# Nonparametric Regression with Response Missing at Random and the Scale Depending on Auxiliary Covariates

Tian Jiang* Sam Efromovich*

**Abstract**

Nonparametric regression with missing at random (MAR) response, univariate regression component of interest, and the scale function depending on both the predictor and auxiliary covariates, is considered. The asymptotic theory suggests that the heteroscedasticity and MAR affect the constant of the sharp minimax MISE convergence. The sharp minimax procedure is based on estimation of unknown nuisance scale function, design density and missing mechanism. The estimator is adaptive to the missing mechanism and unknown smoothness of the estimated regression function. The procedure is tested on simulated data and real examples, and the results justify practical feasibility of the proposed method for this complex regression setting.

**Key Words:** Adaptation, Availability likelihood, Curse of multidimensionality, Heteroscedasticity, MAR, Sharp minimaxity

## 1. Introduction

Nonparametric regression analysis explores the association between response $Y$ and predictor $X$ with almost no assumption about shape of an underlying regression function, which is defined by conditional expectation of response $Y$ given predictor $X$, $m(x) = \mathbb{E}\{Y|X = x\}$. Consider a heteroscedastic regression model

$$Y = m(X) + \sigma(X)\epsilon, \tag{1}$$

where $\sigma(x)$ is called scale or volatility function and $\epsilon$ independent of predicator is a mean zero random error with unit variance. Without loss of generality, we assume predictor is supported on unit interval. There is a vast literature devoted to nonparametric regression using a variety of approaches such as local polynomial, kernel, spline, tree-based method, wavelet and so on; see Efromovich (1999), Wasserman (2006) and Tsybakov (2009) for more details. It is popular to study minmax risk of estimators for functions from Sobolev class and one exciting result is that efficient nonparametric regression estimation is possible even without estimating the scale function, that is, Efromovich and Pinsker (1996) proposed such an estimator $\hat{m}$ achieves optimal convergence in the sense of mean integrated squared error (MISE), MISE$(\hat{m}, m) = \mathbb{E}\{\int_0^1 [\hat{m}(x) - m(x)]^2 dx\}$. We can say the nonparametric regression problem is not sensitive to heteroscedasticity. However, the situation is changing drastically when volatility (scale function) is also driven by some auxiliary process besides predictor of interest,

$$Y = m(X) + \sigma(X, \mathbf{Z})\epsilon, \tag{2}$$

where scale function also involves a $D$-dimensional random vector of covariates, $\mathbf{Z} := (Z_1, Z_2, \ldots, Z_D)$ independent of random error term. Then estimators ignoring heteroscedasticity may still be rate optimal but will no longer be sharp minimax because optimal constant of convergence is inflated. Using scale function in weights explicitly, Efromovich (2013) proposed a data-driven estimator that preserves asymptotic efficiency.

*Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, TX 75080, USA

In this paper, we will study the above nonparametric regression problem in a more general setting when response cannot be observed directly. Missing data is a common challenge for statistical analysis, especially in social and health sciences. Enders (2010), Molenberghs et al. (2014) and Little and Rubin (2019) give a nice discussion about conventional approaches such as likelihood-based parametric inference and weighting based semiparametric methods. However, nonparametric methods is not well developed for missing and modified data analysis and some attempts were made in kernel estimation, series estimation and Bayesian inference in Chu and Cheng (1995), Boente et al. (2009), Mitra and Müller (2015), Efromovich (2018) and Sun, Wang and Han (2019). In the seminal paper of Rubin (1976), the classic hierarchy of missing mechanism is introduced, that is, missing completely at random (MCAR) when missing is not related to observation, missing at random (MAR) when missing depends on observation and missing not at random (MNAR) when missing is also related to unobserved sample. We will focus on the case of response missing at random. Historically, missing data comes from controlled experiments where design factors are fixed and data with missing response is of great interest. Under certain assumption, MAR justifies complete-case approach in parametric estimation.

The main aim of this article is to study nonparametric regression estimation under complication due to both heteroscedasticity and MAR mechanism. To be more specific, regression model (2) generates an underlying or hidden sample (H-sample) of size $n$, $\{(X_l, \mathbf{Z}_l, Y_l), l = 1, 2, \dots, n\}$, where continuous covariates $(X, \mathbf{Z})$ are supported on unit $(D+1)$-dimensional cube. Instead of H-sample, we can only observe a M-sample with missing response, $\{(X_l, \mathbf{Z}_l, A_l Y_l, A_l), l = 1, 2, \dots, n\}$, where realizations of Bernoulli random variable $A$ defining availability of response, that is, underlying response $Y$ is available (observable) when $A = 1$ while it is missed when $A = 0$. Here, a continuous response $Y$ is considered so that we will not distinguish missing ($A = 0$) and zero-value response ($Y = 0$) since $\mathbb{P}(Y = 0) = 0$. Then the missing mechanism is characterized by success probability of Bernoulli distribution depending only on always observable covariates, $\mathbb{P}(A = 1|X = x, \mathbf{Z} = \mathbf{z}, Y = y) = \mathbb{P}(A = 1|X = x, \mathbf{Z} = \mathbf{z}) =: w(x, \mathbf{z})$, where function $w(x, \mathbf{z})$ is referred to as availability likelihood. Then complete-case subsample only consists of observations with $A = 1$, $\{(X_{(l)}, \mathbf{Z}_{(l)}, A_{(l)} Y_{(l)}, A_{(l)} = 1), l = 1, 2, \dots, N\}$, where $N := \sum_{l=1}^{n} A_l$ is the number of complete cases. Note that MCAR is just a special case with constant availability likelihood, $w(x, \mathbf{z}) = w_0$ for some $0 < w_0 < 1$. The observed quadruplet $(X, \mathbf{Z}, AY, A)$ has a mixed distribution with joint mixed density

$$
\begin{aligned}
f^{X,\mathbf{Z},AY,A}(x, \mathbf{z}, ay, a) &= \mathbb{P}(A = a|X = x, \mathbf{Z} = z) f^{X,\mathbf{Z},AY}(x, \mathbf{z}, ay) \\
&= \Big[ w(x, \mathbf{z}) f^{X,\mathbf{Z}}(x, \mathbf{z}) f^{Y|X,\mathbf{Z}}(y|x, \mathbf{z}) \Big]^a \Big[ [1 - w(x, \mathbf{z})] f^{X,\mathbf{Z}}(x, \mathbf{z}) \Big]^{1-a},
\end{aligned}
$$

for $(x, \mathbf{z}) \in [0, 1]^{D+1}$, $y \in \mathbb{R}$ and $a \in \{0, 1\}$. Then the conditional density of complete-case subsample

$$
\begin{aligned}
f^{Y|X,\mathbf{Z},A}(y|x, \mathbf{z}, 1) &= \frac{f^{X,\mathbf{Z},Y|A}(x, \mathbf{z}, y, 1)}{f^{X,\mathbf{Z}}(x, \mathbf{z})} = \frac{w(x, \mathbf{z}) f^{X,\mathbf{Z}}(x, \mathbf{z}) f^{Y|X,\mathbf{Z}}(y|x, \mathbf{z})}{f^{X,\mathbf{Z}}(x, \mathbf{z}) \mathbb{P}(A = 1)} \\
&= \frac{w(x, \mathbf{z})}{\int_{[0,1]^{D+1}} w(u, \mathbf{v}) f^{X,\mathbf{Z}}(u, \mathbf{v}) du d\mathbf{v}} f^{Y|X,\mathbf{Z}}(y|x, \mathbf{z}), \qquad (3)
\end{aligned}
$$

is biased from its underlying counterpart $f^{Y|X,\mathbf{Z}}(y|x, \mathbf{z})$ and so is the regression function of interest based on complete cases (a conditional expectation with $Z$ integrated out). Fortunately, without auxiliary covariates $\mathbf{Z}$, efficient complete-case nonparametric estimation with MAR response is established for univariate model (1) in Efromovich (2011, 2012) and we will use E-estimator of Efromovich (2018) as a pure univariate benchmark.

The rest of this paper is organized as follows. In Section 2 we establish large sample theory of minimax risk bound and then propose adaptive estimators that achieve asymptotic efficiency. Section 3 is devoted to numerical studies. We suggest modified estimators for sample data sets and compare them with univariate benchmarks in intensive Monte Carlo simulations and real examples. Proofs are omitted due to space constraints.

Here, we briefly discuss the terminology of a minimax approach and related estimators that will be used throughout the paper. We will develop asymptotic theory based on a minimax game with four players, that is, nature, the oracle, the dealer and the statistician. The rules are defined by our regression model (2) with MAR mechanism, a class of regression functions of interest, assumptions about nuisance functions and a risk criterion (specifically, MISE). In the minimax game, dealer shows nature the chosen parameters of functional class and nuisance functions. Then nature picks a regression function from the dealt class to maximize its payoff in terms of the risk criterion MISE and generates M-sample while other players propose estimators based on M-sample to minimize MISE. The equilibrium between dealer and nature gives the minimax risk and any estimator achieves this bound is called efficient or sharp minimax. Our goal is a data driven sharp minimax procedure, that is, a statistician's estimator base solely on data without knowing nuisance functions and parameters of considered class. Oracle even knows the chosen estimand besides everything dealer knows and it can suggest a sharp estimator for statistician to mimic its performance. Interested readers can refer to Berger (1985) and Lehmann and Casella (1998) for more details about statistical decision theory and game theory.

