Nonparametric regression with discrete covariate and missing values

SONG XI CHEN AND CHENG YONG TANG

1. INTRODUCTION

Nonparametric regression with continuous covariate has been extensively explored in the last two decades; see [19] and [13] for comprehensive overviews. In practice, discrete covariate from qualitative observations often arises from various scenarios including many social and economic studies. Discrete kernel smoothing for nonparametric distribution estimation was originally proposed in [1]. Its theoretical properties are evaluated in [16] and [18]. Recently, there are increasing interests of nonparametric inference with mixed continuous and discrete covariates. These include [20] for distribution function estimation, [17] for conditional density estimation and variable selection; see [21] for overviews in econometrics.

Missing values are often encountered in practice. Adequately incorporating data missing mechanism is important for valid and efficient inferences [22]. For parametric inference, the impact of missing data is well documented. For overviews and recent developments, see [22, 26, 32] and [25] and references therein. For nonparametric inference with missing values, existing works are mainly confined to situations with only continuous random variables. These include [30] on various imputation schemes for density estimation, [7, 8] and [15] on the conditional mean imputation in estimating regression function and overall mean of the response variable; see also [23] and [10] for recent developments.

In this paper, we consider nonparametric regression with both continuous and discrete covariates, say $X$, where the response variable $Y$ may be missing. We start from studying a nonparametric regression that ignores the portion of data with missing values, which we call complete-cases based estimator. Here “complete” means non-missing values which is a convention in the missing value literature. Although this estimator is still consistent, its efficiency is affected by an inflation of the variance due to its ignoring the incomplete portion of data with missing values. To improve the estimation efficiency, we propose an imputation based nonparametric regression estimator that accommodates both continuous and discrete covariates. Although the variances of both estimators have the same order of magnitude, the imputation based estimator has a smaller variance in the leading order than the complete-cases based estimator, which relates to a feature in approaches of [5] and [6]. We also conduct a second order asymptotic analysis, which reveals that smoothing the discrete covariates leads to a desired variance reduction in the second order. A comprehensive application of the proposed approach has been implemented in the case study reported in [4] for the US Census dual system accuracy and coverage evaluation where the main concern is on the population size estimation. The focus of this paper, however, is on properties of nonparametric regression with mixed covariates in the presence of missing values.

The paper is organized as follows. Section 2 outlines the nonparametric kernel regression estimators with mixed covariates. The issue of missing values and their imputation are considered in Section 3. Section 4 reports a theoretical analysis on the imputation based nonparametric regression. Construction of confidence band and data adaptive smoothing are also considered. Simulation results are reported in Section 5 followed by data analyses in Section 6. All the technical proofs are deferred to the Appendix.

* We thank the Editor, AE and the referee for their insightful comments and constructive suggestions. We thank US Census Bureau for support and allowing access to the ACE data. The authors acknowledge support by National Science Foundation grants SES-0518904 and DMS-0604563. Tang also acknowledges support from National University of Singapore Academic research grants.

†Corresponding author.
2. NONPARAMETRIC REGRESSION WITH MIXED COVARIATES

Let \( \{(X_i, Y_i)\}_{i=1}^n \) be independent and identically distributed random vectors such that \( X_i = (X_{i1}, \ldots, X_{id}) \) is a \( d \)-dimensional explanatory variable and \( Y_i \) is a univariate response. Without loss of generality, we write \( X_i = (X_i^c, X_i^u) \) where \( X_i^c \) consists of \( d_c \) continuous covariates while \( X_i^u \) consists of \( d_u \) unordered categorical ones, where \( d_c + d_u = d \). Our target is to estimate the nonparametric regression function

\[
m(x) = E(Y_i | X_i = x),
\]

with heteroscedastic conditional variance \( \sigma^2(x) = \text{var}(Y_i | X_i = x) \). Here \( x = (x^c, x^u) \) inherits the same partition of \( X_i \).

To smooth continuous covariates, we apply a \( d_c \)-dimensional kernel \( K \) which is a radially symmetric probability density function in \( \mathbb{R}^d \). We define \( K_h(u) = h^{-d_c} K(u/h) \). Here \( h \) is a bandwidth that controls the amount of smoothness. For simplicity in presentation, we consider the product kernel in our studies, i.e. \( K(u) = \prod_{i=1}^{d_c} K_i(u_i) \) for \( u = (u_1, \ldots, u_{d_c}) \), where \( K_i (\cdot) \) is some symmetric univariate density function. The product kernel can be generalized to a multivariate kernel by the formulation in [28].

Smoothing categorical variables is designed to efficiently utilize data information from some neighborhood that may share similar characteristics with the target; see [1, 16] and [17]. For smoothing unordered categorical covariates, we use the discrete kernel originally proposed by [1]. Other formulation of discrete kernels is available in [21]. Suppose \( X_{ij}^u \), the \( j \)th component of \( X_i^u \), takes \( c_j \) discrete values in \( \{0, 1, \ldots, c_j - 1\} \). The bandwidth for smoothing \( X_i^u \) is \( \lambda_j \) and the kernel weight at a \( x^u_j \) is

\[
\lambda_j I(X_{ij}^u = x^u_j) + \frac{1 - \lambda_j}{c_j - 1} I(X_{ij}^u \neq x^u_j)
\]

where \( I(\cdot) \) is the indicator function. As shown in the Appendix, this way of discrete kernel smoothing has an advantage in quantifying the second order reduction in variance. The bandwidth \( \lambda_j \) takes values within \([c_j^{-1}, 1]\). Assigning \( \lambda_j = c_j^{-1} \) leads to a uniform weight irrespective to the difference between \( X_{ij}^u \) and \( x^u_j \), whereas \( \lambda_j = 1 \) gives a kernel weight of 1 if \( X_{ij}^u = x^u_j \) and zero otherwise which coincides with the standard frequency weight. The other \( \lambda_j \) values between \( c_j^{-1} \) and 1 offer a range of choices for information combining for efficiency improvement. The kernel used to smooth the entire categorical component \( X_i^u = (X_{i1}^u, \ldots, X_{id_u}^u) \) at \( x^u = (x_1^u, \ldots, x_{d_u}^u) \) is

\[
L(x^u, X^u_i; \lambda) = \prod_{j=1}^{d_u} \left\{ \lambda_j I(X_{ij}^u = x^u_j) + \frac{1 - \lambda_j}{c_j - 1} I(X_{ij}^u \neq x^u_j) \right\},
\]

where \( \lambda = (\lambda_1, \ldots, \lambda_{d_u}) \) is a bandwidth vector. Finally, the overall kernel weight drawn from \( X_i = (X_i^c, X_i^u) \) for local estimation at \( x = (x^c, x^u) \) is

\[
K_h(x^c - X_i^c)L(x^u, X_i^u; \lambda).
\]

The kernel estimator of \( m(x) \) in the absence of missing values is then given by

\[
\hat{m}(x) = \frac{\sum_{i=1}^{n} K_h(x^c - X_i^c) L(x^u, X_i^u; \lambda) Y_i}{\sum_{i=1}^{n} K_h(x^c - X_i^c) L(x^u, X_i^u; \lambda)}.
\]

This is a Nadaraya-Watson type estimator carrying out the weighted average of the responses \( Y_i \) locally, which allows estimating \( m(x) \) without assuming a parametric model.

