Dissertation

Model selection and nonparametric estimation for regression models

Submitted by

Zonglin He

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2014

Doctoral Committee:

Advisor: Jean Opsomer

F. Jay Breidt
Mary Meyer
John Elder

Abstract

Model selection and nonparametric estimation for regression models

In this dissertation, we deal with two different topics in statistics. The first topic in survey sampling deals with variable selection for linear regression model from which we will sample with a possibly informative design. Under the assumption that the finite population is generated by a multivariate linear regression model from which we will sample with a possibly informative design, we particularly study the variable selection criterion named predicted residual sums of squares in the sampling context theoretically. We examine the asymptotic properties of weighted and unweighted predicted residual sums of squares under weighted least squares regression estimation and ordinary least squares regression estimation. One simulation study for the variable selection criteria are provided, with the purpose of showing their ability to select the correct model in the practical situation.

For the second topic, we are interested in fitting a nonparametric regression model to data for the situation in which some of the covariates are categorical. In the univariate case where the covariate is a ordinal variable, we extend the local polynomial estimator, which normally requires continuous covariates, to a local polynomial estimator that allows for ordered categorical covariates. We derive the asymptotic conditional bias and variance for the local polynomial estimator with ordinal covariate, under the assumption that the categories correspond to quantiles of an unobserved continuous latent variable. We conduct a simulation study with two patterns of ordinal data to evaluate our estimator. In the multivariate case where the covariates contain a mixture of continuous, ordinal, and nominal variables, we use a Nadaraya-Watson estimator with generalized product kernel. We derive the asymptotic conditional bias and variance for the Nadaraya-Watson estimator with continuous, ordinal, and nominal covariates, under the assumption that the categories of the ordinal covariate correspond to quantiles of an unobserved continuous latent variable. We conduct a multivariate simulation study to evaluate our Nadaraya-Watson estimator with generalized product kernel.

Acknowledgements

I would like to express the deepest appreciation to my dissertation advisor, Prof. Jean Opsomer, who is an outstanding researcher, profound scholar, excellent teacher, and considerate mentor. He not only advises me academically but also inspires me by his integrity. Without his continual guidance, persistent help, and personal influence, I could not have become a independent researcher and this dissertation could not have been completed. I would also like to thank Prof. F. Jay Breidt and Prof. Mary Meyer in my committee. Their excellent research, teaching and demeanor have impacted me in a significant way for the past years with them.

Next, I'd like to proclaim my sincerest love to my mother, Benying Fan, from the bottom of my heart. My father passed away in 2004. As the only child of my parents, my mother has supported the family from then on with a strong mind, a tolerant personality, and an integrated demeanour, which have positively influenced all the time. She has always been diligent at work while maintaining a nice and clean house, cooking savory food, and supporting me through college. It is such a heavy burden for her but she definitely has done a perfect job. Her love warms my heart, cheers me up, and encourages me struggling toughly during one and one hard times. It is impossible for me to have gone so far without her decades of educating, accompanying, and tolerating.

I would also like to thank my girlfriend Lihua Wan. We came from China, met in CSU, and have been together for four years. My life is greatly happier after being with you. You have given me tons of sweet, romantic, and unforgettable memories. Your love, your care, your tolerance, your understanding, and your unwavering support have inspired me all these years.

Finally, I gratefully acknowledge the National Science Foundation and the Statistics Department for supporting me finish my program all these years. What is more, I would also like to express my gratitude to my fellow graduate students, the faculty, and the staff in the Statistics Department at Colorado State University for their help over the years.

This dissertation is typset in LaTeX using a document class designed by Leif Anderson.

Table of Contents

List of Tables

CHAPTER 1

Introduction

1.1. Variable selection for linear regression under informative sampling

We receive a large proportion of quantitative information about our economy and our community from sample surveys. National statistical agencies report estimates for items such as unemployment rate, crime rate, crop production, mortgage rate, and median family income. Some of them may come from censuses, but estimates based on a sample of the relevant population are more common. In a fixed finite population, one cannot establish any limit properties. One common approach to establish large-sample properties of sample designs and estimators is to define sequences of finite populations and associated probability samples. For simplicity, we usually assume that the Nth finite population contains N elements. Thus the set of indices for the Nth finite population is $U_N = \{1, 2, \ldots, N\}$. There are column vectors of characteristics with indices U_N for the Nth finite population. The column vectors are often called simply the Nth finite population. We can also assume them to be random vectors. For example the vector of characteristics can be the first N elements of the sequence $\{y_i\}$ of independent and identically distributed (i.i.d.) random variables such that $\mathrm{E}(y_i) = \mu$, where and $\mathrm{Var}(y_i) = \sigma^2$.

We can assume that the finite population is composed of vectors $(y_i, \boldsymbol{x}_i^T)$ that are realizations of i.i.d. random vectors with distribution function F satisfying the model

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i,$$

where ϵ_i's are independent of $\boldsymbol{x}_i$'s with mean 0 and variance σ^2. For example, in a survey concerning the house value of Fort Collins, for the ith element in the population, y_i can be the current house value of the ith family, and the covariate $\boldsymbol{x}_i = (1, x_{1i}, x_{2i})^T$ can be the intercept, the family size of the ith family, and the family income. If one is interested in the relationship between current house value with family size and family income, it will involve estimating $\boldsymbol{\beta}$ using the data in the sample from the finite population. If $\mathrm{E}(\boldsymbol{x}_i \pi_i \epsilon_i) = \mathbf{0}$ where π_i is the probability of including the ith element into the sample, an unbiased estimator of $\boldsymbol{\beta}$ can be the ordinary least squares estimator. If $\mathrm{E}(\boldsymbol{x}_i \pi_i \epsilon_i) \neq \mathbf{0}$, it is said that the design is informative for the model. In such cases it becomes necessary to incorporate the sampling weights, i.e. $1/\pi_i$, into the analysis, for example, using weighted least squares estimator. Problems will arise when the length of $\boldsymbol{x}_i$ is big, i.e. there are many covariates, and a number of elements in $\boldsymbol{\beta}$ are 0, i.e. a number of the covariates are not related with the response y_i. Then it is advisable to select a subset of covariates from $\boldsymbol{x}_i$. In the non-sampling context, the study of this topic is well developed. Many variable selection criteria are proposed and proved effective asymptotically. But when these criteria are applied at the sample stage, they may not have the same ability to select the true model effectively. This motivated us to study the variable selection criteria from the sample stage to the finite population stage and to the asymptotic stage. In Chapter 2, we assume that the finite population is generated by a multivariate linear regression model with informative design and define the relevant statistics. We examine the asymptotic properties of weighted predicted residual sums of squares (wPRESS) under weighted least squares regression estimation (WLS), unweighted PRESS (PRESS) under ordinary least squares regression estimation (OLS), PRESS under WLS, and unweighted PRESS under OLS. We provide new insights on the asymptotic ability of the four variable selection criteria to select the candidate model that reflects the true model as much as possible. We conduct a simulation study to show the properties what was.

1.2. Nonparametric regression with different types of covariates

To do inference to a given dataset with a column of response and columns of covariates, one classical way is to assume a parametric model for the underlying model that relates the covariates with the response. Then appropriate statistics can be calculated and corresponding parameters can be estimated. This is call parametric modelling. However, the strength of parametric modelling is also its weakness. By doing inference to a specified model, great gains in accuracy and stability are possible, but only if the assumed model is true or at least approximately true. If the assumed model is not close to the correct one, inferences can be useless, or even worse, can lead to misleading interpretations of the data. Nonparametric smoothing or regression provide an approach to link making no assumptions on a specific model to making very strong assumptions. For example, consider the simple linear regression model,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, \ldots, n,$$

where the ϵ_i's are i.i.d. with zero mean and variance σ^2. If this model is the true model or very close to the reality, estimates of β_0, β_1 can be calculated and one can use it to do inference and prediction. But if the linear model is not appropriate, for example if the true model is

$$y_i = \beta_0 + e^{\beta_1 x_i} + \epsilon_i,\ i = 1, \ldots, n,$$

fitting a linear model to a non-linear relationship can result in a degree of certainty that is not realistic. A more general approach is the nonparametric regression model

$$y_i = m(x_i) + \epsilon_i.$$

The function $m(x)$ is the conditional expectation $m(x) = \mathrm{E}(Y|X = x)$ with $\mathrm{E}(\epsilon_i|X = x) = 0$, and $\mathrm{Var}(\epsilon_i|X = x) = \sigma_x^2$ not necessarily constant. The covariate X is continuous. The univariate nonparametric model can be generalized to a multivariate model with continuous covariates. Estimators such as Nadaraya-Watson estimator, local polynomial estimator and spline smoothing are well studied.

Less is known about the situation that the observed covariates are not continuous, for example nominal or ordinal. We put our sight on the case that there is only one ordinal covariate in Chapter 3. We assume that y_i is the function of a latent continuous covariate plus some random error and then connect the latent continuous covariate with the ordinal covariate. We use a symmetric and continuous kernel function in our local polynomial estimator. We study the asymptotic bias and variance of this estimator and conduct simulation studies. We consider the case that there are continuous, ordinal and nominal covariates in Chapter 4. We use a Nadaraya-Watson (NW) estimator with product kernel here allowing for both categorical and continuous covariates. We derive the asymptotic properties of the estimator and conduct a simulation study.

CHAPTER 2

Variable Selection For Linear Regression Under Informative Sampling

2.1. Introduction

Multiple linear regression analysis is one of the most widely used of all statistical methods. Given a data set, the assumption underlying linear regression is that the response is a linear function of a number of covariates plus random error. However, even when this assumption is correct, it is often unknown how many covariates are in the true model, as well as which covariates are in the true model. This leads to studies of model selection in multiple linear regression. This is a well-developed area and many model selection criteria, such as predicted residual sums of squares (PRESS) (Allen 1974), Akaike's information criteria (AIC) (Akaike 1974), and Adjusted R-squared (AdjR^2) (Theil 1961), are in use today.

When the data come from a survey, a lot of the results for independent and identically distributed data do not apply, and both the finite population context and the sampling design need to be accounted for. Basically, there are two approaches in the theory of inference for finite populations. One is called design-based, which means the primary source of randomness is the probability ascribed by the sampling design to the various subsets of the finite population. The other is called model-based, which assumes that the values associated with the population units are realized outcomes of random variables. Regression estimation techniques are usually used in the design-based approach to make efficient use of auxiliary information to estimate the population parameters like the population mean. The auxiliary information is comprised of a number of covariates that are assumed to be linearly correlated with the response. We do not necessarily believe that the relation holds but just use it to for increasing the precision of survey estimators (Särndal et al. 1978).

If our interest is in the coefficient of a multiple linear regression, we can assume the finite population to be truly generated by the multiple linear regression model (Royall 1970). This approach is called model-based and can lead to, for example, the best linear unbiased estimators (Brewer 1963; Royall 1970). In this approach, since we believe the relation between the auxiliary information and the response is true but the model that generates the finite population is unknown, we need to decide which covariates in the auxiliary information of the survey should be included in the model. This brings about our interest of model selection in the sampling context.

Nascimento Silva and Skinner (1997) consider the selection of auxiliary variables in the regression estimation of finite population means under simple random sampling. Clark and Chambers (2008) develop an "adaptive calibration" approach, where the auxiliary variables to be used in weighting are selected using sample data. Wang and Wang (2011) propose a variable selection method for the additive model-assisted survey sampling based on the using Bayes information criterion (BIC) based on a comprehensive Monte Carlo study. Koralik and Opsomer (2010) (Master's Project) conducted a simulation study using prediction sum of squares (PRESS), Akaike's information criteria (AIC), and Adjusted R-squared (AdjR^2), with design weights and without design weights, as the model selection criteria in the sampling context. Their results showed that ordinary PRESS (without design weights) was the most accurate ordinary statistic in all but one simulation. The survey design-weighted PRESS (wPRESS) and ordinary PRESS worked almost as well as each other while ordinary

PRESS was slightly more accurate than wPRESS. Both the wPRESS statistic and the ordinary PRESS statistic had the smallest number of variables selected when compared to the design-weighted AIC and AdjR^2 and ordinary AIC and AdjR^2, respectively. These results were unexpected since wPRESS includes information of the sample design with unequal inclusion probability and it would be expected that wPRESS will be more accurate then the ordinary PRESS at least when the sampling was clearly informative.

This motivated our interest in studying the design-weighted and unweighted PRESS in the sampling context theoretically. In Section 2.2, we assume that the finite population is generated by a multivariate linear regression model and define the relevant statistics in this chapter. In Section 2.3, we derive the asymptotic properties of PRESS and wPRESS. We provide new insights in the conditions under which PRESS and wPRESS work as well as each other, and the conditions under which wPRESS works better than PRESS, asymptotically. In Section 2.4, we conduct a simulation study to show the properties we discuss in Section 2.3.

2.2. Problem Statement and Definitions

Let $\{(y_i, \boldsymbol{x}_i^T, \pi_i)\}$ be a sequence of independent and identically distributed (i.i.d.) random vectors of dimension $k+3$ with bounded fourth moments, where y_i is the response, $\boldsymbol{x}_i$ is the the covariate vector in a model we are interested in evaluating, and π_i is the inclusion probability. Let $\boldsymbol{x}_{t,i}$ be the covariates in $\boldsymbol{x}_i$ that are actually related to y_i, $\boldsymbol{x}_{m,i}$ be the candidate covariates in $\boldsymbol{x}_i$ that we are interested in. We further let $\mathcal{M}$ be the set of all candidate models we are interested in. The y_i are related to the $\boldsymbol{x}_{t,i}$ through the model

$$y_i = \boldsymbol{x}_{t,i}^T\boldsymbol{\beta}_t + \epsilon_i,$$

where ϵ_i's are i.i.d. random errors with mean 0 and variance σ^2.

Let $\{\mathcal{F}_N\}$, $N = k+3, k+4, \ldots$, be a sequence of finite populations, where $\mathcal{F}_N$ is composed of the first N elements of $\{(y_i, \boldsymbol{x}_i^T, \pi_i)\}$. Let U_N be the set of indices of the units that are in the Nth population, N be the size of the Nth population, s_N be the set of indices of the units that are in the sample from the Nth population, and n_N be the sample size. Let $\boldsymbol{Y}_N = (y_i)_{i \in U_N}$, $\boldsymbol{X}_{t,N} = (\boldsymbol{x}_{t,i}^T)_{i \in U_N}$, $\boldsymbol{X}_{m,N} = (\boldsymbol{x}_{m,i}^T)_{i \in U_N}$, and $\boldsymbol{X}_N = (\boldsymbol{x}_i^T)_{i \in U_N}$ be the matrices of the relevant covariates from the Nth population. Then, we can divide $\boldsymbol{X}_N$ into five parts $\boldsymbol{X}_N = (\boldsymbol{1}_N, \boldsymbol{X}_{R_1,N}, \boldsymbol{X}_{R_2,N}, \boldsymbol{X}_{R_3,N}, \boldsymbol{X}_{R_4,N})$:

$$\begin{cases} \boldsymbol{X}_{R_1,N} & \boldsymbol{X}_{R_1,N} \subseteq \boldsymbol{X}_{t,N},\ \boldsymbol{X}_{R_1,N} \subseteq \boldsymbol{X}_{m,N} \\ \boldsymbol{X}_{R_2,N} & \boldsymbol{X}_{R_2,N} \cap \boldsymbol{X}_{t,N} = \emptyset,\ \boldsymbol{X}_{R_2,N} \subseteq \boldsymbol{X}_{m,N} \\ \boldsymbol{X}_{R_3,N} & \boldsymbol{X}_{R_3,N} \subseteq \boldsymbol{X}_{t,N},\ \boldsymbol{X}_{R_3,N} \cap \boldsymbol{X}_{m,N} = \emptyset \\ \boldsymbol{X}_{R_4,N} & \boldsymbol{X}_{R_4,N} \cap \boldsymbol{X}_{t,N} = \emptyset,\ \boldsymbol{X}_{R_4,N} \cap \boldsymbol{X}_{m,N} = \emptyset \end{cases}$$

where $\boldsymbol{X}_{R_p,N}$ is $N \times k_p$ matrix and $\sum_{p=1}^{4} k_p = k$. Under this classification, we can rewrite the model as

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i,$$

where $\boldsymbol{\beta}^T = (\beta_0, \boldsymbol{\beta}_{R_1}^T, \boldsymbol{\beta}_{R_2}^T, \boldsymbol{\beta}_{R_3}^T, \boldsymbol{\beta}_{R_4}^T)^T$.

The following relationships hold: $\boldsymbol{X}_{t,N} = (\boldsymbol{1}_N, \boldsymbol{X}_{R_1,N}, \boldsymbol{X}_{R_3,N})$, $\boldsymbol{X}_{m,N} = (\boldsymbol{1}_N, \boldsymbol{X}_{R_1,N}, \boldsymbol{X}_{R_2,N})$, $\boldsymbol{\beta}_t^T = (\beta_0, \boldsymbol{\beta}_{R_1}^T, \boldsymbol{\beta}_{R_3}^T)^T$, $\boldsymbol{\beta}_m^T = (\beta_0, \boldsymbol{\beta}_{R_1}^T, \boldsymbol{\beta}_{R_2}^T)^T$, and $\boldsymbol{\beta}_{R_2} = \boldsymbol{\beta}_{R_4} = \boldsymbol{0}$.

For the sample drawn from the Nth finite population, we consider two estimators of the regression parameters: ordinary least squares (OLS) and weighted least squares (WLS). The coefficient estimators are defined as:

$$\widehat{\boldsymbol{\beta}}_{1,m} = (\boldsymbol{X}_m^T \boldsymbol{W} \boldsymbol{X}_m)^{-1} \boldsymbol{X}_m^T \boldsymbol{W} \boldsymbol{Y}$$

$$\widehat{\boldsymbol{\beta}}_{2,m} = (\boldsymbol{X}_m^T \boldsymbol{X}_m)^{-1} \boldsymbol{X}_m^T \boldsymbol{Y},$$

where

$$\boldsymbol{Y} = (y_i)_{i \in s_N},$$

$$\boldsymbol{X}_m = (\boldsymbol{x}_{m,i}^T)_{i \in s_N}$$

and

$$\boldsymbol{W} = \mathrm{diag}(\frac{1}{\pi_i})_{i \in s_N}.$$

The corresponding "hat" matrices are defined as:

$$\boldsymbol{H}_{1,m} = \boldsymbol{X}_m (\boldsymbol{X}_m^T \boldsymbol{W} \boldsymbol{X}_m)^{-1} \boldsymbol{X}_m^T \boldsymbol{W}$$

$$\boldsymbol{H}_{2,m} = \boldsymbol{X}_m (\boldsymbol{X}_m^T \boldsymbol{X}_m)^{-1} \boldsymbol{X}_m^T.$$

For a certain estimate $\widehat{\boldsymbol{\beta}}_{j,m}$, we are going to use two model selection criteria:

$$\widehat{CV}_{1,j,m} = \frac{1}{N} \sum_{s_N} \frac{1}{\pi_i} (\frac{y_i - \boldsymbol{x}_{m,i}^T \widehat{\boldsymbol{\beta}}_{j,m}}{1 - \boldsymbol{H}_{j,m}(ii)})^2$$

$$\widehat{CV}_{2,j,m} = \frac{1}{N} \sum_{s_N} (\frac{y_i - \boldsymbol{x}_{m,i}^T \widehat{\boldsymbol{\beta}}_{j,m}}{1 - \boldsymbol{H}_{j,m}(ii)})^2,$$

where $\boldsymbol{H}_{j,m}(ii)$ is ith diagonal element of $\boldsymbol{H}_{j,m}$.

On the Nth finite population level, we define

$$\boldsymbol{\beta}_{1,m,N} = (\boldsymbol{X}_{m,N}^T \boldsymbol{X}_{m,N})^{-1} \boldsymbol{X}_{m,N}^T \boldsymbol{Y}_N$$

$$\boldsymbol{\beta}_{2,m,N} = (\boldsymbol{X}_{m,N}^T \boldsymbol{\Pi}_N \boldsymbol{X}_{m,N})^{-1} \boldsymbol{X}_{m,N}^T \boldsymbol{\Pi}_N \boldsymbol{Y}_N,$$

$$\boldsymbol{H}_{1,m,N} = \boldsymbol{X}_{m,N} (\boldsymbol{X}_{m,N}^T \boldsymbol{X}_{m,N})^{-1} \boldsymbol{X}_{m,N}^T$$

$$\boldsymbol{H}_{2,m,N} = \boldsymbol{X}_{m,N} (\boldsymbol{X}_{m,N}^T \boldsymbol{\Pi}_N \boldsymbol{X}_{m,N})^{-1} \boldsymbol{X}_{m,N}^T \boldsymbol{\Pi}_N,$$

and

$$CV_{1,j,m,N} = \frac{1}{N} \sum_{U_N} (\frac{y_i - \boldsymbol{x}_{m,i}^T \boldsymbol{\beta}_{j,m,N}}{1 - \boldsymbol{H}_{j,m,N}(ii)})^2$$

$$CV_{2,j,m,N} = \frac{1}{N} \sum_{U_N} \pi_i (\frac{y_i - \boldsymbol{x}_{m,i}^T \boldsymbol{\beta}_{j,m,N}}{1 - \boldsymbol{H}_{j,m,N}(ii)})^2,$$

where $\boldsymbol{H}_{j,m,N}(ii)$ is ith diagonal element of $\boldsymbol{H}_{j,m,N}$, and

$$\boldsymbol{\Pi}_N = \mathrm{diag}\big(\pi_i\big)_{i \in U_N}.$$

For future reference, we also define $\mu_{t,i} = \boldsymbol{x}_{t,i} \boldsymbol{\beta}_t$ and $\mu_{R_3,i} = \boldsymbol{x}_{R_3,i} \boldsymbol{\beta}_{R_3}$.

Right now, notice there are four model selection options for a given candidate model at the sampling level, which are $\widehat{CV}_{1,1,m}$, $\widehat{CV}_{1,2,m}$, $\widehat{CV}_{2,1,m}$ and $\widehat{CV}_{2,2,m}$. And $CV_{1,1,m,N}$, $CV_{1,2,m,N}$, $CV_{2,1,m,N}$ and $CV_{2,2,m,N}$ are the four model selection options at the finite population level. The model selection option $\widehat{CV}_{1,1,m}$ is wPRESS applying WLS, $\widehat{CV}_{1,2,m}$ is wPRESS applying OLS, $\widehat{CV}_{2,1,m}$ is PRESS applying WLS and $\widehat{CV}_{2,2,m}$ is PRESS applying OLS. We will study the design consistency of $\widehat{CV}_{i,j,m}$ for $CV_{i,j,m,N}$, i.e. $\widehat{CV}_{i,j,m} - CV_{i,j,m,N} = o_p(1)$ (Fuller 2009, p.41), as well as the asymptotic properties of $CV_{i,j,m,N}$ in Section 2.3. We will also use simulations to evaluate the theoretical results in Section 2.4.

2.3. Theoretical Results

Let $\boldsymbol{z}_{m,i}^T = (y_i, \boldsymbol{x}_{m,i}^T)^T$, $\widehat{\boldsymbol{M}}_{1,m} = N^{-1}\boldsymbol{X}_m^T\boldsymbol{W}\boldsymbol{X}_m$, $\widehat{\boldsymbol{M}}_{2,m} = N^{-1}\boldsymbol{X}_m^T\boldsymbol{X}_m$, $\boldsymbol{M}_{1,m,N} = N^{-1}\boldsymbol{X}_{m,N}^T\boldsymbol{X}_{m,N}$, $\boldsymbol{M}_{2,m,N} = N^{-1}\boldsymbol{X}_{m,N}^T\boldsymbol{\Pi}_N\boldsymbol{X}_{m,N}$ and $\boldsymbol{M}_{3,m,N} = N^{-1}\boldsymbol{X}_{m,N}^T\boldsymbol{\Pi}_N\boldsymbol{\Pi}_N\boldsymbol{X}_{m,N}$.

To prove our theoretical results, we use a set of assumptions that are closely related to those in Fuller (2009, p.108-p.111).

A1. *The sample design is such that for any $\boldsymbol{z}_m$ with bounded fourth moments*

$$Var\{\bar{\boldsymbol{z}}_{m,HT} - \bar{\boldsymbol{z}}_{m,N}|\mathcal{F}_N\} = O_p(n_N^{-1}),$$

where

$$\bar{\boldsymbol{z}}_{m,HT} = N^{-1}\sum_{i\in s_N}\pi_i^{-1}\boldsymbol{z}_{m,i},$$

and $\bar{\boldsymbol{z}}_{m,N}$ is the finite population mean of $\boldsymbol{z}_m$.

A2. *There exist constants λ, λ_1, and λ_2 such that for all N, $0 < \lambda < \lambda_1\frac{n_N}{N} < \pi_i < \lambda_2\frac{n_N}{N}$, $\forall i \in U_N$, and $\frac{n_N}{N} \not\to 0$ as $N \to \infty$.*

A3. *$(\boldsymbol{x}_i^T, \pi_i, \epsilon_i)$ are i.i.d. random vectors having uniformly bounded fourth moments, $E(\epsilon_i|\boldsymbol{x}_i) = 0$, $Var(\epsilon_i|\boldsymbol{x}_i) < \infty$ for all i.*

A4. *The matrices $(\boldsymbol{M}_{1,m,N}, \boldsymbol{M}_{2,m,N}, , \boldsymbol{M}_{3,m,N})$ satisfy*

$$\lim_{N\to\infty}(\boldsymbol{M}_{1,m,N}, \boldsymbol{M}_{2,m,N}, \boldsymbol{M}_{3,m,N}) = (\boldsymbol{M}_{1,m}, \boldsymbol{M}_{2,m}, \boldsymbol{M}_{3,m})\, a.s. \quad \forall m \in \mathcal{M},$$

where $\boldsymbol{M}_{1,m}$, $\boldsymbol{M}_{2,m}$, and $\boldsymbol{M}_{3,m}$ are positive definite and $\mathcal{M}$ is the set that consists of all subset of covariates in $\boldsymbol{x}_i$ that we are interested in evaluating.

A5. *The sequences* $\{(\widehat{\boldsymbol{M}}_{1,m}, \widehat{\boldsymbol{M}}_{2,m})\}$ *and* $\{(\boldsymbol{M}_{1,m,N}, \boldsymbol{M}_{2,m,N})\}$ *satisfy*

$$(\widehat{\boldsymbol{M}}_{1,m}, \widehat{\boldsymbol{M}}_{2,m}) - (\boldsymbol{M}_{1,m,N}, \boldsymbol{M}_{2,m,N})|\mathcal{F}_N = O_p(n_N^{-1/2})$$

element-wise, and $\widehat{\boldsymbol{M}}_{1,m}$ *and* $\widehat{\boldsymbol{M}}_{2,m}$ *are positive definite* $\forall m \in \mathcal{M}$.

A6. *There exist constants* c, N_0 *such that for all* $N > N_0$, $|H_{k,m}(ij)| < \frac{c}{n_N}$, $|H_{k,m,N}(ij)| < \frac{c}{N}$, $\forall i \in U_N$, $k = 1, 2$.