## 2. Asymptotic Theory

Since minimax MISE convergence of a nonparametric estimator depends on the smoothness of estimand function, Sobolev class or Sobolev ellipsoid is usually considered in nonparametric literature,

$$\mathcal{E}(\alpha, Q) := \left\{ m : m(x) = \sum_{j=0}^{\infty} \theta_j \varphi_j(x), \sum_{j=0}^{\infty} [1 + (\pi j)^{2\alpha}] \theta_j^2 \leq Q \right\}, \tag{4}$$

where $\theta_j := \int_0^1 m(x)\varphi_j(x)dx$ is the Fourier coefficient for regression function $m(x)$ with respect to $j$th cosine basis function on unit interval [0,1],

$$\varphi_0(x) = 1, \qquad \varphi_j(x) = 2^{1/2}\cos(\pi j x), \qquad j = 1, 2, 3, ... \tag{5}$$

The positive number $Q$ bounds power or energy of member functions and parameter $\alpha > 1$ determines their smoothness. Sometimes it may be reasonable to confine the estimation problem in some vicinity of a particular function $m_0$ of interest instead of a global solution, which is the so called local minimax approach introduced by Golubev (1991) or shrinking class minimax if a sequence of local classes converges to the pivot function as we get more observations. We introduce a more general family of function classes covering all the above discussed classes,

$$\begin{aligned} \mathcal{F} &:= \mathcal{F}(m_0, \rho_n, M_n, \alpha, Q) \\ &:= \left\{ m(x) : m(x) = \sum_{j=0}^{M_n - 1} \theta_{0,j}\varphi_j(x) I(M_n > 0) + \sum_{j \geq M_n} \theta_j \varphi_j(x), x \in [0,1], \right. \\ &\qquad \left. \sum_{j \geq M_n} [1 + (\pi j)^{2\alpha}] \theta_j^2 \leq Q < \infty, \sup_{x \in [0,1]} \left| \sum_{j \geq M_n} \theta_j \varphi_j(x) \right| < \rho_n \right\}. \end{aligned} \tag{6}$$

for a pivot function $m_0$ satisfying squared integrability $\int_0^1 m_0^2(x)dx < \infty$ and finite sup norm $\sup_{x\in[0,1]} |m_0(x)| < \infty$, an integer lower frequency cutoff $M_n \leq n^{1/(2\alpha+1)}/\ln^2(n)$ and a tail bound $\rho_n > n^{-1/(2\alpha+1)}\ln(n)$. Here, $\theta_{0,j} := \int_0^1 m_0(u)\varphi_j(u)du$ is the Fourier coefficient of $m_0$ and $I(\cdot)$ stands for indicator throughout the paper.

Let us make some comments about the above family of function classes. All the member functions from this family share the same low-frequency part as the pivot regression $m_0$ in Fourier frequency domain. Further, the number (cardinality) of low frequencies $M_n$ controls $L_2$ norm of tail series while parameter $\rho_n$ controls its sup norm. It is also easy to see that classic Sobolev class $\mathcal{E}(\alpha, Q)$ corresponds to the case $\mathcal{F}(0, \infty, 0, \alpha, Q)$ without pivot function and additional constraint on tail series.

Then we can formally give the setting and assumptions for our regression model (2) $Y = m(X) + \sigma(X, \mathbf{Z})\epsilon$, where univariate nonparametric regression function $m(x)$ is our estimand of interest. The following moment conditions are imposed on error term,

$$\mathbb{E}\{\epsilon|X, \mathbf{Z}\} = 0, \qquad \mathbb{E}\{\epsilon^2|X, \mathbf{Z}\} = 1, \qquad \mathbb{E}\{\epsilon^4|X, \mathbf{Z}\} < C < \infty \qquad \text{a.s.} \tag{7}$$

Note that nonparametric approach relaxes distribution assumption about error term but finite fourth moment condition is required for adaptation. In order to obtain dealer's lower bound for minimax MISE risk, we employ stronger assumptions of normality and independence in asymptotic analysis. Some mild regularity assumptions on nuisance functions are also necessary.

**Assumption 2.1.** *The error term $\epsilon$ follows standard normal distribution and is independent of covariates $(X, \mathbf{Z})$.*

**Assumption 2.2.** *The joint design density $f^{X,\mathbf{Z}}(x, \mathbf{z})$ and availability likelihood $w(x, \mathbf{z})$ are supported on $[0, 1]^{D+1}$. Nuisance functions $f^{X,\mathbf{Z}}(x, \mathbf{z})$, $w(x, \mathbf{z})$ and $\sigma(x, \mathbf{z})$ are bounded below from zero and also bounded above. In addition, $w(x, z)$ can not exceed 1. The quantity $\mathcal{I}(x) := \int_{[0,1]^D} f^{X,\mathbf{Z}}(x, \mathbf{z})w(x, \mathbf{z})\sigma^{-2}(x, \mathbf{z})d\mathbf{z}$ is Riemann integrable on $[0, 1]$.*

The above assumptions are not very restricted since dealer can directly use them to establish lower bound. Additional assumptions about smoothness are required when we consider statistician's adaptation for unknown nuisance functions.

Denote M-sample $(X, \mathbf{Z}, AY, A)^n := \{(X_1, \mathbf{Z}_1, A_1Y_1, A_1), \ldots, (X_n, \mathbf{Z}_n, A_nY_n, A_n)\}$ and introduce a notation

$$\tilde{m}^*(x) := \tilde{m}^*(x; (X, \mathbf{Z}, AY, A)^n, m_0, \rho_n, M_n, \alpha, Q, f^{X,\mathbf{Z}}, \sigma, w) \tag{8}$$

for a representative dealer's estimator that exploit the privileged knowledge of the dealt class $\mathcal{F}$ of regression functions and the chosen nuisance functions.

**Theorem 2.1.** *Let Assumptions 2.1 and 2.2 hold for the regression model (2). Then we have the following lower minimax bound for dealer's estimators*

$$\inf_{\tilde{m}^*} \sup_{m\in\mathcal{F}(m_0,\rho_n,M_n,\alpha,Q)} \text{MISE}(\tilde{m}^*, m) \geq P(\alpha, Q)\left(n^{-1}d\right)^{2\alpha/(2\alpha+1)}(1 + o_n(1)), \tag{9}$$

*where the infimum is taken over all possible dealer's estimator $\tilde{m}^*$, $d$ is the coefficient of difficulty*

$$d := \int_0^1 \frac{dx}{\int_{[0,1]^D} f^{X,\mathbf{Z}}(x, \mathbf{z})w(x, \mathbf{z})\sigma^{-2}(x, \mathbf{z})d\mathbf{z}} \tag{10}$$

*and Pinsker constant $P(\alpha, Q) := [\alpha/\pi(\alpha + 1)]^{2\alpha/(2\alpha+1)}[Q(2\alpha + 1)]^{1/(2\alpha+1)}$.*

*Furthermore, there exists a dealer's estimator $\check{m}^*$ that achieves the above lower bound,*

$$\sup_{m\in\mathcal{F}(m_0,\rho_n,M_n,\alpha,Q)} \mathrm{MISE}(\check{m}^*, m) \le P(\alpha, Q)\left(n^{-1}d\right)^{2\alpha/(2\alpha+1)}(1+o_n(1)). \tag{11}$$

Let us make some comments about the above result. The first part establishes a lower bound for minimax risk of dealer's estimators and the second part verifies its sharpness, which means certain dealer's estimator indeed attains the lower bound. The factor $1+o_n(1)$ indicates the sharp lower bound is an asymptotic result. The renown constant $P(\alpha, Q)$ is discovered in the seminal work Pinsker (1980) and it is determined by the implied Sobolev ellipsoid in class $\mathcal{F}$ while the coefficient of difficulty is a functional of only nuisance functions of our regression model, $d := d(f^{X,\mathbf{Z}}, w, \sigma) = \int_0^1 \mathcal{I}^{-1}(x)dx$, where exponent -1 denotes reciprocal of $\mathcal{I}(x)$. The form of lower bound is general over a variety of nonparametric estimation problem and $d$ characterizes the model setting. It is easy to show that if we ignore its heteroscedasticity depending on auxiliary variable and estimate it as univariate model (1), an actually inflated coefficient of difficulty implies a loss of efficiency.

Our goal is a data-driven estimation procedure adapted to unknown smoothness of regression function of interest, missing mechanism and underlying nuisance functions. In Theorem 2.1, we verify sharpness by a Pinsker type linear minimax estimator, which is difficult for further adaptation. So we resort to a blockwise shrinkage procedure and begin with an oracle one using the classic idea dating back to Efromovich and Pinsker (1984). Let $\{B_k, k = 1, 2, \ldots\}$ denote the set of ordered blocks partitioning nonnegative integers such that $\max\{j : j \in B_k\} < \min\{j : j \in B_{k+1}\}$ and $L_k$ denote its length (cardinality), the number of frequencies in the block $B_k$. Our blockwise shrinkage oracle is a smoothed series estimator using equal smoothing weight for frequencies in the same block, that is,

$$\hat{m}^*(x) := \sum_{k=1}^{K_n} \mu_k \sum_{j\in B_k} \hat{\theta}_j \varphi_j(x), \tag{12}$$

where $\hat{\theta}_j$ estimates $j$th Fourier coefficient of the regression function $\theta_j = \int_0^1 m(x)\varphi_j$, cutoff $K_n$ is some nondecreasing positive sequence and smoothing coefficient

$$\mu_k := \frac{\Theta_k}{\Theta_k + dn^{-1}} \tag{13}$$

is defined by the coefficient of difficulty $d$ (10) and the Sobolev functional

$$\Theta_k := L_k^{-1} \sum_{j\in B_k} \theta_j^2. \tag{14}$$

We can show the above oracle's estimator (12) attains the dealer's lower bound under some assumptions about cutoff $K_n$ and $\hat{\theta}_j$, which is reasonable since the use of underlying Fourier coefficients reflects oracle's privileged information.