3. INCORPORATING MISSING VALUES

Now we consider the situation when the response \( Y_i \) can be missing. Let \( \delta_i \) be the missing indicator of \( Y_i \) such that \( \delta_i = 0(1) \) for missing (observed) \( Y_i \). We concentrate in this study on the fact that the \( Y_i \) is missing at random (MAR) [27]. MAR is an important notion in missing data analysis. It means that conditioning on the covariate \( X_i \), the missing mechanism of \( Y_i \) is independent of \( Y_i \). In other words, \( \delta_i \), and \( Y_i \) are conditionally independent given \( X_i \), i.e.

\[
P(\delta_i = 1 | Y_i, X_i) = P(\delta_i = 1 | X_i) =: w(X_i).
\]

Where, \( w(\cdot) \) is called the missing propensity function of \( Y_i \). MAR is weaker and more general than the so called missing completely at random which assumes that the propensity \( w(x) \) is a constant function. The statistical inference in case of not missing at random often relies on some parametric model assumption, see for instance the study in [24]. In this paper, we do not rely on any parametric model assumption on the missing propensity function. Instead, we discuss practical remedy to make the MAR assumption reasonable based on available information from data.

Here we extend the MAR assumption (3) to the following form to reflect the reality of applications by assuming that in addition to \( X_i \), some extra covariate \( Z_i \) contributes to the missingness of \( Y_i \). Specifically, we assume that \( Y_i \) is missing at random given \( (X_i, Z_i) \), i.e.

\[
P(\delta_i = 1 | Y_i, X_i, Z_i) = P(\delta_i = 1 | X_i, Z_i) =: w(X_i, Z_i).
\]

The rationale for (4) is based on practical considerations to make (3) more accommodative. As a concrete example in the US Census dual system surveys, in addition to a set of co-ordinates \( X_i \), which consists of racial origin, age, sex, housing tenure and region, an operational variable \( Z_i \) called match coding group collected during the non-response follow-up process is also predictive to the data missingness as shown in [2]. We note that the results in the paper stay and become simpler if (4) is valid without \( Z_i \), namely the MAR in (3) is valid.
In addition, we assume that the extra covariate Z does not have any predicting power on the conditional mean function of Y. This assumption is not restrictive because if a part of Z possesses such power, it should be included in X. In particular, we assume
\[
E(Y|X,Z) = m(X) \quad \text{and} \quad \text{var}(Y|X,Z) = \sigma^2(X).
\]

Though $Z_i$ is invisible in the final results because of the homogeneous assumption (5), the missing mechanism (4) is important to ensure consistency of the nonparametric estimators. The key implication of (4) is that $Y_i$ and $\delta_i$ are not conditional independent without given $Z_i$, i.e. $E(Y_i|X) \neq m(X)w(X)$ where $w(x) = E(\delta|X = x)$.

We will consider two estimators of $m(x)$. The first estimator uses only the complete observations (those with no missing values),
\[
\hat{m}_c(x) = \frac{\sum_{i=1}^{n} K_h(x^c - X_i^c)L(x^c, X_i^c; \lambda) \delta_i Y_i}{\sum_{i=1}^{n} K_h(x^c - X_i^c)L(x^c, X_i^c; \lambda) \delta_i}
\]
We call it the complete case estimator. As shown in the Appendix, despite the impact of a selection bias due to the missingness as specified by $w(x)$, the estimator is still consistent. The intuitive rationale of the consistency is that (6) is a ratio estimator so that the impacts due to selection bias arising in the numerator and denominator cancel each other. However, it does not fully utilize data information in $X$. For improvement, in the second estimator, we impute each missing $Y_i$ by $\tilde{m}_c(X_i)$, which leads to the proposed imputation based estimator
\[
\hat{m}_i(x) = \frac{\sum_{i=1}^{n} K_h(x^c - X_i^c)L(x^c, X_i^c; \lambda)(\delta_i Y_i + (1 - \delta_i)\tilde{m}_c(X_i))}{\sum_{i=1}^{n} K_h(x^c - X_i^c)L(x^c, X_i^c; \lambda)}
\]
In a case that $m(x)$ is constant over the sample space of the covariate $X$, one can easily show that $\hat{m}_c(x)$ and $\hat{m}_i(x)$ are equivalent. In the following discussion, we shall focus on non-degenerate conditional mean function, i.e. the covariate $X$ is relevant in predicting the conditional mean function. We note here that as the target of analysis is a conditional mean function, there is no need to carry out a nonparametric multiple imputation as those in [15] and [32], where the missing values are imputed for multiple times from a nonparametric estimator of the conditional distribution function. The effect of such a multiple imputation for nonparametric regression estimation with only continuous covariates is evaluated in [15], and they find that doing so does not lead to any improvement in efficiency.