Theorem 2.3.1. *Suppose A1-A6 hold. Then*

$$\widehat{CV}_{l,j,m} - CV_{l,j,m,N}|\mathcal{F}_N = O_p(n_N^{-1/2}) \quad l = 1, 2,\ j = 1, 2,\ \forall N > N_0, \forall m \in \mathcal{M}.$$

Proof. By A5

$$\widehat{\boldsymbol{\beta}}_{j,m} - \boldsymbol{\beta}_{j,m,N}|\mathcal{F}_N = O_p(n_N^{-1/2}) \tag{1}$$

holds, which means $\widehat{\boldsymbol{\beta}}_{j,m}$ is design consistent with $\boldsymbol{\beta}_{j,m,N}$. Since

$$\begin{aligned}
\widehat{CV}_{l,j,m} - CV_{l,j,m,N} &= \frac{1}{N}\sum_{s_N}\frac{1}{\pi_i}\big(\frac{y_i - \boldsymbol{x}_{m,i}^T\widehat{\boldsymbol{\beta}}_{j,m}}{1 - \boldsymbol{H}_{j,m}(ii)}\big)^2 - \frac{1}{N}\sum_{U_N}\big(\frac{y_i - \boldsymbol{x}_{m,i}^T\boldsymbol{\beta}_{j,m,N}}{1 - \boldsymbol{H}_{j,m,N}(ii)}\big)^2 \\
&= \frac{1}{N}\sum_{s_N}\frac{1}{\pi_i}\big(y_i - \boldsymbol{x}_{m,i}^T\widehat{\boldsymbol{\beta}}_{j,m} - \boldsymbol{H}_{j,m}(ii)\frac{y_i - \boldsymbol{x}_{m,i}^T\widehat{\boldsymbol{\beta}}_{j,m}}{1 - \boldsymbol{H}_{j,m}(ii)}\big)^2 \\
&\quad - \frac{1}{N}\sum_{U_N}\big(y_i - \boldsymbol{x}_{m,i}^T\boldsymbol{\beta}_{j,m,N} - \boldsymbol{H}_{j,m,N}(ii)\frac{y_i - \boldsymbol{x}_{m,i}^T\boldsymbol{\beta}_{j,m,N}}{1 - \boldsymbol{H}_{j,m,N}(ii)}\big)^2
\end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{1}{N}\sum_{s_N}\frac{1}{\pi_i}(y_i - \boldsymbol{x}_{m,i}^T\widehat{\boldsymbol{\beta}}_{j,m})^2 - \frac{1}{N}\sum_{U_N}(y_i - \boldsymbol{x}_{m,i}^T\boldsymbol{\beta}_{j,m,N})^2\right) \\
&\quad - \left(\frac{1}{N}\sum_{s_N}2\boldsymbol{H}_{j,m}(ii)\frac{(y_i - \boldsymbol{x}_{m,i}^T\widehat{\boldsymbol{\beta}}_{j,m})^2}{1-\boldsymbol{H}_{j,m}(ii)}\right. \\
&\quad \left. - \frac{1}{N}\sum_{U_N}2\boldsymbol{H}_{j,m,N}(ii)\frac{(y_i - \boldsymbol{x}_{m,i}^T\boldsymbol{\beta}_{j,m,N})^2}{1-\boldsymbol{H}_{j,m,N}(ii)}\right) \\
&\quad + \left(\frac{1}{N}\sum_{s_N}\boldsymbol{H}_{j,m}^2(ii)\frac{(y_i - \boldsymbol{x}_{m,i}^T\widehat{\boldsymbol{\beta}}_{j,m})^2}{(1-\boldsymbol{H}_{j,m}(ii))^2}\right. \\
&\quad \left. - \frac{1}{N}\sum_{U_N}\boldsymbol{H}_{j,m,N}^2(ii)\frac{(y_i - \boldsymbol{x}_{m,i}^T\boldsymbol{\beta}_{j,m,N})^2}{(1-\boldsymbol{H}_{j,m,N}(ii))^2}\right) \\
&= D_1 - D_2 + D_3.
\end{aligned}
$$

For the leading term D_1,

$$
\begin{aligned}
\frac{1}{N}\sum_{s_N}\frac{1}{\pi_i}(y_i - \boldsymbol{x}_{m,i}^T\widehat{\boldsymbol{\beta}}_{j,m})^2 - \frac{1}{N}\sum_{U_N}(y_i - \boldsymbol{x}_{m,i}^T\boldsymbol{\beta}_{j,m,N})^2 &= \Big(\frac{1}{N}\sum_{s_N}\frac{1}{\pi_i}y_i^2 - \frac{1}{N}\sum_{U_N}y_i^2\Big) \\
&\quad - 2\Big(\frac{1}{N}\sum_{s_N}\frac{1}{\pi_i}y_i\boldsymbol{x}_{m,i}^T\widehat{\boldsymbol{\beta}}_{j,m} - \frac{1}{N}\sum_{U_N}y_i\boldsymbol{x}_{m,i}^T\boldsymbol{\beta}_{j,m,N}\Big) \\
&\quad + \Big(\frac{1}{N}\sum_{s_N}\frac{1}{\pi_i}(\boldsymbol{x}_{m,i}^T\widehat{\boldsymbol{\beta}}_{j,m})^2 - \frac{1}{N}\sum_{U_N}(\boldsymbol{x}_{m,i}^T\boldsymbol{\beta}_{j,m,N})^2\Big) \\
&= \Big(\frac{1}{N}\sum_{s_N}\frac{1}{\pi_i}y_i^2 - \frac{1}{N}\sum_{U_N}y_i^2\Big)
\end{aligned}
$$

$$
\begin{aligned}
&- 2\big(\frac{1}{N}\sum_{s_N}\frac{1}{\pi_i}y_i\boldsymbol{x}_{m,i}^T(\widehat{\boldsymbol{\beta}}_{j,m}-\boldsymbol{\beta}_{j,m,N})\\
&+ (\frac{1}{N}\sum_{s_N}\frac{1}{\pi_i}y_i\boldsymbol{x}_{m,i}^T-\frac{1}{N}\sum_{U_N}y_i\boldsymbol{x}_{m,i}^T)\boldsymbol{\beta}_{j,m,N}\big)\\
&+ \Bigg(\frac{1}{N}\sum_{s_N}\frac{1}{\pi_i}\big(\boldsymbol{x}_{m,i}^T(\widehat{\boldsymbol{\beta}}_{j,m}-\boldsymbol{\beta}_{j,m,N})\big)^2\\
&- 2\big(\frac{1}{N}\sum_{s_N}\frac{1}{\pi_i}\boldsymbol{x}_{m,i}^T(\widehat{\boldsymbol{\beta}}_{j,m}-\boldsymbol{\beta}_{j,m,N})\big)\\
&+ \big((\frac{1}{N}\sum_{s_N}\frac{1}{\pi_i}\boldsymbol{x}_{m,i}^T-\frac{1}{N}\sum_{U_N}\boldsymbol{x}_{m,i}^T)\boldsymbol{\beta}_{j,m,N}\big)^2\Bigg)\\
=\;& O_p(n_N^{-1/2})
\end{aligned}
$$

by (1), A1, and A3.

In the second part D_2,

$$
\begin{aligned}
\left|\frac{1}{N}\sum_{s_N}\frac{1}{\pi_i}\boldsymbol{H}_{j,m}(ii)\frac{(y_i-\boldsymbol{x}_{m,i}^T\widehat{\boldsymbol{\beta}}_{j,m})^2}{1-\boldsymbol{H}_{j,m}(ii)}\right| &\leq \frac{1}{N}\sum_{s_N}\lambda_1\frac{N}{n_N}\frac{c}{n_N}\frac{(y_i-\boldsymbol{x}_{m,i}^T\widehat{\boldsymbol{\beta}}_{j,m})^2}{1+\frac{c}{n_N}}\\
&= \frac{\lambda_1 c}{n_N+c}\frac{1}{n_N}\sum_{s_N}(y_i-\boldsymbol{x}_{m,i}^T\widehat{\boldsymbol{\beta}}_{j,m})^2\\
&\leq \frac{2\lambda_1 c}{n_N+c}\frac{1}{n_N}\sum_{s_N}\big(y_i^2+(\boldsymbol{x}_{m,i}^T\widehat{\boldsymbol{\beta}}_{j,m})^2\big)\\
&= \frac{2\lambda_1 c}{n_N+c}\frac{1}{n_N}\sum_{s_N}\big(y_i^2+(\sum_{s_N}\boldsymbol{H}_{j,m}(ik)y_k)^2\big)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{2\lambda_1 c}{n_N + c} \frac{1}{n_N} \sum_{s_N} \left(y_i^2 + (\sum_{s_N} \frac{c}{n_N} y_k)^2 \right) \\
&\leq \frac{2\lambda_1 c}{n_N + c} \frac{1}{n_N} \sum_{s_N} \left(y_i^2 + \frac{c^2}{n_N} \sum_{s_N} y_k^2 \right) \\
&\leq \frac{2\lambda_1 c(1+c^2)}{n_N + c} \frac{1}{n_N} \sum_{s_N} y_i^2 \\
&= O_p(n_N^{-1})
\end{aligned}
$$

by A1 and A6.

Similarly,

$$\frac{1}{N} \sum_{U_N} \boldsymbol{H}_{j,m,N}(ii) \frac{(y_i - \boldsymbol{x}_{m,i}^T \boldsymbol{\beta}_{j,m,N})^2}{1 - \boldsymbol{H}_{j,m,N}(ii)} = O_p(N^{-1}),$$

$$\frac{1}{N} \sum_{s_N} \frac{1}{\pi_i} \boldsymbol{H}_{j,m}^2(ii) \frac{(y_i - \boldsymbol{x}_{m,i}^T \widehat{\boldsymbol{\beta}}_{j,m})^2}{(1 - \boldsymbol{H}_{j,m}(ii))^2} = O_p(n_N^{-2}),$$

and

$$\frac{1}{N} \sum_{U_N} \boldsymbol{H}_{j,m,N}^2(ii) \frac{(y_i - \boldsymbol{x}_{m,i}^T \boldsymbol{\beta}_{j,m,N})^2}{(1 - \boldsymbol{H}_{j,m,N}(ii))^2} = O_p(N^{-2})$$

in D_3.

Therefore,

$$\widehat{CV}_{1,j,m} - CV_{1,j,m,N} | \mathcal{F}_N = O_p(n_N^{-1/2}) \quad j = 1, 2, \ \forall m \in \mathcal{M}$$

follows.

In the same way, we can prove that

$$\widehat{CV}_{2,j,m} - CV_{2,j,m,N}|\mathcal{F}_N = O_p(n_N^{-1/2}) \quad j = 1, 2, , \ \forall N > N_0, \ \forall m \in \mathcal{M}.$$

□

We use the following theorem to show the asymptotic property of $CV_{k,j,m,N}\, j = 1, 2,\ k = 1, 2$, which are the weighted and unweighted PRESS.

Theorem 2.3.2. *Suppose A1-A6 hold. Then*

$$CV_{1,1,m,N} \xrightarrow{P} E(\epsilon_i^2) + E(\mu_{R_3,i}^2) - (E(\boldsymbol{x}_{m,i}\mu_{R_3,i}))^T \boldsymbol{M}_{1,m} \left(E(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right),$$

$$\begin{aligned} CV_{1,2,m,N} \xrightarrow{P} & E(\epsilon_i^2) + E(\mu_{R_3,i}^2) - (E(\boldsymbol{x}_{m,i}\mu_{R_3,i}))^T \boldsymbol{M}_{2,m} \left(E(\pi_i\boldsymbol{x}_{m,i}\mu_{R_3,i})\right) \\ & - 2\left(E(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T \boldsymbol{M}_{2,m} \left(E\left(\pi_i\boldsymbol{x}_{m,i}\epsilon_i\right)\right), \end{aligned}$$

In this way, we've proved that

$$\begin{aligned} CV_{2,1,m,N} \xrightarrow{P} & E(\pi_i\epsilon_i^2) + E(\pi_i\mu_{R_3,i}^2) - (E(\boldsymbol{x}_{m,i}\mu_{R_3,i}))^T \boldsymbol{M}_{1,m}\boldsymbol{M}_{2,m}\boldsymbol{M}_{1,m} \left(E(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right) \\ & + 2E(\pi_i\mu_{R_3,i}\epsilon_i), \end{aligned}$$

and

$$\begin{aligned} CV_{2,2,m,N} \xrightarrow{P} & E(\pi_i\epsilon_i^2) + E(\pi_i\mu_{R_3,i}^2) - (E(\boldsymbol{x}_{m,i}\mu_{R_3,i}))^T \boldsymbol{M}_{2,m}\boldsymbol{M}_{3,m}\boldsymbol{M}_{2,m} \left(E(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right) \\ & + 2E(\pi_i\mu_{R_3,i}\epsilon_i) - 2\left(E(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T \boldsymbol{M}_{3,m} \left(E\left(\pi_i\boldsymbol{x}_{m,i}\epsilon_i\right)\right). \end{aligned}$$

Proof. From the proof of Theorem 2.3.1, it is straightforward that

$$CV_{1,j,m,N} = \frac{1}{N}\sum_{U_N}(y_i - \boldsymbol{x}_{m,i}^T\boldsymbol{\beta}_{j,m,N})^2 + O_p(N^{-1}).$$

For term i in the leading summation,

$$\begin{aligned}(y_i - \boldsymbol{x}_{m,i}^T\boldsymbol{\beta}_{j,m,N})^2 &= \big(\boldsymbol{x}_{R_1,i}^T\boldsymbol{\beta}_{R_1} + \boldsymbol{x}_{R_3,i}^T\boldsymbol{\beta}_{R_3} + \epsilon_i \\ &\quad - \boldsymbol{H}_{j,m,N}^T(i,)\left(\boldsymbol{X}_{R_1,N}\boldsymbol{\beta}_{R_1} + \boldsymbol{X}_{R_3,N}\boldsymbol{\beta}_{R_3} + \boldsymbol{\epsilon}_N\right)\big)^2 \\ &= \Big(\boldsymbol{x}_{R_1,i}^T\boldsymbol{\beta}_{R_1} + \boldsymbol{x}_{R_3,i}^T\boldsymbol{\beta}_{R_3} + \epsilon_i - \boldsymbol{x}_{R_1,i}^T\boldsymbol{\beta}_{R_1} - \boldsymbol{x}_{m,i}^T\left(\boldsymbol{X}_{m,N}^T\boldsymbol{X}_{m,N}\right)^{-1}\boldsymbol{X}_{m,N}^T\boldsymbol{\epsilon}_N\Big)^2 \\ &= \left(\boldsymbol{x}_{R_3,i}^T\boldsymbol{\beta}_{R_3} + \epsilon_i - \boldsymbol{H}_{j,m,N}^T(i,)\boldsymbol{X}_{R_3,N}\boldsymbol{\beta}_{R_3} - \boldsymbol{H}_{j,m,N}^T(i,)\boldsymbol{\epsilon}_N\right)^2,\end{aligned}$$

where $\boldsymbol{H}_{j,m,N}^T(i,)$ is row i in $\boldsymbol{H}_{j,m,N}$. Then

$$\begin{aligned}\frac{1}{N}\sum_{U_N}(y_i - \boldsymbol{x}_{m,i}^T\boldsymbol{\beta}_{1,m,N})^2 = \frac{1}{N}(&\boldsymbol{\mu}_{R_3}^T\boldsymbol{\mu}_{R_3} + \boldsymbol{\epsilon}_N^T\boldsymbol{\epsilon}_N \\ &+ 2\boldsymbol{\mu}_{R_3}^T\boldsymbol{\epsilon}_N - 2\boldsymbol{\mu}_{R_3}^T\boldsymbol{H}_{j,m,N}\boldsymbol{\epsilon}_N \\ &- \boldsymbol{\mu}_{R_3}^T\boldsymbol{H}_{j,m,N}\boldsymbol{\mu}_{R_3} - \boldsymbol{\epsilon}_N^T\boldsymbol{H}_{j,m,N}\boldsymbol{\epsilon}_N).\end{aligned}$$

By A3

$$\frac{1}{N}(\boldsymbol{\mu}_{R_3}^T\boldsymbol{\mu}_{R_3}) \xrightarrow{\text{P}} \text{E}(\mu_{R_3,i}^2),$$

$$\frac{1}{N}(\boldsymbol{\epsilon}_N^T\boldsymbol{\epsilon}_N) \xrightarrow{\text{P}} \text{E}(\epsilon_i^2),$$

and

$$\frac{1}{N}(2\boldsymbol{\mu}_{R_3}^T\boldsymbol{\epsilon}_N) \xrightarrow{\text{P}} 0.$$

Next, since

$$\begin{aligned}\frac{1}{N}\boldsymbol{\epsilon}_N^T\boldsymbol{H}_{j,m,N}\boldsymbol{\epsilon}_N &= \frac{1}{N}\boldsymbol{\epsilon}_N^T\boldsymbol{X}_{m,N}\left(\boldsymbol{X}_{m,N}^T\boldsymbol{X}_{m,N}\right)^{-1}\boldsymbol{X}_{m,N}^T\boldsymbol{\epsilon}_N \\ &= \frac{1}{N}\boldsymbol{\epsilon}_N^T\boldsymbol{X}_{m,N}\left(\frac{1}{N}\boldsymbol{X}_{m,N}^T\boldsymbol{X}_{m,N}\right)^{-1}\frac{1}{N}\boldsymbol{X}_{m,N}^T\boldsymbol{\epsilon}_N\end{aligned}$$

when $j = 1$ and

$$\frac{1}{N}\boldsymbol{\epsilon}_N^T\boldsymbol{X}_{m,N}\left(\frac{1}{N}\boldsymbol{X}_{m,N}^T\boldsymbol{\Pi}\boldsymbol{X}_{m,N}\right)^{-1}\frac{1}{N}\boldsymbol{X}_{m,N}^T\boldsymbol{\Pi}\boldsymbol{\epsilon}_N$$

when $j = 2$, is a product of three matrices of means, each of which has a probability limit. The middle one converges to a constant, the last one converges to a constant (since we do not know that $\boldsymbol{X}_{m,N}^T\boldsymbol{\Pi}\boldsymbol{\epsilon}_N$ has mean zero because of the correlation between $\boldsymbol{\Pi}$ and $\boldsymbol{\epsilon}_N$), but the first one converges to 0 since we assume that $\mathrm{E}(\epsilon_i|\boldsymbol{x}_i) = 0$ and $\mathrm{Var}(\epsilon_i|\boldsymbol{x}_i) < \infty$. Hence, the whole expression converges to 0 in probability. Following the same idea, it is straightforward to show that

$$\boldsymbol{\mu}_{R_3}^T\boldsymbol{H}_{1,m,N}\boldsymbol{\mu}_{R_3} \xrightarrow{\mathrm{P}} \left(\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T\boldsymbol{M}_{1,m}\left(\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right),$$

$$\boldsymbol{\mu}_{R_3}^T\boldsymbol{H}_{2,m,N}\boldsymbol{\mu}_{R_3} \xrightarrow{\mathrm{P}} \left(\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T\boldsymbol{M}_{2,m}\left(\mathrm{E}(\pi_i\boldsymbol{x}_{m,i}\mu_{R_3,i})\right),$$

$$\boldsymbol{\mu}_{R_3}^T\boldsymbol{H}_{1,m,N}\boldsymbol{\epsilon}_N \xrightarrow{\mathrm{P}} 0,$$

and

$$\boldsymbol{\mu}_{R_3}^T\boldsymbol{H}_{2,m,N}\boldsymbol{\epsilon}_N \xrightarrow{\mathrm{P}} \left(\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T\boldsymbol{M}_{2,m}\left(\mathrm{E}\left(\pi_i\boldsymbol{x}_{m,i}\epsilon_i\right)\right).$$

In this way, we've proved that

$$CV_{1,1,m,N} \xrightarrow{\text{P}} \mathrm{E}(\epsilon_i^2) + \mathrm{E}(\mu_{R_3,i}^2) - \left(\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T \boldsymbol{M}_{1,m} \left(\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right),$$

and

$$CV_{1,2,m,N} \xrightarrow{\text{P}} \mathrm{E}(\epsilon_i^2) + \mathrm{E}(\mu_{R_3,i}^2) - \left(\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T \boldsymbol{M}_{2,m} \left(\mathrm{E}(\pi_i\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)$$

$$- 2\left(\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T \boldsymbol{M}_{2,m} \left(\mathrm{E}\left(\pi_i\boldsymbol{x}_{m,i}\epsilon_i\right)\right).$$

Similarly, it holds that

$$CV_{2,j,m,N} = \frac{1}{N}\sum_{U_N} \pi_i(y_i - \boldsymbol{x}_{m,i}^T\boldsymbol{\beta}_{j,m,N})^2 + O_p(N^{-1}),$$

and

$$\frac{1}{N}\sum_{U_N} \pi_i(y_i - \boldsymbol{x}_{m,i}^T\boldsymbol{\beta}_{j,m,N})^2 = \frac{1}{N}(\boldsymbol{\mu}_{R_3}^T\boldsymbol{\Pi}\boldsymbol{\mu}_{R_3} + \boldsymbol{\epsilon}_N^T\boldsymbol{\Pi}\boldsymbol{\epsilon}_N$$

$$+ 2\boldsymbol{\mu}_{R_3}^T\boldsymbol{\Pi}\boldsymbol{\epsilon}_N - 2\boldsymbol{\mu}_{R_3}^T\boldsymbol{H}_{j,m,N}\boldsymbol{\Pi}\boldsymbol{H}_{j,m,N}\boldsymbol{\epsilon}_N$$

$$- \boldsymbol{\mu}_{R_3}^T\boldsymbol{H}_{j,m,N}\boldsymbol{\Pi}\boldsymbol{H}_{j,m,N}\boldsymbol{\mu}_{R_3} - \boldsymbol{\epsilon}_N^T\boldsymbol{H}_{j,m,N}\boldsymbol{\Pi}\boldsymbol{H}_{j,m,N}\boldsymbol{\epsilon}_N).$$

By A3

$$\frac{1}{N}(\boldsymbol{\mu}_{R_3}^T\boldsymbol{\Pi}\boldsymbol{\mu}_{R_3}) \xrightarrow{\text{P}} \mathrm{E}(\pi_i\mu_{R_3,i}^2),$$

$$\frac{1}{N}(\boldsymbol{\epsilon}_N^T\boldsymbol{\Pi}\boldsymbol{\epsilon}_N) \xrightarrow{\text{P}} \mathrm{E}(\pi_i\epsilon_i^2),$$

and

$$\frac{1}{N}(2\boldsymbol{\mu}_{R_3}^T\boldsymbol{\Pi}\boldsymbol{\epsilon}_N) \xrightarrow{\text{P}} 2\text{E}(\pi_i\mu_{R_3,i}\epsilon_i).$$

Next, since

$$\begin{aligned}\frac{1}{N}\boldsymbol{\epsilon}_N^T\boldsymbol{H}_{j,m,N}\boldsymbol{\Pi}\boldsymbol{H}_{j,m,N}\boldsymbol{\epsilon}_N &= \frac{1}{N}\boldsymbol{\epsilon}_N^T\boldsymbol{X}_{m,N}\left(\boldsymbol{X}_{m,N}^T\boldsymbol{X}_{m,N}\right)^{-1}\boldsymbol{X}_{m,N}^T\boldsymbol{\Pi}\boldsymbol{X}_{m,N}\\ &\quad\times\left(\boldsymbol{X}_{m,N}^T\boldsymbol{X}_{m,N}\right)^{-1}\boldsymbol{X}_{m,N}^T\boldsymbol{\Pi}\boldsymbol{\epsilon}_N\\ &= \frac{1}{N}\boldsymbol{\epsilon}_N^T\boldsymbol{X}_{m,N}\left(\frac{1}{N}\boldsymbol{X}_{m,N}^T\boldsymbol{X}_{m,N}\right)^{-1}\left(\frac{1}{N}\boldsymbol{X}_{m,N}^T\boldsymbol{\Pi}\boldsymbol{X}_{m,N}\right)\\ &\quad\times\left(\frac{1}{N}\boldsymbol{X}_{m,N}^T\boldsymbol{X}_{m,N}\right)^{-1}\frac{1}{N}\boldsymbol{X}_{m,N}^T\boldsymbol{\epsilon}_N\end{aligned}$$

when $j = 1$ and

$$\begin{aligned}&\frac{1}{N}\boldsymbol{\epsilon}_N^T\boldsymbol{X}_{m,N}\left(\frac{1}{N}\boldsymbol{X}_{m,N}^T\boldsymbol{\Pi}\boldsymbol{X}_{m,N}\right)^{-1}\left(\frac{1}{N}\boldsymbol{X}_{m,N}^T\boldsymbol{\Pi}\boldsymbol{\Pi}\boldsymbol{X}_{m,N}\right)\\ &\times\left(\frac{1}{N}\boldsymbol{X}_{m,N}^T\boldsymbol{\Pi}\boldsymbol{X}_{m,N}\right)^{-1}\frac{1}{N}\boldsymbol{X}_{m,N}^T\boldsymbol{\Pi}\boldsymbol{\epsilon}_N\end{aligned}$$

when $j = 2$, the whole expression converges to 0 in probability for the same reason as we dealt with $CV_{1,j,m,N}$ above.