For further adaptation, our strategy is to replace functionals of underlying model with statistics and then analyze requirements for a sharp minimax adaptive estimator mimicking oracle (12). Specifically, we will use a general blockwise shrinkage estimator framework

$$\hat{m}(x) := \sum_{k=1}^{K_n} \frac{\widehat{\Theta}_k}{\widehat{\Theta}_k + \hat{d}n^{-1}} I(\widehat{\Theta}_k > (b_n n)^{-1}) \sum_{j\in B_k} \hat{\theta}_j \varphi_j(x). \tag{15}$$

Here, hardthresholding is also employed for better performance. Let us introduce some notations for increasing sequence at different rates, $b_n := \lfloor \ln(n + 20) \rfloor$, $c_n := \lfloor \ln(b_n) \rfloor$

and $r := \lfloor n/(7c_n) \rfloor$, where $\lfloor x \rfloor$ denotes the greatest integer that is less than or equal to $x$. From now on, it is assumed that $n$ is large enough such that $r > 3$. Set block length $L_k := 1$ for low frequency range $k = 1, 2, \ldots, b_n$ and $L_k := \lfloor (1 + b_n^{-1})^k \rfloor$ for $k > b_n$. It is also easy to show the total number of blocks $K_n$ is of order $O(\ln^2(n))$. Now we can propose sharp minimax estimators with different level of adaptation.

### 2.1 Known Design, Availability Likelihood and Scale

In this case, parameters of functional class $\mathcal{F}(m_0, \rho_n, M_n, \alpha, Q)$ are the only unknown components of model setting. Note that it is weaker than the dealer's information in Theorem 2.1. The suggested estimator of this section can be regarded as an adaptive estimator under controlled experiment. However, it is also an eligible dealer's estimator when nuisance function are unknown for statistician. Later we will also use it as a benchmark in numerical experiments. We begin with component estimates of blockwise shrinkage estimator (15). Define a U-statistic to estimate Sobolev functional,

$$\widehat{\Theta}_k := \frac{2}{r(r-1)} \sum_{r+1 \le l_1 < l_2 \le 2r} L_k^{-1} \sum_{j \in B_k} \prod_{t=1}^{2} \frac{A_{l_t} Y_{l_t} \varphi_j(X_{l_t})}{f^{X,\mathbf{Z}}(X_{l_t}, \mathbf{Z}_{l_t}) w(X_{l_t}, \mathbf{Z}_{l_t})}, \tag{16}$$

and a scale function weighted Fourier coefficient estimate,

$$\hat{\theta}_j := \frac{1}{n-2r} \sum_{l=2r+1}^{n} \frac{A_l [Y_l - \tilde{m}_{-j}(X_l)] \sigma^{-2}(X_l, \mathbf{Z}_l) \varphi_j(X_l)}{\mathcal{I}(X_l)}, \tag{17}$$

with underlying quantity $\mathcal{I}(x) := \int_{[0,1]^D} f^{X,\mathbf{Z}}(x, \mathbf{z}) w(x, \mathbf{z}) \sigma^{-2}(x, \mathbf{z}) d\mathbf{z}$ and variation reduction term

$$\tilde{m}_{-j}(x) := r^{-1} \sum_{l=1}^{r} \sum_{i \in \mathcal{N}_{-j}} \frac{A_l Y_l \varphi_i(X_l)}{f^{X,\mathbf{Z}}(X_l, \mathbf{Z}_l) w(X_l, \mathbf{Z}_l)} \varphi_i(x), \tag{18}$$

where $\mathcal{N}_{-j} = \{0, 1, \ldots, b_n\} \backslash \{j\}$ is the set of irrelevant frequencies. Separate subsamples are used for sequences $\hat{\theta}_j$'s and $\widehat{\Theta}_k$'s, which would simplify proofs with the help of independence. Note that major part of M-sample are used for Fourier coefficient estimates $\hat{\theta}_j$'s, the backbone of series estimator.

**Proposition 2.1.** *Consider the regression model (2) with regression error satisfying (7). Design density, scale and availability likelihood are given and Assumption 2.2 holds. Then the blockwise shrinkage regression estimator (15) with estimated Fourier coefficient $\hat{\theta}_j$ defined in (17), estimated Sobolev functional $\widehat{\Theta}_k$ defined in (16) and the coefficient of difficulty with its underlying value (10), $\hat{d} = d$, is adaptive to the studied functional class $\mathcal{F}(m_0(x), \rho_n, M_n, \alpha, Q)$ and sharp minimax, namely*

$$\sup_{m \in \mathcal{F}(m_0, \rho_n, M_n, \alpha, Q)} \text{MISE}(\hat{m}, m) \le P(\alpha, Q)(d/n)^{2\alpha/(2\alpha+1)}(1 + o_n(1)). \tag{19}$$

### 2.2 Unknown Nuisance Functions

Next we consider the case of fully adaptation, that is, a statistician's estimator solely based on data without knowing any nuisance functions. We will employ a standard plug-in technique for joint design density $f^{X,\mathbf{Z}}(x, \mathbf{z})$, availability likelihood $w(x, \mathbf{z})$, scale $\sigma(x, \mathbf{z})$ and related quantity $\mathcal{I}(x)$ and the coefficient of difficulty $d$ based on the previous estimator.

Before proposing estimates for multivariate nuisance functions, let's introduce some notations for multivariate series estimation. Define a tensor-product cosine basis on $[0,1]^D$, $\psi_{\mathbf{s}}(\mathbf{v}) := \prod_{k=1}^{D} \varphi_{s_k}(v_k)$ for frequency index $\mathbf{s} := (s_1, \ldots, s_D) \in \{0, 1, \ldots\}^D$ and $\mathbf{v} \in [0,1]^D$. We also use sup norm notation to denote the max index $\|\mathbf{s}\|_\infty := \max(s_1, \ldots, s_D)$.

Although no assumption on smoothness of a underlying scale function $\sigma(x, \mathbf{z})$ is imposed, boundedness is required in adaptation, $c_* \le \sigma^2(x, \mathbf{z}) \le c^*$ for some postive constants $c_*$ and $c^*$ known by statistician. To preserve the same upper bound of MISE risk, we have to impose some regularity conditions on joint design density and availability likelihood. For example, let us introduce a $(D+1)$-dimensional analytic class $\mathcal{A} := \mathcal{A}(\beta_0, \ldots, \beta_D, Q)$ with finite positive numbers $Q$ and $\beta_k$, $k = 0, 1, \ldots, D$,

$$\mathcal{A} := \left\{ f : f(x, \mathbf{z}) := \sum_{(i,\mathbf{s})} \theta_{i\mathbf{s}} \varphi_i(x) \psi_{\mathbf{s}}(\mathbf{z}), |\theta_{i\mathbf{s}}| \le Q \Big[ e^{\beta_0 i} + \sum_{k=1}^{D} e^{\beta_k s_k} \Big]^{-1} \right\}. \tag{20}$$

Then projection estimators for design density and availability likelihood from this analytic class will have good approximation properties. It puts a relative strong constraint on Fourier coefficients of member functions for large $D$ due to the curse of dimensionality for multivariate function estimation. Let us introduce nine pairs of independent truncated projection estimators, that is, for $t = 1, \ldots, 9$,

$$\hat{f}_t^{X,\mathbf{Z}}(x, \mathbf{z}) := r^{-1} \sum_{l=(t-1)r+1}^{tr} \sum_{\|(i,\mathbf{s})\|_\infty \le N_a} \varphi_i(X_l) \psi_{\mathbf{s}}(\mathbf{Z}_l) \varphi_i(x) \psi_{\mathbf{s}}(\mathbf{z}), \tag{21}$$

$$\tilde{f}_t^{X,\mathbf{Z}}(x, \mathbf{z}) := \max(c_n^{-1}, \hat{f}_t^{X,\mathbf{Z}}(x, \mathbf{z})), \tag{22}$$

and

$$\hat{w}_t(x, \mathbf{z}) := r^{-1} \sum_{l=(t+8)r+1}^{(t+9)r} \sum_{\|(i,\mathbf{s})\| \le N_a} A_l \varphi_i(X_l) \psi_{\mathbf{s}}(\mathbf{Z}_l) \varphi_i(x) \psi_{\mathbf{s}}(\mathbf{z}), \tag{23}$$

$$\tilde{w}_t(x, \mathbf{z}) := \max(c_n^{-1}, \hat{w}_t(x)), \tag{24}$$

where $N_a := \lfloor b_n c_n \rfloor$ is the cutoff for estimated frequencies. Truncation is used to avoid divide-by-zero problem since joint design density and availability likelihood appear in the denominator of many component statistics.

Then we can propose an estimate for the coefficient of difficulty $d$,

$$\tilde{d} := \int_0^1 \frac{dx}{\int_{[0,1]^D} \tilde{f}_3^{X,\mathbf{Z}}(x, \mathbf{z}) \tilde{w}_3(x, \mathbf{z}) \tilde{\sigma}_1^{-2}(x, \mathbf{z}) d\mathbf{z}} \tag{25}$$

with truncated projection estimator of squared scale function

$$\tilde{\sigma}_1^2(x, \mathbf{z}) := \max\left( c_*, \min\left( c^*, \sum_{\|(i,\mathbf{s})\|_\infty < b_n} \tilde{\sigma}_{1i\mathbf{s}} \varphi_i(x) \psi_{\mathbf{s}}(\mathbf{z}) \right) \right), \tag{26}$$

where $\tilde{\sigma}_{1i\mathbf{s}}$ is a sample mean estimate for Fourier coefficient of $\sigma^2(x, \mathbf{z})$

$$\tilde{\sigma}_{1i\mathbf{s}} := r^{-1} \sum_{l=19r+1}^{20r} \frac{A_l (Y_l - \tilde{m}_1(X_l))^2}{\tilde{f}_2^{X,\mathbf{Z}}(X_l, \mathbf{Z}_l) \tilde{w}_2(X_l, \mathbf{Z}_l)} \varphi_i(X_l) \psi_{\mathbf{s}}(\mathbf{Z}_l),$$

with a truncated estimate for regression function

$$\tilde{m}_1(x) := \max\left( -b_n, \min\left( b_n, r^{-1} \sum_{l=18r+1}^{19r} \sum_{i=0}^{b_n-1} \frac{A_l Y_l \varphi_i(X_l)}{\tilde{f}_1^{X,\mathbf{Z}}(X_l, \mathbf{Z}_l)\tilde{w}_1(X_l, Z_l)} \varphi_i(x) \right) \right).$$