4. MAIN RESULTS

4.1 Effects of imputation and discrete smoothing

Let $\tilde{f}(x,z)$ be the probability density function of $(X_i, Z_i)$, $f(x) = \int f(x,z)dz$ be the marginal density of $X_i$, $\tilde{w}(x) = \int w(x,z)\tilde{f}(x,z)dz$ be the marginally weighted average propensity. We note that $\tilde{w}(x) \neq w(x)$ in general, which reflects the MAR assumption (4). We define the following with respect to the continuous kernel $K$: $R(K) = \int K^2(u)du$ and $\sigma_K^2 = \int u^2 K(u)du$. The following quantities are needed to describe effects of smoothing the discrete covariates. Let $\mathcal{C}_x = \{s^u: \sum_{i=1}^{d_u} I(x^c_i \neq s^u) = 1\}$ be the nearest strata whose discrete component differs from $x^u$ only in one component. For a $s^u \in \mathcal{C}_x$, define
\[
\alpha(x^u, s^u) = \sum_{k=1}^{d_u} c_k I(x^c_k = s^u),
\]
\[
\beta(x^u, s^u) = \sum_{k=1}^{d_u} \lambda_k I(x^c_k = s^u)
\]
where $c_k$ is the total number of the levels of $X_k^c$. For $\hat{m}_c(x)$, the bias induced by the smoothing of the discrete variables is quantified by
\[
b_{1,u}(x; \lambda) = \sum_{s^u \in \mathcal{C}_x} \left( 1 - \frac{1}{\alpha(x^u, s^u)} \right) \left( \frac{\tilde{w}(x^c, s^u)}{\tilde{w}(x)} \left( m(x^c, s^u) - m(x) \right) \right),
\]
and the bias by smoothing the continuous variables is
\[
b_{2,c}(x; h) = \frac{1}{2} h^2 \sigma_K^2 \tilde{w}^{-1}(x) \left( \text{tr}[\nabla^2 \{m(x, \tilde{w}(x))\}] - m(x)\text{tr}[\nabla^2 \{f(x)\}] \right),
\]
where $\text{tr}$ and $\nabla$ are the trace and differentiation (with respect to $x^c$) operators. Similarly, the biases of $\hat{m}_i(x)$ from smoothing discrete and continuous variables are given by
\[
b_{2,u}(x; \lambda) = \frac{f(x) - \tilde{w}(x)}{f(x)} b_{1,u}(x; \lambda) + \sum_{s^u \in \mathcal{C}_x} \left( 1 - \frac{1}{\alpha(x^u, s^u)} \right) \left( \frac{f(x^c, s^u)}{f(x)} \left( m(x^c, s^u) - m(x) \right) \right),
\]
\[
b_{2,c}(x; h) = \frac{f(x) - \tilde{w}(x)}{f(x)} b_{1,c}(x; h) + \frac{1}{2} h^2 \sigma_K^2 f^{-1}(x) \left( \text{tr}[\nabla^2 \{m(x, f(x))\}] - m(x)\text{tr}[\nabla^2 \{f(x)\}] \right).
\]
Furthermore, let $V_i^c(x) = R(K)\sigma^2(x)/\tilde{w}(x)$,
\[
V_{1,i}^c(x) = \frac{\sigma^2(x)}{f^2(x)} \left[ R(K)\tilde{w}(x) + 2R_2(K)\{f(x) - \tilde{w}(x)\} \right] + R_3(K) \frac{(f(x) - \tilde{w}(x))^2}{\tilde{w}(x)^2}
\]
\[
V_{2,i}^c(x) = \frac{\sigma^2(x)}{f^2(x)} \left[ R(K)\tilde{w}(x) + 3R_2(K)\{f(x) - \tilde{w}(x)\} \right] \text{Nonparametric regression with missing values 465}
\]
where \( R_3(K) = \int K^{(2)}(u)K(u)du \), \( R_4(K) = \int K^{(4)}(u)K(u)du \) and \( K^{(j)}(t) \) is the \( j \)th convolution of \( K(t) \). The following theorem, whose proof is deferred to the appendix, reports the properties of \( \hat{\mu}_c(x) \) and \( \hat{\mu}_I(x) \).

**Theorem 4.1.** Under the assumptions given in the Appendix, let \( \lambda = \min_{i=1}^{d_v} \{ \lambda_i \} \) and \( A(\tilde{\lambda}) = 2 \sum_{j=1}^{d_v} (1 - \lambda_j) \),

\[
E(\hat{\mu}_c(x)) = m(x) + b_{1,a}(x; \tilde{\lambda}) + b_{1,c}(x; h) \\
+ O(h^2(1 - \lambda)^2),
\]

\[
\text{var}(\hat{\mu}_c(x)) = \frac{1}{nh^{d_v}} V_1^c(x) = \frac{A(\tilde{\lambda})}{nh^{d_v}} V_1^c(x) + \frac{1}{nh^{d_v}-2} V_2^c(x) \\
+ O(n^{-1}h^{-d_v-2}\{ h^2 + (1 - \lambda)^2 \}),
\]

\[
E(\hat{\mu}_I(x)) = m(x) + b_{2,a}(x; \tilde{\lambda}) + b_{2,c}(x; h) \\
+ O(h^2(1 - \lambda)^2),
\]

\[
\text{var}(\hat{\mu}_I(x)) = \frac{1}{nh^{d_v}} V_1^I(x) = \frac{A(\tilde{\lambda})}{nh^{d_v}} V_1^I(x) + \frac{1}{nh^{d_v}-2} V_3^I(x) \\
+ O(n^{-1}h^{-d_v-2}\{ h^2 + (1 - \lambda)^2 \}),
\]

where \( V_2^c(x) \) and \( V_3^I(x) \) are bounded terms associated with second order variances given by (21) and (28) in the Appendix.

The implications of Theorem 1 are the following:

**Remark 1.** Theorem 1 shows that the imputation based estimator \( \hat{\mu}_I(x) \) has smaller variance than \( \hat{\mu}_c(x) \). This can be appreciated by comparing \( V_1^c(x) \) and \( V_1^I(x) \), which define the leading order variances of the two estimators. For a commonly used symmetric kernel, 0 is the unique maximizer of \( K \) and its convolution \( K^{(2)} \) and we have \( R_2(K) < R(K) \) and \( R_3(K) < R(K) \); see also [5] and [6]. Therefore,

\[
V_1^I(x) \leq \frac{R(K)\sigma^2(x)}{f^2(x)\bar{w}(x)} \{ \bar{w}(x) + f(x) - \bar{w}(x) \}^2 = V_1^c(x).
\]

We note here that \( f(x) - \bar{w}(x) = \int \{1 - w(x, z)\}f(x, z)dz \geq 0 \). When there is no missing value, i.e. \( w(x, z) \equiv 1 \), the theorem implies that the leading variance of the oracle (who knows all missing \( Y_i \)) estimator \( \hat{\mu}_c(x) \) in (2) is

\[
(nh^{d_v})^{-1} R(K)\sigma^2(x)/f(x).
\]

Thus, \( \hat{\mu}_c(x) \) which ignores the missing values endures a variance inflation by a factor \( f(x)/\bar{w}(x) \). The proposed \( \hat{\mu}_I(x) \) removes part of the variance inflation by utilizing missing value information. Our finding here is more explicit and general than that given in [8] who considers the case where all covariates are continuous.

As an alternative, one could replace \( Y_i \) in the numerator of (7) also by \( \bar{\mu}_c(X_i) \), resulting in a smoothing twice estimator. In this case, the variance of the resulting estimator can be further reduced. However, one can show that extra bias arises due to smoothing twice. This is related to the finding in [5] and [6], although their approaches were proposed for nonparametric regression with no missing data.

**Remark 2.** By smoothing the categorical variables, both \( \hat{\mu}_c(x) \) and \( \hat{\mu}_I(x) \) enjoy variance reductions as shown by the terms involving \( A(\tilde{\lambda}) \). This is a result of combining data within neighboring cells defined by the discrete variables. Although the variance reductions are at the second order \( (1 - \lambda)/(nh^{d_v}) \), the realized reduction in finite samples can be substantial, especially when large number of categorical variables in presence resulting in some cells with sparse observations which is the case for many applications. The second order result on the effects of smoothing the discrete variables can be viewed as extensions of [16] for density estimation for purely discretely-valued random variables for nonparametric regression with mixed covariates in the absence of missing values. If the regression is absent of the discrete covariates, the implication of Theorem 1 agrees to that in [8]. And we note that a second order variance reduction is the distinctive benefit of smoothing the discrete covariate. In finite sample performance, such reduction may be substantial.