Following the same idea, it is straightforward to show that

$$\boldsymbol{\mu}_{R_3}^T\boldsymbol{H}_{1,m,N}\boldsymbol{\Pi}\boldsymbol{H}_{1,m,N}\boldsymbol{\mu}_{R_3} \xrightarrow{\text{P}} \left(\text{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T\boldsymbol{M}_{1,m}\boldsymbol{M}_{2,m}\boldsymbol{M}_{1,m}\left(\text{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right),$$

$$\boldsymbol{\mu}_{R_3}^T\boldsymbol{H}_{2,m,N}\boldsymbol{\Pi}\boldsymbol{H}_{2,m,N}\boldsymbol{\mu}_{R_3} \xrightarrow{\text{P}} \left(\text{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T\boldsymbol{M}_{2,m}\boldsymbol{M}_{3,m}\boldsymbol{M}_{2,m}\left(\text{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right),$$

$$\boldsymbol{\mu}_{R_3}^T \boldsymbol{H}_{1,m,N} \boldsymbol{\Pi} \boldsymbol{H}_{1,m,N} \boldsymbol{\epsilon}_N \xrightarrow{\text{P}} \left(\text{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T \boldsymbol{M}_{2,m} \left(E(\boldsymbol{x}_{m,i}\epsilon_i)\right) = 0,$$

and

$$\boldsymbol{\mu}_{R_3}^T \boldsymbol{H}_{2,m,N} \boldsymbol{\Pi} \boldsymbol{H}_{2,m,N} \boldsymbol{\epsilon}_N \xrightarrow{\text{P}} \left(\text{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T \boldsymbol{M}_{3,m} \left(\text{E}\left(\pi_i \boldsymbol{x}_{m,i}\epsilon_i\right)\right).$$

In this way, we've proved that

$$\begin{aligned} CV_{2,1,m,N} &\xrightarrow{\text{P}} \text{E}(\pi_i \epsilon_i^2) + \text{E}(\pi_i \mu_{R_3,i}^2) - \left(\text{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T \boldsymbol{M}_{1,m} \boldsymbol{M}_{2,m} \boldsymbol{M}_{1,m} \left(\text{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right) \\ &\quad + 2\text{E}(\pi_i \mu_{R_3,i}\epsilon_i), \end{aligned}$$

and

$$\begin{aligned} CV_{2,2,m,N} &\xrightarrow{\text{P}} \text{E}(\pi_i \epsilon_i^2) + \text{E}(\pi_i \mu_{R_3,i}^2) - \left(\text{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T \boldsymbol{M}_{2,m} \boldsymbol{M}_{3,m} \boldsymbol{M}_{2,m} \left(\text{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right) \\ &\quad + 2\text{E}(\pi_i \mu_{R_3,i}\epsilon_i) - 2\left(\text{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T \boldsymbol{M}_{3,m} \left(\text{E}\left(\pi_i \boldsymbol{x}_{m,i}\epsilon_i\right)\right). \end{aligned}$$

□

By combining Theorem 2.3.1 and Theorem 2.3.2, it is straightward to claim that

$$\begin{aligned} \widehat{CV}_{1,1,m} &\xrightarrow{\text{P}} \text{E}(\epsilon_i^2) + \text{E}(\mu_{R_3,i}^2) - \left(\text{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T \boldsymbol{M}_{1,m} \left(\text{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right) \\ &= T_1 + T_{1,1,m}^{(1)}, \end{aligned}$$

$$
\begin{aligned}
\widehat{CV}_{1,2,m} &\xrightarrow{\mathrm{P}} \mathrm{E}(\epsilon_i^2) + \mathrm{E}(\mu_{R_3,i}^2) - (\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i}))^T \boldsymbol{M}_{2,m} (\mathrm{E}(\pi_i \boldsymbol{x}_{m,i}\mu_{R_3,i})) \\
&\quad - 2 (\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i}))^T \boldsymbol{M}_{2,m} (\mathrm{E}(\pi_i \boldsymbol{x}_{m,i}\epsilon_i)) \\
&= T_1 + T_{1,2,m}^{(1)} + T_{1,2,m}^{(2)},
\end{aligned}
$$

$$
\begin{aligned}
\widehat{CV}_{2,1,m} &\xrightarrow{\mathrm{P}} \mathrm{E}(\pi_i\epsilon_i^2) + \mathrm{E}(\pi_i\mu_{R_3,i}^2) - (\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i}))^T \boldsymbol{M}_{1,m}\boldsymbol{M}_{2,m}\boldsymbol{M}_{1,m} (\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})) \\
&\quad + 2\mathrm{E}(\pi_i\mu_{R_3,i}\epsilon_i) \\
&= T_2 + T_{2,1,m}^{(1)} + T_{2,1,m}^{(2)},
\end{aligned}
$$

and

$$
\begin{aligned}
\widehat{CV}_{2,2,m} &\xrightarrow{\mathrm{P}} \mathrm{E}(\pi_i\epsilon_i^2) + \mathrm{E}(\pi_i\mu_{R_3,i}^2) - (\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i}))^T \boldsymbol{M}_{2,m}\boldsymbol{M}_{3,m}\boldsymbol{M}_{2,m} (\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})) \\
&\quad + 2\mathrm{E}(\pi_i\mu_{R_3,i}\epsilon_i) - 2 (\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i}))^T \boldsymbol{M}_{3,m} (\mathrm{E}(\pi_i \boldsymbol{x}_{m,i}\epsilon_i)) \\
&= T_2 + T_{2,2,m}^{(1)} + T_{2,2,m}^{(2)},
\end{aligned}
$$

where

$$T_1 = \mathrm{E}(\epsilon_i^2),$$

$$T_2 = \mathrm{E}(\pi_i\epsilon_i^2),$$

$$T^{(1)}_{1,1,m} = \mathrm{E}(\mu^2_{R_3,i}) - \left(\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T \boldsymbol{M}_{1,m} \left(\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right),$$

$$T^{(1)}_{1,2,m} = \mathrm{E}(\mu^2_{R_3,i}) - \left(\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T \boldsymbol{M}_{2,m} \left(\mathrm{E}(\pi_i\boldsymbol{x}_{m,i}\mu_{R_3,i})\right),$$

$$T^{(2)}_{1,2,m} = -2\left(\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T \boldsymbol{M}_{2,m} \left(\mathrm{E}\left(\pi_i\boldsymbol{x}_{m,i}\epsilon_i\right)\right),$$

$$T^{(1)}_{2,1,m} = \mathrm{E}(\pi_i\mu^2_{R_3,i}) - \left(\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T \boldsymbol{M}_{1,m}\boldsymbol{M}_{2,m}\boldsymbol{M}_{1,m} \left(\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right),$$

$$T^{(2)}_{2,1,m} = 2\mathrm{E}(\pi_i\mu_{R_3,i}\epsilon_i),$$

$$T^{(1)}_{2,2,m} = \mathrm{E}(\pi_i\mu^2_{R_3,i}) - \left(\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T \boldsymbol{M}_{2,m}\boldsymbol{M}_{3,m}\boldsymbol{M}_{2,m} \left(\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right),$$

and

$$T^{(2)}_{2,2,m} = 2\mathrm{E}(\pi_i\mu_{R_3,i}\epsilon_i) - 2\left(\mathrm{E}(\boldsymbol{x}_{m,i}\mu_{R_3,i})\right)^T \boldsymbol{M}_{3,m} \left(\mathrm{E}\left(\pi_i\boldsymbol{x}_{m,i}\epsilon_i\right)\right).$$

Note that since we use $CV_{l,j,m,N}$ to connect $\widehat{CV}_{l,j,m}$ to their asymptotic limits, the "P" is with respect to both the sampling design and the regression model, considered jointly.

If the sampling is uninformative, $T^{(2)}_{1,2,m}$, $T^{(2)}_{2,1,m}$ and $T^{(2)}_{2,2,m}$ are 0. Terms $T^{(1)}_{1,2,m}$, $T^{(1)}_{2,1,m}$ and $T^{(1)}_{2,2,m}$ reflect how much $\boldsymbol{X}_{R_3}$ is explained by $\boldsymbol{X}_m$. It is smaller when $\boldsymbol{X}_{R_3}$ is explained more, resulting in smaller criteria. On the one hand, if $\boldsymbol{X}_{R_3} \neq \emptyset$ but $\boldsymbol{X}_{R_3}$ is fully explained by $\boldsymbol{X}_m$, $T^{(1)}_{1,2,m}$, $T^{(1)}_{2,1,m}$ and $T^{(1)}_{2,2,m}$ are 0 and the four criteria are minimized. One the other hand, the asymptotic limits of the four criteria are also minimized when $\boldsymbol{X}_m = \boldsymbol{X}_t$ or as long as

$\boldsymbol{X}_t \subset \boldsymbol{X}_m$. Therefore, the four criteria will tend to select the candidate model that reflects the true model as much as possible and their ability to select such model is approximately the same.

If the sampling is informative, $T^{(2)}_{1,2,m}$, $T^{(2)}_{2,1,m}$ and $T^{(2)}_{2,2,m}$ are not 0. For $\widehat{CV}_{1,1,m}$, note that the asymptotic limit of $\widehat{CV}_{1,1,m}$ is minimized when $\boldsymbol{X}_t \subset \boldsymbol{X}_m$ or $\boldsymbol{X}_{R_3}$ is fully explained by $\boldsymbol{X}_m$, and it will converge to T_1 in probability. So $\widehat{CV}_{1,1,m}$ will still tend to select the candidate model that reflects the true model as much as possible.

For $\widehat{CV}_{1,2,m}$, $\widehat{CV}_{2,1,m}$ and $\widehat{CV}_{2,2,m}$, first, the asymptotic limit of them are minimized when $\boldsymbol{X}_t \subset \boldsymbol{X}_m$ or $\boldsymbol{X}_{R_3}$ is fully explained by $\boldsymbol{X}_m$ if $T^{(1)}_{1,2,m}+T^{(2)}_{1,2,m}$, $T^{(1)}_{2,1,m}+T^{(2)}_{2,1,m}$ and $T^{(1)}_{2,2,m}+T^{(2)}_{2,2,m}$ are greater than or equal to 0 for all the candidate models. Then, they will tend to have the same ability as $\widehat{CV}_{1,1,m}$ to select the select the candidate model that reflects the true model as much as possible. However, the asymptotic limits are not minimized when $\boldsymbol{X}_m = \boldsymbol{X}_t$ if $T^{(1)}_{1,2,m} + T^{(2)}_{1,2,m}$, $T^{(1)}_{2,1,m} + T^{(2)}_{2,1,m}$ and $T^{(1)}_{2,2,m} + T^{(2)}_{2,2,m}$ are less than 0 for some $\boldsymbol{X}_m$, which means they may tend to select some $\boldsymbol{X}_m$ that minimize them but not fully reflect the true model. As a result, it will increase the chance of making wrong decision.

However, notice the scenarios that make $T^{(1)}_{1,2,m} + T^{(2)}_{1,2,m}$, $T^{(1)}_{2,1,m} + T^{(2)}_{2,1,m}$ and $T^{(1)}_{2,2,m} + T^{(2)}_{2,2,m}$ less than 0 are very hard to reach. In order to get the effect of the three terms, we need a model with extremely poor predictive power, for example, we need π_i's very small for some elements and very big for others, and we need ϵ_i to have a very large variance. In other words, these are models that are mostly composed of noise and hence, variable selection will be extremely challenging. As a result, the asymptotic advantage of $\widehat{CV}_{1,1,m}$ over the other three criteria cannot be detected in the simulations. In this sense, we will not present a extreme scenario simulation study of that in Section 2.4.

2.4. Simulation

A population of 1000 different values for 8 different x variables are randomly generated using uniform$[0,1]$ distribution, and 1000 different values for ϵ are generated using $N(0,\sigma^2)$. Four different responses were generated where:

$$\boldsymbol{Y}_1 = 1 + \boldsymbol{X}(0,-1,0,0,0,0,0,0)^T + \boldsymbol{\epsilon},$$

$$\boldsymbol{Y}_2 = 1 + \boldsymbol{X}(0,-1,0,1.5,0,0,0,-1)^T + \boldsymbol{\epsilon},$$

$$\boldsymbol{Y}_3 = 1 + \boldsymbol{X}(-0.5,-1,0,1.5,2.5,0,0,-1)^T + \boldsymbol{\epsilon},$$

$$\boldsymbol{Y}_4 = 1 + \boldsymbol{X}(-0.5,-1,-0.8,1.5,2.5,2,-1,-1)^T + \boldsymbol{\epsilon}.$$

The design of stratified simple random sampling without replacement uses 4 strata with each stratum containing 250 elements, and the stratification is based on a random variable z_i and ratio r. First, we generate v_i from a standard normal distribution $N(0,\sigma_v^2)$ with σ_v^2 satisfying $r = \frac{\sigma^2}{\sigma^2+\sigma_v^2}$. Then z_i's are constructed as follows:

$$z_i = \begin{cases} v_i - x_{2,i}\epsilon_i & \text{if} \quad 0 < r < 1 \\ v_i & \text{if} \quad r = 0 \\ -x_{2,i}\epsilon_i & \text{if} \quad r = 1. \end{cases}$$

After sorting by z_i $(i = 1,\ldots,N)$, the population is separated into 4 strata with boundaries given by equally-spaced quantiles of z. Then, simulations are conducted with the stratum sample sizes (35, 30, 20, 15). For different values of r, this will cause the z_i to be correlated with $x_{2,i}\epsilon_i$. In other words, the constant r controls the extent of informativeness. That is, when $r = 0$, the sampling is uninformative. As r increases, the extent of informativeness of the sampling increases.

In this simulation, we use $\sigma^2 = 0.01$. In addition, we use five different choices of r, which are 0, 0.25, 0.5, 0.75 and 1.

One thousand samples of size 100 are drawn from this population using stratified random sampling without replacement(STSI). Since there are eight covariates in the full model, there are 256 possible candidate models to compare. For each sample we compute ordinary least square regression(OLS) and weighted least square regression(WLS) on the all 256 possible models. The four model selection criteria $\widehat{CV}_{l,j,m}$ are calculated for each candidate model $\boldsymbol{X}_m$.

The results are shown in Tables 2.1. In those tables, the "Correct" column in the tables is the Fraction of Correct Selection (FCS), which is the percentage of $\widehat{CV}_{l,j,m}$ picking the correct model from the 256 candidate models over the one thousand samples. The "Picked" column in the tables is the number of Picked Covariates (NPC), which is the average number of covariates $\widehat{CV}_{l,j,m}$ picked over the one thousand samples.

In Table 2.1, the random error ϵ_i generating the finite population follows $N(0, .1)$. Based on the population model and sampling strategy specified above, $T^{(1)}_{1,2,m} + T^{(2)}_{1,2,m}$, $T^{(1)}_{2,1,m} + T^{(2)}_{2,1,m}$ and $T^{(1)}_{2,2,m} + T^{(2)}_{2,2,m}$ are greater than 0 when $\boldsymbol{X}_{R_3} \neq \emptyset$.

First, since $T^{(1)}_{1,2,m} + T^{(2)}_{1,2,m}$, $T^{(1)}_{2,1,m} + T^{(2)}_{2,1,m}$ and $T^{(1)}_{2,2,m} + T^{(2)}_{2,2,m}$ are greater than 0 when $\boldsymbol{X}_{R_3} \neq \emptyset$, $\widehat{CV}_{1,1,m}$, $\widehat{CV}_{1,2,m}$, $\widehat{CV}_{2,1,m}$ and $\widehat{CV}_{2,2,m}$ tend to have similar ability to select the correct model, and we can find that FCS and NPC of them are very close. Next, since the FCS relies on this ability, given a certain r, we find that as the number of variables in the model increases, NPC increased and FCS improve. What is more, when the true model is the full model, all $\widehat{CV}_{l,j,m}$ choose the correct model in every sample. In conclusion, these facts indicate that the ability to select the candidate models that contain $\boldsymbol{X}_t$ is approximately the same across the four criteria in this scenario.

Table 1. Results of the simulation study for STSI design with $n = 100$, and $\sigma^2 = 0.01$. NPC and FCS of each criteria are shown.

		1 Var. Model		3 Var. Model		5 Var. Model		8 Var. Model	
r	Criterion	Picked	Correct	Picked	Correct	Picked	Correct	Picked	Correct
0	$\widehat{CV}_{1,1,m}$	2.124	0.312	3.842	0.412	5.436	0.622	8.000	1.000
0	$\widehat{CV}_{1,2,m}$	2.151	0.285	3.799	0.427	5.469	0.604	8.000	1.000
0	$\widehat{CV}_{2,1,m}$	2.073	0.339	3.769	0.455	5.379	0.666	8.000	1.000
0	$\widehat{CV}_{2,2,m}$	2.095	0.307	3.747	0.445	5.453	0.608	8.000	1.000
0.25	$\widehat{CV}_{1,1,m}$	2.212	0.280	3.847	0.414	5.473	0.608	8.000	1.000
0.25	$\widehat{CV}_{1,2,m}$	2.266	0.262	3.835	0.410	5.519	0.577	8.000	1.000
0.25	$\widehat{CV}_{2,1,m}$	2.123	0.326	3.779	0.437	5.400	0.661	8.000	1.000
0.25	$\widehat{CV}_{2,2,m}$	2.120	0.313	3.778	0.444	5.479	0.604	8.000	1.000
0.5	$\widehat{CV}_{1,1,m}$	2.169	0.274	3.835	0.394	5.510	0.574	8.000	1.000
0.5	$\widehat{CV}_{1,2,m}$	2.245	0.263	3.843	0.388	5.557	0.543	8.000	1.000
0.5	$\widehat{CV}_{2,1,m}$	2.118	0.297	3.807	0.422	5.495	0.584	8.000	1.000
0.5	$\widehat{CV}_{2,2,m}$	2.140	0.297	3.783	0.419	5.509	0.575	8.000	1.000
0.75	$\widehat{CV}_{1,1,m}$	2.143	0.292	3.831	0.408	5.509	0.580	8.000	1.000
0.75	$\widehat{CV}_{1,2,m}$	2.226	0.258	3.898	0.370	5.537	0.569	8.000	1.000
0.75	$\widehat{CV}_{2,1,m}$	2.138	0.296	3.773	0.432	5.483	0.591	8.000	1.000
0.75	$\widehat{CV}_{2,2,m}$	2.111	0.299	3.805	0.414	5.470	0.609	8.000	1.000
1	$\widehat{CV}_{1,1,m}$	2.157	0.297	3.811	0.412	5.523	0.560	8.000	1.000
1	$\widehat{CV}_{1,2,m}$	2.362	0.227	3.966	0.342	5.570	0.533	8.000	1.000
1	$\widehat{CV}_{2,1,m}$	2.223	0.249	3.847	0.399	5.520	0.560	8.000	1.000
1	$\widehat{CV}_{2,2,m}$	2.159	0.290	3.816	0.400	5.482	0.593	8.000	1.000

2.5. Conclusion

In this chapter, the study was motiviated by the surprising simulation results from Koralik and Opsomer (2010) (Master's Project). In that project, they found that wPRESS and PRESS using OLS and WLS in the sampling context work approximately as well as each other with a informative design. We studied the criteria in the sampling context theoretically. We assumed that the finite population is generated by a multivariate linear regression model with a possibly informative design. We derived the asymptotic properties of PRESS and wPRESS using OLS and WLS. We provided new insights in the conditions under which wPRESS using WLS worked as well as the others, and the conditions wPRESS using WLS worked better than the others, asymptotically. We found that in order to reach the conditions under which wPRESS using WLS worked better, we need a model with extremely poor predictive power, i.e., we are basically fitting noise so that all of the criteria performed poorly in selecting. As a result, the asymptotic advantage of wPRESS using WLS over the others could not be detected. This theoretically confirmed the simulation studies in Koralik and Opsomer (2010) and indicated that the four criteria worked as well as each other in the sampling context practically. In Section 2.4, we conducted a simulation study to show the facts we discussed in Section 2.3 in the conditions under which wPRESS using WLS worked as well as the others.

CHAPTER 3

Local Polynomial Regression With An Ordinal Covariate

3.1. Introduction

Nonparametric methods have attracted much attention among statisticians in the last several decades. There have been several landmark papers and monographs on the topic (Nadaraya 1964; Watson 1964; Stone 1977; Cleveland 1979; Gasser and Müller 1979; Gasser and Müller 1984; Müller 1987; Cleveland and Devlin 1988; Eubank 1988; Härdle 1990; Wahba 1990; Fan 1992; Fan 1993; Ruppert and Wand 1994; Wand and Jones 1995; Fan and Gijbels 1996; Simonoff 1996) that have shown that nonparametric regression techniques have much to apply to a range of problem domains.

Nadaraya (1964) and Watson (1964) proposed the Nadaraya-Watson kernel estimator. Gasser and Müller (1979, p.23-68) and Gasser and Müller (1984) originated the Gasser-Müller estimator. Härdle (1990) gave a book-length discussion on different kernel-type approaches to regression estimation. Wahba (1990) gave general descriptions of smoothing splines.

Stone (1977) examined the consistency properties of many nonparametric regression estimators, including local polynomial estimators. Cleveland (1979) and Cleveland and Devlin (1988) showed that local polynomial regression techniques are applicable to a wide range of problems. Müller (1987), Fan (1992), Fan (1993), Ruppert and Wand (1994) examined the asymptotic properties of local polynomial estimators. Eubank (1988), Wand and Jones (1995), Fan and Gijbels (1996) and Simonoff (1996) gave book-length general descriptions of local polynomial estimation.

The work of Aitchison and Aitken (1976) has initiated the literature on the kernel smoothing of categorical variables. Afterwards, the estimation of the (conditional) probability distribution of categorical variables using nonparametric techniques has been greatly developed. Titterington (1980) gave a comparative study of kernel-based density estimates for categorical data. Hall (1981) studied nonparametric multivariate binary discrimination. Wang and van Ryzin (1981) investigated a class of smooth estimators for discrete distributions. Bowman, Hall, and Titterington (1984) studied the cross-validation in nonparametric estimation of probabilities and probability densities, Hall and Wand (1988) studied nonparametric discrimination using density differences. Grund and Hall (1993) discussed the performance of kernel estimators for high-dimensional sparse binary data. Scott (1992), Fahrmeir and Tutz (1994), and Simonoff (1996) gave descriptions of this topic in their books.

Notice that though it is not uncommon to encounter regression situations in which covariates are categorical, much less effort has been devoted to this situation in the nonparametric context, especially in which covariates are ordered categorical (ordinal). Bierens (1983) began the consideration of kernel regression with mixed continuous and discrete covariates. Li and Racine (2004), Racine and Li (2004), Hall, Racine, and Li (2004), Hall, Li, and Racine (2007), Li and Racine (2008), and Ouyang, Li, and Racine (2009) have considered nonparametric estimation of regression functions, conditional density, and distribution functions, and quantile functions containing a mix of categorical and continuous covariates.

Let Y_i be the continuous response and C_i be the ordinal covariate having M categories with a natural ordering, e.g., preferences (disagree, indifference, agree), professional class (assistant professor, associate professor, professor), etc. For simplicity, we will let it take values in $\{1, 2, \ldots, M\}$, where lower order of C_i has smaller value. Ouyang, Li, and Racine

(2009) assume that Y_i is directly the function of C_i plus some random error:

$$Y_i = f(C_i) + \epsilon_i.$$

They proposed a kernel:

$$K_\lambda(C_i, t) = \begin{cases} 1 & \text{if } C_i = t \\ \lambda^{|C_i - t|} & \text{if } C_i \neq t, \end{cases}$$

where $t = 1, 2, \ldots, M$.

They proved that the Nadaraya-Watson estimator using this kernel

$$\hat{m}_{OLR}(t, \lambda) = \frac{\sum_{i=1}^{n} K_\lambda(C_i, t) Y_i}{\sum_{i=1}^{n} K_\lambda(C_i, t)} \tag{2}$$

is consistent with $f(\cdot)$ and the optimized smoothing parameter $\hat{\lambda} = O_p(n^{-1})$ under the leave-one-out cross validation (CV) criterion. However, since neither C_i nor the kernel is continuous, one cannot do the usual Taylor-based arguments that rely on C_i having a density (Ruppert and Wand 1994) to derive the asymptotic bias and variance.

In this paper, we assume that Y_i is the function of a latent continuous covariate X_i plus some random error:

$$Y_i = m(X_i) + \epsilon_i, \tag{3}$$

and then connect X_i with C_i. We use a symmetric and continuous kernel function in our estimator. We will propose our estimator in Section 3.2, study the asymptotic bias and variance of this estimator in Section 3.3, and conduct simulation in Section 3.4 compared with the estimator proposed by Ouyang, Li, and Racine (2009).

3.2. Proposed Estimator

Let $(C_1, Y_1), \ldots, (C_n, Y_n)$ be a set of independent and identically distributed (i.i.d.) random vectors following a joint distribution F, where the Y_i are scalar response variables and C_i are univariate ordinal covariates having M categories and taking values in $\{1, 2, \ldots, M\}$, where lower order of C_i has smaller value. In our latent variable model (3), X_i are iid latent continuous covariates with density function $f_X(x)$ and known bounded support $\text{supp}(f_X)$, and ϵ_i's are iid, independent of X_i's with $\text{E}(\epsilon_i) = 0$ and $\text{Var}(\epsilon_i) = \sigma^2$. We can be more specific: we assume X is on [0,1], we create a grid of boundary points, and we assume that there X has a density such that $P(X \in U_j) = P(C = j)$, where $U_j = (\frac{j-1}{M}, \frac{1}{M}]$, $j = 1, 2, \ldots, M$.

We first briefly review local polynomial regression for the case where the continuous covariate is observed directly. Then the nonparametric regression problem is that of estimating

$$m(x) = \text{E}(Y|X = x)$$

at a point $x \in supp(f_X)$ without the imposition of m belonging to a parametric family of functions. The local polynomial estimator $\hat{m}(x; p, h)$ is obtained by fitting the polynomial

$$\beta_0 + \beta_1(. - x) + \ldots + \beta_p(. - x)^p$$

to the (X_i, Y_i) using weighted least squares with kernel weights $K_h(X_i - x) = \frac{1}{h}K(\frac{X_i - x}{h})$. The value of $\hat{m}(x; p, h)$ is the intercept of the fit $\hat{\beta}_0$, where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^T$ minimizes

$$\sum_{i=1}^{n} \{Y_i - \beta_0 - \ldots - \beta_p (X_i - x)^p\}^2 K_h(X_i - x).$$

Assuming the invertibility of $\boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{X}_x$, standard weighted least squares theory leads to the solution

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{X}_x)^{-1} \boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{Y},$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is the vector of responses,

$$\boldsymbol{X}_x = \begin{pmatrix} 1 & X_1 - x & \cdots & (X_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n - x & \cdots & (X_n - x)^p \end{pmatrix}$$

is an $n \times (p+1)$ design matrix and

$$\boldsymbol{W}_x = diag(K_h(X_1 - x), \ldots, K_h(X_n - x))$$

is an $n \times n$ diagonal matrix of weights, $K_h(\cdot) = 1/hK(\cdot/h)$ and $K(\cdot)$ is a symmetric and continuous kernel function having compact support $[-1, 1]$. We also define the moments of $K(\cdot)$ as $\mu_k = \int_{-1}^{1} z^k \, K(z) \mathrm{d}z$ and $R_k(K) = \int_{-1}^{1} z^k \, K^2(z) \mathrm{d}z$. Since the estimator of $m(x)$ is the intercept coefficient, we obtain

$$\hat{m}(x; p, h) = \boldsymbol{e}_1^T (n^{-1} \boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{X}_x)^{-1} n^{-1} \boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{Y},$$

where $\boldsymbol{e}_1^T = (1, 0, \dots, 0)$. This estimator has been well-studied (Cleveland 1979; Cleveland and Devlin 1988; Fan 1992; Fan 1993; Ruppert and Wand 1994) and is widely used in statistics.

In our situation, the main target will be estimating

$$\begin{aligned} m_t &= E(Y_i | C_i = t) \\ &= E(Y_i | X_i \in U_t) \\ &= E(m(X_i) | X_i \in U_t) \\ &= \frac{\int_{U_t} m(x) f_X(x)\,\mathrm{d}x}{\int_{U_t} f_X(x)\,\mathrm{d}x} \\ &= \frac{m(x_t) \int_{U_t} f_X(x)\,\mathrm{d}x}{\int_{U_t} f_X(x)\,\mathrm{d}x} \\ &= m(x_t), \end{aligned}$$

for some $x_t \in U_t$, $t \in 1, \dots, M$. Next, we will estimate m_t at the point $\phi_t = \frac{t-0.5}{M}$ in cell U_t, and our estimator is

$$\hat{m}(t; p, h) = e^T(\boldsymbol{\Phi}_t^T \boldsymbol{W}_t \boldsymbol{\Phi}_t)^{-1} \boldsymbol{\Phi}_t^T \boldsymbol{W}_t \boldsymbol{Y},$$

where

$$\boldsymbol{\Phi}_t = \begin{pmatrix} 1 & \sum_{j=1}^{M} (\phi_j - \phi_t) I_{\{x_1 \in U_j\}} & \cdots & \sum_{j=1}^{M} (\phi_j - \phi_t)^p I_{\{X_1 \in U_j\}} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \sum_{j=1}^{M} (\phi_j - \phi_t) I_{\{x_n \in U_j\}} & \cdots & \sum_{j=1}^{M} (\phi_j - \phi_t)^p I_{\{X_n \in U_j\}} \end{pmatrix},$$

$$
\boldsymbol{W}_t = \begin{pmatrix} \sum_{j=1}^{M} K_h(\phi_j - \phi_t) I_{\{x_1 \in U_j\}} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sum_{j=1}^{M} K_h(\phi_j - \phi_t) I_{\{x_n \in U_j\}} \end{pmatrix},
$$

and $\phi_j = \frac{j-0.5}{M}$, $j = 1, 2, \ldots, M$.

For conciseness in what follows, let $\boldsymbol{B} = \boldsymbol{\Phi}_t^T \boldsymbol{W}_t \boldsymbol{\Phi}_t$ and $\boldsymbol{c} = \boldsymbol{\Phi}_t^T \boldsymbol{W}_t \boldsymbol{Y}$. Then,

$$
\hat{m}(t; p, h) = \boldsymbol{e}_1^T \boldsymbol{B}^{-1} \boldsymbol{c}.
$$

Here, $\boldsymbol{B}$ is $(p+1) \times (p+1)$ matrix having the (k, l)th entry equal to

$$
\sum_{i=1}^{n} \sum_{j=1}^{M} K_h(\phi_j - \phi_t)(\phi_j - \phi_t)^{k+l-2} I_{\{X_i \in U_j\}}
$$

and $\boldsymbol{c}$ is $(p+1) \times 1$ vector having the kth entry equal to

$$
\sum_{i=1}^{n} \sum_{j=1}^{M} K_h(\phi_j - \phi_t)(\phi_j - \phi_t)^{k-1} Y_i I_{\{X_i \in U_j\}}.
$$

3.3. Conditional Mean and Variance Properties

In this section we investigate the asymptotic properties of $\hat{m}(t; p, h)$.