For Sobolev functional in the smoothing coefficient, we have a plug-in estimate

$$\widehat{\Theta}_k := \frac{2}{r(r-1)} \sum_{20r+1 \le l_1 < l_2 \le 21r} L_k^{-1} \sum_{j \in B_k} \prod_{t=1}^{2} \frac{A_{l_t} Y_{l_t} \varphi_j(X_{l_t})}{\tilde{f}_{3+t}^{X,\mathbf{Z}}(X_{l_t}, \mathbf{Z}_{l_t})\tilde{w}_{3+t}(X_{l_t}, \mathbf{Z}_{l_t})}. \tag{27}$$

Then we consider adaptation of estimated Fourier coefficient $\hat{\theta}_j$. The term (18) for variation reduction in $\hat{\theta}_j$ becomes

$$\tilde{m}_{-j}(x) := r^{-1} \sum_{l=21r+1}^{22r} \sum_{i \in \mathcal{N}_{a,-j}} \frac{A_l Y_l \varphi_i(X_l)}{\tilde{f}_6^{X,\mathbf{Z}}(X_l, \mathbf{Z}_l)\tilde{w}_6(X_l, \mathbf{Z}_l)} \varphi_i(x), \tag{28}$$

where index set $\mathcal{N}_{a,-j} = \{0, 1, \ldots, b_n\} \backslash \{j\}$ with subscript '$a$' for the setting using analytic density and availability likelihood. We also need another estimate $\tilde{\sigma}^2(x, \mathbf{z})$ for scale function independent of that in $\tilde{d}$ following the steps of $\tilde{\sigma}_1^2(x, \mathbf{z})$. Since Fourier coefficient estimates $\hat{\theta}_j$'s deserve better accuracy, Fejér approximation is employed to improve the smoothness of estimated inverse squared scale, which is defined by

$$\tilde{\sigma}_{b_n}^{-2}(x, \mathbf{z}) := b_n^{-1} \sum_{t=0}^{b_n-1} \sum_{\|(i,\mathbf{s})\|_\infty \le t} \left[ \int_{[0,1]^{D+1}} \tilde{\sigma}^{-2}(u, \mathbf{v}) \varphi_i(u) \psi_{\mathbf{s}}(\mathbf{v}) du d\mathbf{v} \right] \varphi_i(x) \psi_{\mathbf{s}}(\mathbf{z}).$$

Similarly, the quantity $\mathcal{I}(x)$ in the denominator also uses this improved scale estimate,

$$\tilde{\mathcal{I}}_{b_n}(x) := \int_{[0,1]^D} \tilde{f}_9^{X,\mathbf{Z}}(x, \mathbf{z}) \tilde{w}_9(x, \mathbf{z}) \tilde{\sigma}_{b_n}^{-2}(x, \mathbf{z}) d\mathbf{z}.$$

Finally, we can propose our adaptive estimators for Fourier coefficient $\theta_j$

$$\hat{\theta}_j := \frac{1}{n - 24r} \sum_{l=24r+1}^{n} \frac{A_l [Y_l - \tilde{m}_{-j}(X_l)] \tilde{\sigma}_{b_n}^{-2}(X_l, \mathbf{Z}_l) \varphi_j(X_l)}{\tilde{\mathcal{I}}_{b_n}(X_l)}. \tag{29}$$

**Proposition 2.2.** *Consider the regression model (2) with regression error satisfying (7). Assumption 2.2 holds and both joint design density $f^{X,\mathbf{Z}}(x, \mathbf{z})$ and availability likelihood $w(x, \mathbf{z})$ belong to analytic class $\mathcal{A}$ (20). In addition, there are two finite positive constants, $c_*$ and $c^*$, such that $c_* \le \sigma^2(x, \mathbf{z}) \le c^*$. Then the blockwise shrinkage regression estimator (15) with $\widehat{\Theta}_k$ defined in (27), $\hat{\theta}_j$ in (29) and $\hat{d} = \tilde{d}$ defined in (25) is adaptive and sharp minimax.*

Besides the above analytic class (20) assumption for both joint design density and availability likelihood function, multivariate Sobolev class is also popular in nonparametric studies. Let's introduce a Sobolev class for $k$-variate functions with isotropic smoothness coefficient $k$ and a finite constant $Q > 0$,

$$\mathcal{S} := \mathcal{S}(k, Q) := \Bigg\{ f : f(x_1, x_2, \ldots, x_k) := \sum_{j_1, j_2, \ldots, j_k = 0}^{\infty} \theta_{j_1, j_2, \ldots, j_k} \prod_{s=1}^{k} \varphi_{j_s}(x_s), \\ \sum_{j_1, j_2, \ldots, j_k = 0}^{\infty} \Big[ 1 + \sum_{s=1}^{k} (2\pi j_s)^{2k} \Big] \theta_{j_1, j_2, \ldots, j_k}^2 \le Q \Bigg\}. \tag{30}$$

Here, smoothness in each coordinate matches the number of variables in order to preserve proper approximation result due to curse of dimensionality. However, it is still a much weaker assumption than analytic class $\mathcal{A}$ (20). Specifically, we assume a joint design density and an availability likelihood function $w(x, \mathbf{z})$ belong to $(D+1)$-varaite Sobolev class $\mathcal{S}(D+1, Q)$. Then with some modifications such as larger cutoffs in component statistics to compensate slow convergence rate about Sobolev class and variation reduction in Sobolev functional estimate $\widehat{\Theta}_k$, we can also establish sharp minimaxity for our data-driven blockwise shrinkage estimator (15) under Sobolev design density and availability likelihood.

### 2.3 Extension: A General Additive Model

Let us consider a natural extension of the regression model (2), a general additive model

$$Y = m(X) + g(\mathbf{Z}) + \sigma(X, \mathbf{Z})\epsilon, \tag{31}$$

where nuisance additive term $g(\mathbf{z})$ is integrated to zero on $[0, 1]^D$ for identification issue. However, it digresses a little from the topic since loss of efficiency using univariate procedure also comes from omitted variable problem when $X$ is not independent of $\mathbf{Z}$. Since oracle and dealer know underlying nuisance functions, they actually estimate the pivot model (2) by subtracting the known $g(\mathbf{Z})$ from observed response $Y$ and previous results (Theorem 2.1 and Proposition 2.1) still hold for general additive model (31). For statistician's estimator, the standard plug-in technique is used in adaptation for $g(\mathbf{z})$ under some regularity conditions such as additive component $g(\mathbf{z})$ belonging to $D$-variate Sobolev class $\mathcal{S}$ (30). Then we can propose sample mean series estimate for $g(z)$ and subtract it in the Fourier coefficient estimate for $\sigma_{i\mathbf{s}}$'s and $\theta_j$'s in addition to estimates $\tilde{m}_{-j}$ or $\tilde{m}$. We can show that the blockwise shrinkage estimator modified for additive component $g(\mathbf{z})$ asymptotically achieves the dealer's lower bound for minimax risk under this general additive model. In next section, the impact of nuisance additive component on the performance of statistician's estimator will also be tested in numerical experiments.

## 3. Numerical Studies

Asymptotic theory indicates that a data-driven estimator can mimic the performance of a dealer's estimator for both the basic regression model (2) and a general additive model (31) under response missing at random. In this section we want to use intensive Monte Carlo study to shed light on its feasibility for small data sets. Scenarios with different sample sizes and levels of heteoscedasticity and missing severity are considered in the numerical experiments. We will modify our estimator for better small sample performance with some corrections for cutoff and estimating component statistics based on whole sample instead of subsamples. Then we denote D-estimator and S-estimator as small sample counterpart of dealer's and statistician's estimator in asymptotic analysis. Simulation results are also compared with two good estimators for pure univariate model (1) under both directly observed data and missing data. We choose E-estimator of Efromovich (2018) as orthogonal series benchmark and Nadaraya-Watson type estimator of Chu and Cheng (1995) (K-estimator) as kernel benchmark. Both are only rate optimal in large sample because of ignoring auxiliary covariates in scale function from previous analysis about the coefficient of difficulty.

Let us describe the specific statistical experiments. We consider just one auxiliary covariate ($D = 1$) in model $Y = m(X) + \sigma(X, Z)\epsilon$. We use two candidate regression functions for $m(x)$ supported on unit interval [0,1], a bell shaped "Normal" corner function and a more complicated "Bimodal" corner of Efromovich (2018). The error term consists of a standard normal error $\epsilon$ independent of $(X, Z)$ and a scale function $\sigma(x, z) = \exp(\lambda z/2)$,

where $\lambda \in \{1, 2, 3\}$ controls the level of heteroscedasticity. Mutually independent predictor $X$ and auxiliary covariate $Z$ follow a joint uniform distribution on the unit square with joint density $f^{X,Z}(x, z) = I((x, z) \in [0, 1]^2)$. The missing at random mechanism is assumed to be driven by the predictor of interest only and define three linear candidates for availability likelihood, $w_1(x) = 0.5+0.4x$, $w_2(x) = 0.4+0.4x$, $w_3(x) = 0.3+0.4x$, to test robustness of our results under different missing severity.