**Remark 3.** The theorem also contains results on the bias of the estimators. The bias terms have contributions from both discrete and continuous components. By a closer look at the bias terms \( b_{j,a}(\cdot) \) and \( b_{j,c}(\cdot) \) where \( j = 1, 2 \), we may find that the feature of the bias has two sources. One is the behavior of \( m(\cdot, s^a) \), \( w(\cdot, s^a) \) and \( f(\cdot, s^a) \) in neighboring cells \( s^a \) around the target cell \( x^a \); and the other is the derivatives of \( m(\cdot, x^a) \), \( w(\cdot, x^a) \) and \( f(\cdot, x^a) \) in the local cell defined by \( x^a \). We note that when discrete covariate is incorporated, the bias in \( \hat{\mu}_I(\cdot) \) may not necessarily worsen. The optimal bandwidths \( (h, 1 - \tilde{\lambda}) \) that minimize the mean square error (MSE) or the mean integrated square error (MISE) satisfy \( h \sim n^{-1/(4+d_v)} \) and \( (1 - \tilde{\lambda}) \sim n^{-2/(4+d_v)} \), which means that \( A(\tilde{\lambda}) \sim n^{-2/(4+d_v)} \) as well. These rates coincide with those obtained in [17] when there is no missing values.

**Remark 4.** When \( d_u \) is fixed and finite, we see from Theorem 1 that the variances of \( \hat{\mu}_I(\cdot) \) and \( \hat{\mu}_c(\cdot) \) are \( O(n^{-1}h^{-d_v}) \) which is coincident with that of a \( d_v \) dimensional continuous covariate smoothing. This illustrates another advantage of a smoothing discrete covariate being less restrictive to the curse of dimensionality [21]. When there is no continuous covariate, the regression function estimates the overall mean of each category formed by the combinations of discrete covariate. In such a case, \( \hat{\mu}_c(x) \) and \( \hat{\mu}_I(x) \) have the same variance of \( O(n^{-1}) \) in the leading order of magnitude.
4.2 Simultaneous Confidence Bands

To assess the level of uncertainty, we study how to obtain simultaneous confidence bands for the nonparametric regression estimates in the presence of discrete covariate and missing values. When all components of data are available, simultaneous confidence bands for kernel smoothing methods have been studied extensively; see for example [3, 11, 33] and [34] and reference therein.

We follow the convention in existing confidence band theory [11, 33] by considering \( d_x = 1 \) and \( X^\infty = (0, 1) \) for simplicity, and extend the theory to the mixed covariate case. Following Theorem 4.1, it can be shown following the approach in [6] that as \( n \to \infty \),

\[
\sqrt{\frac{nh}{V_1^T(x)}} \left( \hat{m}_I(x) - m(x) - b_{2,u}(x, \bar{\lambda}) - b_{2,c}(x, h) \right) \overset{d}{\to} N(0,1).
\]

Thus the confidence band for \( m(., x^u) \) can be constructed for each \( x^u \in X^u \). Because \( d_u \) is fixed and finite, for each \( x^u \in X^u \),

\[
\lim_{n \to \infty} P \left( \sup_{x^u \in (0,1)} \left\{ \left\| \sqrt{\frac{nh}{V_1^T(x)}} \left( \hat{m}_I(x) - m(x) - b_{2,u}(x, \bar{\lambda}) - b_{2,c}(x, h) \right) \right\| \leq L_n(\alpha) \right\} = 1 - \alpha
\]

where \( L_n(\alpha) = \sqrt{-2 \log(\log(n)) + (A - z_\alpha)/\sqrt{2 \log(h)}} \), \( z_\alpha = \log \log \{ \{ 1 - \alpha \}^{-1/2} \} \) and \( A = \log \{ \int K^2(u)du/\int K^2(u)du \}^{1/2} \). When \( \bar{\lambda} = 1 \), i.e. without smoothing discrete covariate, the proof of (10) follows the conventional approaches in [3, 33] and [34]. When \( \bar{\lambda} \) is chosen by the cross-validation method, we observe that the impact on the variance only occurs at the second order, so that the conventional arguments carry over for (10) as well. A similar result for \( \hat{m}(x) \) is also valid. In applications, \( V_1^T(x) \), \( b_{2,u}(x, \bar{\lambda}) \) and \( b_{2,c}(x, h) \) are substituted by their consistent estimators respectively.

4.3 Locally Adaptive Smoothing

In practice, the smoothing parameter \( (h, \bar{\lambda}) \) can be chosen by minimizing the cross-validation score function \( CV(h, \lambda) = \sum_{i=1}^{n} \left\{ Y_i - \hat{m}_{h,\bar{\lambda}}(X_i) \right\}^2 \delta_i \) where \( \hat{m}_{h,\bar{\lambda}}(\cdot) \) is the nonparametric estimator excluding the \( i \)th using observation using \( (h, \bar{\lambda}) \). The \( h \) and \( \bar{\lambda} \) chosen asymptotically minimize the mean integrated squared error (MISE) [17]. For smoothing continuous covariate, the choice of bandwidth locally adaptive to the smoothness of the underlying regression function was studied in [12] and [14].

In our study, we use the method proposed in [14] to choose the bandwidth that is locally adaptive with respect to the continuous covariate. According to Theorem 1, the optimal bandwidth \( h \) is \( O(n^{-1/5}) \) when \( d_x = 1 \). Suppose there is a collection of functions \( G \) and for each \( g(\cdot) \in G \), the corresponding bandwidth is given by \( h(x) = n^{-1/5}g(x) \). Then the locally adaptive bandwidth can be chosen by

\[
\min_{g \in G} CV(g) = \sum_{i=1}^{n} \left\{ Y_i - \hat{m}_{h,\lambda}^{(-i)}(X_i) \right\}^2 \delta_i.
\]

The \( G \) is chosen to be a class of continuous function with certain degrees of derivatives. According to the theoretical results in [14], the minimization of the CV function over a large enough \( G \) achieves the optimal bandwidth that is adaptive to the smoothness of \( m(x) \). In our paper, we follow the suggestion in [14] to formulate \( G \) by cubic spline interpolation in simulation study and data analysis.

5. SIMULATION STUDIES

We conduct simulation studies to examine the finite sample performance of the methods. A natural extension of existing nonparametric regression such as in [21] to a situation with missing data is corresponding to the \( \hat{m}_I(x) \) in our paper. Therefore in the simulation we compare the performance of \( m(., x)^u \) and \( m_I(x) \).