We will need the following assumptions:

A7. $m^{(p+2)}(\cdot)$ *is continuous.*

A8. f_X *has bounded support* $[0, 1]$, $f_X > 0$ *and has two derivatives.*

A9. *The kernel K satisfies $\int_{-1}^{1} K(z)\mathrm{d}z = 1$, $\mu_k = 0$ for k odd, $\mu_k \neq 0$ for k even, and $R_l(K) > 0$, $k = 1, \ldots, 2p+2$, $l = 1, \ldots, 2p$.*

A10. *$h \to 0$, $h^3 n \to \infty$ and $h^{p+2} M \to \infty$.*

Lemma 3.3.1. *Suppose A7-A10 hold. Then $n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{M} K_h(\phi_j - \phi_t)(\phi_j - \phi_t)^k I_{\{X_i \in U_j\}}$ is equal to*

$$h^k \mu_k\, f_X(\phi_t) + o_p(h^k)$$

for k even and

$$h^{k+1}\mu_{k+1}\, f_X^{'}(\phi_t) + o_p(h^{k+1})$$

for k odd. And $n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{M} K_h^2(\phi_j - \phi_t)(\phi_j - \phi_t)^k I_{\{X_i \in U_j\}}$ is equal to

$$h^{k-1} R_{k-1}(K) f_X(\phi_t) + o_p(h^{k-1})$$

in either case.

Proof. Since

$$n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{M} K_h(\phi_j - \phi_t)(\phi_j - \phi_t)^k I_{\{X_i \in U_j\}} = \mathrm{E}\left(\sum_{j=1}^{M} K_h(\phi_j - \phi_t)(\phi_j - \phi_t)^k I_{\{X_i \in U_j\}}\right)$$
$$+ O_p\left(\sqrt{\frac{\mathrm{Var}\left(\sum_{j=1}^{M} K_h(\phi_j - \phi_t)(\phi_j - \phi_t)^k I_{\{X_i \in U_j\}}\right)}{n}}\right),$$

for the expectation,

$$\mathrm{E}\left(\sum_{j=1}^{M} K_h(\phi_j-\phi_t)(\phi_j-\phi_t)^k I_{\{X_i\in U_j\}}\right) = h^{-1}\sum_{j=1}^{M} K(\frac{\phi_j-\phi_t}{h})(\phi_j-\phi_t)^k P(X_i\in U_j)$$

$$= h^{-1}\int K\left(\frac{l-\phi_t}{h}\right)(l-\phi_t)^k\ f_X(l)\,dl + O(M^{-1})$$

$$= h^k\int z^k K(z) f_X\left(zh+\phi_t\right)\,\mathrm{d}z + O(M^{-1})$$

$$= h^k\int z^k K(z)\left(\left(f_X(\phi_t)+zhf_X^{'}(\phi_t)+o(1)\right)\,\mathrm{d}z\right) + O(M^{-1}).$$

For k even, it is equal to

$$h^k\mu_k f_X(\phi_t)+o(h^k),$$

and for k odd, it is equal to

$$h^{k+1}\mu_{k+1}f_X^{'}(\phi_t)+o(h^{k+1}).$$

The leading term for odd k and even k are different because $\mu_k=\int z^k K(z)\,\mathrm{d}z$ is 0 when k is odd and we need to expand $f_X(zh+\phi_t)$ at $z=0$ one more order to get a non-zero leading term.

For the variance,

$$\mathrm{Var}\left(\sum_{j=1}^{M} K_h(\phi_j-\phi_t)(\phi_j-\phi_t)^k I_{\{X_i\in U_j\}}\right) = \mathrm{E}\left(\left(\sum_{j=1}^{M} K_h(\phi_j-\phi_t)(\phi_j-\phi_t)^k\ I_{\{X_i\in U_j\}}\right)^2\right)$$

$$-\left(\mathrm{E}\left(\sum_{j=1}^{M} K_h(\phi_j-\phi_t)(\phi_j-\phi_t)^k\ I_{\{X_i\in U_j\}}\right)\right)^2.$$

The first term is

$$\mathrm{E}\left(\left(\sum_{j=1}^{M} K_h(\phi_j-\phi_t)(\phi_j-\phi_t)^k I_{\{X_i\in U_j\}}\right)^2\right) = h^{-2}\sum_{j=1}^{M} K^2(\frac{\phi_j-\phi_t}{h})(\phi_j-\phi_t)^{2k}P(X_i\in U_j)$$

$$= h^{-2}\int K^2\left(\frac{l-\phi_t}{h}\right)(l-\phi_t)^{2k}\ f_X(l)\,dl + O(M^{-1})$$

$$= h^{2k-1}\int z^{2k}K^2(z)f_X(zh+\phi_t)\,\mathrm{d}z + O(M^{-1})$$

$$= h^{2k-1}\int z^{2k}K^2(z)\,\mathrm{d}z\,(f_X(\phi_t)+o(1)) + O(M^{-1})$$

$$= h^{2k-1}R_{2k}f_X(\phi_t) + o(h^{2k-1}).$$

Since the second term is at most $O(h^{2k})$,

$$\mathrm{Var}\left(\sum_{j=1}^{M} K_h(\phi_j-\phi_t)(\phi_j-\phi_t)^k I_{\{X_i\in U_j\}}\right) = O(h^{2k-1}).$$

By A10,

$$O_p\left(\sqrt{\frac{\mathrm{Var}\left(\sum_{j=1}^{M} K_h(\phi_j-\phi_t)(\phi_j-\phi_t)^k I_{\{X_i\in U_j\}}\right)}{n}}\right) = o_p(h^{k+1}).$$

Then, the first two results follow.

For the third result, since

$$n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{M} K_h^2(\phi_j-\phi_t)(\phi_j-\phi_t)^k I_{\{X_i\in U_j\}} = \mathrm{E}\left(\sum_{j=1}^{M} K_h^2(\phi_j-\phi_t)(\phi_j-\phi_t)^k I_{\{X_i\in U_j\}}\right)$$
$$+\, O_p\left(\sqrt{\frac{\mathrm{Var}\left(\sum_{j=1}^{M} K_h^2(\phi_j-\phi_t)(\phi_j-\phi_t)^k I_{\{X_i\in U_j\}}\right)}{n}}\right),$$

we can derive the leading term of

$$\mathrm{E}\left(\sum_{j=1}^{M} K_h^2(\phi_j-\phi_t)(\phi_j-\phi_t)^k I_{\{X_i\in U_j\}}\right)$$

and

$$\mathrm{Var}\left(\sum_{j=1}^{M} K_h^2(\phi_j-\phi_t)(\phi_j-\phi_t)^k I_{\{X_i\in U_j\}}\right)$$

and determine the order of $O_p\left(\sqrt{\frac{\mathrm{Var}\left(\sum_{j=1}^{M} K_h^2(\phi_j-\phi_t)(\phi_j-\phi_t)^k I_{\{X_i\in U_j\}}\right)}{n}}\right)$, which is $O_p(2k-2)$, in the same way. Then the third result follows. □

THEOREM 3.3.2. *Suppose that $t \in (1,2,\ldots,M)$, and that A7-A10 hold. Let $\hat{m}(t;p,h) = \boldsymbol{e}_1^T\boldsymbol{B}^{-1}\boldsymbol{c}$, $\boldsymbol{H}_p = diag(1,h,\ldots,h^p)$, $\boldsymbol{N}_p$ be the $(p+1)\times(p+1)$ matrix having the (i,j)th entry equal to μ_{i+j-2}, $\boldsymbol{Q}_p$ be the $(p+1)\times(p+1)$ matrix having the (i,j)th entry equal to μ_{i+j-1}, $\boldsymbol{T}_p$ is the $(p+1)\times(p+1)$ matrix having the (k,l)th element equal to $\int u^{k+l-2}K^2(u)\,du$, and*

$\boldsymbol{C} = (C_1, \ldots, C_n)^T$. *Then for* p *odd,*

$$E\{\hat{m}(t; p, h) - m_t|\boldsymbol{C}\} = \left\{\sum_{j=1}^{p+1}((\boldsymbol{N}_p^{-1})_{1,j}\,\mu_{p+j})\right\}\frac{m^{(p+1)}(\phi_t)}{(p+1)!}h^{p+1} + o_p(h^{p+1}),$$

and for p *even,*

$$E\{\hat{m}(t; p, h) - m_t|\boldsymbol{C}\} = \left\{\sum_{j=1}^{p+1}((\boldsymbol{N}_p^{-1})_{1,j}\,\mu_{p+j+1})\right\}$$
$$\times\{\frac{m^{(p+1)}(\phi_t)}{(p+1)!}\frac{f_X'(\phi_t)}{f_X(\phi_t)} + \frac{m^{(p+2)}(\phi_t)}{(p+2)!}\}h^{p+2} + o_p(h^{p+2}).$$

In either case

$$Var\{\hat{m}(t; p, h)|\boldsymbol{C}\} = (n^{-1}h^{-1}\frac{\sigma^2}{f_X(\phi_t)}\boldsymbol{e}_1^T\boldsymbol{N}_p^{-1}\boldsymbol{T}_p\boldsymbol{N}_p^{-1}\boldsymbol{e}_1)(1 + o_p(1)).$$

Remark 1. The order of the asymptotic bias is $O_p(h^{p+1})$ for p odd and, $O_p(h^{p+2})$ for p even. The order of the asymptotic variance is $O_p(\frac{1}{nh})$ in either case. This is the same as the results for local polynomial estimator with a continuous covariate.

Remark 2. The asymptotic bias of the local polynomial estimator with an continuous covariate is

$$\mathrm{E}\{\hat{m}(x; p, h) - m_t|\boldsymbol{X}\} = \left\{\sum_{j=1}^{p+1}((\boldsymbol{N}_p^{-1})_{1,j}\,\mu_{p+j})\right\}\frac{m^{(p+1)}(x)}{(p+1)!}h^{p+1} + o_p(h^{p+1})$$

when p is odd, and

$$\mathrm{E}\{\hat{m}(x; p, h) - m_t|\boldsymbol{X}\} = \left\{\sum_{j=1}^{p+1}((\boldsymbol{N}_p^{-1})_{1,j}\,\mu_{p+j+1})\right\}$$
$$\times\{\frac{m^{(p+1)}(x)}{(p+1)!}\frac{f_X'(x)}{f_X(x)} + \frac{m^{(p+2)}(x)}{(p+2)!}\}h^{p+2} + o_p(h^{p+2})$$

when p is even. The asymptotic variance of the local polynomial estimator with an continuous covariate is

$$\mathrm{Var}\{\hat{m}(x;p,h)-m_t|\boldsymbol{x}\}=(n^{-1}h^{-1}\frac{\sigma^2}{f_X(x)}\boldsymbol{e}_1^T\boldsymbol{N}_p^{-1}\boldsymbol{T}_p\boldsymbol{N}_p^{-1}\boldsymbol{e}_1)(1+o_p(1))$$

For either odd or even p. Therefore, the leading terms are exactly the same form as ours. The difference is that the position of the target and the values of X_i are known in the case of local polynomial estimator with an continuous covariate. The conditional bias caused by the uncertain values of X_i and the target's position x_t is in a smaller order in probability than the leading terms and is absorbed in the remnant terms. Similar thing happens to the conditional variance. This can be found clearly in the proof.

Remark 3. To prove the theorem, we are following the approach used in Ruppert and Wand (1994).

Proof. The conditional expectation

$$\mathrm{E}\{\hat{m}(t;p,h)-m_t|\boldsymbol{C}\}=\boldsymbol{e}_1^T(n^{-1}\boldsymbol{B})^{-1}\mathrm{E}(n^{-1}\boldsymbol{c}|\boldsymbol{C})-m(\phi_t)+m(\phi_t)-m_t.$$

First, since $|x_t-\phi_t|\leq\frac{0.5}{M}$ and by A7, $m(\phi_t)-m_t=O(\frac{1}{M})$. Now, we only need to focus on the first difference, which is

$$\boldsymbol{e}_1^T(n^{-1}\boldsymbol{B})^{-1}\mathrm{E}(n^{-1}\boldsymbol{c}|\boldsymbol{C})-m(\phi_t)$$

$$= \boldsymbol{e}_1^T (n^{-1}\boldsymbol{B})^{-1} \begin{pmatrix} n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{M} K_h(\phi_j - \phi_t) m_j I_{\{X_i \in U_j\}} \\ \vdots \\ n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{M} K_h(\phi_j - \phi_t)(\phi_j - \phi_t)^p m_j I_{\{X_i \in U_j\}} \end{pmatrix}$$

$$- m(\phi_t)$$

$$= \boldsymbol{e}_1^T (n^{-1}\boldsymbol{B})^{-1} \begin{pmatrix} n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{M} K_h(\phi_j - \phi_t) \left(m(\phi_t) + \dots \right. \\ \left. + \frac{m^{(p+2)}(\phi_t)}{(p+2)!} (\phi_j - \phi_t)^{p+2} + r_t + O(\frac{1}{M}) \right) I_{\{X_i \in U_j\}} \\ \vdots \\ n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{M} K_h(\phi_j - \phi_t)(\phi_j - \phi_t)^p \left(m(\phi_t) + \dots \right. \\ \left. + \frac{m^{(p+2)}(\phi_t)}{(p+2)!} (\phi_j - \phi_t)^{p+2} + r_t + O(\frac{1}{M}) \right) I_{\{X_i \in U_j\}} \end{pmatrix}$$

$$- m(\phi_t)$$

$$= \boldsymbol{e}_1^T (n^{-1}\boldsymbol{B})^{-1} (n^{-1}\boldsymbol{B}) \begin{pmatrix} m(\phi_t) \\ \vdots \\ \frac{m^{(p)}(\phi_t)}{(p)!} \end{pmatrix} + \boldsymbol{e}_1^T (n^{-1}\boldsymbol{B})^{-1} (\boldsymbol{S}_t + \boldsymbol{R}_t) - m(\phi_t)$$

$$= \boldsymbol{e}_1^T (n^{-1}\boldsymbol{B})^{-1} (\boldsymbol{S}_t + \boldsymbol{R}_t),$$

where

$$\boldsymbol{S}_t = n^{-1}\left\{\frac{m^{(p+1)}(\phi_t)}{(p+1)!}\begin{pmatrix}\sum_{i=1}^{n}\sum_{j=1}^{M} K_h(\phi_j-\phi_t)(\phi_j-\phi_t)^{p+1}I_{\{X_i\in U_j\}}\\ \vdots \\ \sum_{i=1}^{n}\sum_{j=1}^{M} K_h(\phi_j-\phi_t)(\phi_j-\phi_t)^{2p+1}I_{\{X_i\in U_j\}}\end{pmatrix}\right.$$

$$\left.+\frac{m^{(p+2)}(\phi_t)}{(p+2)!}\begin{pmatrix}\sum_{i=1}^{n}\sum_{j=1}^{M} K_h(\phi_j-\phi_t)(\phi_j-\phi_t)^{p+2}I_{\{X_i\in U_j\}}\\ \vdots \\ \sum_{i=1}^{n}\sum_{j=1}^{M} K_h(\phi_j-\phi_t)(\phi_j-\phi_t)^{2p+2}I_{\{X_i\in U_j\}}\end{pmatrix}\right\},$$

$\boldsymbol{R}_t$ is a vector of Taylor series remainder terms plus $O(\frac{1}{M})$, $m_j = E(Y_i|C_i = j)$ for some $x_j \in U_j$, and noting that

$$\text{sup}_{1\leq j\leq M}|m_j - m(\phi_j)| \leq C\frac{1}{M}$$

for a positive real constant C using A7.

By Lemma 3.3.1,

$$(n^{-1}\boldsymbol{B}) = \boldsymbol{H}_p\{f_X(\phi_t)\boldsymbol{N}_p + hf_X^{'}(\phi_t)\boldsymbol{Q}_p\}\boldsymbol{H}_p(\boldsymbol{I}_p + o_p(\boldsymbol{I}_p)).$$

Then,

$$\boldsymbol{e}_1^T(n^{-1}\boldsymbol{B})^{-1} = \frac{1}{f_X(\phi_t)}\{\boldsymbol{e}_1^T\boldsymbol{N}_p^{-1} - h\frac{f_X^{'}(\phi_t)}{f_X(\phi_t)}\boldsymbol{e}_1^T\boldsymbol{N}_p^{-1}\boldsymbol{Q}_p\boldsymbol{N}_p^{-1}\}\boldsymbol{H}_p^{-1}(\boldsymbol{I}_p + o_p(\boldsymbol{I}_p),.$$

where $\boldsymbol{I}_p$ is the $(p+1)\times(p+1)$ identity matrix.

Again, by Lemma 3.3.1,

$$\boldsymbol{S}_t = \boldsymbol{H}_p \left\{ \frac{m^{(p+1)}(\phi_t)}{(p+1)!} \left(h^{p+1} f_X(\phi_t) \begin{pmatrix} \mu_{p+1} \\ \vdots \\ \mu_{2p+1} \end{pmatrix} + h^{p+2} f_X^{'}(\phi_t) \begin{pmatrix} \mu_{p+2} \\ \vdots \\ \mu_{2p+2} \end{pmatrix} \right) \right.$$

$$\left. + \frac{m^{(p+2)}(\phi_t)}{(p+2)!} \left(h^{p+2} f_X(\phi_t) \begin{pmatrix} \mu_{p+2} \\ \vdots \\ \mu_{2p+2} \end{pmatrix} + h^{p+3} f_X^{'}(\phi_t) \begin{pmatrix} \mu_{p+3} \\ \vdots \\ \mu_{2p+3} \end{pmatrix} \right) \right\} (1 + o_p(1)).$$

Then,

$$\mathrm{E}\{\hat{m}(t;p,h) - m(t)|\boldsymbol{C}\} = \frac{1}{f_X(\phi_t)} \left\{ \frac{m^{(p+1)}(\phi_t)}{(p+1)!} \boldsymbol{e}_1^T \boldsymbol{N}_p^{-1} h^{p+1} f_X(\phi_t) \begin{pmatrix} \mu_{p+1} \\ \vdots \\ \mu_{2p+1} \end{pmatrix} \right.$$

$$+ \frac{m^{(p+1)}(\phi_t)}{(p+1)!} \boldsymbol{e}_1^T \boldsymbol{N}_p^{-1} h^{p+2} f_X^{'}(\phi_t) \begin{pmatrix} \mu_{p+2} \\ \vdots \\ \mu_{2p+2} \end{pmatrix}$$

$$+ \frac{m^{(p+2)}(\phi_t)}{(p+2)!} \boldsymbol{e}_1^T \boldsymbol{N}_p^{-1} h^{p+2} f_X(\phi_t) \begin{pmatrix} \mu_{p+2} \\ \vdots \\ \mu_{2p+2} \end{pmatrix}$$

$$+ \frac{m^{(p+2)}(\phi_t)}{(p+2)!}\boldsymbol{e}_1^T\boldsymbol{N}_p^{-1}h^{p+3}f_X'(\phi_t)\begin{pmatrix}\mu_{p+3}\\ \vdots\\ \mu_{2p+3}\end{pmatrix}$$

$$- \frac{m^{(p+1)}(\phi_t)}{(p+1)!}h\frac{f_X'(\phi_t)}{f_X(\phi_t)}\boldsymbol{e}_1^T\boldsymbol{N}_p^{-1}\boldsymbol{Q}_p\boldsymbol{N}_p^{-1}h^{p+1}f_X(\phi_t)\begin{pmatrix}\mu_{p+1}\\ \vdots\\ \mu_{2p+1}\end{pmatrix}$$

$$- \frac{m^{(p+1)}(\phi_t)}{(p+1)!}h\frac{f_X'(\phi_t)}{f_X(\phi_t)}\boldsymbol{e}_1^T\boldsymbol{N}_p^{-1}\boldsymbol{Q}_p\boldsymbol{N}_p^{-1}h^{p+2}f_X'(\phi_t)\begin{pmatrix}\mu_{p+2}\\ \vdots\\ \mu_{2p+2}\end{pmatrix}$$

$$- \frac{m^{(p+2)}(\phi_t)}{(p+2)!}h\frac{f_X'(\phi_t)}{f_X(\phi_t)}\boldsymbol{e}_1^T\boldsymbol{N}_p^{-1}\boldsymbol{Q}_p\boldsymbol{N}_p^{-1}h^{p+2}f_X(\phi_t)\begin{pmatrix}\mu_{p+2}\\ \vdots\\ \mu_{2p+2}\end{pmatrix}$$

$$- \frac{m^{(p+2)}(\phi_t)}{(p+2)!}h\frac{f_X'(\phi_t)}{f_X(\phi_t)}\boldsymbol{e}_1^T\boldsymbol{N}_p^{-1}\boldsymbol{Q}_p\boldsymbol{N}_p^{-1}h^{p+3}f_X'(\phi_t)\begin{pmatrix}\mu_{p+3}\\ \vdots\\ \mu_{2p+3}\end{pmatrix}\Bigg\}$$

$$\times (1 + o_p(1)).$$

First, $\boldsymbol{e}_1^T \boldsymbol{N}_p^{-1} \boldsymbol{Q}_p = \boldsymbol{0}$ since the last p columns of $\boldsymbol{N}_p$ is the first p columns of $\boldsymbol{Q}_p$. This implies that the last four terms are all 0. So,

$$
\begin{aligned}
\mathrm{E}\{\hat{m}(t;p,h) - m_t|\boldsymbol{C}\} = \frac{1}{f_X(\phi_t)} \Bigg\{ & \frac{m^{(p+1)}(\phi_t)}{(p+1)!} \boldsymbol{e}_1^T \boldsymbol{N}_p^{-1} h^{p+1} f_X(\phi_t) \begin{pmatrix} \mu_{p+1} \\ \vdots \\ \mu_{2p+1} \end{pmatrix} \\
& + \frac{m^{(p+1)}(\phi_t)}{(p+1)!} \boldsymbol{e}_1^T \boldsymbol{N}_p^{-1} h^{p+2} f_X'(\phi_t) \begin{pmatrix} \mu_{p+2} \\ \vdots \\ \mu_{2p+2} \end{pmatrix} \\
& + \frac{m^{(p+2)}(\phi_t)}{(p+2)!} \boldsymbol{e}_1^T \boldsymbol{N}_p^{-1} h^{p+2} f_X(\phi_t) \begin{pmatrix} \mu_{p+2} \\ \vdots \\ \mu_{2p+2} \end{pmatrix} \\
& + \frac{m^{(p+2)}(\phi_t)}{(p+2)!} \boldsymbol{e}_1^T \boldsymbol{N}_p^{-1} h^{p+3} f_X'(\phi_t) \begin{pmatrix} \mu_{p+3} \\ \vdots \\ \mu_{2p+3} \end{pmatrix} \Bigg\} \\
& \times (1 + o_p(1)).
\end{aligned}
$$

Second, for p even, the 1st and 4th term are both 0. Therefore,

$$
\begin{aligned}
\mathrm{E}\{\hat{m}(t;p,h) - m_t|\boldsymbol{C}\} = & \left\{ \sum_{j=1}^{p+1} ((\boldsymbol{N}_p^{-1})_{1,j}\, \mu_{p+j+1}) \right\} \\
& \times \{ \frac{m^{(p+1)}(\phi_t)}{(p+1)!} \frac{f_X'(\phi_t)}{f_X(\phi_t)} + \frac{m^{(p+2)}(\phi_t)}{(p+2)!} \} h^{p+2} + o_p(h^{p+2}).
\end{aligned}
$$

For p odd, the 2st and 3th term are both 0. Therefore,

$$\mathrm{E}\{\hat{m}(t;p,h) - m_t|\boldsymbol{C}\} = \left\{\sum_{j=1}^{p+1}((\boldsymbol{N}_p^{-1})_{1,j}\,\mu_{p+j})\right\}\frac{m^{(p+1)}(\phi_t)}{(p+1)!}h^{p+1} + o_p(h^{p+1}).$$

In this result, term 4 is absorbed in $o_p(h^{p+1})$.

If we write $\hat{m}(t;p,h)$ as $\boldsymbol{e}_1^T\boldsymbol{B}\boldsymbol{D}^T\boldsymbol{Y}$, then

$$\mathrm{Var}(\hat{m}(t;p,h)|\boldsymbol{C}) = n^{-1}\boldsymbol{e}_1^T(n^{-1}\boldsymbol{B})^{-1}(n^{-1}\boldsymbol{D}^T\mathrm{Var}(\boldsymbol{Y}|\boldsymbol{C})\boldsymbol{D})(n^{-1}\boldsymbol{B})^{-1}\boldsymbol{e}_1,$$

where

$$\boldsymbol{D} = \begin{pmatrix} \sum_{j=1}^{M} K_h(\phi_j - \phi_t)I_{\{X_1\in U_j\}} & \cdots & \sum_{j=1}^{M} K_h(\phi_j - \phi_t)(\phi_j - \phi_t)^p I_{\{X_1\in U_j\}} \\ \vdots & \ddots & \vdots \\ \sum_{j=1}^{M} K_h(\phi_j - \phi_t)I_{\{X_n\in U_j\}} & \cdots & \sum_{j=1}^{M} K_h(\phi_j - \phi_t)(\phi_j - \phi_t)^p I_{\{X_n\in U_j\}} \end{pmatrix}.$$

Since

$$\begin{aligned}\mathrm{Var}(Y_i|\boldsymbol{C}) &= \mathrm{Var}(\mathrm{E}(Y_i|X_i)|\boldsymbol{C}) + \mathrm{E}(\mathrm{Var}(Y_i|X_i)|\boldsymbol{C}) \\ &= \mathrm{Var}(m(X_i)|\boldsymbol{C}) + \mathrm{E}(\sigma^2|\boldsymbol{C}) \\ &= O(\frac{1}{M}) + \sigma^2.\end{aligned}$$

Then, $(n^{-1}\boldsymbol{D}^T\mathrm{Var}(\boldsymbol{Y}|\boldsymbol{C})\boldsymbol{D}) = (\sigma^2 + O(\frac{1}{M}))(n^{-1}\boldsymbol{D}^T\boldsymbol{D})$ and $(n^{-1}\boldsymbol{D}^T\boldsymbol{D})$ is the $(p+1)\times(p+1)$ matrix having the (k,l)th element equal to $n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{M}K_h^2(\phi_j - \phi_t)(\phi_j - \phi_t)^{k+l-2}I_{\{X_i\in U_j\}}$.

By lemma 3.3.1,

$$(n^{-1}\boldsymbol{D}^T\boldsymbol{D}) = h^{-1}f_X(\phi_t)\boldsymbol{H}_p\boldsymbol{T}_p\boldsymbol{H}_p(\boldsymbol{I}_p + o_p(\boldsymbol{I}_p)).$$

Then,

$$\begin{aligned}
&\mathrm{Var}(\hat{m}(t;p,h)|\boldsymbol{C})\\
&= n^{-1}\frac{\sigma^2 + O(\frac{1}{M})}{f_X(\phi_t)}\{\boldsymbol{e}_1^T\boldsymbol{N}_p^{-1} - h\frac{f_X^{'}(\phi_t)}{f_X(\phi_t)}\boldsymbol{e}_1^T\boldsymbol{N}_p^{-1}\boldsymbol{Q}_p\boldsymbol{N}_p^{-1}\}\boldsymbol{H}_p^{-1}\\
&\quad \times h^{-1}f_X(\phi_t)\boldsymbol{H}_p\boldsymbol{T}_p\boldsymbol{H}_p\\
&\quad \times \frac{1}{f_X(\phi_t)}\boldsymbol{H}_p^{-1}\{\boldsymbol{N}_p^{-1}\boldsymbol{e}_1 - h\frac{f_X^{'}(\phi_t)}{f_X(\phi_t)}\boldsymbol{N}_p^{-1}\boldsymbol{Q}_p\boldsymbol{N}_p^{-1}\boldsymbol{e}_1\}(1 + o_p(1))\\
&= (n^{-1}h^{-1}\frac{\sigma^2}{f_X(\phi_t)}\boldsymbol{e}_1^T\boldsymbol{N}_p^{-1}\boldsymbol{T}_p\boldsymbol{N}_p^{-1}\boldsymbol{e}_1)(1 + o_p(1)).
\end{aligned}$$

□

3.4. Simulation

In this section, we do 2 patterns of simulations based on a latent covariate model:

$$Y_i = \sin(\frac{\pi}{2}X_i) + \epsilon_i,\ i = 1, \ldots, 100,$$

where X_i are generated from $U(0,1)$, and ϵ_i are generated from $N(0, 0.1)$. Each pattern is simulated 1000 times.