Figure 1 gives a particular simulation for "Normal" regression function with sample size $n = 100$, scale function $\sigma(x, z) = \exp(z)$ and availability likelihood function $w_2(x) = 0.4 + 0.4x$ defined above. Top diagrams show scattergrams of the underlying H-sample and the corresponding M-sample. It is not easy to imagine the regression function without the guide of the solid curve of underlying "Normal" corner function. Some points look like outliers in the left $XY$-scattergram, for example, the point near top left corner. But we know there is no outliers and heteroscedasticity driven by auxiliary covariate causes this illusion. From the middle $ZY$-scattergram, the highest point around $Z = 0.8$ corresponds the top left suspect outlier in the $XY$-scattergram and its high value is reasonable since scale function is increasing in $Z$. So ignoring volatility from auxiliary covariate will harm those good univariate estimation procedures. Comparing the three scattergrams, we can see the impact of missing may be two edged. On the one hand missing mechanism reduces the available sample size for estimation, but on the other hand it may also remove those misleading fake outliers for univariate procedure. Note that $N = 61$ complete cases shown by circles is a very small sample size for nonparametric estimation but our estimators still did a good job. Two scattergrams in the bottom of Figure 1 are overlaid by underlying regression function and its estimates (see the description of line type for each estimate in the caption). Subtitles exhibit their integrated squared errors (ISE), calculated as $\int_0^1 [\tilde{m}(x) - m(x)]^2 dx$, where $\tilde{m}(x)$ is a particular estimate for the underlying regression function $m(x)$. For this particular simulation, all the estimates succeed to catch the symmetric bell shape of "Normal" corner function. The dot-dashed curve of S-estimate closely follows the long dashed curve of D-estimate. We can also see heteroscedasticity from auxiliary covariate inflates left tails of E-estimate and K-estimate. In the right diagram for M-sample, we add one more long-short dashed curve for S-estimate based on complete cases, which is clearly shifted to right to match the available data. We want to remind readers of randomness in simulation and a large variability in outcomes can be expected for small samples.

Figure 2 presents a similar experiment for "Bimodal" corner function, which is even more challenging to visualize the underlying curve from the data. In this particular simulation, all estimates successfully identify the pattern of two modes. However, it's very common to obtain a high left mode or just an oversmoothed single modal curve in simulation since nonparametric estimator can only learn what data tell us. The relative locations of estimated curves and their changes under missing data are more complicated. We can see D-estimate and S-estimate give a little higher and more accurate left mode under H-sample while those missed points help all estimates to locate the left mode. It is surprising that nonparametric estimates can recover such a complicated bimodal function with just $N = 58$ complete cases.

Those diagrams shed some light on the excellent performance of our shrinkage estimators but the relative performance of considered estimates may change drastically during simulations. So we want to use intensive Monte Carlo studies to test the performance of above defined four pseudo or adaptive estimates (D-estimate, S-estimate, E-estimate, K-estimate) under five sample sizes $n \in \{100, 200, 400, 600, 800\}$ for underlying H-samples and corresponding modified samples with MAR mechanism. Relative performance are measured by average ratios of their ISEs over 1000 simulations. Table 1 presents results for "Normal" regression function. According to combinations of sample sizes, scale func-

tions and availability likelihood function including the case of no missing mechanism, there are 60 scenario blocks with number of observations given on the top, where we have underlying sample size $n$ for H-sample and average number of complete cases $N$ for three M-samples below its H-sample from the case with slight missing introduced by availability likelihood $w_1(x)$ to severer missing case with $w_3(x)$. The first cell in each block is the ratio of S-estimate with respect to D-estimate and it decreases as sample size increases along each row, which supports previous asymptotic analysis that statistician's estimator mimics the performance of dealer's estimator. The remaining two numbers in the first row of each block are average ratios of pure univariate procedures, series E-estimator and kernel K-estimator over S-estimate. Our S-estimate dominates both univariate estimates and those ratios seem to increase with sample size since more observations can be used to improve accuracy of its complicated component statistics and then present its advantage of taking into account heteroscedasticity from auxiliary covariate. For corresponding ratios under different $\lambda$, heteroscedasticity mitigates relative performance of E-estimate and K-estimate, which also supports our asymptotic theory. The relation between relative efficiency and severity of missing is not clear since S-estimate has more component statistics exposed to missing mechanism than simpler univariate estimates but the slight loss of relative efficiency indicates that S-estimate does a good job in adaptation to missing mechanism.

In Section 2.3 we extend the efficient statistician's estimator for a general additive model (31), $Y = m(X) + g(Z) + \sigma(X, Z)\epsilon$ and the second row in each block of Table 1 corresponds results of S-estimator with respect to D-estimate for nuisance additive components, $g_1(z) = z - 1/2$, $g_2(z) = z^2 - 1/3$ and $g_3(z) = z^3 + z - 3/4$, which are polynomials integrated to zero on [0,1] with degree one to three. Three ratios in the second row of each block also indicate the desired trend and we can say S-estimator succeeds to adapt to unknown nuisance additive component in spite of their larger values due to errors from additional estimation steps about $g(z)$. The ratios increase as heteroscedasticity gets severe, which shows that dealer's knowledge of nuisance functions is really valuable in more complicated scenarios under small sample. It is not easy to quantitatively explain the effect of different availability likelihood functions because of their indirect impact on data generation process through random error term and Bernouli availability variable.

Table 2 gives similar results for estimating a "Bimodal" regression function. From previous graphic example, we know it is a more difficult problem for nonparametric estimator. The underlying curve has two closely located modes and a random sample may pronounce the left mode or even present a single mode pattern. We can see all the ratios in Table 2 are smaller than those in Table 1, which indicates it is a challenge even for dealer's estimator. The overall results still support the efficiency of our estimation procedure that estimates scale function when auxiliary covariate affects heteroscedasticity and adaptive S-estimator succeeds to mimic good performance of D-estimator.

In the last part of numerical studies, let us consider application of the proposed methodology for the analysis of a real data with missing response and scale function depending on auxiliary covariate. Ozone is an important trace gas in the atmosphere. While stratosphere ozone (the ozone layer) helps to protect the earth's surface by absorbing most of harmful ultraviolet radiation of the sun, low level ozone (tropospheric ozone) is a harmful pollutant involved in the chemical reaction of photochemical smog. So monitoring and controlling ozone level is a hot environmental topic in both mass media and scientific researches. We will analyze a small data set of ozone level at 147 sites in the midwestern region from Harezlak, Ruppert and Wand (2018), which is a subset of data used in the cooperative research program of the National Institute of Statistical Sciences (NISS) and the U.S. Environmental Protection Agency (EPA); see Nychka, Piegorsch and Cox (1998). Interested readers may check the EPA air quality data base for more data and information. We will model the data

by (2), where the predictor of interest $X$ is the latitude of an agency station and auxiliary covariate $Z$ is the corresponding longitude. Response $Y$ is the 8-hour (from 9AM-4PM) average (surface) ozone concentration measured in parts per billion (PPB). The top left diagram in Figure 3 presents the distribution of observations $\{(X_l, Z_l, Y_l), l = 1, 2, \ldots, 147\}$. We can see ozone level roughly increases towards north, which also justifies our model setting. The top right diagram exhibits a scattergram of $(X, Y)$ overlaid by univariate linear regression estimate, E-estimate and S-estimate. We also give 95% pointwise and simultaneous confidence bands of Efromovich (2018) for S-estimate by dashed and dotted curves, respectively. The result supports higher latitude region has a higher ozone level and all estimate are acceptable with respect to simultaneous confidence band. Because of data points spreading out in the high latitude region, the right tail behavior of E-estimate differs from our heteroscedasticity corrected S-estimate and even moves out of pointwise confidence band while the their left tails in the low latitude region are almost the same.

Next, we consider the impact of missing mechanism on regression estimation. In the bottom diagrams, we consider two scenarios modifying response by Bernoulli availability variable with availability likelihood $w(x) = 0.3 + 0.6x$ and $w(x) = 0.6$, which will imply the same expected number of complete cases when predictor $X$ is uniformly distributed. Here, MAR sample has 94 complete cases while MCAR sample has 95 complete cases, that is, about 40% observations are not available. According to missed cases shown by crosses, MAR setting suffers severer missing in the low latitude region than the MCAR one in the right as suggested by availability likelihood. For this particular MAR realization, it seems that missing mitigates heteroscedasticity and E-estimate and S-estimate agree on a larger interval in the low availability likelihood region comparing with the top right H-sample diagram. Note that MCAR is a special case of MAR and complete-case estimator is recommended in practice and theory analysis under MCAR. We can see S-estimate does a good job and the three curves in the bottom diagrams look almost the same, which indicates a successful adaptation to missing mechanism and also supports complete-case practice for MCAR setting. You can check the slight change of S-estimate in the two diagrams of the MCAR sample keeping mind that data points and other estimates are fixed. Althought randomness affects particular realization, Figure 3 reflects nice performance of asymptotically efficient statistician's estimator in small sample.

## 4. Conclusion and Future Work

In this paper, we consider a nonparametric regression problem with complexity due to both heteroscedasticity and missing at random response. It is known that efficient nonparametric estimator for univariate regression does not require knowledge of scale function and there also exists a corresponding sharp minimax complete-case estimator for the MAR response setting. However, the use of scale function is indispensable for sharp minimax estimation procedure when heteroscedasticity involves auxiliary covariates, that is, a univariate regression with a multivariate scale function. MAR response also introduces additional bias in data and asymptotic theory is developed with a lower bound of minimax risk for this heteroscedastic regression setting with missing data, which also shows univariate procedure ignoring this complicated heteroscedasticity can not be optimal. We show that it is still possible to propose a data-driven sharp minimax estimator adapted to heteroscedasticity and missing mechanism that mimics the performance of efficient pseudo estimators. A general additive model (31) is also considered as an extension and a data-driven estimator adapted to unknown additive component succeeds to attain the same sharp minimax risk bound as model (2) under some mild regularity assumptions on nuisance functions. Monte Carlo simulations and real data examples shed light on the feasibility of asymptotic theory

for small sample data sets. It would also be of great interest to consider a more challenging problem of MAR predictor for this auxiliary covariate involved heteroscedastic regression in the future, where consistent complete-case approach is impossible even for the pure univariate setting.

## REFERENCES


Berger, J. O., (1985), *Statistical Decision Theory and Bayesian Analysis*, New York, NY: Springer.

Boente, G., González–Manteiga, W., and Pérez–González, A., (2009) "Robust nonparametric estimation with missing data," *Journal of Statistical Planning and Inference*, 139 (2), 571–592.

Chu, C. and Cheng, P., (1995), "Nonparametric regression estimation with missing data," *Journal of Statistical Planning and Inference*, 48 (1): 85–99.