We consider three variables in the covariate with \( X_1 \) continuous following a uniform distribution on the interval \((0, 1)\), \( X_2, X_3 \in \{0, 1\} \) with \( P(X_2 = 0) = P(X_3 = 0) = 0.4 \). The conditional mean function was chosen to be

\[
E(Y|X) = m(X) = \beta_0 + \beta_1 X_1 + \beta_2 \sin^2(2\pi(X_1 - 0.5))
\]

where \( l = 2X_0 + X_1 + 1 \) is a one to one mapping from \((0, 1)^2 \) to \((0, 4)\). The vector \( \beta = (\beta_0, \beta_1, \beta_2)^T \) reflects different features with respect to the continuous variable \( X_1 \) among combinations of discrete variables. The \( \beta_l \) was generated from \( \mathcal{N}(\mu_1, \Sigma_1) \) and fixed throughout the simulation, where \( \mu = (1, 0.5, 3.5)^T \) and \( \Sigma = \text{diag}(0.04, 0.04 \times 0.5, 0.04 \times 3.5) \). The \( Y \) is generated by \( Y = m(X) + \epsilon \) where \( \epsilon \sim \mathcal{N}(0, \sigma^2) \). We attempted two noise levels \( \sigma = 0.3 \) and \( \sigma = 0.6 \) in the simulation study and the sample sizes were set to be \( n = 50, 100, 200 \). The missing propensity (no extra \( Z \) was specified by

\[
P(\delta = 1|X) = \{ \exp(-b(X)) + 1 \}^{-1},
\]

and \( \phi(x) \) is the density function of the standard normal distribution. The \( \theta_l = (\theta_{0l}, \ldots, \theta_{ul}) \) was generated from \( \mathcal{N}(\mu_2, \Sigma_2) \) and kept fixed during the simulation, and \( \mu_2 = (1, 0.5, -0.3, 0.5, 0.1)^T \) and \( \Sigma_2 = \text{diag}(0.01, \ldots, 0.01) \). Using this propensity function, about 30–35% response variables are missing.

The bandwidths for smoothing discrete covariate were chosen by the Cross Validation method as in [17]. The bandwidth in the simulation with respect to the continuous covariate is calculated by (11) which is adaptive to the smoothness of the continuous covariate. The cubic spline interpolation was conducted by choosing ten equal spaced grid points.

Nonparametric regression with missing values 467
Table 1. Simulation results of the averaged squared bias (ABias), Variance (AVar) and MSE (AMSE) of \(\hat{m}_c(x)\) and \(\hat{m}_l(x)\), and the empirical coverages of the confidence bands for nominal levels \(\alpha = 0.1\) and \(\alpha = 0.05\)

<table>
<thead>
<tr>
<th></th>
<th>ABias(^2)</th>
<th>AVar</th>
<th>AMSE</th>
<th>(\alpha = 0.1) Coverage</th>
<th>(\alpha = 0.05) Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\hat{m}_c)</td>
<td>0.048</td>
<td>0.500</td>
<td>0.546</td>
<td>0.863</td>
<td>0.587</td>
</tr>
<tr>
<td>(\hat{m}_l)</td>
<td>0.051</td>
<td>0.443</td>
<td>0.494</td>
<td>0.871</td>
<td>0.901</td>
</tr>
<tr>
<td>(\hat{m}_c)</td>
<td>0.042</td>
<td>0.521</td>
<td>0.564</td>
<td>0.840</td>
<td>0.878</td>
</tr>
<tr>
<td>(\hat{m}_l)</td>
<td>0.050</td>
<td>0.465</td>
<td>0.515</td>
<td>0.848</td>
<td>0.899</td>
</tr>
<tr>
<td>(n = 100)</td>
<td>0.017</td>
<td>0.135</td>
<td>0.152</td>
<td>0.904</td>
<td>0.953</td>
</tr>
<tr>
<td>(\hat{m}_c)</td>
<td>0.019</td>
<td>0.114</td>
<td>0.133</td>
<td>0.892</td>
<td>0.951</td>
</tr>
<tr>
<td>(\hat{m}_l)</td>
<td>0.023</td>
<td>0.175</td>
<td>0.198</td>
<td>0.877</td>
<td>0.926</td>
</tr>
<tr>
<td>(\hat{m}_l)</td>
<td>0.028</td>
<td>0.151</td>
<td>0.179</td>
<td>0.889</td>
<td>0.943</td>
</tr>
<tr>
<td>(n = 200)</td>
<td>0.009</td>
<td>0.045</td>
<td>0.054</td>
<td>0.918</td>
<td>0.956</td>
</tr>
<tr>
<td>(\hat{m}_c)</td>
<td>0.011</td>
<td>0.038</td>
<td>0.049</td>
<td>0.916</td>
<td>0.959</td>
</tr>
<tr>
<td>(\hat{m}_l)</td>
<td>0.017</td>
<td>0.083</td>
<td>0.010</td>
<td>0.909</td>
<td>0.945</td>
</tr>
<tr>
<td>(\hat{m}_l)</td>
<td>0.021</td>
<td>0.070</td>
<td>0.091</td>
<td>0.903</td>
<td>0.941</td>
</tr>
</tbody>
</table>

Then (11) is optimized with respect to the ten bandwidths. On each combination of the discrete covariate, we estimated the conditional mean function on 50 equally spaced grid points. By repeating the simulation 1,000 times, we summarized the bias, variance and mean squared error (MSE) for each point. In Table 1, we report the averaged squared bias (ABias\(^2\)), variance (AVar) and MSE (AMSE) for \(\hat{m}_c\) and \(\hat{m}_l\) for each noise level and sample size. From Table 1, we observe that \(\hat{m}_l(x)\) consistently had smaller AMSE than \(\hat{m}_c(x)\) because of its smaller variance, especially when the sample size is small. This confirmed our finding in Theorem 1.

To obtain the simultaneous confidence band by (10), we follow the approaches in [33] and [34] to estimate \(\sigma^2(x)\), \(\hat{w}(x)\) and \(f(x)\) using kernel smoothing method with bandwidths chosen by minimizing corresponding cross-validation methods. The \(m^{\prime\prime}(x)\) and \(f'(x)\) were also estimated by kernel smoothing method whose bandwidths were chosen by the reference rule [19]. Two confidence levels \(\alpha = 0.1\) and \(\alpha = 0.05\) were studied in the simulation whose results are reported in Table 1. It is seen from the simulation that the confidence bands had empirical coverage close to the nominal level when the sample size is reasonably large. When sample size was small, \(n = 50\), the confidence bands had coverage below the nominal level. This is reasonable considering the amount of data missing and the fact of a slow convergent rate of simultaneous confidence bands [34].

6. DATA ANALYSIS

6.1 AIDS clinical trials group study 175 data

We demonstrate the application of a proposed approach by an application in the ACGT 175 data considered in [9]. The data consist of the CD4 counts along with other variables for subjects receiving ZDV (control) and three other therapies (treatments). We consider the logarithm of the CD4 counts at the 96 week as the response variable, where 1,342 out of 2,139 were complete cases and the rest were missing, because of death, dropout or other reasons. The logarithm of interim CD4 counts at 20 weeks is used as the continuous covariate, and the other covariates are the indicators of symptoms, prior antiretroviral therapy and Karnofsky score. Although Karnofsky score is a continuous measurement, it takes only four values (70, 80, 90, 100) where the first two were rarely observed. So we treat it as a binary variable by dichotomizing it at 100. It is documented that the covariate can impact the missing mechanism of the response variable [9].

