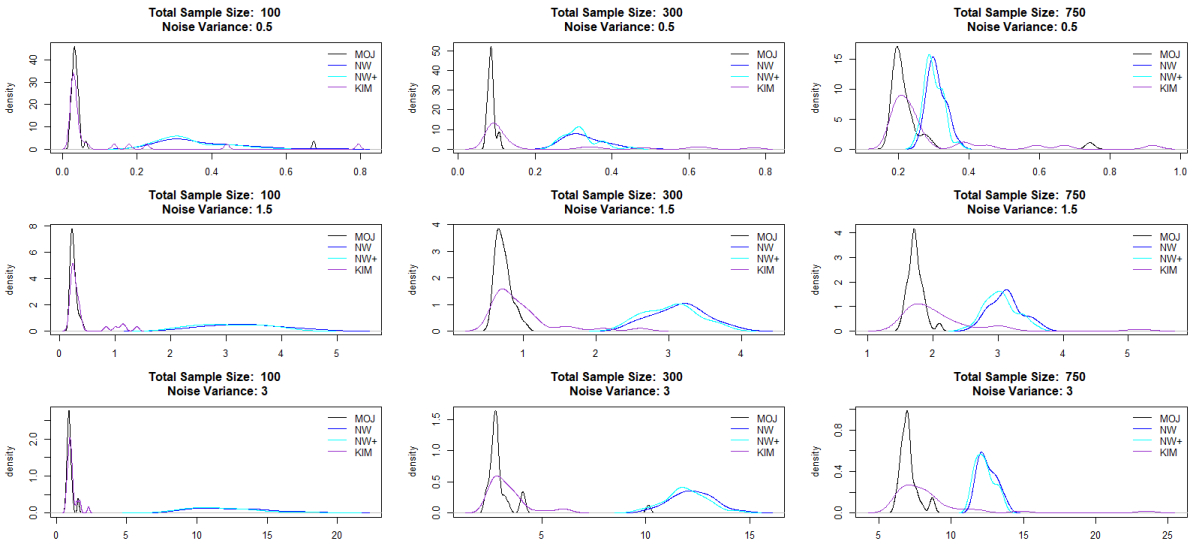


Figure 6

Table 1 summarizes the results shown with the densities using 95% Confidence Interval based on a normal approximation. Once again, it can be seen that both MOJ and KY outperform the simple NW estimators.

Using bootstrap, also the follow-up sub-sample was studied. From Figure 7 it appears that the sub-sampled follow-up is indeed correctly sampled from where the most missingness occurs. Also, with respect of the sample size of $n = 1001$, the follow up sub sample introduce a significantely small number of units ranging from 1 to at most 12. This can be considered a fairly small sub sample, which makes more evident the strength of the estimator, that with such few values manages to significantely improve the fit also with respect to the NW+ estimator.

One last consideration where the convergence rate for the various model used. By computing the values described above, we can obtain the value displayed in Table 2. As mentioned previously, fixing $p_n = c \in [0, 1]$ highly improves the convergence rate.

Table 1: Average MSE and 95% corresponding Confidence Interval

N	Noise	NW	NW+	MOJ	KIM
100	0.5	0.376 (0.12, 0.63)	0.320 (0.17, 0.47)	0.031 (0.02, 0.05)	0.042 (-0.08, 0.17)
300	0.5	0.327 (0.24, 0.42)	0.307 (0.23, 0.39)	0.086 (0.05, 0.12)	0.090 (-0.03, 0.21)
750	0.5	0.307 (0.25, 0.36)	0.296 (0.25, 0.34)	0.210 (0.15, 0.26)	0.205 (0.13, 0.28)
100	1.5	3.235 (1.90, 4.57)	2.915 (1.75, 4.08)	0.239 (0.16, 0.32)	0.231 (0.14, 0.32)
300	1.5	3.140 (2.54, 3.74)	3.002 (2.43, 3.57)	0.703 (0.57, 0.83)	0.694 (0.54, 0.85)
750	1.5	3.181 (2.74, 3.62)	3.101 (2.68, 3.52)	1.785 (1.44, 2.13)	1.750 (1.46, 2.04)
100	3.0	12.761 (7.24, 18.28)	11.876 (6.79, 16.96)	0.922 (0.60, 1.25)	0.911 (0.57, 1.25)
300	3.0	12.673 (9.91, 15.44)	12.222 (9.61, 14.83)	2.823 (2.24, 3.41)	2.784 (2.27, 3.30)
750	3.0	12.517 (10.90, 14.14)	12.302 (10.72, 13.89)	7.020 (5.88, 8.16)	6.951 (6.10, 7.80)