Efromovich, S., (1999), *Nonparametric Curve Estimation: Method, Theory, and Applications*, New York, NY: Springer.

Efromovich, S., (2011), "Nonparametric regression with responses missing at random," *Journal of Statistical Planning and Inference*, 141 (12): 3744–3752.

Efromovich, S., (2012), "Sequential analysis of nonparametric heteroscedastic regression with missing responses," *Sequential Analysis*, 31 (3): 351–367.

Efromovich, S., (2013), "Nonparametric regression with the scale depending on auxiliary variable," *The Annals of Statistics*, 41 (3): 1542–1568.

Efromovich, S., (2018), *Missing and Modified Data in Nonparametric Estimation: With R Examples*, Boca Raton, FL: Chapman and Hall/CRC.

Efromovich, S., and Pinsker, M., (1984), "A learning algorithm of non-parametric filtering," *Automation and Remote Control*, 45 (11): 58–65.

Efromovich, S., and Pinsker, M., (1996), " Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression," *Statistica Sinica*, 6 (4): 925–942, 1996.

Enders, C. K., (2010) *Applied Missing Data Analysis*, New York, NY: Guilford Publications.

Golubev, G. K., (1991), "LAN in problems of nonparametric estimation of functions and lower bounds for quadratic risks," *Theory of Probability & Its Applications*, 36 (1): 152–157.

Harezlak, J., Ruppert, D., and Wand, M. P., (2018), *Semiparametric Regression with R*, New York, NY: Springer.

Lehmann, E. L., and Casella, G., (1998), *Theory of Point Estimation*, New York, NY: Springer.

Little, R., and Rubin, D., (2019), *Statistical Analysis with Missing Data, Third Edition* (3rd ed.), Hoboken, NJ: Wiley.

Mitra, R., and Müller, P., (2015), *Nonparametric Bayesian Inference in Biostatistics*, Springer International Publishing.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A. A., and Verbeke, G., (2014), *Handbook of Missing Data Methodology*, Boca Raton, FL: Chapman and Hall/CRC.

Nychka, D., Piegorsch, W. W., and Cox, L. H., (1998), *Case Studies in Environmental Statistics*, New York, NY: Springer.

Pinsker, M., (1980), "Optimal filtering of square-integrable signals in gaussian noise," *Problems of Information Transmission*, 16 (2): 52–68.

Rubin, D. B., (1976), "Inference and missing data," *Biometrika*, 63 (3): 581–592.

Sun, Y., Wang, L., and Han, P., (2019), "Multiply robust estimation in nonparametric regression with missing data," *Journal of Nonparametric Statistics*, 32 (1): 73–92.

Tsybakov, A. B., (2009), *Introduction to Nonparametric Estimation*, New York, NY: Springer.

Wasserman, L., (2006) *All of Nonparametric Statistics*, New York, NY: Springer.

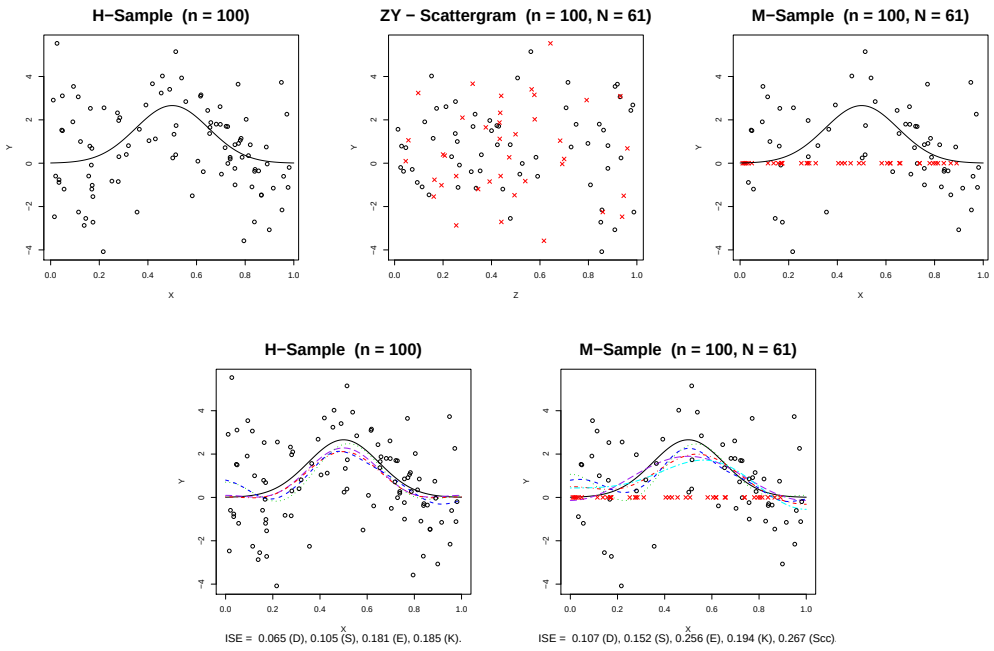


**Figure 1**: Analysis of a data simulated according to model (2) with "Normal" regression function. Scale function $\sigma(x,z) = e^z$, joint design density $f^{X,Z}(x,z) = I((x,z) \in [0,1]^2)$ and availability likelihood $w(x,z) = 0.4 + 0.4x$. The top left diagram shows the $XY$-scattergram for underlying hidden sample (H-sample) of size $n = 100$ while the top right shows the corresponding $XY$-scattergram for M-sample with $N = 61$ complete cases. Observations are shown by circles and missed incomplete cases $(X_l, Z_l, A_lY_l, A_l)$ with $A_l = 0$ are shown by crosses for M-sample. The underlying regression function $m(x)$ is also shown by a solid line. The middle diagram gives $ZY$-scattergram. The bottom scattergrams are overlaid by corresponding estimation results of D-estimate by (purple) long dashed lines, S-estimate by (red) dot-dashed lines, E-estimate by (green) dotted lines, K-estimate by (blue) short dashed lines and an additional S-estimate based on complete-case (Scc) by a (cyan) long-short dashed line for M-sample. Subtitles under diagrams give corresponding intergrated squared errors (ISE).

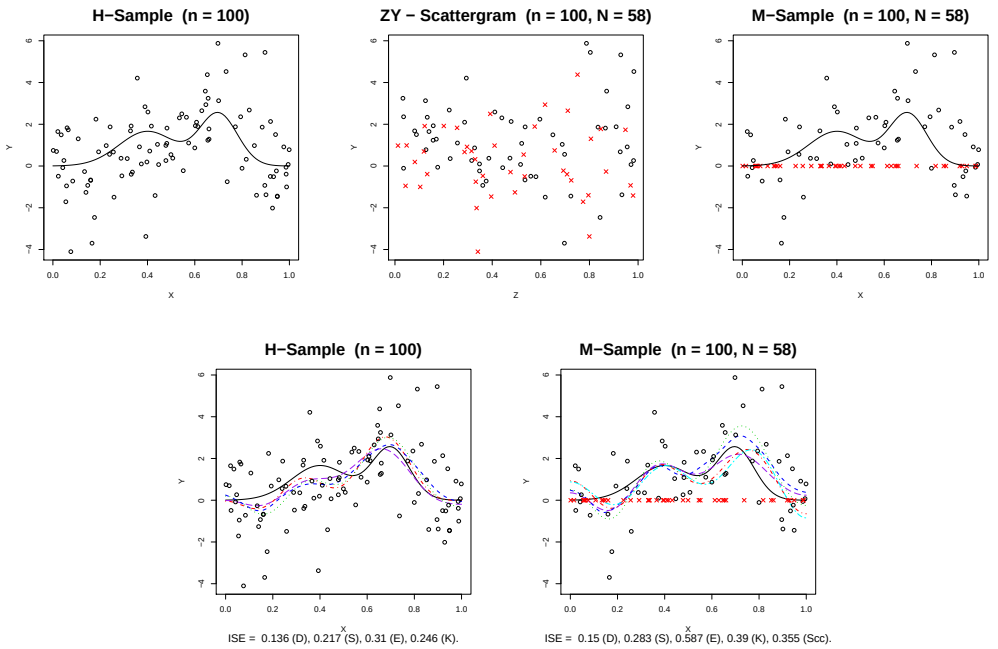


**Figure 2**: Analysis of a data simulated according to model (2) with "Bimodal" regression function. Scale function $\sigma(x,z) = e^z$, joint design density $f^{X,Z}(x,z) = I((x,z) \in [0,1]^2)$ and availability likelihood $w(x,z) = 0.4 + 0.4x$. The top left diagram shows the $XY$-scattergram for underlying hidden sample (H-sample) of size $n = 100$ while the top right shows the corresponding $XY$-scattergram for M-sample with $N = 61$ complete cases. Observations are shown by circles and missed incomplete cases $(X_l, Z_l, A_lY_l, A_l)$ with $A_l = 0$ are shown by crosses for M-sample. The underlying regression function $m(x)$ is also shown by a solid line. The middle diagram gives $ZY$-scattergram. The bottom scattergrams are overlaid by corresponding estimation results of D-estimate by (purple) long dashed lines, S-estimate by (red) dot-dashed lines, E-estimate by (green) dotted lines, K-estimate by (blue) short dashed lines and an additional S-estimate based on complete-case (Scc) by a (cyan) long-short dashed line for M-sample. Subtitles under diagrams give corresponding intergrated squared errors (ISE).