Biweight kernel \(K(x) = 15/16(1-x^2)^2I(|x| \leq 1)\) is applied for the continuous covariate, and we chose the smoothing bandwidths by the cross-validation, which led to \(\hat{m}_c(x)\) and \(\hat{m}_l(x)\). The adaptive bandwidth selected for the continuous variable is close to a constant, due to the lack of a local feature as seen from the estimated regression function. We then estimated the density function \(f(x)\), the marginal propensity function \(\hat{w}(x)\), the conditional variance function \(\sigma^2(x)\) and the relevant derivatives by the kernel smoothing methods respectively as in the simulation study. Based on these estimated quantities, we obtained the confidence bands from (10). We can observe from the estimations that heterogeneous patterns of the conditional mean function with respect to the interim CD4 counts for different groups of patients. Figure 1 displays the nonparametric regression estimates \(\hat{m}_c(x)\) and \(\hat{m}_l(x)\) with 95% confidence bands respectively for controlled and treated patients from a specific group identified by the patient’s characteristic. From

468 S. X. Chen and C. Y. Tang
features illustrated by Figure 1, we can see that the estimated regression functions differ between the control and treatment groups, as well as among different combinations of the other discrete covariate. The $\hat{m}_c(x)$ and $\hat{m}_I(x)$ estimates share some similar patterns, though there are some differences at various places. These results found by the non-parametric method is complementary to the additive parametric model with quadratic predictors considered in [9]. We also note that the width of the confidence bands for $\hat{m}_I(x)$ is smaller than that of $\hat{m}_c(x)$, which is expected from our theoretical results. From the confidence band estimates, we observed a widening trend at the two ends for the two estimates. This is mainly due to the sparsity of observations towards the both ends.

6.2 Estimating US Census enumeration probability function

We apply the proposed kernel estimators to an ACE revision II research data [31] of the 2000 US Census. The ACE study consists of two independent samples from the US population, namely the E and P samples respectively. The capture-recapture design is implemented to identify matches between the two samples. The response variable $Y$ is the match indicator and its conditional mean given the covariate is the important enumeration probability function. A significant portion of enumerations in the US Census ACE study has missing match indicator, due to the fact that a definite match or no-match can not be established between the records in the P and E samples. In our analysis, the covariates $X_i$ include age ($X_{i1}$), sex ($X_{i2}$, 2 levels), housing tenure ($X_{i3}$, 2 levels: owner and renter), and racial origins ($X_{i4}$, 7 levels: American Indian or Alaska Natives on Reservation, Off-Reservation American Indian or Alaska Native, Hispanic, Non-Hispanic Black, Native Hawaiian or Pacific Islander, Non-Hispanic Asian and Non-Hispanic white or other races) and Geographical region ($X_{i5}$, 4 levels: Northeast, Midwest, South and West). Additional covariates may be included without changing the tune of the analysis. The response variable $Y_i$ is the match status between the P sample $\mathcal{P}$ and the E sample $\mathcal{E}$, where

$$Y_i = \begin{cases} 1 & \text{(a match)} \\ 0 & \text{(not a match)} \\ ? & \text{(missing value)} \end{cases} \quad \text{if } i \in \mathcal{P} \cap \mathcal{E};$$
$$Y_i = \begin{cases} 1 & \text{(a match)} \\ 0 & \text{(not a match)} \\ ? & \text{(missing value)} \end{cases} \quad \text{if } i \in \mathcal{P} \text{ but not in } \mathcal{E};$$
$$Y_i = \begin{cases} 1 & \text{(a match)} \\ 0 & \text{(not a match)} \\ ? & \text{(missing value)} \end{cases} \quad \text{if } i \text{ is an unresolved case}.$$
Figure 2. Kernel estimates of the enumeration probability $m(x)$ based on $m_1(x)$. Bandwidths used are $h = 5.5$ and $\lambda = 0.8$.

The bandwidths prescribed were $h = 5.5$ and $\lambda = 0.8$, which were the bandwidths used in the imputation based estimates for $m(x)$ in Figure 2.

It is observed from Figure 2 that the covariate in the analysis contributes to the heterogeneity in the enumeration probability function $m(x)$. The age effect was quite apparent in the estimates for $m(x)$. At the same time, the kernel estimates change substantially with respect to the other categorical variables. Figure 2 indicates that Northeast White Male Owners had an overall higher enumeration probability than Northeast Hispanic Female Owners and Midwest Black Male Renters, which might be expected. While these confirm the effects of these covariates, they do reveal the difficulty in capturing the underlying forms of the functions with respect to these discrete covariates. The wave-like pattern in the $m(x)$ estimates in some cells suggests some age-heaping in a multiple of 5 or 10 years in ages beyond 30. Figure 3 displays the kernel estimates for the missing propensity score $w(x)$, which was as interesting as Figure 2. For instance, the White Male owners had a very small chance of being missing. In contrast, the Hispanic Female owners endured larger missingness while the Black Male renters experienced the highest missing values among the three.

Both Figures 2 and 3 also reveal challenges that one would face in proposing reasonable parametric regression models. There are 112 post-strata based on the four discrete covariates. The sample size within some of these 112 post-strata can be very small, for instance the Native Hawaiian or Pacific Islander. Getting a workable model for each stratum is quite a task. The task will only grow when more covariates are included. At the same time, the figures show that the proposed kernel estimation is flexible and adaptive to varying functional forms in $m(x)$. As shown by our theoretical investigation, the kernel estimates are consistent and reflective to the underlying model structure without imposing any subjective assumptions.

APPENDIX A. TECHNICAL DETAILS

Technical assumptions

Let covariate $X_i = (X_i^c, X_i^u)$ where $X_i^c$ is a $d_c$-dimensional continuous covariate and $X_i^u$ is a $d_u$-dimensional unordered categorical covariate, $X = (X^c, X^u)$ be the support of $X_i$, where $X^c$ and $X^u$ are the supports of $X_i^c$ and $X_i^u$ respectively. We assume the model (5) for independent and
Figure 3. Kernel estimates of the census missing propensity function \(w(x)\). Bandwidths used are \(h = 5.0\) and \(\lambda = 0.8\).

identically distributed data pairs \(\{(X_i, Z, Y_i)\}_{i=1}^{n}\). And the following conditions are assumed in Theorem 1.

C.1 Let \(K(\cdot)\) be a \(d_v\) variates nonnegative, bounded and symmetric probability density function with bounded second derivative. The smoothing bandwidths satisfy that \(h \to 0\) and \(\max_{1 \leq j \leq d_v} \{(1 - \lambda_j)\} \to 0\) and \(nh^{d_v}/\log(n) \to \infty\) as \(n \to \infty\).

C.2 We assume missing at random in \(Y\), namely \(P(\delta = 1|Y, X, Z) = P(\delta = 1|X, Z) := w(X, Z)\), where \(w(x^c, x^u, z) \geq C_w > 0\) for a constant \(C_w\) and \(w(x^c, x^u, z)\) has bounded continuous second partial derivative with respect to \(x^c\).

C.3 For given \(x^u \in \mathcal{X}^u\), \(m(x^c, x^u)\) and the probability density of the covariate \(f(x^c, x^u)\) have bounded continuous second partial derivatives with respect to \(x^c\), and there exist \(C_f > 0\) such that \(f(x^c, x^u) \geq C_f\).