In the first pattern, the number of observations that fall in different categories is set to be the same. In the second pattern, the number of observations that fall in different categories is set to be different, where the ratio is proportional to $(5, 6, 7, 8, 10, 5, 3, 2, 2, 2)$ for each category. The targets are $\{0.078, 0.233, 0.382, 0.522, 0.649, 0.760, 0.852, 0.923, 0.971, 0.996\}$.

We apply local linear estimation for 6 different bandwidth values $h = 0.15$, 0.2, 0.25, 0.3, 0.35, 0.4. We estimate m_t at $\{\phi_1, \ldots, \phi_{10}\} = \{\frac{1-0.5}{10}, \ldots, \frac{10-0.5}{10}\}$. For comparison, we also estimate the response with $h = 0.1$, that is, weighted averaging the response in each category. At a fixed h and ϕ_t, using the estimation $\hat{m}_{t,h}$, we approximate the pointwise bias $B(\hat{m}_{t,h}) = \mathrm{E}(\hat{m}_{t,h} - m_t)$, the pointwise standard deviation $SD(\hat{m}_{t,h}) = \sqrt{\mathrm{Var}(\hat{m}_{t,h})}$, the pointwise mean squared errors $MSE(\hat{m}_{t,h}) = \mathrm{E}(\hat{m}_{t,h} - m_t)^2$, and the mean sum of squared errors $MSSE_h = \mathrm{E}\sum_t(\hat{m}_{t,h} - m_t)^2$ by averaging the simulation results. For bandwidth selection, we apply the approaches of CV and GCV. That is, at a certain point t, we use $h_{CV} \in (0.1, 1)$ that minimizes $CV = \sum_i \left(\frac{\hat{m}_{t,h} - y_i}{1 - S_{ii}}\right)^2$, and $h_{GCV} \in (0.1, 1)$ that minimizes $GCV = \sum_i \left(\frac{\hat{m}_{t,h} - y_i}{1 - \bar{S}}\right)^2$, where S_{ii} is diagonal element of the projection matrix $\boldsymbol{S}$ and $\bar{S} = 1/100 \sum_i S_{ii}$ (Rudemo 1982; Bowman, Hall, and Titterington 1984; Hall, Marron, and Park 1992). The actual averages of the CV and GCV bandwidths are given denoted by ave(h_{CV}) and ave(h_{GCV}), respectively. The projection matrix $\boldsymbol{S}$ is 100×100 such that $\hat{m}_{t,h} = \boldsymbol{h}_i \boldsymbol{Y}$ if $X_i \in U_t$, where $\boldsymbol{h}_i$ is row i of $\boldsymbol{S}$. We also estimate the variance by

$$\hat{V}(\hat{m}_{t,h}) = \boldsymbol{e}_1^T \boldsymbol{B}^{-1} \boldsymbol{D}^T \hat{V}(\boldsymbol{Y}|\boldsymbol{C}) \boldsymbol{D} \boldsymbol{B}^{-1} \boldsymbol{e}_1,$$

where

$$\hat{V}(\boldsymbol{Y}|\boldsymbol{C}) = \begin{pmatrix} \hat{\sigma}_1^2 & & & & & \\ & \ddots & & & & \\ & & \hat{\sigma}_1^2 & & & \\ & & & \hat{\sigma}_2^2 & & \\ & & & & \ddots & \\ & & & & & \hat{\sigma}_{10}^2 \end{pmatrix}$$

where $\hat{\sigma}_t^2$ is the simulation variance for the Y_i such that $X_i \in U_t$, i.e.

$$\hat{\sigma}_t^2 = \frac{1}{\sum_i I_{\{X_i \in U_t\}} - 1} \sum_i I_{\{X_i \in U_t\}} (Y_i - \bar{Y}_t)^2,$$

where $\bar{Y}_t = \frac{1}{\sum_i I_{\{X_i \in U_t\}} - 1} \sum_i I_{\{X_i \in U_t\}} Y_i$. Finally, the relative bias of variance estimator $\mathrm{E}\left(\frac{\hat{V}(\hat{m}_{t,h}) - \mathrm{Var}(\hat{m}_{t,h})}{\mathrm{Var}(\hat{m}_{t,h})}\right)$ is given by averaging the simulation results.

For comparision, we estimate the same targets using the estimator (2) Ouyang, Li, and Racine (2009) proposed with $\lambda = 1e-13, 1e-12, \ldots, 1e-7$, and λ_{CV} and λ_{GCV} that minimizes CV and GCV similarly defined above except replacing $\hat{m}_{t,h}$ by $\hat{m}_{OLR}(t, \lambda)$.

From the results, we can find that the local linear estimator works well in both of the two patterns. Although the bias of the estimator in the 10 categories become bigger when h increases, the standard deviations becomes smaller, that is, incorporating data from more neighbour cells for local linear regression. The MSE first decrease and then increase when h increases. The same thing happens to the MSSE. The bandwidth h selected by CV and GCV matches in both of the two patterns as well as all of them are greater than 0.1. Notice that the bias, MSSE and standard deviation in the boundary categories, i.e. $\hat{m}_1$ and $\hat{m}_{10}$, are bigger than those in other categories. This is because the local linear regression incorporates one less cell of data at the boundary. The bias, MSSE and standard deviation are similar between the two patterns, which means our estimator works well for both balance spaced and unbalance spaced ordinal data. The relative bias between the estimated variance and true variance is very small, which validates our variance estimator.

In the results of $\hat{m}_{OLR}(t, \lambda)$, we also estimate the targets using $\lambda = 0$, that is, averaging the response in each category. Notice that the estimators are exactly the same when $h = 0.1$ and $\lambda = 0$. Therefore the simulation results are same as well. Similar to h, when $\lambda = 0$, the bias of $\hat{m}_{OLR}(t, \lambda)$ are the smallest and when λ increases the bias become bigger. The standard deviations become smaller when λ increases. Similarly, The MSE first decrease and then increase when λ increases. The same thing happens to MSSE. In both of the two patterns, using the bandwidths selected by CV and GCV, our estimator has smaller variance and mean squared errors at all of the 10 targets compared to the $\hat{m}_{OLR}(t, \lambda)$ estimator except m_{10}, and our estimator has smaller mean sum of squared errors.

Table 2. Simulation bias (in percent) $B(\hat{m}_{t,h}) = \mathrm{E}(\hat{m}_{t,h} - m_t)$ for pattern 1 for the 10 values of the ordinal covariate for 7 fixed bandwidth as well as at bandwidths selected by CV and GCV, with ave$(h_{CV}) = 0.2689$ and ave$(h_{GCV}) = 0.2664$.

h	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}
0.1	0.068	-0.19	0.16	-0.028	0.026	-0.054	-0.021	0.0097	0.087	-0.22
0.15	0.068	0.12	0.27	0.37	0.41	0.47	0.53	0.62	0.62	-0.22
0.2	0.068	0.17	0.28	0.43	0.47	0.54	0.61	0.71	0.69	-0.22
0.25	-0.070	0.26	0.63	0.87	1.1	1.2	1.4	1.5	0.97	-0.46
0.3	-0.093	0.29	0.76	1.0	1.3	1.5	1.7	1.8	1.0	-0.49
0.35	-0.24	0.32	1.0	1.6	2.0	2.4	2.6	2.3	1.1	-0.93
0.4	-0.29	0.32	1.1	1.9	2.4	2.8	3.1	2.5	1.1	-1.1
CV	-0.12	0.20	0.67	0.95	1.20	1.40	1.60	1.40	0.82	-0.54
GCV	-0.12	0.20	0.66	0.93	1.10	1.40	1.50	1.40	0.81	-0.53

Table 3. Simulation bias (in percent) $B(\hat{m}_{OLR}(t,\lambda)) = \mathrm{E}(\hat{m}_{OLR}(t,\lambda) - m_t)$ for pattern 1 for the 10 values of the ordinal covariate for 7 fixed bandwidth as well as at bandwidths selected by CV and GCV, with ave$(\lambda_{CV}) = 2.50e-8$ and ave$(\lambda_{GCV}) = 2.29e-8$.

λ	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}
0	0.068	-0.19	0.16	-0.028	0.026	-0.054	-0.021	0.0097	0.087	-0.22
1e-13	-0.759	-0.203	0.185	0.0538	0.109	0.0552	0.0961	0.138	0.208	-0.0688
1e-12	-0.987	-0.231	0.189	0.0764	0.133	0.0869	0.13	0.176	0.246	-0.025
1e-11	-1.28	-0.281	0.192	0.106	0.166	0.129	0.177	0.227	0.299	0.0332
1e-10	-1.67	-0.365	0.19	0.143	0.211	0.187	0.24	0.297	0.371	0.112
1e-09	-2.18	-0.504	0.177	0.191	0.274	0.266	0.328	0.397	0.475	0.22
1e-08	-2.86	-0.735	0.138	0.249	0.361	0.379	0.456	0.541	0.626	0.372
1e-07	-3.79	-1.11	0.044	0.314	0.485	0.545	0.648	0.759	0.853	0.596
CV	-1.96	-0.513	0.146	0.163	0.243	0.235	0.319	0.372	0.459	0.175
GCV	-1.86	-0.482	0.144	0.155	0.229	0.227	0.297	0.355	0.436	0.158

Table 4. Simulation standard deviation $SD(\hat{m}_{t,h})$ for pattern 1 for the 10 values of the ordinal covariate for 7 fixed bandwidth as well as at bandwidths selected by CV and GCV, with $\text{ave}(h_{CV}) = 0.2689$ and $\text{ave}(h_{GCV}) = 0.2664$.

h	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}
0.1	0.035	0.035	0.034	0.034	0.033	0.033	0.032	0.032	0.032	0.032
0.15	0.035	0.021	0.021	0.020	0.020	0.020	0.020	0.019	0.019	0.032
0.2	0.035	0.020	0.020	0.020	0.019	0.019	0.019	0.019	0.019	0.032
0.25	0.032	0.019	0.016	0.016	0.016	0.016	0.016	0.015	0.018	0.029
0.3	0.032	0.019	0.016	0.016	0.015	0.015	0.015	0.015	0.017	0.029
0.35	0.030	0.019	0.015	0.014	0.014	0.014	0.013	0.014	0.017	0.027
0.4	0.029	0.019	0.015	0.013	0.013	0.013	0.013	0.014	0.017	0.027
CV	0.032	0.022	0.021	0.020	0.020	0.019	0.019	0.019	0.021	0.030
GCV	0.032	0.023	0.021	0.020	0.020	0.020	0.020	0.020	0.021	0.030

Table 5. Simulation standard deviation $SD(\hat{m}_{OLR}(t, \lambda))$ for pattern 1 for the 10 values of the ordinal covariate for 7 fixed bandwidth as well as at bandwidths selected by CV and GCV, with $\text{ave}(\lambda_{CV}) = 2.50e-8$ and $\text{ave}(\lambda_{GCV}) = 2.29e-8$.

λ	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}
0	0.035	0.035	0.034	0.034	0.033	0.033	0.032	0.032	0.032	0.032
1e-13	0.032	0.032	0.031	0.03	0.03	0.03	0.029	0.03	0.028	0.03
1e-12	0.032	0.031	0.03	0.029	0.029	0.03	0.028	0.029	0.027	0.03
1e-11	0.032	0.03	0.029	0.028	0.029	0.029	0.027	0.028	0.027	0.03
1e-10	0.031	0.029	0.028	0.027	0.028	0.028	0.026	0.027	0.026	0.029
1e-09	0.03	0.028	0.027	0.026	0.026	0.026	0.025	0.026	0.025	0.028
1e-08	0.029	0.027	0.025	0.025	0.025	0.025	0.024	0.024	0.024	0.027
1e-07	0.028	0.025	0.024	0.023	0.023	0.023	0.022	0.023	0.022	0.026
CV	0.032	0.029	0.028	0.027	0.027	0.027	0.026	0.027	0.026	0.029
GCV	0.033	0.03	0.029	0.027	0.028	0.028	0.026	0.027	0.026	0.029

Table 6. Simulation square root of mean squared errors $MSE(\hat{m}_{t,h})$ and square root of mean sum of squared errors $MSSE_h = \mathrm{E}\sum_t(\hat{m}_{t,h} - m_t)^2$ for pattern 1 for the 10 values of the ordinal covariate for 7 fixed bandwidth as well as at bandwidths selected by CV and GCV, with ave(h_{CV}) = 0.2689 and ave(h_{GCV}) = 0.2664.

h	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}	$\sqrt{MSSE_h}$
0.1	0.035	0.035	0.034	0.034	0.033	0.033	0.032	0.032	0.032	0.032	0.1
0.15	0.035	0.021	0.021	0.021	0.021	0.02	0.02	0.02	0.02	0.032	0.074
0.2	0.035	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.032	0.073
0.25	0.032	0.019	0.018	0.018	0.019	0.02	0.021	0.022	0.02	0.03	0.07
0.3	0.032	0.019	0.017	0.019	0.02	0.021	0.022	0.023	0.02	0.029	0.072
0.35	0.03	0.019	0.018	0.021	0.024	0.027	0.03	0.027	0.021	0.029	0.078
0.4	0.029	0.019	0.019	0.023	0.027	0.031	0.034	0.028	0.021	0.029	0.083
CV	0.032	0.022	0.022	0.022	0.023	0.024	0.025	0.024	0.022	0.03	0.082
GCV	0.032	0.023	0.022	0.022	0.023	0.024	0.025	0.024	0.022	0.03	0.082

Table 7. Simulation square root of mean squared errors $MSE(\hat{m}_{OLR}(t, \lambda))$ and square root of mean sum of squared errors $MSSE_h = \mathrm{E}\sum_t(\hat{m}_{OLR}(t, \lambda) - m_t)^2$ for pattern 1 for the 10 values of the ordinal covariate for 7 fixed bandwidth as well as at bandwidths selected by CV and GCV, with ave(λ_{CV}) = $2.50e - 8$ and ave(λ_{GCV}) = $2.29e - 8$.

λ	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}	$\sqrt{MSSE_h}$
0	0.035	0.035	0.034	0.034	0.033	0.033	0.032	0.032	0.032	0.032	0.105
1e-13	0.033	0.032	0.031	0.03	0.03	0.03	0.029	0.03	0.028	0.03	0.0963
1e-12	0.034	0.031	0.03	0.029	0.03	0.03	0.028	0.029	0.027	0.03	0.0944
1e-11	0.034	0.031	0.03	0.028	0.029	0.029	0.027	0.028	0.027	0.03	0.0923
1e-10	0.035	0.03	0.028	0.027	0.028	0.028	0.026	0.027	0.026	0.029	0.09
1e-09	0.037	0.028	0.027	0.026	0.026	0.026	0.025	0.026	0.025	0.028	0.088
1e-08	0.041	0.028	0.026	0.025	0.025	0.025	0.024	0.025	0.024	0.028	0.0865
1e-07	0.047	0.027	0.024	0.023	0.024	0.024	0.023	0.024	0.024	0.027	0.087
CV	0.038	0.03	0.028	0.027	0.028	0.027	0.026	0.027	0.026	0.029	0.091
GCV	0.038	0.03	0.029	0.027	0.028	0.028	0.026	0.027	0.026	0.029	0.092

TABLE 8. Simulation relative bias (in percent) of variance estimator $\mathrm{E}\left(\frac{\hat{V}(\hat{m}_{t,h})-\mathrm{Var}(\hat{m}_{t,h})}{\mathrm{Var}(\hat{m}_{t,h})}\right)$ for pattern 1 for the 10 values of the ordinal covariate for 7 fixed bandwidth as well as at bandwidths selected by CV and GCV, with ave(h_{CV}) = 0.2689 and ave(h_{GCV}) = 0.2664.

h	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}
0.1	-2.30	-1.42	0.496	1.21	1.68	1.98	-0.551	-1.81	1.15	-1.13
0.15	-2.30	-1.23	0.257	1.16	1.64	1.45	-0.289	-1.01	0.143	-1.13
0.2	-2.30	-1.16	0.164	1.14	1.63	1.25	-0.188	-0.701	-0.245	-1.13
0.25	-2.20	-1.45	0.0792	1.04	1.47	1.03	0.00457	-0.480	-0.334	-0.996
0.3	-2.16	-1.50	0.0162	0.947	1.31	0.87	0.172	-0.331	-0.397	-0.943
0.35	-2.1	-1.55	-0.265	0.802	1.11	0.775	0.266	-0.284	-0.431	-0.744
0.4	-2.04	-1.56	-0.38	0.648	0.904	0.723	0.315	-0.257	-0.453	-0.684
CV	-2.24	-2.00	0.25	0.58	1.61	0.55	0.45	-0.64	-0.29	-1.18
GCV	-2.19	-1.91	-0.64	0.75	1.38	0.13	0.48	-0.27	0.75	-1.17

TABLE 9. Simulation bias (in percent) $B(\hat{m}_{t,h}) = \mathrm{E}(\hat{m}_{t,h} - m_t)$ for pattern 2 for the 10 values of the ordinal covariate for 7 fixed bandwidth as well as at bandwidths selected by CV and GCV, with ave(h_{CV}) = 0.1674 and ave(h_{GCV}) = 0.1703.

h	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}
0.1	0.068	-0.16	0.057	0.07	-0.033	-0.039	0.062	0.058	-0.36	-0.18
0.15	0.068	0.1	0.25	0.38	0.31	0.46	0.57	0.59	0.43	-0.18
0.2	0.068	0.14	0.28	0.42	0.37	0.54	0.65	0.66	0.54	-0.18
0.25	-0.079	0.23	0.62	0.74	0.76	0.98	1.4	1.6	0.9	-0.53
0.3	-0.1	0.25	0.74	0.86	0.92	1.2	1.7	1.9	0.97	-0.59
0.35	-0.32	0.23	0.87	1.2	1.3	1.7	2.5	2.4	0.98	-1.2
0.4	-0.38	0.22	0.94	1.4	1.6	2.1	3	2.5	0.98	-1.3
CV	-0.011	-0.048	0.25	0.39	0.33	0.47	0.7	0.67	0.26	-0.51
GCV	-0.025	-0.037	0.25	0.41	0.34	0.49	0.72	0.7	0.25	-0.51

Table 10. Simulation bias (in percent) $B(\hat{m}_{OLR}(t, \lambda)) = \mathrm{E}(\hat{m}_{OLR}(t, \lambda) - m_t)$ for pattern 2 for the 10 values of the ordinal covariate for 7 fixed bandwidth as well as at bandwidths selected by CV and GCV, with ave$(\lambda_{CV}) = 1.17e-8$ and ave$(\lambda_{GCV}) = 1.13e-8$.

λ	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}
0	0.068	-0.16	0.057	0.07	-0.033	-0.039	0.062	0.058	-0.36	-0.18
1e-13	-0.93	-0.441	-0.136	-0.0835	0.219	0.774	0.667	0.386	-0.175	-0.0467
1e-12	-1.21	-0.538	-0.194	-0.112	0.291	0.979	0.849	0.493	-0.114	-0.00731
1e-11	-1.57	-0.675	-0.272	-0.143	0.384	1.23	1.09	0.642	-0.028	0.0467
1e-10	-2.04	-0.87	-0.377	-0.172	0.508	1.55	1.41	0.854	0.0969	0.123
1e-09	-2.67	-1.15	-0.521	-0.195	0.674	1.95	1.85	1.16	0.284	0.235
1e-08	-3.52	-1.57	-0.721	-0.206	0.899	2.45	2.44	1.62	0.575	0.407
1e-07	-4.68	-2.18	-1	-0.192	1.21	3.09	3.25	2.32	1.05	0.693
CV	-1.59	-0.734	-0.309	-0.0995	0.391	1.22	1.11	0.737	0.074	0.0366
GCV	-1.57	-0.727	-0.303	-0.0989	0.386	1.2	1.1	0.724	0.0662	0.0346

Table 11. Simulation standard deviation $SD(\hat{m}_{t,h})$ for pattern 2 for the 10 values of the ordinal covariate for 7 fixed bandwidth as well as at bandwidths selected by CV and GCV, with ave$(h_{CV}) = 0.1674$ and ave$(h_{GCV}) = 0.1703$.

h	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}
0.1	0.035	0.032	0.029	0.027	0.024	0.033	0.042	0.051	0.05	0.05
0.15	0.035	0.019	0.018	0.016	0.016	0.02	0.026	0.029	0.03	0.05
0.2	0.035	0.019	0.017	0.015	0.016	0.02	0.025	0.028	0.029	0.05
0.25	0.031	0.018	0.014	0.013	0.014	0.017	0.02	0.022	0.028	0.046
0.3	0.031	0.018	0.014	0.013	0.014	0.016	0.019	0.021	0.027	0.046
0.35	0.028	0.018	0.013	0.012	0.013	0.015	0.017	0.02	0.027	0.042
0.4	0.028	0.018	0.013	0.012	0.013	0.014	0.016	0.02	0.027	0.042
CV	0.033	0.027	0.025	0.023	0.021	0.028	0.035	0.042	0.043	0.048
GCV	0.033	0.027	0.025	0.023	0.021	0.028	0.035	0.042	0.043	0.048

Table 12. Simulation standard deviation $SD(\hat{m}_{OLR}(t,\lambda))$ for pattern 2 for the 10 values of the ordinal covariate for 7 fixed bandwidth as well as at bandwidths selected by CV and GCV, with ave(λ_{CV}) $= 1.17e - 8$ and ave(λ_{GCV}) $= 1.13e - 8$.

λ	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}
0	0.035	0.032	0.029	0.027	0.024	0.033	0.042	0.051	0.05	0.05
1e-13	0.032	0.029	0.027	0.024	0.021	0.03	0.037	0.044	0.046	0.048
1e-12	0.032	0.028	0.026	0.024	0.021	0.029	0.036	0.042	0.045	0.047
1e-11	0.031	0.027	0.025	0.023	0.021	0.028	0.035	0.041	0.043	0.046
1e-10	0.03	0.026	0.024	0.022	0.02	0.026	0.033	0.039	0.042	0.045
1e-09	0.029	0.025	0.023	0.021	0.019	0.025	0.031	0.037	0.04	0.044
1e-08	0.028	0.024	0.022	0.02	0.019	0.023	0.029	0.034	0.038	0.043
1e-07	0.026	0.022	0.02	0.019	0.018	0.022	0.026	0.031	0.035	0.041
CV	0.034	0.028	0.026	0.023	0.021	0.029	0.037	0.042	0.044	0.047
GCV	0.034	0.028	0.026	0.023	0.021	0.029	0.037	0.042	0.044	0.047

Table 13. Simulation square root of mean squared errors $MSE(\hat{m}_{t,h})$ and square root of mean sum of squared errors $MSSE_h = \mathrm{E}\sum_t(\hat{m}_{t,h} - m_t)^2$ for pattern 2 for the 10 values of the ordinal covariate for 7 fixed bandwidth as well as at bandwidths selected by CV and GCV, with ave(h_{CV}) $= 0.1674$ and ave(h_{GCV}) $= 0.1703$.

h	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}	$\sqrt{MSSE_h}$
0.1	0.035	0.032	0.029	0.027	0.024	0.033	0.042	0.051	0.05	0.05	0.12
0.15	0.035	0.019	0.018	0.016	0.017	0.021	0.026	0.029	0.031	0.05	0.088
0.2	0.035	0.019	0.017	0.016	0.016	0.02	0.026	0.029	0.03	0.05	0.087
0.25	0.031	0.018	0.015	0.015	0.016	0.02	0.024	0.027	0.029	0.046	0.081
0.3	0.031	0.018	0.015	0.016	0.017	0.02	0.025	0.029	0.029	0.046	0.082
0.35	0.029	0.018	0.016	0.017	0.019	0.023	0.03	0.031	0.029	0.044	0.085
0.4	0.028	0.018	0.016	0.018	0.02	0.025	0.034	0.032	0.029	0.044	0.087
CV	0.033	0.027	0.025	0.023	0.021	0.028	0.036	0.043	0.043	0.049	0.11
GCV	0.033	0.027	0.025	0.023	0.021	0.028	0.036	0.043	0.043	0.048	0.11

TABLE 14. Simulation square root of mean squared errors $MSE(\hat{m}_{OLR}(t, \lambda))$ and square root of mean sum of squared errors $MSSE_h = \mathrm{E}\sum_t(\hat{m}_{OLR}(t, \lambda) - m_t)^2$ for pattern 2 for the 10 values of the ordinal covariate for 7 fixed bandwidth as well as at bandwidths selected by CV and GCV, with ave$(\lambda_{CV}) = 1.17e - 8$ and ave$(\lambda_{GCV}) = 1.13e - 8$.

λ	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}	$\sqrt{MSSE_h}$
0	0.035	0.032	0.029	0.027	0.024	0.033	0.042	0.051	0.05	0.05	0.121
1e-13	0.033	0.029	0.027	0.024	0.021	0.031	0.038	0.044	0.046	0.048	0.111
1e-12	0.034	0.029	0.026	0.024	0.021	0.03	0.037	0.043	0.045	0.047	0.109
1e-11	0.035	0.028	0.026	0.023	0.021	0.03	0.036	0.041	0.043	0.046	0.108
1e-10	0.036	0.028	0.025	0.022	0.021	0.031	0.036	0.04	0.042	0.045	0.106
1e-09	0.04	0.028	0.024	0.021	0.021	0.032	0.036	0.038	0.04	0.044	0.105
1e-08	0.045	0.028	0.023	0.02	0.021	0.034	0.038	0.038	0.038	0.043	0.107
1e-07	0.054	0.031	0.023	0.019	0.022	0.038	0.042	0.039	0.037	0.041	0.114
CV	0.038	0.029	0.026	0.023	0.021	0.032	0.038	0.043	0.044	0.047	0.111
GCV	0.037	0.029	0.026	0.023	0.021	0.032	0.038	0.043	0.044	0.047	0.111

TABLE 15. Simulation relative bias (in percent) of variance estimator $\mathrm{E}\left(\frac{\hat{V}(\hat{m}_{t,h}) - \mathrm{Var}(\hat{m}_{t,h})}{\mathrm{Var}(\hat{m}_{t,h})}\right)$ for pattern 2 for the 10 values of the ordinal covariate for 7 fixed bandwidth as well as at bandwidths selected by CV and GCV, with ave$(h_{CV}) = 0.1674$ and ave$(h_{GCV}) = 0.1703$.

h	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}
0.1	-2.3	-1.92	2.35	1.59	0.51	-0.64	-0.0903	0.0377	-3.03	4.05e-05
0.15	-2.3	-1.35	1.3	1.6	0.438	-0.39	-0.117	-0.755	-1.87	4.05e-05
0.2	-2.3	-1.12	0.885	1.6	0.403	-0.289	-0.128	-1.03	-1.41	4.05e-05
0.25	-2.19	-1.45	0.503	1.27	0.419	-0.172	-0.418	-1.04	-1.18	-0.197
0.3	-2.15	-1.49	0.242	1.04	0.431	-0.118	-0.663	-0.969	-1.08	-0.273
0.35	-2.15	-1.44	-0.00721	0.845	0.452	-0.114	-0.678	-0.979	-1.03	-0.558
0.4	-2.08	-1.42	-0.143	0.698	0.445	-0.181	-0.673	-0.948	-0.995	-0.628
CV	-1.85	-2.89	0.298	0.134	0.0228	-1.84	1.81	0.074	-4.17	-0.265
GCV	-1.69	-3.08	0.362	0.105	-0.139	-1.73	1.84	-0.0104	-4.46	-0.107

3.5. Conclusion

In this chapter, we fitted a nonparametric regression model to data for the situation in which the covariate is an ordered categorical variable. We extended the local polynomial estimator, which normally requires continuous covariates, to a local polynomial estimator that allowed for ordered categorical covariates. We derived the asymptotic conditional bias and variance for the local polynomial estimator with ordinal covariate, under the assumption that the categories correspond to quantiles of an unobserved continuous latent variable. The form of the leading terms of our asymptotic bias and variance were exactly the same as those in Ruppert and Wand (1994). The difference is that the position of the target and the values of X_i are known in the case of local polynomial estimator with an continuous covariate. The conditional bias caused by the uncertain values of X_i and the target's position are in a smaller order in probability than the leading terms and are absorbed in the remnant terms. Similar thing happened to the conditional variance. We conducted two patterns of simulations and the results were in accordance with the asymptotic properties we had proved. We also compared our simulation results to these using the estimator proposed in Ouyang, Li, and Racine (2009), and our estimator had smaller variance and mean squared errors at most of the targets using the bandwidths selected by CV and GCV.