**Table 1**: Results for "Normal" Regression function

| $\lambda$ | $n=100$ | | | $n=200$ | | | $n=400$ | | | $n=600$ | | | $n=800$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1.22 | 1.58 | 1.34 | 1.18 | 1.61 | 1.71 | 1.09 | 1.60 | 1.96 | 1.05 | 1.60 | 2.11 | 1.04 | 1.62 | 2.22 |
| | 1.36 | 1.34 | 1.35 | 1.28 | 1.29 | 1.30 | 1.17 | 1.18 | 1.17 | 1.11 | 1.11 | 1.13 | 1.10 | 1.10 | 1.12 |
| | $N=70.05$ | | | $N=140.57$ | | | $N=279.70$ | | | $N=420.08$ | | | $N=559.62$ | | |
| | 1.26 | 1.49 | 1.35 | 1.27 | 1.52 | 1.57 | 1.26 | 1.34 | 1.74 | 1.07 | 1.38 | 1.96 | 1.04 | 1.36 | 2.04 |
| | 1.46 | 1.45 | 1.49 | 1.40 | 1.41 | 1.40 | 1.35 | 1.35 | 1.35 | 1.14 | 1.14 | 1.15 | 1.09 | 1.09 | 1.10 |
| 1 | $N=60.05$ | | | $N=120.33$ | | | $N=240.17$ | | | $N=360.05$ | | | $N=478.70$ | | |
| | 1.32 | 1.38 | 1.30 | 1.14 | 1.52 | 1.58 | 1.30 | 1.35 | 1.65 | 1.17 | 1.37 | 1.91 | 1.04 | 1.37 | 2.00 |
| | 1.49 | 1.48 | 1.52 | 1.28 | 1.27 | 1.26 | 1.41 | 1.39 | 1.38 | 1.25 | 1.25 | 1.26 | 1.10 | 1.10 | 1.11 |
| | $N=50.06$ | | | $N=99.93$ | | | $N=199.63$ | | | $N=300.44$ | | | $N=399.47$ | | |
| | 1.34 | 1.34 | 1.36 | 1.18 | 1.58 | 1.54 | 1.28 | 1.38 | 1.57 | 1.18 | 1.35 | 1.82 | 1.08 | 1.35 | 1.96 |
| | 1.49 | 1.48 | 1.52 | 1.31 | 1.31 | 1.30 | 1.40 | 1.40 | 1.39 | 1.25 | 1.26 | 1.27 | 1.14 | 1.15 | 1.17 |
| | $n=100$ | | | $n=200$ | | | $n=400$ | | | $n=600$ | | | $n=800$ | | |
| | 1.37 | 1.64 | 1.32 | 1.46 | 2.09 | 1.74 | 1.41 | 2.08 | 2.09 | 1.26 | 1.95 | 2.19 | 1.22 | 1.98 | 2.44 |
| | 1.56 | 1.56 | 1.55 | 1.72 | 1.71 | 1.64 | 1.68 | 1.66 | 1.58 | 1.42 | 1.42 | 1.44 | 1.39 | 1.38 | 1.38 |
| | $N=70.05$ | | | $N=140.57$ | | | $N=279.70$ | | | $N=420.08$ | | | $N=559.62$ | | |
| | 1.49 | 1.46 | 1.29 | 1.90 | 1.71 | 1.51 | 1.55 | 1.80 | 1.76 | 1.26 | 1.80 | 2.02 | 1.23 | 1.69 | 2.16 |
| | 1.64 | 1.65 | 1.73 | 2.09 | 1.99 | 2.02 | 1.77 | 1.77 | 1.70 | 1.41 | 1.42 | 1.41 | 1.39 | 1.39 | 1.39 |
| 2 | $N=60.05$ | | | $N=120.33$ | | | $N=240.17$ | | | $N=360.05$ | | | $N=478.70$ | | |
| | 1.76 | 1.46 | 1.29 | 1.40 | 1.78 | 1.67 | 1.62 | 1.72 | 1.67 | 1.44 | 1.82 | 1.98 | 1.23 | 1.70 | 2.05 |
| | 1.88 | 1.89 | 1.93 | 1.61 | 1.58 | 1.62 | 1.86 | 1.84 | 1.80 | 1.61 | 1.59 | 1.56 | 1.38 | 1.37 | 1.37 |
| | $N=50.06$ | | | $N=99.93$ | | | $N=199.63$ | | | $N=300.44$ | | | $N=399.47$ | | |
| | 1.58 | 1.47 | 1.40 | 1.43 | 1.64 | 1.53 | 1.55 | 1.68 | 1.58 | 1.40 | 1.80 | 1.88 | 1.28 | 1.83 | 2.07 |
| | 1.78 | 1.78 | 1.92 | 1.62 | 1.64 | 1.64 | 1.76 | 1.76 | 1.70 | 1.57 | 1.59 | 1.56 | 1.44 | 1.45 | 1.43 |
| | $n=100$ | | | $n=200$ | | | $n=400$ | | | $n=600$ | | | $n=800$ | | |
| | 1.93 | 1.80 | 1.44 | 2.52 | 1.97 | 1.54 | 2.68 | 2.33 | 1.92 | 2.29 | 2.38 | 2.14 | 2.24 | 2.55 | 2.41 |
| | 2.36 | 2.30 | 2.44 | 2.81 | 2.93 | 3.03 | 3.17 | 3.15 | 3.01 | 2.69 | 2.74 | 2.68 | 2.64 | 2.60 | 2.52 |
| | $N=70.05$ | | | $N=140.57$ | | | $N=279.70$ | | | $N=420.08$ | | | $N=559.62$ | | |
| | 2.18 | 1.70 | 1.37 | 2.53 | 1.74 | 1.51 | 2.36 | 2.02 | 1.77 | 2.14 | 2.16 | 1.84 | 2.37 | 2.23 | 2.08 |
| | 2.36 | 2.36 | 2.45 | 2.92 | 2.89 | 2.98 | 2.83 | 2.76 | 2.70 | 2.50 | 2.47 | 2.42 | 2.70 | 2.76 | 2.63 |
| 3 | $N=60.05$ | | | $N=120.33$ | | | $N=240.17$ | | | $N=360.05$ | | | $N=478.70$ | | |
| | 2.99 | 1.68 | 1.31 | 2.29 | 1.82 | 1.58 | 2.32 | 2.01 | 1.68 | 2.30 | 2.13 | 1.86 | 2.15 | 2.19 | 2.05 |
| | 2.98 | 2.95 | 3.12 | 2.63 | 2.57 | 2.80 | 2.77 | 2.71 | 2.70 | 2.78 | 2.73 | 2.58 | 2.51 | 2.52 | 2.47 |
| | $N=50.06$ | | | $N=99.93$ | | | $N=199.63$ | | | $N=300.44$ | | | $N=399.47$ | | |
| | 2.44 | 1.73 | 1.43 | 2.29 | 1.94 | 1.62 | 2.69 | 1.75 | 1.57 | 2.56 | 2.07 | 1.82 | 2.21 | 2.32 | 2.04 |
| | 2.68 | 2.62 | 2.65 | 2.61 | 2.62 | 2.76 | 3.47 | 3.39 | 3.08 | 3.03 | 2.96 | 2.86 | 2.68 | 2.67 | 2.59 |

Consider a general additive regression model $Y = m(X)+g(Z)+\sigma(X,Z)\epsilon$ with "Normal" corner function $m(x)$, uniform joint design $f^{X,Z}(x,z) = I((x,z) \in [0,1]^2)$, scale function $\sigma(x,z) = \exp(\lambda z/2)$. For fixed $\lambda$, four rows of blocks in the table correspond scenario under H-sample with sample size $n$ and then M-samples with availability likelihood functions $w_1(x) = 0.5 + 0.4x$, $w_2(x) = 0.4 + 0.4x$ and $w_3(x) = 0.3 + 0.4x$, respectively. Average number of complete cases denoted by $N$ is listed above each missing scenario. In each block, three ratios in the first row are average values of $\mathrm{ISE_S}/\mathrm{ISE_D}$, $\mathrm{ISE_E}/\mathrm{ISE_S}$ and $\mathrm{ISE_K}/\mathrm{ISE_S}$ under base model without additive component ($g(z) = 0$) while three ratios in second row are average values of $\mathrm{ISE_S}/\mathrm{ISE_D}$ under general additive model with $g_1(z) = z - 1/2$, $g_2(z) = z^2 - 1/3$ and $g_3(z) = z^3 + z - 3/4$, respectively. Here, ISE's stand for integrated squared errors of D-estimate, S-estimate, E-estimate and K-estimate according to their subscripts.