**Sketched Proof of Theorem 1.** We sketch the proof of Theorem 1 here where the complete proof with more details is available in [29]. Let \(K_{h,\lambda}(u, v) = K_h(u^c - v^c) L(u^v, v^u; \lambda)\) and define \(\hat{f}_c(x) = n^{-1} \sum_{i=1}^{n} K_{h,\lambda}(x, X_i) \delta_i\) and \(\hat{\phi}_c(x) = n^{-1} \sum_{i=1}^{n} K_{h,\lambda}(x, X_i) \delta_i Y_i\). From (6), \(\hat{m}_c(x) = \hat{\phi}_c(x)/\hat{f}_c(x)\).

We show that

\[
E(\hat{f}_c(x)) = \tilde{w}(x) + \frac{1}{2} h^2 \sigma_K^2 \text{tr} [\nabla^2 \{ \tilde{w}(x) f(x) \}]
- \tilde{w}(x) \sum_{j=1}^{p_x} (1 - \lambda_j)
+ \sum_{s^u \in \mathcal{C}^u_2} (1 - \beta_s)(x^c, s^u) \tilde{w}(x^c, s^u)
+ O(h^2(1 - \lambda)^2);
\]

\[
\text{var}(\hat{f}_c(x)) = \frac{1 - A(\lambda)}{nh^{d_v}} R(K) \tilde{w}(x)
+ \frac{1}{2nh^{d_v-2}} S_2(K) \text{tr} [\nabla^2 \{ \tilde{w}(x) f(x) \}]
+ O(n^{-1}h^{-d_v+2} \{ k^2 + (1 - \lambda)^2 \}),
\]

\[
E(\hat{\phi}_c(x)) = m(x) \tilde{w}(x) + \frac{1}{2} h^2 \sigma_K^2 \text{tr} [\nabla^2 \{ m(x) \tilde{w}(x) \}]
- m(x) \tilde{w}(x) \sum_{j=1}^{p_x} (1 - \lambda_j)
\]

*Nonparametric regression with missing values* 471
\[+ \sum_{s^u \in \mathbb{C}^u_2} \frac{1 - \beta_3(x^u, s^u)}{\alpha(x^u, s^u)} m(x^c, s^u) \hat{w}(x^c, s^u) + O(h^2(1 - \lambda)^2),\]

\[\text{(17)}\]
\[\text{var}\{\hat{\phi}_c(x)\} = \frac{1 - A(\bar{\lambda})}{nh^{d_c}} R(K) \hat{w}(x) \{m^2(x) + \sigma^2(x)\}
+ \frac{S_2(K)\text{tr}\{\nabla^2 \{m^2(x) + \sigma^2(x)\} \hat{w}(x)f(x)\}}{2nh^{d-2}},
+ O(n^{-1}h^{-d_c+2}\{h^2 + (1 - \lambda)^2\}),\]

where \(S_1(K) = \int u^t K^2(\hat{u})du\). Results in (14)–(17) indicate that based on the complete data only, the estimation of the density function \(f(x)\) by \(\hat{f}_c(x)\) is actually biased and so is the estimation of \(\phi(x) = m(x)f(x)\) by \(\hat{\phi}_c(x)\). Similarly, we have

\[\text{(18)}\]
\[\text{cov}\{\hat{\phi}_c(x), \hat{f}_c(x)\} = \frac{1 - A(\bar{\lambda})}{nh^{d_c}} R(K) m(x) \hat{w}(x)
+ \frac{S_2(K)\text{tr}\{\nabla^2 \{m(x)\hat{w}(x)f(x)\}\}}{2nh^{d-2}},
+ O(n^{-1}h^{-d_c+2}\{h^2 + (1 - \lambda)^2\}).\]

Define \(f_c(x) = \hat{w}(x)\) and \(\phi_c(x) = m(x)\hat{w}(x)\). An expansion of \(\hat{m}_c(x)\) is

\[\text{(19)}\]
\[\hat{m}_c(x) = \frac{\hat{\phi}_c(x)}{f_c(x)} + \frac{\{\hat{\phi}_c(x) - \phi_c(x)\}}{f_c(x)}
- \phi_c(x) \frac{\{\hat{f}_c(x) - f_c(x)\}}{f_c^2(x)}(1 + o_p(1)).\]

Taking expectation on (19), and from (14) and (16),

\[\text{(20)}\]
\[E\{\hat{m}_c(x)\} = m(x) + b_{1, u}(x; \bar{\lambda}) + b_{1, c}(x; h) + O\{h^2(1 - \lambda)^2\}.\]

By applying the variance operation on (19) and substituting equations (15), (17) and (18), we summarize that

\[\text{(21)}\]
\[\text{var}\{\hat{m}_c(x)\} = \frac{1}{nh^{d_c}} V_1(x) - \frac{A(\bar{\lambda})}{nh^{d_c}} V_1^c(x) + \frac{1}{nh^{d-2}} V_2^c(x)
+ O(n^{-1}h^{-d_c+2}\{h^2 + (1 - \lambda)^2\}),\]

\[V_2^c(x) = \frac{1}{2} \hat{w}^2(x) \{S_2(K) \{\text{tr}\{m^2(x)\nabla^2 \{\hat{w}(x)f(x)\}\} \nabla^2 \{m^2(x) + \sigma^2(x)\} \hat{w}(x)f(x)\}\}
- 2m(x) \nabla^2 \{m(x)\hat{w}(x)f(x)\} \}.\]

These establish the first part of Theorem 1.

We now evaluate the imputation based estimator \(\hat{m}_I(x)\). Let \(\hat{f}(x) = n^{-1} \sum_{i=1}^n K_{h,r}(x, X_i)\) and \(\hat{\phi}_I(x) = n^{-1} \sum_{i=1}^n K_{h,r}(x, X_i)(1 - \delta_i)\hat{m}_I(X_i)\). From (7), by letting \(\hat{\phi}_I(x) = \hat{\phi}_c(x) + \hat{\phi}_I(x)\),

\[\text{(22)}\]
\[\hat{m}_I(x) = \{\hat{\phi}_c(x) + \hat{\phi}_I(x)\}/\hat{f}(x) = \hat{\phi}_I(x)/\hat{f}(x).\]

Similar to (19), we may establish an expansion for (22). Let \(\phi_I(x) = m(x)f(x)\), we have

\[\text{(23)}\]
\[\hat{m}_I(x) = \frac{\phi_I(x)}{f(x)} + \frac{\{\hat{\phi}_I(x) - \phi_I(x)\}}{f(x)}
- \frac{\phi_I(x)\{\hat{f}(x) - f(x)\}}{f^2(x)}(1 + o_p(1)).\]