CHAPTER 4

Generalized Product Kernel Smoothing

4.1. Introduction

Multivariate nonparametric regression has been proved to be very useful in practice. Stone (1980) and Stone (1982) have shown that the local regression estimators having optimal rates of convergence, and Cleveland and Devlin (1988) have proved that they are very useful in modelling data. Ruppert and Wand (1994) derived the asymptotic properties of the multivariate local linear and local quadratic estimators. No systematic study of bandwidth matrix choice has been made for these estimators until Wand and Jones (1994) studied multivariate plug-in bandwidth selection and Herrmann, Wand, Engel, and Gasser (1995) proposed plug-in approaches for bivariate convolution kernel estimator.

Eubank (1988, p.286-292), Wahba (1990, p.30-39), and Green and Silverman (1994, Chapter 7) described thin plate smoothing splines and gave earlier references to the method. Green and Silverman (1994, p.155-159) also discussed the possibility of constructing multivariate regression estimators based on univariate splines using tensor products, and Simonoff (1996, Chapter 5) gave a overall description on nonparametric regression.

However, these works are based on the assumption that the covariates in the model are continuous. Much less efforts have been contributed to the situations that there are categorical (nominal and/or ordinal) covariates in the regression model. Bierens (1983) began the consideration of kernel regression with mixed continuous and categorical covariates. Li and Racine (2004), Racine and Li (2004), Hall, Racine, and Li (2004), Hall, Li, and Racine

(2007), Li and Racine (2008), and Ouyang, Li, and Racine (2009) have considered nonparametric estimation of regression functions, conditional density, and distribution functions, and quantile functions containing a mix of categorical and continuous covariates.

In this paper, we consider a Nadaraya-Watson (NW) estimator with product kernel, allowing for both categorical and continuous covariates. For conciseness, we simplify this problem by assuming that there is a continuous response (Y) and three covariates in this problem. The three covariates are continuous (X) ordinal (C) and nominal (D), respectively. We will propose an estimator for this problem in Section 4.2, derive the asymptotic properties of the estimator in Section 4.3, and conduct a simulation study in Section 4.4.

4.2. Proposed Estimator

Let $(X_1, C_1, D_1, Y_1), \ldots, (X_n, C_n, D_n, Y_n)$ be a set of independent and identically distributed (i.i.d.) 4-dimensional random vectors, where the Y_i are scalar response variables and X_i, C_i and D_i are three univariate covariates. We only consider one univariate covariate of each type for simplicity, but the approach generalizes to higher dimensions. The first covariate X_i is continuous. The ordinal covariate C_i takes values in $(1, 2, \ldots, M)$, where lower order of C_i has smaller value. For example, it could be a typical five-level Likert item such that C_i takes values in 1 to 5, 1 for strongly disagree, 2 for disagree, 3 for neither disagree nor agree, 4 for agree, and 5 for strongly agree (Likert 1932). There exists a latent continuous covariate Z_i such that (X_i, Z_i) are $\mathbb{R}^2$-valued covariates having common density $f_{X,Z}$ with bounded support supp($f_{X,Z}$) and their own bounded marginal support supp(X) and supp(Z). We can be more specific: we assume Z is on [0,1], we create a grid of boundary points, and we assume that there Z has a density such that $P(Z \in U_j) = P(C = j)$, where $U_j = [\frac{j-1}{M}, \frac{1}{M}]$, $j = 1, 2, \ldots, M$. The third covariate D_i is a nominal variable and for simplicity again, we will assume that there are only two possible values. There is no order of the two categories and their values can be literal, for example, (New York City, Boston), (Dogs, Cats), etc. The random vector (X_i, Z_i, D_i) has a joint distribution. The conditional distribution of D_i given X_i and C_i, i.e. $D_i|X_i, C_i$ follows a Bernoulli distribution where

$$P(D_i = d_1|X_i, C_i) = \sum_{j=1}^{M} I_{\{Z_i \in U_j\}} \int_{U_j} p(X_i, z) f_{X,Z}(X_i, z)\,\mathrm{d}z,$$

$$P(D_i = d_2|X_i, C_i) = 1 - P(D_i = d_1|X_i, C_i),$$

where d_1 and d_2 are the 2 categories.

We plan to use a product kernel, i.e. we will smooth local values around the target point, where "local" will be defined by kernels in each dimension. Since the variables are of different types, different types of kernels will be used.

For the first covariate, i.e. X, we will use a kernel function $K_{1,h_1}(\cdot)$ with bandwidth h_1 such that $K_{1,h_1}(\cdot) = \frac{1}{h_1} K_1(\frac{\cdot}{h_1})$, and K_1 has compact support $[-1, 1]$, is symmetric and continuous, and satisfies $\int_{-1}^{1} K_1(u_1) du_1 = 1$.

For the second covariate, i.e. C, we will use a kernel function $K_{2,h_2}(\cdot)$ with bandwidth h_2 such that $K_{2,h_2}(\cdot) = \frac{1}{h_2} K_2(\frac{\cdot}{h_2})$, and K_2 has compact support $[-1, 1]$, is symmetric and continuous, and satisfies $\int_{-1}^{1} K_2(u_2) du_2 = 1$. While we will apply this kernel to the ordinal covariate in a manner that is similar to the continuous covariate, the interpretation will be different, since it will rely on the latent variable underlying the ordinal covariate. We have done an in-depth study of this approach in Chapter 3. This will be made clearer below.

For these first two kernels, we define the moments of $K_{l,h_i}(\cdot)$ as

$$\mu_j(K_l) = \int_{-1}^{1} z_i^j K_l(z_i) dz_i,\ l = 1, 2,\ j = 0, 1, 2, \ldots,$$

and

$$R_j(K_l) = \int_{-1}^{1} z_i^j K_l^2(z_i) dz_i, \; l = 1, 2, \; j = 0, 1, 2, \ldots.$$

For the third covariate, i.e. D, with a slight abuse of notation, we define a kernel function $K_{3,\lambda}(D_i, d)$ with penalty parameter λ such that

$$K_{3,\lambda}(D_i, d) = \begin{cases} 1 & \text{if } D_i = d \\ \\ \lambda & \text{otherwise,} \end{cases}$$

$i = 1, \ldots, n$ and $d \in (d_1, d_2)$ (Ouyang, Li, and Racine 2009, p.3). For example, if the two categories are different pets, like (Dogs, Cats), and d is Cats, then if D_i is Dogs, $K_{3,\lambda}(D_i, d) = \lambda$. If D_i is Cats, $K_{3,\lambda}(D_i, d) = 1$.

$$m_{\boldsymbol{\psi}} = \mathrm{E}(Y_i | (X_i, C_i, D_i)^T = \boldsymbol{\psi})$$

We will assume the model

$$\begin{aligned} Y_i &= m(X_i, Z_i, D_i) + \epsilon_i \\ &= \begin{cases} m_1(X_i, Z_i) + \epsilon_i & \text{if } D_i = d_1 \\ \\ m_2(X_i, Z_i) + \epsilon_i & \text{if } D_i = d_2, \end{cases} \end{aligned}$$

$i = 1, \ldots, n$, where the ϵ_i are i.i.d with zero mean and σ^2 variance and are independent of X_i, Z_i and D_i.

While we are not currently specifying anything about how m_1 and m_2 are related to each other, we are assuming that m_1 and m_2 are similar and not far apart here. For an simple example, $m_2(X_i, Z_i)$ can be $m_1(X_i, Z_i) + \delta$, where δ is a finite real number. But the method will be applicable more broadly than to just an intercept shift between m_1 and m_2.

The multivariate nonparametric regression problem is that of estimating

$$\begin{aligned} m_{\boldsymbol{\psi}} &= \mathrm{E}(Y_i|(X_i, C_i, D_i)^T = \boldsymbol{\psi}) \\ &= \mathrm{E}(Y_i|(X_i = x, Z_i \in U_t, D_i = d)) \\ &= \mathrm{E}(m(X_i, Z_i, D_i)|(X_i = x, Z_i \in U_t, D_i = d)). \end{aligned}$$

When $d = d_1$,

$$\begin{aligned} m_{\boldsymbol{\psi}} &= \mathrm{E}\left(m_1(X_i, Z_i)|(X_i = x, Z_i \in U_t)\right) \\ &= \frac{\int_{U_t} m_1(x, v) f_{X,Z}(x, v)\,\mathrm{d}v}{\int_{U_t} f_{X,Z}(x, v)\,\mathrm{d}v} \\ &= \frac{m_1(x, z_{1,t}) \int_{U_t} f_{X,Z}(x, v)\,\mathrm{d}v}{\int_{U_t} f_{X,Z}(x, v)\,\mathrm{d}v} \\ &= m_1(x, z_{1,t}), \end{aligned}$$

and similarly when $d = d_2$,

$$m_{\boldsymbol{\psi}} = m_2(x, z_{2,t}),$$

at a vector $\boldsymbol{\psi} = (x, t, d)^T$ without imposing that m belongs to a parametric family of functions for some $z_{1,t}\, z_{2,t} \in U_t$. But since $z_{1,t}$ and $z_{2,t}$ are unknown. Next, we will estimate $m_1(x, z_{1,t})$ and $m_2(x, z_{2,t})$ at $m_{\boldsymbol{\psi}}$, which is equal to $m_1(x, \phi_t)$ when $d = d_1$ and $m_2(x, \phi_t)$ when $d = d_2$, where $\phi_t = \frac{t-0.5}{M}$. We will use a NW estimator with product kernel.

Let $W_{\boldsymbol{\psi},i} = (K_{1,h_1}(X_i - x)) \left(\sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t) I_{\{Z_i \in U_j\}} \right) (K_{3,\lambda}(D_i, d))$, $i = 1, \ldots, n$, our NW estimator with product kernel for $m_{\boldsymbol{\psi}}$ is

$$\hat{m}(\boldsymbol{\psi}; h_1, h_2, \lambda) = \frac{\frac{1}{n}\sum_{i=1}^{n} W_{\boldsymbol{\psi},i} Y_i}{\frac{1}{n}\sum_{i=1}^{n} W_{\boldsymbol{\psi},i}}.$$

4.3. Conditional Mean and Variance Properties

In this section we investigate the asymptotic properties of the conditional bias and variance of the NW estimator with product kernel.

We make the following assumptions:

A11. *The second-order partial derivatives of m_q with respect to X and Z are continuous for $q = 1, 2$.*

A12. *The point (x, ϕ_t) is in the interior of supp(X, Z). The second-order partial derivatives of $f_{X,Z}$ with respect to X and Z are continuous and $f_{X,Z}(X, Z) > 0$ in supp(X, Z). The second-order partial derivatives of $p(X, Z)$ with respect to X and Z are continuous.*

A13. *For the kernel function K_l, $\mu_j(K_l) = 0$ for all nonnegative integer j such that it is odd, $\mu_j(K_l) \neq 0$ for all nonnegative integer j such that it is even, and $R_j(K_l) > 0$ for all nonnegative integer j, $l = 1, 2$.*

A14. *The bandwidths $h_1, h_2, \lambda \to 0$, $h_1^3 h_2^3 n$, $h_1^2 M$, $h_2^2 M$, and $h_1 h_2 M \to \infty$. We also assume that $0 \leq \lambda \leq 1$.*

We use the following lemma to prove Theorem 4.3.2.

LEMMA 4.3.1. *Suppose A12-A14 hold. Let*

$$A_{a,b,i}(x,\phi_t,d) = (K_{1,h_1}(X_i - x)(X_i - x)^a)\left(\sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t)(\phi_j - \phi_t)^b I_{\{Z_i \in U_j\}}\right)$$

$$\times \left(K_{3,\lambda}(D_i, d)\right),$$

where a and b are arbitrary nonnegative integers,

$$G_1(r,s) = p(r,s) f_{X,Z}(r,s),$$

and

$$G_2(r,s) = (1 - p(r,s))\, f_{X,Z}(r,s).$$

Then, when $d = d_q$, for a, b even,

$$\frac{1}{n}\sum_{i=1}^{n} A_{a,b,i}(x,\phi_t,d) = h_1^a h_2^b \mu_a(K_1)\mu_b(K_2) G_q(x,\phi_t) + o_p(h_1^a h_2^b), \tag{4}$$

for a even and b odd,

$$\frac{1}{n}\sum_{i=1}^{n} A_{a,b,i}(x,\phi_t,d) = h_1^a h_2^{b+1} \mu_a(K_2)\mu_{b+1}(K_2) \frac{\partial}{\partial s} G_q(x,\phi_t) + o_p(h_1^a h_2^{b+1}), \tag{5}$$

for a odd and b even,

$$\frac{1}{n}\sum_{i=1}^{n} A_{a,b,i}(x,\phi_t,d) = h_1^{a+1} h_2^b \mu_{a+1}(K_1)\mu_b(K_2) \frac{\partial}{\partial r} G_q(x,\phi_t) + o_p(h_1^{a+1} h_2^b), \tag{6}$$

for a odd and b odd,

$$(7) \quad \frac{1}{n}\sum_{i=1}^{n} A_{a,b,i}(x,\phi_t,d) = h_1^{a+1}h_2^{b+1}\mu_{a+1}(K_2)\mu_{b+1}(K_2)\frac{\partial^2}{\partial r\partial s}G_q(x,\phi_t) + o_p(h_1^{a+1}h_2^{b+1}),$$

and for all a and b,

$$(8) \quad \frac{1}{n}\sum_{i=1}^{n} A^2_{a,b,i}(x,\phi_t,d) = h_1^{2a-1}h_2^{2b-1}R_{2a}(K_1)R_{2b}(K_2)G_q(x,\phi_t) + o_p(h_1^{2a-1}h_2^{2b-1}),\ q = 1,2.$$

Remark 1. Notice that although the smoothing parameter λ appears in $A_{a,b,i}(x,\phi_t,d)$, it is not involved in the leading terms and is absorbed in the remnant terms. This can be found clearly in the proof below.

Proof. By definition, $\frac{1}{n}\sum_{i=1}^{n} A_{a,b,i}(x,\phi_t,d)$ is

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} A_{a,b,i}(x,\phi_t,d) &= \frac{1}{n}\sum_{i=1}^{n}\left(K_{1,h_1}(X_i-x)(X_i-x)^a\right)\\ &\qquad \times\left(\sum_{j=1}^{M} K_{2,h_2}(\phi_j-\phi_t)(\phi_j-\phi_t)^b I_{\{Z_i\in U_j\}}\right)\left(K_{3,\lambda}(D_i,d)\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(K_{1,h_1}(X_i-x)(X_i-x)^a\right)\left(\sum_{j=1}^{M} K_{2,h_2}(\phi_j-\phi_t)(\phi_j-\phi_t)^b I_{\{Z_i\in U_j\}}\right) I_{\{D_i=d\}}\\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\lambda\left(K_{1,h_1}(X_i-x)(X_i-x)^a\right)\left(\sum_{j=1}^{M} K_{2,h_2}(\phi_j-\phi_t)(\phi_j-\phi_t)^b I_{\{Z_i\in U_j\}}\right) I_{\{D_i\neq d\}}\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(A_{1,a,b,i}(x,\phi_t,d)+\lambda A_{2,a,b,i}(x,\phi_t,d)\right).\end{aligned}$$

Since

$$\frac{1}{n}\sum_{i=1}^{n} A_{a,b,i}(x,\phi_t,d) = \mathrm{E}(A_{a,b,i}(x,\phi_t,d)) + O_p(\frac{\mathrm{Var}(A_{a,b,i}(x,\phi_t,d))}{\sqrt{n}}),$$

when $d = d_1$, the expectation of $A_{a,b,i}(x, \phi_t, d)$ is

$$\begin{aligned}
\mathrm{E}(A_{a,b,i}(x, \phi_t, d)) &= \mathrm{E}\left(\mathrm{E}(A_{1,a,b,i}(x, \phi_t, d) + \lambda A_{2,a,b,i}(x, \phi_t, d)|X_i, C_i)\right) \\
&= \mathrm{E}\Bigg(P(D_i = d_1|X_i, C_i) K_{1,h_1}(X_i - x)(X_i - x)^a \\
&\quad \times \sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t)(\phi_j - \phi_t)^b I_{\{Z_i \in U_j\}} \Bigg) \\
&\quad + \lambda \mathrm{E}\Bigg(\left(1 - P(D_i = d_1|X_i, C_i)\right) K_{1,h_1}(X_i - x)(X_i - x)^a \\
&\quad \times \sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t)(\phi_j - \phi_t)^b I_{\{Z_i \in U_j\}} \Bigg) \\
&= E_1 + E_2.
\end{aligned}$$

The first part can be derived as

$$\begin{aligned}
E_1 &= \mathrm{E}\Bigg(P(D_i = d_1|X_i, C_i) K_{1,h_1}(X_i - x)(X_i - x)^a \\
&\quad \times \sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t)(\phi_j - \phi_t)^b I_{\{Z_i \in U_j\}} \Bigg) \\
&= \mathrm{E}\Bigg(\mathrm{E}\Bigg(P(D_i = d_1|X_i, C_i) K_{1,h_1}(X_i - x)(X_i - x)^a \\
&\quad \times \sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t)(\phi_j - \phi_t)^b I_{\{Z_i \in U_j\}} \Bigg) |X_i \Bigg) \\
&= \mathrm{E}\Bigg(K_{1,h_1}(X_i - x)(X_i - x)^a \\
&\quad \times \sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t)(\phi_j - \phi_t)^b \mathrm{E}\Bigg(P(D_i = d_1|X_i, C_i) I_{\{Z_i \in U_j\}} |X_i \Bigg) \Bigg)
\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}\Bigg(K_{1,h_1}(X_i - x)(X_i - x)^a \\
&\quad \times \sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t)(\phi_j - \phi_t)^b \int_{Z \in U_j} p(X_i, Z) f_{X,Z|X}(X_i, Z)\,\mathrm{d}Z \Bigg) \\
&= \mathrm{E}\Bigg(K_{1,h_1}(X_i - x)(X_i - x)^a \\
&\quad \times \left(\frac{1}{h_2} \int K_2\left(\frac{Z - \phi_t}{h_2}\right)(Z - \phi_t)^b \right. \\
&\quad \left. \times\, p(X_i, Z) \frac{f_{X,Z}(X_i, Z)}{f_X(X_i)}\,\mathrm{d}Z + O(\frac{1}{M}) \right) \Bigg).
\end{aligned}
$$

Let

$$v = \frac{Z - \phi_t}{h_2},$$

then Z can be written as

$$Z = vh_2 + \phi_t.$$

After changing variable, for a even and b even, E_1 is

$$
\begin{aligned}
E_1 = \mathrm{E}\Bigg(& K_{1,h_1}(X_i - x)(X_i - x)^a \\
& \times \left(h_2^b \int v^b K_2(v) p\left(X_i, (vh_2 + \phi_t)\right) \right. \\
& \left. \times \frac{f_{X,Z}\left(X_i, (vh_2 + \phi_t)\right)}{f_X(X_i)}\,\mathrm{d}v + O(\frac{1}{M}) \right) \Bigg)
\end{aligned}
$$

$$
\begin{aligned}
&= \mathrm{E}\Bigg(K_{1,h_1}(X_i - x)(X_i - x)^a h_2^b \int v^b K_2(v)\,\mathrm{d}v\,(p(X_i,\phi_t) \\
&\quad \times \frac{f_{X,Z}(X_i,\phi_t)}{f_X(X_i)} + o(1)\Bigg) + O(\frac{1}{M})\Bigg) \\
&= h_2^b \mu_b(K_2)\mathrm{E}\Bigg(K_{1,h_1}(X_i - x)(X_i - x)^a p(X_i,\phi_t)\frac{f_{X,Z}(X_i,\phi_t)}{f_X(X_i)} + o(1)\Bigg) \\
&= h_2^b \mu_b(K_2)\Bigg(\frac{1}{h_1}\int K_1(\frac{X-x}{h_1})(X_i - x)^a p(X,\phi_t) \\
&\quad \times \frac{f_{X,Z}(X,\phi_t)}{f_X(X)} f_X(X)\,\mathrm{d}X + o(1)\Bigg) \\
&= h_2^b \mu_b(K_2)\Bigg(\frac{1}{h_1}\int K_1(\frac{X-x}{h_1})(X - x)^a p(X,\phi_t) f_{X,Z}(X,\phi_t)\,\mathrm{d}X + o(1)\Bigg).
\end{aligned}
$$

Notice that $O(\frac{1}{M})$ disappears because it is also $o(1)$ by A14.

Let

$$u = \frac{X-x}{h_1},$$

then X can be written as

$$X = uh_1 + x.$$

After changing variable, E_1 is

$$
\begin{aligned}
\mathrm{E}_1 &= h_2^b \mu_b(K_2)\Bigg(\int h_1^a u^a K_1(u) p(uh_1 + x,\phi_t) f_{X,Z}(uh_1 + x,\phi_t)\,\mathrm{d}u + o(1)\Bigg) \\
&= h_2^b \mu_b(K_2)\Bigg(\int u^a K(u)\,\mathrm{d}u\,(p(x,\phi_t) f_{X,Z}(x,\phi_t) + o(1)) + o(1)\Bigg)
\end{aligned}
$$

$$= h_1^a h_2^b \mu_a(K_1)\mu_b(K_2)p(x,\phi_t)f_{X,Z}(x,\phi_t) + o(h_1^a h_2^b)$$

$$= h_1^a h_2^b \mu_a(K_1)\mu_b(K_2)G_1(x,\phi_t) + o(h_1^a h_2^b).$$

Similarly, the second part is

$$\mathrm{E}_2 = h_1^a h_2^b \lambda \mu_a(K_1)\mu_b(K_2)\left(1 - p(x,\phi_t)\right) f_{X,Z}(x,\phi_t) + o(h_1^a h_2^b \lambda)$$

$$= o(h_1^a h_2^b).$$

Therefore, when a and b are even,

$$E(A_{a,b,i}(x,\phi_t,d)) = h_1^a h_2^b \mu_a(K_1)\mu_b(K_2)G_1(x,\phi_t) + o(h_1^a h_2^b)$$

holds.

For a even and b odd, after changing variable, the first part is

$$E_1 = \mathrm{E}\Bigg(K_{1,h_1}(X_i - x)(X_i - x)^a \Bigg(h_2^b \int v^b K_2(v) p\left(X_i, vh_2 + \phi_t\right) \times \frac{f_{X,Z}\left(X_i, vh_2 + \phi_t\right)}{f_X(X_i)} \,\mathrm{d}v + O(\frac{1}{M})\Bigg)\Bigg)$$

$$\begin{aligned}
&= \mathrm{E}\Bigg(K_{1,h_1}(X_i - x)(X_i - x)^a h_2^b \int v^b K(v) \left(\Bigg(p(X_i, \phi_t) \frac{f_{X,Z}(X_i, \phi_t)}{f_X(X_i)} \right. \\
&\quad + v h_2 \left(\frac{\partial p}{\partial v}(X_i, \phi_t) \frac{f_{X,Z}(X_i, \phi_t)}{f_X(X_i)} \right. \\
&\quad \left. + p(X_i, \phi_t) \frac{\frac{\partial f_{X,Z}}{\partial v}(X_i, \phi_t)}{f_X(X_i)} \right) + o(1) \Bigg) \, \mathrm{d}v + O(\frac{1}{M}) \Bigg) \Bigg) \\
&= h_2^{b+1} \mu_{b+1}(K_2) \mathrm{E}\Bigg(K_{h_1}(X_i - x)(X_i - x)^a \left(\frac{\partial p}{\partial v}(X_i, \phi_t) \frac{f_{X,Z}(X_i, \phi_t)}{f_X(X_i)} \right. \\
&\quad \left. + p(X_i, \phi_t) \frac{\frac{\partial f_{X,Z}}{\partial v}(X_i, \phi_t)}{f_X(X_i)} \right) + o(1) \Bigg) \\
&= h_2^{b+1} \mu_{b+1}(K_2) \Bigg(\frac{1}{h_1} \int K(\frac{X - x}{h_1})(X - x)^a \left(\frac{\partial p}{\partial v}(X, \phi_t) \frac{f_{X,Z}(X, \phi_t)}{f_X(X)} \right. \\
&\quad \left. + p(X, \phi_t) \frac{\frac{\partial f_{X,Z}}{\partial v}(X, \phi_t)}{f_X(X)} \right) f_X(X) \, \mathrm{d}X + o(1) \Bigg).
\end{aligned}$$

After changing variable and following the same technique used before, E_1 is

$$\begin{aligned}
E_1 &= h_1^a h_2^{b+1} \mu_a K_1 \mu_{b+1}(K_2) \Bigg(\left(\frac{\partial p}{\partial v}(x, \phi_t) f_{X,Z}(x, \phi_t) \right. \\
&\quad \left. + p(x, \phi_t) \frac{\partial f_{X,Z}}{\partial v}(x, \phi_t) \right) \Bigg) + o(h_1^a h_2^{b+1}) \\
&= h_1^a h_2^{b+1} \mu_a(K_2) \mu_{b+1}(K_2) \frac{\partial}{\partial s} G_1(x, \phi_t) + o(h_1^a h_2^{b+1}).
\end{aligned}$$

Similarly, $E_2 = o(h_1^a h_2^{b+1})$, so when a is even and b is odd,

$$E_1 = h_1^a h_2^{b+1} \mu_a(K_2) \mu_{b+1}(K_2) \frac{\partial}{\partial s} G_1(x, \phi_t) + o(h_1^a h_2^{b+1})$$

holds.

For a odd and b even, the first part is

$$E_1 = h_2^b \mu_b(K_2)\left(\frac{1}{h_1}\int K(\frac{X-x}{h_1})(X-x)^a p(X,\phi_t) f_{X,Z}(X,\phi_t)\,\mathrm{d}X + o(1)\right).$$

Let

$$u = \frac{X-x}{h_1},$$

then X can be written as

$$X = uh_1 + x.$$

After changing variable, E_1 is

$$\begin{aligned}
E_1 &= h_2^b \mu_b(K_2)\Bigg(u^a \int K(u)u^a p(x,\phi_t) f_{X,Z}(x,\phi_t) + uh_1\left(\frac{\partial p}{\partial u}(x,\phi_t) f_{X,Z}(x,\phi_t)\right. \\
&\quad \left. + p(x,\phi_t)\frac{\partial f_{X,Z}}{\partial u}(x,\phi_t) + o(1)\right)\,\mathrm{d}u + o(1)\Bigg) \\
&= h_1^{a+1}\mu_{a+1}(K_1) h_2^b \mu_b(K_2)\left(\frac{\partial p}{\partial u}(x,\phi_t) f_{X,Z}(x,\phi_t) + p(x,\phi_t)\frac{\partial f_{X,Z}}{\partial u}(x,\phi_t)\right) \\
&\quad + o(h_1^{a+1}h_2^b) \\
&= h_1^{a+1} h_2^b \mu_{a+1}(K_1)\mu_b(K_2)\frac{\partial}{\partial r} G_1(x,\phi_t) + o(h_1^{a+1}h_2^b).
\end{aligned}$$

Similarly, $E_2 = o(h_1^{a+1}h_2^b)$, so when a is odd and b is even,

$$E(A_{a,b,i}(x,\phi_t,d)) = h_1^{a+1}h_2^b \mu_{a+1}(K_1)\mu_b(K_2)\frac{\partial}{\partial r}G_1(x,\phi_t) + o(h_1^{a+1}h_2^b)$$

holds.