**Table 2**: Results for "Biomodal" Regression function

| $\lambda$ | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $n = 100$ | | | $n = 200$ | | | $n = 400$ | | | $n = 600$ | | | $n = 800$ | | |
| | 1.06 | 0.98 | 0.69 | 1.02 | 0.95 | 0.68 | 1.03 | 0.90 | 0.97 | 1.01 | 0.92 | 1.08 | 1.01 | 0.95 | 1.14 |
| | 1.09 | 1.08 | 1.09 | 1.04 | 1.04 | 1.05 | 1.06 | 1.06 | 1.07 | 1.03 | 1.04 | 1.04 | 1.03 | 1.03 | 1.04 |
| | $N = 70.05$ | | | $N = 140.57$ | | | $N = 279.70$ | | | $N = 420.08$ | | | $N = 559.62$ | | |
| | 1.08 | 1.10 | 0.80 | 1.04 | 1.05 | 0.75 | 0.80 | 1.33 | 0.90 | 1.00 | 0.94 | 1.01 | 1.01 | 0.95 | 1.09 |
| | 1.11 | 1.11 | 1.11 | 1.07 | 1.07 | 1.07 | 0.82 | 0.83 | 0.83 | 1.03 | 1.03 | 1.04 | 1.03 | 1.03 | 1.03 |
| 1 | $N = 60.05$ | | | $N = 120.33$ | | | $N = 240.17$ | | | $N = 360.05$ | | | $N = 478.70$ | | |
| | 1.10 | 1.12 | 0.85 | 1.04 | 1.06 | 0.78 | 0.85 | 1.27 | 0.89 | 0.72 | 1.46 | 1.00 | 1.00 | 0.95 | 1.09 |
| | 1.14 | 1.15 | 1.16 | 1.07 | 1.08 | 1.08 | 0.87 | 0.87 | 0.87 | 0.74 | 0.74 | 0.75 | 1.03 | 1.03 | 1.04 |
| | $N = 50.06$ | | | $N = 99.93$ | | | $N = 199.63$ | | | $N = 300.44$ | | | $N = 399.47$ | | |
| | 1.07 | 1.16 | 0.93 | 1.04 | 1.10 | 0.83 | 0.88 | 1.22 | 0.89 | 0.76 | 1.37 | 0.98 | 0.97 | 1.02 | 1.06 |
| | 1.15 | 1.17 | 1.16 | 1.08 | 1.08 | 1.08 | 0.91 | 0.91 | 0.91 | 0.79 | 0.79 | 0.79 | 1.00 | 1.00 | 1.01 |
| | $n = 100$ | | | $n = 200$ | | | $n = 400$ | | | $n = 600$ | | | $n = 800$ | | |
| | 1.14 | 1.13 | 0.81 | 1.11 | 1.12 | 0.79 | 1.14 | 1.11 | 0.99 | 1.09 | 1.12 | 1.12 | 1.10 | 1.12 | 1.21 |
| | 1.21 | 1.21 | 1.19 | 1.16 | 1.16 | 1.15 | 1.21 | 1.22 | 1.21 | 1.14 | 1.14 | 1.15 | 1.16 | 1.15 | 1.16 |
| | $N = 70.05$ | | | $N = 140.57$ | | | $N = 279.70$ | | | $N = 420.08$ | | | $N = 559.62$ | | |
| | 1.19 | 1.24 | 0.93 | 1.14 | 1.25 | 0.87 | 0.97 | 1.40 | 0.96 | 1.10 | 1.23 | 1.05 | 1.11 | 1.20 | 1.12 |
| | 1.29 | 1.29 | 1.26 | 1.20 | 1.19 | 1.19 | 1.03 | 1.03 | 1.02 | 1.15 | 1.15 | 1.15 | 1.16 | 1.16 | 1.16 |
| 2 | $N = 60.05$ | | | $N = 120.33$ | | | $N = 240.17$ | | | $N = 360.05$ | | | $N = 478.70$ | | |
| | 1.28 | 1.26 | 0.99 | 1.14 | 1.27 | 0.93 | 1.00 | 1.40 | 0.97 | 0.88 | 1.49 | 1.05 | 1.09 | 1.24 | 1.13 |
| | 1.32 | 1.32 | 1.34 | 1.19 | 1.20 | 1.20 | 1.07 | 1.07 | 1.05 | 0.92 | 0.92 | 0.92 | 1.15 | 1.15 | 1.15 |
| | $N = 50.06$ | | | $N = 99.93$ | | | $N = 199.63$ | | | $N = 300.44$ | | | $N = 399.47$ | | |
| | 1.21 | 1.30 | 1.09 | 1.13 | 1.30 | 0.99 | 1.03 | 1.36 | 0.98 | 0.91 | 1.44 | 1.04 | 1.07 | 1.27 | 1.10 |
| | 1.29 | 1.28 | 1.30 | 1.19 | 1.19 | 1.19 | 1.09 | 1.08 | 1.08 | 0.96 | 0.96 | 0.96 | 1.12 | 1.12 | 1.12 |
| | $n = 100$ | | | $n = 200$ | | | $n = 400$ | | | $n = 600$ | | | $n = 800$ | | |
| | 1.44 | 1.34 | 0.96 | 1.43 | 1.29 | 0.89 | 1.62 | 1.35 | 0.97 | 1.54 | 1.38 | 1.10 | 1.57 | 1.36 | 1.19 |
| | 1.57 | 1.55 | 1.58 | 1.54 | 1.54 | 1.53 | 1.78 | 1.77 | 1.74 | 1.66 | 1.65 | 1.62 | 1.68 | 1.67 | 1.67 |
| | $N = 70.05$ | | | $N = 140.57$ | | | $N = 279.70$ | | | $N = 420.08$ | | | $N = 559.62$ | | |
| | 1.54 | 1.43 | 1.12 | 1.49 | 1.40 | 1.00 | 1.40 | 1.50 | 1.02 | 1.50 | 1.51 | 1.07 | 1.55 | 1.51 | 1.13 |
| | 1.63 | 1.62 | 1.64 | 1.60 | 1.59 | 1.59 | 1.52 | 1.52 | 1.49 | 1.62 | 1.62 | 1.58 | 1.67 | 1.67 | 1.64 |
| 3 | $N = 60.05$ | | | $N = 120.33$ | | | $N = 240.17$ | | | $N = 360.05$ | | | $N = 478.70$ | | |
| | 1.74 | 1.41 | 1.16 | 1.51 | 1.47 | 1.08 | 1.38 | 1.54 | 1.07 | 1.26 | 1.62 | 1.11 | 1.53 | 1.55 | 1.14 |
| | 1.80 | 1.74 | 1.80 | 1.57 | 1.58 | 1.63 | 1.50 | 1.50 | 1.49 | 1.36 | 1.35 | 1.32 | 1.68 | 1.67 | 1.64 |
| | $N = 50.06$ | | | $N = 99.93$ | | | $N = 199.63$ | | | $N = 300.44$ | | | $N = 399.47$ | | |
| | 1.66 | 1.43 | 1.28 | 1.52 | 1.46 | 1.13 | 1.43 | 1.51 | 1.10 | 1.27 | 1.59 | 1.13 | 1.46 | 1.58 | 1.16 |
| | 1.69 | 1.67 | 1.76 | 1.61 | 1.61 | 1.65 | 1.54 | 1.54 | 1.53 | 1.41 | 1.40 | 1.38 | 1.59 | 1.59 | 1.56 |

Consider a general additive regression model $Y = m(X)+g(Z)+\sigma(X,Z)\epsilon$ with "Bimodal" corner function $m(x)$, uniform joint design $f^{X,Z}(x,z) = I((x,z) \in [0,1]^2)$, scale function $\sigma(x,z) = \exp(\lambda z/2)$. For fixed $\lambda$, four rows of blocks in the table correspond scenario under H-sample with sample size $n$ and then M-samples with availability likelihood functions $w_1(x) = 0.5 + 0.4x$, $w_2(x) = 0.4 + 0.4x$ and $w_3(x) = 0.3 + 0.4x$, respectively. Average number of complete cases denoted by $N$ is listed above each missing scenario. In each block, three ratios in the first row are average values of $\mathrm{ISE_S}/\mathrm{ISE_D}$, $\mathrm{ISE_E}/\mathrm{ISE_S}$ and $\mathrm{ISE_K}/\mathrm{ISE_S}$ under base model without additive component ($g(z) = 0$) while three ratios in second row are average values of $\mathrm{ISE_S}/\mathrm{ISE_D}$ under general additive model with $g_1(z) = z - 1/2$, $g_2(z) = z^2 - 1/3$ and $g_3(z) = z^3 + z - 3/4$, respectively. Here, ISE's stand for integrated squared errors of D-estimate, S-estimate, E-estimate and K-estimate according to their subscripts.

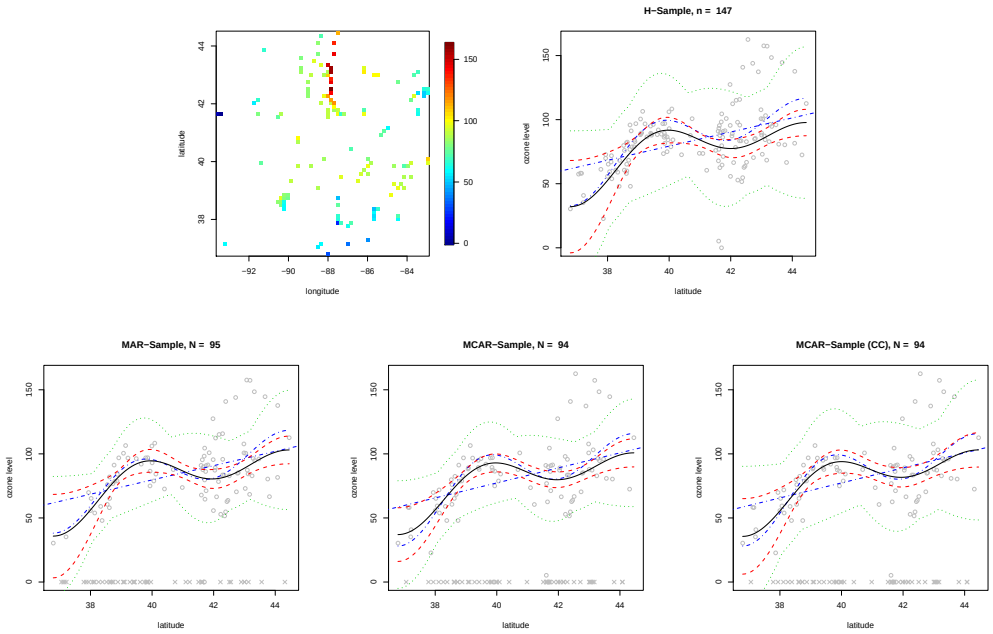


**Figure 3**: Analysis of the ozone data set with missing response. Consider the model (2) $Y = m(X) + \sigma(X, Z)\epsilon$, where respons $Y$ is ozone measure, predictor of interest $X$ is the latitude and auxiliary covariate $Z$ is the longitude of a monitoring station. Top left diagram presents the triplet $(X, Z, Y)$ of sample size $n = 147$. In the top right diagram, observations are shown by circles and the (blue) dot-dashed straight line and curve, (black) solid, (red) dashed, (green) dotted curves are the linear regression, E-estimate, S-estimate and its 95% pointwise and simultaneous confidence bands. In the bottom diagrams, we use availability likelihood function $w(x) = 0.3 + 0.6x$ to generate MAR sample with $N = 95$ complete cases and constant function $w(x) = 0.6$ to generate MCAR sample with $N = 94$ complete cases, where missed cases are showed by crosses at horizon. The S-estimate in the bottom right diagram is based on complete cases as linear regression and E-estimate.