We establish that

\[\text{(24)}\]
\[E\{\hat{f}(x)\} = f(x) + \frac{1}{2} h^2 \sigma^2(K)\text{tr}\{\nabla^2 \{f(x)\}\} - f(x) \sum_{j=1}^p \lambda_j
+ \sum_{s^u \in \mathbb{C}^u_2} \frac{1 - \beta_3(x^u, s^u)}{\alpha(x^u, s^u)} f(x^c, s^u) + O(h^2(1 - \lambda)^2),\]

\[\text{var}\{\hat{f}(x)\}
= \frac{1 - A(\bar{\lambda})}{nh^{d_c}} R(K) f(x) + \frac{1}{2nh^{d-2}} S_2(K)\text{tr}\{\nabla^2 \{f(x)\}\},\]

\[+ O(n^{-1}h^{-d_c+2}\{h^2 + (1 - \lambda)^2\}).\]

By taking expectation on \(\hat{\phi}_I(x)\), we have

\[\text{(25)}\]
\[E\{\hat{m}_I(x)\} = m(x) + b_{2, u}(x; \bar{\lambda}) + b_{2, c}(x; h) + O\{h^2(1 - \lambda)^2\}.\]

The following notations of functions are introduced for simplification. Let \(g_1 = \frac{1 - \bar{w}}{\bar{w}^2}, g_2 = m\bar{w}, g_3 = \frac{1 - \bar{w}}{\bar{w}}, g_4 = \{m^2 + \sigma^2\} \bar{w} f, g_5 = \bar{w} f, g_6(x) = \frac{(1 - \bar{w}) m}{\bar{w}}\) and \(g_7(x) = \frac{(1 - \bar{w}) m}{\bar{w}^2}\). Further, it can be shown that
where

\[ \xi_{ij}(K) = \int w^j K^{(i)}(u)K(u)du \quad \text{and} \]

\[ \zeta_i(K) = \int u_1 u_2 K^{(i)}(u_1 + u_2)K(u_2)du_1du_2, \]

\[ C_1(x) = \text{tr}\left(\nabla^2[\{1 - \tilde{\omega}\}\{m^2(x)f\}]\right), \]

\[ C_2(x) = (1 - \tilde{\omega})\sigma^2(x)f\text{tr}[\nabla^2\{g_3\}] + \frac{1}{2}g_3^2\text{tr}[\nabla^2\{g_4\}]
+ \frac{1}{2}g_6^2\text{tr}[\nabla^2\{g_6\}] - g_3g_6\text{tr}[\nabla^2\{g_5\}], \]

\[ C_3(x) = 2g_6\nabla^T\{g_3\}J\nabla\{g_4\} + g_6\nabla^T\{g_6\}J\nabla\{g_5\}
+ g_4\nabla^T\{g_3\}J\nabla\{g_5\} + g_6\nabla^T\{g_6\}J\nabla\{g_5\}
+ g_4\nabla^T\{g_4\}J\nabla\{g_5\}
- 2g_6\nabla^T\{g_3\}J\nabla\{g_5\} + g_4\nabla^T\{g_3\}J\nabla\{g_6\}
+ g_3\nabla^T\{g_3\}J\nabla\{g_6\}]. \]

Here \( J = 11^T \) for \( 1 = (1, \ldots, 1) \) in a similar fashion, we establish that

\[ \text{cov}\{\hat{\phi}_i(x), \hat{\phi}_j(x)\} = \left\{1 - 1.5A(\tilde{\lambda})\right\} R_2(K)\left\{f(x) - \tilde{\omega}(x)\right\}\sigma^2(x)
+ \frac{1}{nh^{d-2}}\left\{\zeta_1(K)C_4 + \frac{1}{2}\xi_{22}(K)C_5\right\}
+ O\left(n^{-1}h^{-d+2}\{h^2 + (1 - \lambda)^2\}\right) \]

where \( C_4(x) = \nabla^T\{g_4\}J\nabla\{g_5\} - \nabla^T\{g_6\}J\nabla\{g_6\} \) and

\[ C_5(x) = g_4\nabla\{g_4\}J\nabla\{g_5\} + g_6\nabla\{g_4\}J\nabla\{g_5\} - g_3g_6\nabla\{g_6\} - g_3g_6\nabla\{g_5\}. \]

Furthermore

\[ \text{cov}\{\hat{\phi}_i(x), \hat{f}(x)\} = \left\{1 - A(\tilde{\lambda})\right\} R_2(K)\left\{m(x)f(x) - \tilde{\omega}(x)\right\} + m(x)\tilde{\omega}(x)
+ \frac{1}{nh^{d-2}}\left\{\zeta_{12}(K)S_0(K)\right\}
+ \frac{1}{nh^{d-2}}\left\{\frac{1}{2}S_2(K)C_6 + \zeta_1(K)C_7 + \frac{1}{2}\xi_{22}(K)C_9\right\}
+ O\left(n^{-1}h^{-d+2}\{h^2 + (1 - \lambda)^2\}\right) \]

where \( C_6(x) = \text{tr}\left(\nabla^2[\{1 - w\}mf]\right) + \text{tr}[\nabla^2\{g_2\}], \)

\[ C_7(x) = \nabla^T\{g_2\}J\nabla\{g_3\} - \nabla^T\{g_6\}J\nabla\{g_6\} \) and \( C_9(x) = g_6\nabla\{g_2\}J\nabla\{g_3\} + g_3\nabla\{g_2\}J\nabla\{g_6\} - g_3g_6\nabla\{g_6\} - g_3g_6\nabla\{g_5\}. \]

Finally, from \((17), (24), (25), (26)\) and \((27)\), we have

\[ \text{var}\{\hat{m}_i(x)\} = \frac{1}{nh^d} V_2^i(x) - \frac{1}{nh^d} V_2^i(x) + \frac{1}{nh^d} V_2^i(x)
+ O\left(n^{-1}h^{-d+2}\{h^2 + (1 - \lambda)^2\}\right), \]

\[ V_2^i(x) = S_2(K)T_1(x) + \xi_{22}(K)T_2(x)
+ \zeta_1(K)T_4(x) + \zeta_2(K)T_5(x) \]

where

\[ T_1(x) = \frac{1}{2}f^{-2}(x)\text{tr}[\nabla^2\{g_3\} + C_1(x)
+ m^2(x)\nabla^2\{f(x)\} - 2m(x)C_6(x)], \]

\[ T_2(x) = \{C_5(x) - (m(x)C_6(x))/f^2(x), \]

\[ T_3(x) = C_2(x)/f^2(x), \]

\[ T_4(x) = 2[C_4(x) - m(x)C_7(x)]/f^2(x) \] and

\[ T_5(x) = C_3(x)/f^2(x). \]

These conclude Theorem 1.

\[ \boxed{} \]

Received 13 June 2011

REFERENCES


Nonparametric regression with missing values 473


Song Xi Chen
Department of Business Statistics and Econometrics
Guanghua School of Management
and Center for Statistical Science
Peking University
Beijing 100871
China

E-mail address: songchen@iastate.edu

Cheng Yong Tang
Department of Statistics
Iowa State University
Ames, IA 50011-1210
USA
E-mail address: tangc@nus.edu.sg