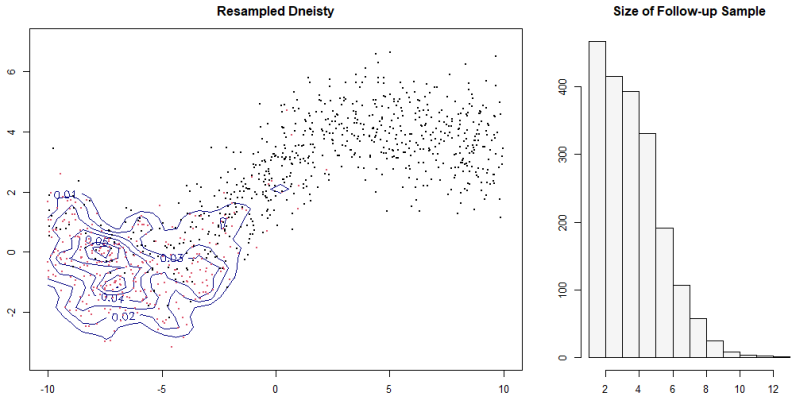


Figure 7: Follow-up Sub-Sample Distribution

Table 2: Rates of convergence for the considered estimators

Model	Bandwidth	Sample Size	Rate
Moj with $p_n = c$	0.7437	700	0.0964
Complete NW	0.6716	1001	0.4525
Moj with $p_n = o(1)$	0.7437	700	0.4816
NW with Kim Values	0.7505	1001	0.5646
NW with Missing Data	0.7516	698	0.5669

4 Conclusion

Mojirsheibani and Khudaverdyan propose a kernel-type semiparametric regression estimator for the case of NMAR non-ignorable response variable. Using a full non-parametric specification both for the regression estimate and the missing probability would lead to identifiability issues. Thus Mojirsheibani and Khudaverdyan resort to the setup of Kim and Yu which specify a parametric model for the missing probability.

The aim of the paper is to generalize the class of functions that can be used to fit the missing probability by defyning an ε-cover of the true function φ. In pratice, a parametric function φ_γ is used, such that the estimation procedure and the optimization procedure boils down to estimating $\hat{\gamma}$ that fully describes φ_γ.

The estimator turned out to improve the fit on the data not only by capturing the missing data values with a small sub sample, but by also achieve smaller confidence interval than the Nadaraya-Watson with and without the follow-up sub-sample.

Unfortunately, no real data example was used due to the specificity of the dataset. Indeed, altough it is fairly common to find datasets with non-ignorable Not Missing At Random response variable, the requirment to have a follow-up sub-sample is not quite feasable in pratice. To overcome this issue Kim and Yu [2011] recreated a real data example by creating a missing function as specified in the course of this report and applied it to a real dataset. Moreover, the choice of the specification of a φ_γ thath satisfy the mathematical properties for the proposed estimator is up to the user.

At last, the proposed estimator is still capable to achieve fast convergence rate under some conditions, thus the estimators seems as a valid tool to use in the case of NMAR data.

References

Majid Mojirsheibani and Arin Khudaverdyan. A kernel-type regression estimator for NMAR response variables with applications to classification. *Stat. Probab. Lett.*, 215(110246): 110246, December 2024.

Roderick J A Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, Inc., Hoboken, NJ, USA, August 2002.

Donald B Rubin. *Multiple Imputation for Nonresponse in Surveys*. Probability & Mathematical Statistics S. John Wiley & Sons, Nashville, TN, 99 edition, July 1987.

Donald B. Rubin. Missing data, imputation, and the bootstrap: Comment. *Journal of the American Statistical Association*, 89(426):475–478, 1994. ISSN 01621459, 1537274X. URL `http://www.jstor.org/stable/2290847`.

Fritz Scheuren. Multiple imputation: How it began and continues. *The American Statistician*, 59(4):315–319, 2005. ISSN 00031305. URL `http://www.jstor.org/stable/27643702`.

Jae Kwang Kim and Cindy Long Yu. A semiparametric estimation of mean functionals with nonignorable missing data. *J. Am. Stat. Assoc.*, 106(493):157–165, March 2011.

Jun Shao and Lei Wang. Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika*, 103(1):175–187, March 2016.

Tristen Hayfield and Jeffrey S. Racine. Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5):1–32, 2008. doi: 10.18637/jss.v027.i05. URL `https://www.jstatsoft.org/index.php/jss/article/view/v027i05`.

Jeff Racine and Qi Li. Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, 119(1):99–130, 2004. ISSN 0304-4076. doi: https://doi.org/10.1016/S0304-4076(03)00157-X. URL `https://www.sciencedirect.com/science/article/pii/S030440760300157X`.