For a odd and b odd, the first part is

$$E_1 = h_2^{b+1}\mu_{b+1}(K_2)\left(\frac{1}{h_1}\int K(\frac{X-x}{h_1})(X-x)^a\left(\frac{\partial p}{\partial v}(X,\phi_t)\frac{f_{X,Z}(X,\phi_t)}{f_X(X)}\right.\right.$$
$$\left.\left.+p(X,\phi_t)\frac{\frac{\partial f_{X,Z}}{\partial v}(X,\phi_t)}{f_X(X)}+o(1)\right)f_X(X)\,\mathrm{d}X+o(1)\right).$$

Let

$$u=\frac{X-x}{h_1},$$

then X can be written as

$$X=uh_1+x.$$

After changing variable, E_1 is

$$\begin{aligned}E_1 &= h_1^a h_2^{b+1}\mu_{b+1}(K_2)\left(\int u^a K(u)\left(\frac{\partial p}{\partial v}(x,\phi_t)f_{X,Z}(x,\phi_t)+p(x,\phi_t)\frac{\partial f_{X,Z}}{\partial v}(x,\phi_t)\right.\right.\\ &\quad +uh_1\left(\frac{\partial^2 p}{\partial u\partial v}(x,\phi_t)f_{X,Z}(x,\phi_t)+\frac{\partial p}{\partial v}(x,\phi_t)\frac{\partial f_{X,Z}}{\partial u}(x,\phi_t)\right.\\ &\quad \left.\left.\left.+\frac{\partial p}{\partial u}(x,\phi_t)\frac{\partial f_{X,Z}}{\partial v}(x,\phi_t)+p(x,\phi_t)\frac{\partial^2 f_{X,Z}}{\partial u\partial v}(x,\phi_t)\right)+o(1)\right)\,\mathrm{d}u+o(1)\right)\\ &= h_1^{a+1}h_2^{b+1}\mu_{a+1}(K_1)\mu_{b+1}(K_2)\frac{\partial^2}{\partial r\partial s}G_1(x,\phi_t)+o(h_1^{a+1}h_2^{b+1}).\end{aligned}$$

Similarly, $E_2=o(h_1^{a+1}h_2^{b+1})$, so when a and b are odd,

$$E(A_{a,b,i}(x,\phi_t,d))=h_1^{a+1}h_2^{b+1}\mu_{a+1}(K_1)\mu_{b+1}(K_2)\frac{\partial^2}{\partial r\partial s}G_1(x,\phi_t)+o(h_1^{a+1}h_2^{b+1})$$

holds.

The expectation of $A^2_{a,b,i}(x, \phi_t, d)$ is

$$\begin{aligned}
\mathrm{E}(A^2_{a,b,i}(x, \phi_t, d)) &= \mathrm{E}\left(\mathrm{E}(A^2_{a,b,i}(x, \phi_t, d)|X_i, C_i)\right) \\
&= \mathrm{E}\Big(P(D_i = d_1|X_i, C_i)K^2_{1,h_1}(X_i - x)(X_i - x)^{2a} \\
&\quad \times \sum_{j=1}^{M} K^2_{2,h_2}(\phi_j - \phi_t)(\phi_j - \phi_t)^{2b} I_{\{Z_i \in U_j\}}\Big) \\
&\quad + \lambda \mathrm{E}\left((1 - P(D_i = d_1|X_i, C_i))\, K^2_{1,h_1}(X_i - x)(X_i - x)^{2a}\right. \\
&\quad \left.\times \sum_{j=1}^{M} K^2_{2,h_2}(\phi_j - \phi_t)(\phi_j - \phi_t)^{2b} I_{\{Z_i \in U_j\}}\right) \\
&= E_3 + E_4.
\end{aligned}$$

For the first part, E_3 can be derived as

$$\begin{aligned}
E_3 &= \mathrm{E}\Bigg(P(D_i = d_1|X_i, C_i)K^2_{1,h_1}(X_i - x)(X_i - x)^{2a} \\
&\quad \times \sum_{j=1}^{M} K^2_{2,h_2}(\phi_j - \phi_t)(\phi_j - \phi_t)^{2b} I_{\{Z_i \in U_j\}}\Bigg) \\
&= \mathrm{E}\Bigg(\mathrm{E}\Bigg(P(D_i = d_1|X_i, C_i)K^2_{1,h_1}(X_i - x)(X_i - x)^{2a} \\
&\quad \times \sum_{j=1}^{M} K^2_{2,h_2}(\phi_j - \phi_t)(\phi_j - \phi_t)^{2b} I_{\{Z_i \in U_j\}}\Bigg)|X_i\Bigg) \\
&= \mathrm{E}\Bigg(K^2_{1,h_1}(X_i - x)(X_i - x)^{2a} \\
&\quad \times \sum_{j=1}^{M} K^2_{2,h_2}(\phi_j - \phi_t)(\phi_j - \phi_t)^{2b} \mathrm{E}\Bigg(P(D_i = d_1|X_i, C_i) I_{\{Z_i \in U_j\}}|X_i\Bigg)\Bigg)
\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}\Bigg((K^2_{1,h_1}(X_i - x)(X_i - x)^{2a} \\
&\quad \times \sum_{j=1}^{M} K^2_{2,h_2}(\phi_j - \phi_t)(\phi_j - \phi_t)^{2b} \int_{Z \in U_j} p(X_i, Z) f_{X,Z|X}(X_i, Z)\, \mathrm{d}Z\Bigg) \\
&= \mathrm{E}\Bigg((K^2_{1,h_1}(X_i - x)(X_i - x)^{2a} \\
&\quad \times \left(\frac{1}{h_2^2} \int K_2^2\left(\frac{Z - \phi_t}{h_2}\right)(Z - \phi_t)^{2b}\right. \\
&\quad \left.\times\, p(X_i, Z) \frac{f_{X,Z}(X_i, Z)}{f_X(X_i)}\, \mathrm{d}Z + O(\frac{1}{M})\right)\Bigg).
\end{aligned}
$$

Let

$$
v = \frac{Z - \phi_t}{h_2},
$$

then Z can be written as

$$
Z = vh_2 + \phi_t.
$$

After changing variable, E_3 is

$$
\begin{aligned}
E_3 &= \mathrm{E}\Bigg(K^2_{1,h_1}(X_i - x)(X_i - x)^{2a} bigg(h_2^{2b-1} \int v^{2b} K_2^2(v) p\left(X_i, vh_2 + \phi_t\right) \\
&\quad \times \frac{f_{X,Z}\left(X_i, vh_2 + \phi_t\right)}{f_X(X_i)}\, \mathrm{d}v + O(\frac{1}{M})\Big)\Bigg) \\
&= \mathrm{E}\Bigg(K^2_{1,h_1}(X_i - x)(X_i - x)^{2a} h_2^{2b-1} \int v^{2b} K_2^2(v)\, \mathrm{d}v (p(X_i, \phi_t) \\
&\quad \times \frac{f_{X,Z}(X_i, \phi_t)}{f_X(X_i)} + o(1)) + O(\frac{1}{M})\Bigg)
\end{aligned}
$$

$$= h_2^{2b-1} R_{2b}(K_2) \mathrm{E}\Bigg(K_{1,h_1}(X_i - x)(X_i - x)^{2a} p(X_i, \phi_t) \frac{f_{X,Z}(X_i, \phi_t)}{f_X(X_i)} + o(1) \Bigg)$$

$$= h_2^{2b-1} R_{2b}(K_2) \Bigg(\frac{1}{h_1^2} \int K_1^2(\frac{X-x}{h_1})(X_i - x)^{2a} p(X, \phi_t)$$

$$\times \frac{f_{X,Z}(X, \phi_t)}{f_X(X)} f_X(X) \, \mathrm{d}X + o(1) \Bigg)$$

$$= h_2^{2b-1} R_{2b}(K_2) \Bigg(\frac{1}{h_1^2} \int K_1^2(\frac{X-x}{h_1})(X - x)^{2a} p(X, \phi_t) f_{X,Z}(X, \phi_t) \, \mathrm{d}X + o(1) \Bigg).$$

Let

$$u = \frac{X-x}{h_1},$$

then X can be written as

$$X = uh_1 + x.$$

After changing variable, E_3 is

$$\mathrm{E}_3 = h_2^{2b-1} R_{2b}(K_2) \Bigg(\int h_1^{2a-1} u^{2a} K_1^2(u) p(uh_1 + x, \phi_t) f_{X,Z}(uh_1 + x, \phi_t) \, \mathrm{d}u + o(1) \Bigg)$$

$$= h_1^{2a-1} h_2^{2b-1} R_{2b}(K_2) \Bigg(\int u^{2a} K_1^2(u) \, \mathrm{d}u \, (p(x, \phi_t) f_{X,Z}(x, \phi_t) + o(1)) + o(1) \Bigg)$$

$$= h_1^{2a-1} h_2^{2b-1} R_{2a}(K_1) R_{2b}(K_2) p(x, \phi_t) f_{X,Z}(x, \phi_t) + o(h_1^{2a-1} h_2^{2b-1})$$

$$= h_1^{2a-1} h_2^{2b-1} R_{2a}(K_1) R_{2b}(K_2) G_1(x, \phi_t) + o(h_1^{2a-1} h_2^{2b-1}).$$

Similarly, the second part is $o(h_1^{2a-1} h_2^{2b-1})$. So,

$$\mathrm{E}(A_{a,b,i}^2(x, \phi_t, d)) = h_1^{2a-1} h_2^{2b-1} R_{2a}(K_1) R_{2b}(K_2) G_1(x, \phi_t) + o(h_1^{2a-1} h_2^{2b-1})$$

holds. In the same way, it is straightforward to show that

$$\mathrm{E}(A^4_{a,b,i}(x,\phi_t,d)) = h_1^{4a-1}h_2^{4b-1}R_{4a}(K_1)R_{4b}(K_2)G_1(x,\phi_t) + o(h_1^{4a-1}h_2^{4b-1}).$$

Combining the results above, when $d = d_1$, the variance of $A_{a,b,i}$ is

$$\begin{aligned}\mathrm{Var}(A_{a,b,i}(x,\phi_t,d)) &= \mathrm{E}\left(A^2_{a,b,i}(x,\phi_t,d)\right) + \left(\mathrm{E}(A_{a,b,i}(x,\phi_t,d))\right)^2 \\ &= h_1^{2a-1}h_2^{2b-1}R_{2a}(K_1)R_{2b}(K_2)G_1(x,\phi_t) + o(h_1^{2a-1}h_2^{2b-1}).\end{aligned}$$

The variance of $A^2_{a,b,i}(x,\phi_t,d)$ is

$$\begin{aligned}\mathrm{Var}(A^2_{a,b,i}(x,\phi_t,d)) &= \mathrm{E}\left(A^4_{a,b,i}(x,\phi_t,d)\right) + \left(\mathrm{E}(A^2_{a,b,i}(x,\phi_t,d))\right)^2 \\ &= h_1^{4a-1}h_2^{4b-1}R_{4a}(K_1)R_{4b}(K_2)G_1(x,\phi_t) + o(h_1^{4a-1}h_2^{4b-1}).\end{aligned}$$

By A14, $O_p(\sqrt{\frac{\mathrm{Var}(A_{a,b,i}(x,\phi_t,d))}{n}}) = o_p(h_1^{a+1}h_2^{b+1})$ and $O_p(\sqrt{\frac{\mathrm{Var}(A^2_{a,b,i}(x,\phi_t,d))}{n}})$ $= o_p(h_1^{2a+1}h_2^{2b+1})$ follows. Combining the results of $\mathrm{E}(A_{a,b,i}(x,\phi_t,d))$, $\mathrm{E}(A^2_{a,b,i}(x,\phi_t,d))$, $\mathrm{Var}(A_{a,b,i}(x,\phi_t,d))$, and $\mathrm{Var}(A^2_{a,b,i}(x,\phi_t,d))$, (4), (6), (5), (7) and (8) follows when $q = 1$.

When $d = d_2$, $A_{a,b,i}(x,\phi_t,d)$ is

$$A_{a,b,i}(x,\phi_t,d) = A_{2,a,b,i}(x,\phi_t,d) + \lambda A_{1,a,b,i}(x,\phi_t,d).$$

The expectation of $A_{a,b,i}(x,\phi_t,d)$ is

$$\mathrm{E}(A_{a,b,i}(x,\phi_t,d)) = \mathrm{E}\left(\mathrm{E}(A_{a,b,i}(x,\phi_t,d)|X_i,C_i)\right)$$

$$
\begin{aligned}
&= \mathrm{E}\left((1 - P(D_i = d_1|X_i, C_i))\, K_{1,h_1}(X_i - x)(X_i - x)^a \right. \\
&\quad \left. \times \sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t)(\phi_j - \phi_t)^b I_{\{Z_i \in U_j\}} \right) \\
&\quad + \lambda \mathrm{E}\left(P(D_i = d_1|X_i, C_i) K_{1,h_1}(X_i - x)(X_i - x)^a \right. \\
&\quad \left. \times \sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t)(\phi_j - \phi_t)^b I_{\{Z_i \in U_j\}} \right).
\end{aligned}
$$

Applying exactly the same technique above, (4), (6), (5), (7) and (8) follows immediately when $q = 2$. □

THEOREM 4.3.2. *Let $\boldsymbol{X} = (X_1, \ldots, X_n)^T$, $\boldsymbol{C} = (C_1, \ldots, C_n)^T$ and $\boldsymbol{D} = (D_1, \ldots, D_n)^T$ and assume that A11-A14 hold. The conditional bias of our NW estimator is*

$$
\begin{aligned}
&E(\hat{m}(\boldsymbol{\psi}; h_1, h_2, \lambda)|\boldsymbol{X}, \boldsymbol{C}, \boldsymbol{D}) - m_{\boldsymbol{\psi}} \\
&= G_q^{-1}(x, \phi_t) \left(h_1^2 \mu_2(K_1) \left(\frac{\partial}{\partial r} G_q(x, \phi_t) \frac{\partial m_q}{\partial X}(x, \phi_t) + \frac{1}{2} \frac{\partial^2}{\partial r^2} G_q(x, \phi_t) \frac{\partial^2 m_q}{\partial X^2}(x, \phi_t) \right) \right. \\
&\quad + h_2^2 \mu_2(K_2) \left(\frac{\partial}{\partial s} G_q(x, \phi_t) \frac{\partial m_q}{\partial \phi}(x, \phi_t) + \frac{1}{2} \frac{\partial^2}{\partial s^2} G_q(x, \phi_t) \frac{\partial^2 m_q}{\partial \phi^2}(x, \phi_t) \right) \\
&\quad \left. + h_1 h_2 \mu_2(K_1) \mu_2(K_2) \left(\frac{\partial^2}{\partial r \partial s} G_q(x, \phi_t) \frac{\partial^2 m_q}{\partial X \partial \phi}(x, \phi_t) \right) \right) \\
&\quad + o_p(h_1^2) + o_p(h_2^2) + o_p(h_1 h_2), q = 1, 2.
\end{aligned}
$$

The conditional variance is

$$Var\left(\hat{m}(\boldsymbol{\psi}; h_1, h_2, \lambda)|\boldsymbol{X}, \boldsymbol{C}, \boldsymbol{D}\right) = \frac{1}{nh_1h_2}\frac{\sigma^2 R_0(K_1)R_0(K_2)}{G_q(x, \phi_t)}(1 + o_p(1)),\ q = 1, 2,$$

where

$$G_1(r, s) = p(r, s)f_{X,Z}(r, s),$$

and

$$G_2(r, s) = (1 - p(r, s))\, f_{X,Z}(r, s)$$

are the same as these defined in Lemma 4.3.1.

Remark 3. Notice that although the smoothing parameter λ appears in $\hat{m}(\boldsymbol{\psi}; h_1, h_2, \lambda)$, it is again not involved in the leading terms and absorbed in the remnant terms. This can be found clearly in the proof below. Not only does λ disappear, the asymptotic bias actually gets "split" into its two components, so that the asymptotic bias for estimating m_1 does not depend at all on m_2 and vice versa. This happens because λ is forced to go to zero. Something similar happens in the asymptotic variance, but it's more subtle: when we estimate m_1, the sample size appearing the denominator is nG_1, which is basically np_1, the number of observations expected to belong to m_1.

Remark 4. The asymptotic bias and variance of our NW estimator has the same rate as when bivariate smoothing using local linear estimator with both continuous covariates (Ruppert and Wand 1994, p.1351). Combining the results of the asymptotic conditional bias and asymptotic conditional variance can give the the asymptotic mean conditional squared error (MSE) for estimation at $\boldsymbol{\psi}$. If we assume h_1 and h_2 converge in the same rate, it is straightforward to show that the optimal rate of h_1 and h_2 to minimize MSE is in the order of $n^{-1/6}$.

Remark 5. The ordinal covariate has a similar effect on the asymptotic bias and variance as when we do local polynomial regression with an ordinal covariate in Chapter 3.

Proof. First, we consider the case that $\boldsymbol{\psi} = (x, t, d_1)^T$. The conditional expectation of $\hat{m}(\boldsymbol{\psi}; h_1, h_2, \lambda)$ given $\boldsymbol{X}$, $\boldsymbol{C}$ and $\boldsymbol{D}$ is

$$\mathrm{E}\left(\hat{m}(\boldsymbol{\psi}; h_1, h_2, \lambda)|\boldsymbol{X}, \boldsymbol{C}, \boldsymbol{D}\right) = \frac{\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} W_{\boldsymbol{\psi},i} Y_i | \boldsymbol{X}, \boldsymbol{C}, \boldsymbol{D}\right)}{\frac{1}{n}\sum_{i=1}^{n} W_{\boldsymbol{\psi},i}}.$$

By Lemma 4.3.1, the denominator of the conditional expectation is

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} W_{\boldsymbol{\psi},i} &= \frac{1}{n}\sum_{i=1}^{n}\left(K_{1,h_1}(X_i - x)\right)\left(\sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t) I_{\{Z_i \in U_j\}}\right)\left(K_{3,\lambda}(D_i, d_1)\right) \\ &= G_1(x, \phi_t) + o_p(1).\end{aligned}$$

Since the numerator of $\hat{m}(\boldsymbol{\psi}; h_1, h_2, \lambda)$ is

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} W_{\boldsymbol{\psi},i} Y_i &= \frac{1}{n}\sum_{i=1}^{n}\left(K_{1,h_1}(X_i - x)\right)\left(\sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t) I_{\{Z_i \in U_j\}}\right)\left(K_{3,\lambda}(D_i, d)\right) Y_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(K_{1,h_1}(X_i - x)\right)\left(\sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t) I_{\{Z_i \in U_j\}}\right) Y_i I_{\{D_i = d_1\}} \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\lambda\left(K_{1,h_1}(X_i - x)\right)\left(\sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t) I_{\{Z_i \in U_j\}}\right) Y_i I_{\{D_i = d_2\}} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(A_{1,0,0,i} Y_i + \lambda A_{2,0,0,i} Y_i\right),\end{aligned}$$

the conditional expectation of the numerator given $\boldsymbol{X}$, $\boldsymbol{C}$, $\boldsymbol{D}$ is

$$\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} W_{\boldsymbol{\psi},i} Y_i|\boldsymbol{X},\boldsymbol{C},\boldsymbol{D}\right) = \mathrm{E}\Bigg(\frac{1}{n}\sum_{i=1} (K_{1,h_1}(X_i - x)) \left(\sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t) I_{\{Z_i \in U_j\}}\right)$$

$$\times (K_{3,\lambda}(D_i, d))\, Y_i|\boldsymbol{X},\boldsymbol{C},\boldsymbol{D}\Bigg)$$

$$= \frac{1}{n}\sum_{i=1}^{n} (K_{1,h_1}(X_i - x)) \left(\sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t) m_1(X_i, z_{1,j}) I_{\{Z_i \in U_j\}}\right) I_{\{D_i = d_1\}}$$

$$+ \lambda \frac{1}{n}\sum_{i=1}^{n} (K_{1,h_1}(X_i - x)) \left(\sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t) m_2(X_i, z_{2,j}) I_{\{Z_i \in U_j\}}\right) I_{\{D_i = d_2\}}$$

$$= \frac{1}{n}\sum_{i=1}^{n} (K_{1,h_1}(X_i - x)) \left(\sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t) m_1(X_i, \phi_j) I_{\{Z_i \in U_j\}}\right) I_{\{D_i = d_1\}} + O(\frac{1}{M})$$

$$+ \lambda \frac{1}{n}\sum_{i=1}^{n} (K_{1,h_1}(X_i - x)) \left(\sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t) m_2(X_i, \phi_j) I_{\{Z_i \in U_j\}}\right) I_{\{D_i = d_2\}} + O(\frac{1}{M})$$

$$= B_1 + B_2,$$

by A11 and A14. Notice that we shorten $A_{1,a,b,i}(x, \phi_t, d)$ and $A_{2,a,b,i}(x, \phi_t, d)$ as $A_{1,a,b,i}$ and $A_{2,a,b,i}$ here for simplicity. They are shortened in the same way in the following proof.

The first part is

$$B_1 = \frac{1}{n}\sum_{i=1}^{n} (K_{1,h_1}(X_i - x)) \Bigg(\sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t)$$

$$\times \Bigg(m_1(x, \phi_t) + \frac{\partial m_1}{\partial X}(x, \phi_t)(X_i - x) + \frac{\partial m_1}{\partial \phi}(x, \phi_t)(\phi_j - \phi_t)$$

$$+ \frac{1}{2}\frac{\partial^2 m_1}{\partial X^2}(x, \phi_t)(X_i - x)^2 + \frac{1}{2}\frac{\partial^2 m_1}{\partial \phi^2}(x, \phi_t)(\phi_j - \phi_t)^2$$

$$+ \frac{\partial^2 m_1}{\partial \phi^2}(x, \phi_t)(x, \phi_t)(X_i - x)(\phi_j - \phi_t)\Bigg) I_{\{Z_i \in U_j\}}\Bigg) I_{\{D_i = d_1\}} + O(\frac{1}{M})$$

$$= \frac{1}{n}\sum_{i=1}^{n}\Bigg(A_{1,0,0,i}m_1(x,\phi_t) + A_{1,1,0,i}\frac{\partial m_1}{\partial X}(x,\phi_t) + A_{1,0,1,i}\frac{\partial m_1}{\partial \phi}(x,\phi_t)$$

$$+\frac{1}{2}A_{1,2,0,i}\frac{\partial^2 m_1}{\partial X^2}(x,\phi_t) + \frac{1}{2}A_{1,0,2,i}\frac{\partial^2 m_1}{\partial \phi^2}(x,\phi_t) + A_{1,1,1,i}\frac{\partial^2 m_1}{\partial \phi^2}(x,\phi_t)\Bigg) + O(\frac{1}{M}).$$

By Lemma 4.3.1, it is

$$B_1 = G_1(x,\phi_t)m_1(x,\phi_t) + h_1^2\mu_2(K_1)\left(\frac{\partial}{\partial r}G_1(x,\phi_t)\frac{\partial m_1}{\partial X}(x,\phi_t) + \frac{1}{2}\frac{\partial^2}{\partial r^2}G_1(x,\phi_t)\frac{\partial^2 m_1}{\partial X^2}(x,\phi_t)\right)$$

$$+ h_2^2\mu_2(K_2)\left(\frac{\partial}{\partial s}G_1(x,\phi_t)\frac{\partial m_1}{\partial X}(x,\phi_t) + \frac{1}{2}\frac{\partial^2}{\partial s^2}G_1(x,\phi_t)\frac{\partial^2 m_1}{\partial \phi^2}(x,\phi_t)\right)$$

$$+ h_1h_2\mu_2(K_1)\mu_2(K_2)\left(\frac{\partial^2}{\partial r\partial s}G_1(x,\phi_t) + \frac{\partial^2 m_1}{\partial X\partial \phi}(x,\phi_t)\right) + o(h_1^2) + o(h_2^2) + o(h_1h_2).$$

Notice that $O(\frac{1}{M})$ is absorbed in $o(h_1^2) + o(h_2^2) + o(h_1h_2)$ by A14.

Similarly, $B_2 = o(h_1^2) + o(h_2^2) + o(h_1h_2)$ follows immediately.

Then, since the inverse of the denominator can be approximated by

$$\left(\frac{1}{n}\sum_{i=1}^{n}W_{\boldsymbol{\psi},i}\right)^{-1} = G_1^{-1}(x,\phi_t) + o_p(1),$$

the conditional expectation of $\hat{m}(\boldsymbol{\psi}; h_1, h_2, \lambda)$ given $\boldsymbol{X}$, $\boldsymbol{C}$, and $\boldsymbol{D}$ is

$$\mathrm{E}\left(\hat{m}(\boldsymbol{\psi}; h_1, h_2, \lambda)|\boldsymbol{X},\boldsymbol{C},\boldsymbol{D}\right) = m_1(x,\phi_t) + G_1^{-1}(x,\phi_t)\Bigg(h_1^2\mu_2(K_1)\Bigg(\frac{\partial}{\partial r}G_1(x,\phi_t)\frac{\partial m_1}{\partial X}(x,\phi_t)$$

$$+\frac{1}{2}\frac{\partial^2}{\partial r^2}G_1(x,\phi_t)\frac{\partial^2 m_1}{\partial X^2}(x,\phi_t)\Bigg)$$

$$+h_2^2\mu_2(K_2)\left(\frac{\partial}{\partial s}G_1(x,\phi_t)\frac{\partial m_1}{\partial \phi}(x,\phi_t)+\frac{1}{2}\frac{\partial^2}{\partial s^2}G_1(x,\phi_t)\frac{\partial^2 m_1}{\partial \phi^2}(x,\phi_t)\right)$$

$$+h_1h_2\mu_2(K_1)\mu_2(K_2)\left(\frac{\partial^2}{\partial r\partial s}G_1(x,\phi_t)\frac{\partial^2 m_1}{\partial X\partial \phi}(x,\phi_t)\right)\Bigg)$$

$$+\,o_p(h_1^2)+o_p(h_2^2)+o_p(h_1h_2).$$

Note that $m_1(x,\phi_t)-m_1(x,z_{1,t})$ is $O(\frac{1}{M})$ by A11. It is again absorbed in $o_p(h_1^2)+o_p(h_2^2)+o_p(h_1h_2)$ by A14. Then the conditional bias of our NW estimator is

$$\begin{aligned}
&\mathrm{E}\left(\hat{m}(\boldsymbol{\psi};h_1,h_2,\lambda)|\boldsymbol{X},\boldsymbol{C},\boldsymbol{D}\right)-m_1(x,z_{1,t})\\
&=G_1^{-1}(x,\phi_t)\Bigg(h_1^2\mu_2(K_1)\left(\frac{\partial}{\partial r}G_1(x,\phi_t)\frac{\partial m_1}{\partial X}(x,\phi_t)+\frac{1}{2}\frac{\partial^2}{\partial r^2}G_1(x,\phi_t)\frac{\partial^2 m_1}{\partial X^2}(x,\phi_t)\right)\\
&\quad+h_2^2\mu_2(K_2)\left(\frac{\partial}{\partial s}G_1(x,\phi_t)\frac{\partial m_1}{\partial \phi}(x,\phi_t)+\frac{1}{2}\frac{\partial^2}{\partial s^2}G_1(x,\phi_t)\frac{\partial^2 m_1}{\partial \phi^2}(x,\phi_t)\right)\\
&\quad+h_1h_2\mu_2(K_1)\mu_2(K_2)\left(\frac{\partial^2}{\partial r\partial s}G_1(x,\phi_t)\frac{\partial^2 m_1}{\partial X\partial \phi}(x,\phi_t)\right)\Bigg)\\
&\quad+o_p(h_1^2)+o_p(h_2^2)+o_p(h_1h_2).
\end{aligned}$$

When $\boldsymbol{\psi}=(x,t,d_2)^T$, the conditional expectation of $\hat{m}(\boldsymbol{\psi};h_1,h_2,\lambda)$ given $\boldsymbol{X}$, $\boldsymbol{C}$ and $\boldsymbol{D}$ is

$$\mathrm{E}\left(\hat{m}(\boldsymbol{\psi};h_1,h_2,\lambda)|\boldsymbol{X},\boldsymbol{C},\boldsymbol{D}\right)=\frac{\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}W_{\boldsymbol{\psi},i}Y_i|\boldsymbol{X},\boldsymbol{C},\boldsymbol{D}\right)}{\frac{1}{n}\sum_{i=1}^{n}W_{\boldsymbol{\psi},i}}.$$

By Lemma 4.3.1, the denominator of the conditional expectation is

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} W_{\boldsymbol{\psi},i} &= \frac{1}{n}\sum_{i=1}^{n}\left(K_{1,h_1}(X_i - x)\right)\left(\sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t) I_{\{Z_i \in U_j\}}\right)\left(K_{3,\lambda}(D_i, d_2)\right) \\ &= G_2(x, \phi_t) + o_p(1).\end{aligned}$$

Since the numerator of $\hat{m}(\boldsymbol{\psi}; h_1, h_2, \lambda)$ is

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} W_{\boldsymbol{\psi},i} Y_i &= \frac{1}{n}\sum_{i=1}^{n}\left(K_{1,h_1}(X_i - x)\right)\left(\sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t) I_{\{Z_i \in U_j\}}\right)\left(K_{3,\lambda}(D_i, d)\right) Y_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(K_{1,h_1}(X_i - x)\right)\left(\sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t) I_{\{Z_i \in U_j\}}\right) Y_i I_{\{D_i = d_2\}} \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\lambda\left(K_{1,h_1}(X_i - x)\right)\left(\sum_{j=1}^{M} K_{2,h_2}(\phi_j - \phi_t) I_{\{Z_i \in U_j\}}\right) Y_i I_{\{D_i = d_1\}} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(A_{2,0,0,i} Y_i + \lambda A_{1,0,0,i} Y_i\right).\end{aligned}$$

Following exactly the same steps derived above, the conditional bias of our NW estimator is

$$\begin{aligned}&\mathrm{E}\left(\hat{m}(\boldsymbol{\psi}; h_1, h_2, \lambda) | \boldsymbol{X}, \boldsymbol{C}, \boldsymbol{D}\right) - m_2(x, z_{2,t}) \\ &= G_2^{-1}(x, \phi_t)\left(h_1^2 \mu_2(K_1)\left(\frac{\partial}{\partial r} G_2(x, \phi_t)\frac{\partial m_2}{\partial X}(x, \phi_t) + \frac{1}{2}\frac{\partial^2}{\partial r^2} G_2(x, \phi_t)\frac{\partial^2 m_2}{\partial X^2}(x, \phi_t)\right)\right. \\ &\quad + h_2^2 \mu_2(K_2)\left(\frac{\partial}{\partial s} G_2(x, \phi_t)\frac{\partial m_2}{\partial \phi}(x, \phi_t) + \frac{1}{2}\frac{\partial^2}{\partial s^2} G_2(x, \phi_t)\frac{\partial^2 m_2}{\partial \phi^2}(x, \phi_t)\right) \\ &\quad \left. + h_1 h_2 \mu_2(K_1)\mu_2(K_2)\left(\frac{\partial^2}{\partial r \partial s} G_2(x, \phi_t)\frac{\partial^2 m_2}{\partial X \partial \phi}(x, \phi_t)\right)\right) \\ &\quad + o_p(h_1^2) + o_p(h_2^2) + o_p(h_1 h_2).\end{aligned}$$

When $d = d_1$, the conditional variance of $\hat{m}(\boldsymbol{\psi}; h_1, h_2, \lambda)$ given $\boldsymbol{X}$, $\boldsymbol{C}$ and $\boldsymbol{D}$ is

$$
\begin{aligned}
\operatorname{Var}\left(\hat{m}(\boldsymbol{\psi}; h_1, h_2, \lambda)\right|\boldsymbol{X}, \boldsymbol{C}, \boldsymbol{D}) &= \frac{\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} W_{\boldsymbol{\psi},i} Y_i | \boldsymbol{X}, \boldsymbol{C}, \boldsymbol{D}\right)}{\left(\frac{1}{n}\sum_{i=1}^{n} W_{\boldsymbol{\psi},i}\right)^2} \\
&= \frac{\frac{1}{n^2}\sum_{i=1}^{n}\left(W_{\boldsymbol{\psi},i}^2 \operatorname{Var}(Y_i I_{\{D_i=d_1\}} + Y_i I_{\{D_i=d_2\}} | \boldsymbol{X}, \boldsymbol{C}, \boldsymbol{D})\right)}{\left(\frac{1}{n}\sum_{i=1}^{n} W_{\boldsymbol{\psi},i}\right)^2} \\
&= \frac{\frac{\sigma^2}{n}\frac{1}{n}\sum_{i=1}^{n}\left(W_{\boldsymbol{\psi},i}^2\right)}{\left(\frac{1}{n}\sum_{i=1}^{n} W_{\boldsymbol{\psi},i}\right)^2} \\
&= \frac{1}{nh_1h_2}\frac{\sigma^2 R_0(K_1) R_0(K_2)}{G_1(x, \phi_t)}(1 + o_p(1)).
\end{aligned}
$$

Similarly, when $d = d_2$, the conditional variance is

$$
\operatorname{Var}\left(\hat{m}(\boldsymbol{\psi}; h_1, h_2, \lambda)\right|\boldsymbol{X}, \boldsymbol{C}, \boldsymbol{D}) = \frac{1}{nh_1h_2}\frac{\sigma^2 R_0(K_1) R_0(K_2)}{G_2(x, \phi_t)}(1 + o_p(1)).
$$

□

4.4. Simulation

In this section, we do three patterns of simulations based on a model with a continuous observed covariate X, a continuous latent covariate Z, and a nominal observed convariate

D:

$$Y_i = \begin{cases} m_1(X_i, Z_i, D_i) + \epsilon_i = X_i^2 + \sin(\frac{\pi}{2} Z_i) + 1 + \epsilon_i & \text{if } D_i = 0 \\ m_2(X_i, Z_i, D_i) + \epsilon_i = (1+\alpha)\left(1.1X_i^2 + 1.2\sin(\frac{\pi}{2} Z_i) + 1.1\right) + \epsilon_i & \text{if } D_i = 1, \end{cases}$$

$i = 1, \ldots, 100$, where X_i are generated from $U(0,1)$, Z_i are generated from $U(0,1)$, and the number of observations that fall in different categories is set to be the same. D_i are generated from a Bernoulli distribution, where $P(D_i = 0) = \frac{1}{2}\left(X_i^2 + (\frac{C_i - 0.5}{10})^2\right)$, $P(D_i = 1) = 1 - P(D_i = 0)$, $C_i = \sum_{j=1}^{10} j I_{\{Z_i \in U_j\}}$, ϵ_i are generated from $N(0, 0.1)$, and α takes values in $0, 0.5, 1$, which makes three patterns of models. While α increases, the difference between m_1 and m_2 becomes bigger. Each pattern is simulated 10000 times.

We apply our NW estimator with product kernel for different bandwidth values choices, where h_1 takes two values $0.2, 0.7$, h_2 takes two values $0.2, 0.7$, and λ takes two values $0.1, 0.9$. Lower value of the bandwidth means smoothing more locally and higher value means smoothing more globally in each dimension. So there are 8 different combinations of bandwidths.

We estimate $m(X, Z, D)$ at different values of (x, t, d), where x takes two values $0.3, 0.7$, t takes three values $3, 5, 8$, and d takes two values $0, 1$. Therefore there are 12 positions to estimate. At a fixed (h_1, h_2, λ) and (x, t, d), using the estimator $\hat{m}(x, \phi_t, d; h_1, h_2, \lambda)$, where $\phi_t = \frac{t-0.5}{10}$, $t = 1, 2, \ldots, 10$, we approximate the bias $B(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda)) = \mathrm{E}(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda) - m_{x,t,d})$, the mean squared errors (MSE) $MSE(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda)) = \mathrm{E}(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda) - m_{x,t,d})^2$, and the mean sum of squared errors (MSSE) $MSSE_{h_1, h_2, \lambda} = \mathrm{E}\sum_{x,t,d}(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda) - m_{x,t,d})^2$ by averaging the simulation results. The simulation standard deviation of each estimation is given, as well.

Table 4.1-4.6 provide the simulation results when $\alpha = 0$, i.e., m_1 and m_2 are very close to each other. Table 4.1-4.2 shows that when λ changes from 0.1 to 0.9, amongst the 48 estimations, 23 biases are reduced. Table 4.5-4.6 shows that when λ changes from 0.1 to 0.9, amongst the 48 estimations, MSE of exactly the same 23 estimations are reduced. The MSSE is also reduced when $(h_1, h_2) = (0.7, 0.7)$. Table 4.3-4.4 shows that when λ changes from 0.1 to 0.9, amongst the 48 estimations, 29 simulation standard errors are reduced and 12 of them are overlapped with the estimations in which bias or MSE are reduced. All these facts indicate that when m_1 and m_2 are close, increasing λ, i.e. smoothing more globally in dimension D may not only reduce bias and MSE but also the variance.

Table 4.7-4.12 provide the simulation results when $\alpha = 0.5$, i.e., m_1 and m_2 are further from each other. Table 4.7-4.8 shows that when λ changes from 0.1 to 0.9, amongst the 48 estimations, 12 biases are reduced. Table 4.11-4.12 shows that when λ changes from 0.1 to 0.9, amongst the 48 estimations, MSE of exactly the same 12 estimations are reduced. But the MSSE are all increased. Table 4.9-4.10 shows that when λ changes from 0.1 to 0.9, amongst the 48 estimations, 13 simulation standard errors are reduced but only one of them are overlapped with the estimations in which bias or MSE are reduced. All these facts indicate that when m_1 and m_2 are not close, increasing λ could sometimes reduce the bias and MSE but will at the same increase the variance.

Table 4.13-4.16 provide the simulation results when $\alpha = 1$, i.e., m_1 and m_2 are even further from each other. Table 4.13-4.14 shows that when λ changes from 0.1 to 0.9, amongst the 48 estimations, 12 biases are reduced. Table 4.17-4.18 shows that when λ changes from 0.1 to 0.9, amongst the 48 estimations, 11 MSE are reduced, which are contained in the

12 estimations in which biases are reduced. But the MSSE are all increased. Table 4.15-4.16 shows that when λ changes from 0.1 to 0.9, amongst the 48 estimations, 11 simulation standard errors are reduced but none of them are overlapped with the estimations in which bias or MSE are reduced. All these facts indicate that when m_1 and m_2 are far from each other, increasing λ could sometimes reduce the bias and MSE but will at the same increase the variance.

In all of the three patterns, increasing h_1 or h_2 can reduce the variance of estimation in most of the times. Increasing h_1 or h_2 can also reduce the MSSE in all of the patterns except when $\alpha = 0$. In that pattern increasing h_2 will slightly increase MSSE. The bandwidth choice of $(h_1, h_2, \lambda) = (0.7, 0.7, 0.1)$ has the smallest variance in each point estimation. And all of them have the smallest MSSE except when $\alpha = 0$. In that pattern the choice of $(0.7, 0.2, 0.1)$ has the smallest MSSE.

Table 16. Simulation bias $B(\hat{m}(x,\phi_t,d;h_1,h_2,\lambda)) = \mathrm{E}(\hat{m}(x,\phi_t,d;h_1,h_2,\lambda) - m_{x,t,d})$ when $\alpha = 0$ and $\lambda = 0.1$.

(h_1,h_2) (x,t,d)	(0.2,0.2)	(0.2,0.7)	(0.7,0.2)	(0.7,0.7)
(0.3,3,0)	0.682	0.812	0.0762	0.197
(0.3,5,0)	0.69	0.632	0.0748	0.0216
(0.3,8,0)	0.718	0.519	0.0817	-0.0963
(0.7,3,0)	-0.468	-0.343	-0.0862	0.0323
(0.7,5,0)	-0.468	-0.519	-0.0874	-0.146
(0.7,8,0)	-0.459	-0.639	-0.0833	-0.264
(0.3,3,1)	0.743	0.999	0.104	0.371
(0.3,5,1)	0.745	0.813	0.0952	0.189
(0.3,8,1)	0.757	0.639	0.0973	-0.00396
(0.7,3,1)	-0.605	-0.289	8.55e-05	0.238
(0.7,5,1)	-0.601	-0.46	-0.0128	0.0497
(0.7,8,1)	-0.565	-0.658	-0.0249	-0.135

Table 17. Simulation bias $B(\hat{m}(x,\phi_t,d;h_1,h_2,\lambda)) = \mathrm{E}(\hat{m}(x,\phi_t,d;h_1,h_2,\lambda) - m_{x,t,d})$ when $\alpha = 0$ and $\lambda = 0.9$.

(h_1,h_2) (x,t,d)	(0.2,0.2)	(0.2,0.7)	(0.7,0.2)	(0.7,0.7)
(0.3,3,0)	0.797	0.987	0.115	0.278
(0.3,5,0)	0.839	0.847	0.135	0.129
(0.3,8,0)	0.928	0.755	0.196	0.0264
(0.7,3,0)	-0.455	-0.294	0.00216	0.172
(0.7,5,0)	-0.439	-0.447	0.0314	0.0275
(0.7,8,0)	-0.386	-0.555	0.103	-0.0691
(0.3,3,1)	0.627	0.825	-0.0619	0.109
(0.3,5,1)	0.618	0.634	-0.0911	-0.0887
(0.3,8,1)	0.652	0.486	-0.0793	-0.245
(0.7,3,1)	-0.678	-0.508	-0.207	-0.0279
(0.7,5,1)	-0.712	-0.711	-0.227	-0.222
(0.7,8,1)	-0.708	-0.873	-0.206	-0.373

Table 18. Simulation Standard Deviation for each estimation when $\alpha = 0$ and $\lambda = 0.1$.

(h_1, h_2) (x, t, d)	(0.2,0.2)	(0.2,0.7)	(0.7,0.2)	(0.7,0.7)
(0.3,3,0)	0.0981	0.0832	0.0415	0.033
(0.3,5,0)	0.0967	0.0815	0.0423	0.031
(0.3,8,0)	0.0995	0.0753	0.0463	0.0292
(0.7,3,0)	0.0595	0.0649	0.0627	0.0431
(0.7,5,0)	0.0564	0.0616	0.0637	0.0413
(0.7,8,0)	0.0557	0.051	0.0691	0.0423
(0.3,3,1)	0.142	0.105	0.0839	0.0567
(0.3,5,1)	0.13	0.0952	0.0762	0.0491
(0.3,8,1)	0.103	0.0826	0.0601	0.0443
(0.7,3,1)	0.0873	0.103	0.11	0.0689
(0.7,5,1)	0.0954	0.0909	0.0996	0.0613
(0.7,8,1)	0.0735	0.077	0.0825	0.0585

Table 19. Simulation Standard Deviation for each estimation when $\alpha = 0$ and $\lambda = 0.9$.

(h_1, h_2) (x, t, d)	(0.2,0.2)	(0.2,0.7)	(0.7,0.2)	(0.7,0.7)
(0.3,3,0)	0.104	0.0815	0.0438	0.0321
(0.3,5,0)	0.108	0.081	0.0459	0.03
(0.3,8,0)	0.109	0.0812	0.0517	0.0307
(0.7,3,0)	0.0604	0.0665	0.0658	0.0426
(0.7,5,0)	0.0613	0.0635	0.0684	0.0412
(0.7,8,0)	0.0684	0.056	0.074	0.0455
(0.3,3,1)	0.107	0.0834	0.0453	0.033
(0.3,5,1)	0.111	0.0825	0.0477	0.0311
(0.3,8,1)	0.109	0.082	0.0527	0.0321
(0.7,3,1)	0.0611	0.068	0.0685	0.0442
(0.7,5,1)	0.0633	0.0654	0.0711	0.0428
(0.7,8,1)	0.0703	0.0586	0.0751	0.0472

Table 20. Simulation squared root of mean squared errors $MSE(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda)) = \mathrm{E}(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda) - m_{x,t,d})^2$ and squared root of mean sum of squared errors $MSSE_{h_1,h_2,\lambda} = \mathrm{E}\sum_{x,t,d}(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda) - m_{x,t,d})^2$ when $\alpha = 0$ and $\lambda = 0.1$.

(h_1, h_2) (x, t, d)	(0.2,0.2)	(0.2,0.7)	(0.7,0.2)	(0.7,0.7)
(0.3,3,0)	0.689	0.816	0.0867	0.2
(0.3,5,0)	0.696	0.637	0.0859	0.0378
(0.3,8,0)	0.725	0.524	0.0939	0.101
(0.7,3,0)	0.471	0.349	0.107	0.0539
(0.7,5,0)	0.471	0.523	0.108	0.152
(0.7,8,0)	0.462	0.641	0.108	0.268
(0.3,3,1)	0.756	1	0.134	0.375
(0.3,5,1)	0.756	0.818	0.122	0.195
(0.3,8,1)	0.764	0.645	0.114	0.0444
(0.7,3,1)	0.612	0.307	0.11	0.247
(0.7,5,1)	0.608	0.469	0.1	0.0789
(0.7,8,1)	0.569	0.663	0.0862	0.147
$\sqrt{MSSE_h}$	2.22	2.24	0.366	0.647

Table 21. Simulation squared root of mean squared errors $MSE(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda)) = \mathrm{E}(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda) - m_{x,t,d})^2$ and squared root of mean sum of squared errors $MSSE_{h_1,h_2,\lambda} = \mathrm{E}\sum_{x,t,d}(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda) - m_{x,t,d})^2$ when $\alpha = 0$ and $\lambda = 0.9$.

(h_1, h_2) (x, t, d)	(0.2,0.2)	(0.2,0.7)	(0.7,0.2)	(0.7,0.7)
(0.3,3,0)	0.804	0.991	0.123	0.28
(0.3,5,0)	0.846	0.851	0.143	0.132
(0.3,8,0)	0.934	0.759	0.202	0.0405
(0.7,3,0)	0.459	0.301	0.0658	0.178
(0.7,5,0)	0.443	0.452	0.0753	0.0495
(0.7,8,0)	0.392	0.558	0.127	0.0827
(0.3,3,1)	0.636	0.829	0.0767	0.114
(0.3,5,1)	0.628	0.639	0.103	0.094
(0.3,8,1)	0.661	0.493	0.0952	0.247
(0.7,3,1)	0.68	0.513	0.218	0.0523
(0.7,5,1)	0.715	0.714	0.238	0.226
(0.7,8,1)	0.711	0.875	0.219	0.376
$\sqrt{MSSE_h}$	2.35	2.4	0.529	0.646

TABLE 22. Simulation bias $B(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda)) = \mathrm{E}(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda) - m_{x,t,d})$ when $\alpha = 0.5$ and $\lambda = 0.1$.

(h_1, h_2) (x, t, d)	(0.2,0.2)	(0.2,0.7)	(0.7,0.2)	(0.7,0.7)
(0.3,3,0)	0.792	0.944	0.0904	0.223
(0.3,5,0)	0.852	0.796	0.102	0.0576
(0.3,8,0)	1.08	0.743	0.156	-0.0438
(0.7,3,0)	-0.463	-0.329	-0.0476	0.0881
(0.7,5,0)	-0.452	-0.497	-0.0264	-0.0736
(0.7,8,0)	-0.401	-0.604	0.0542	-0.165
(0.3,3,1)	0.936	1.39	-0.189	0.302
(0.3,5,1)	0.967	1.12	-0.145	0.0693
(0.3,8,1)	1.06	0.881	-0.0125	-0.183
(0.7,3,1)	-1.44	-0.799	-0.222	0.187
(0.7,5,1)	-1.36	-0.99	-0.21	-0.0708
(0.7,8,1)	-1.1	-1.24	-0.146	-0.324

TABLE 23. Simulation bias $B(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda)) = \mathrm{E}(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda) - m_{x,t,d})$ when $\alpha = 0.5$ and $\lambda = 0.9$.

(h_1, h_2) (x, t, d)	(0.2,0.2)	(0.2,0.7)	(0.7,0.2)	(0.7,0.7)
(0.3,3,0)	1.3	1.62	0.226	0.47
(0.3,5,0)	1.5	1.59	0.334	0.385
(0.3,8,0)	1.94	1.64	0.632	0.372
(0.7,3,0)	-0.416	-0.187	0.264	0.535
(0.7,5,0)	-0.329	-0.285	0.411	0.471
(0.7,8,0)	-0.0699	-0.318	0.768	0.486
(0.3,3,1)	0.353	0.7	-0.758	-0.492
(0.3,5,1)	0.349	0.455	-0.845	-0.777
(0.3,8,1)	0.575	0.284	-0.738	-0.999
(0.7,3,1)	-1.68	-1.43	-0.952	-0.659
(0.7,5,1)	-1.79	-1.73	-1	-0.926
(0.7,8,1)	-1.72	-1.97	-0.844	-1.12

Table 24. Simulation Standard Deviation for each estimation when $\alpha = 0.5$ and $\lambda = 0.1$.

(h_1, h_2) (x, t, d)	(0.2,0.2)	(0.2,0.7)	(0.7,0.2)	(0.7,0.7)
(0.3,3,0)	0.129	0.104	0.0432	0.0338
(0.3,5,0)	0.168	0.106	0.0462	0.0317
(0.3,8,0)	0.304	0.121	0.0628	0.0315
(0.7,3,0)	0.0606	0.0662	0.0659	0.0454
(0.7,5,0)	0.0627	0.0632	0.0705	0.0436
(0.7,8,0)	0.0983	0.0545	0.0951	0.0475
(0.3,3,1)	0.338	0.172	0.204	0.105
(0.3,5,1)	0.3	0.155	0.177	0.0898
(0.3,8,1)	0.2	0.139	0.112	0.0826
(0.7,3,1)	0.256	0.248	0.218	0.113
(0.7,5,1)	0.333	0.219	0.19	0.101
(0.7,8,1)	0.259	0.2	0.139	0.0976

Table 25. Simulation Standard Deviation for each estimation when $\alpha = 0.5$ and $\lambda = 0.9$.

(h_1, h_2) (x, t, d)	(0.2,0.2)	(0.2,0.7)	(0.7,0.2)	(0.7,0.7)
(0.3,3,0)	0.318	0.218	0.0849	0.0651
(0.3,5,0)	0.361	0.217	0.113	0.0693
(0.3,8,0)	0.379	0.238	0.158	0.0868
(0.7,3,0)	0.0933	0.107	0.145	0.0978
(0.7,5,0)	0.15	0.12	0.172	0.0992
(0.7,8,0)	0.248	0.145	0.205	0.118
(0.3,3,1)	0.332	0.221	0.0956	0.0724
(0.3,5,1)	0.368	0.217	0.125	0.0766
(0.3,8,1)	0.365	0.233	0.165	0.0939
(0.7,3,1)	0.103	0.119	0.16	0.105
(0.7,5,1)	0.168	0.133	0.185	0.105
(0.7,8,1)	0.263	0.161	0.206	0.123

Table 26. Simulation squared root of mean squared errors $MSE(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda)) = \mathrm{E}(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda) - m_{x,t,d})^2$ and squared root of mean sum of squared errors $MSSE_{h_1,h_2,\lambda} = \mathrm{E}\sum_{x,t,d}(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda) - m_{x,t,d})^2$ when $\alpha = 0.5$ and $\lambda = 0.1$.

(h_1, h_2) (x, t, d)	(0.2,0.2)	(0.2,0.7)	(0.7,0.2)	(0.7,0.7)
(0.3,3,0)	0.803	0.95	0.1	0.225
(0.3,5,0)	0.869	0.803	0.112	0.0658
(0.3,8,0)	1.12	0.753	0.168	0.0539
(0.7,3,0)	0.466	0.336	0.0813	0.0991
(0.7,5,0)	0.457	0.501	0.0753	0.0855
(0.7,8,0)	0.413	0.607	0.109	0.171
(0.3,3,1)	0.995	1.4	0.278	0.32
(0.3,5,1)	1.01	1.14	0.229	0.113
(0.3,8,1)	1.08	0.892	0.113	0.201
(0.7,3,1)	1.46	0.837	0.311	0.219
(0.7,5,1)	1.4	1.01	0.283	0.123
(0.7,8,1)	1.13	1.25	0.202	0.339
$\sqrt{MSSE_h}$	3.44	3.19	0.66	0.661

Table 27. Simulation squared root of mean squared errors $MSE(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda)) = \mathrm{E}(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda) - m_{x,t,d})^2$ and squared root of mean sum of squared errors $MSSE_{h_1,h_2,\lambda} = \mathrm{E}\sum_{x,t,d}(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda) - m_{x,t,d})^2$ when $\alpha = 0.5$ and $\lambda = 0.9$.

(h_1, h_2) (x, t, d)	(0.2,0.2)	(0.2,0.7)	(0.7,0.2)	(0.7,0.7)
(0.3,3,0)	1.34	1.64	0.242	0.474
(0.3,5,0)	1.54	1.6	0.353	0.392
(0.3,8,0)	1.98	1.65	0.651	0.382
(0.7,3,0)	0.427	0.215	0.301	0.544
(0.7,5,0)	0.362	0.309	0.445	0.481
(0.7,8,0)	0.258	0.35	0.795	0.5
(0.3,3,1)	0.484	0.734	0.764	0.498
(0.3,5,1)	0.507	0.504	0.855	0.781
(0.3,8,1)	0.681	0.367	0.756	1
(0.7,3,1)	1.68	1.43	0.965	0.667
(0.7,5,1)	1.8	1.73	1.02	0.931
(0.7,8,1)	1.74	1.98	0.869	1.13
$\sqrt{MSSE_h}$	4.3	4.26	2.48	2.4

TABLE 28. Simulation bias $B(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda)) = \mathrm{E}(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda) - m_{x,t,d})$ when $\alpha = 1$ and $\lambda = 0.1$.

(h_1, h_2) / (x, t, d)	(0.2,0.2)	(0.2,0.7)	(0.7,0.2)	(0.7,0.7)
(0.3,3,0)	0.902	1.08	0.105	0.248
(0.3,5,0)	1.02	0.961	0.13	0.0936
(0.3,8,0)	1.43	0.967	0.231	0.0088
(0.7,3,0)	-0.458	-0.316	-0.00904	0.144
(0.7,5,0)	-0.437	-0.476	0.0345	-0.0012
(0.7,8,0)	-0.343	-0.57	0.192	-0.0649
(0.3,3,1)	1.13	1.77	-0.482	0.233
(0.3,5,1)	1.19	1.44	-0.384	-0.0499
(0.3,8,1)	1.36	1.12	-0.122	-0.363
(0.7,3,1)	-2.28	-1.31	-0.443	0.137
(0.7,5,1)	-2.12	-1.52	-0.408	-0.191
(0.7,8,1)	-1.64	-1.82	-0.267	-0.514

TABLE 29. Simulation bias $B(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda)) = \mathrm{E}(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda) - m_{x,t,d})$ when $\alpha = 1$ and $\lambda = 0.9$.

(h_1, h_2) / (x, t, d)	(0.2,0.2)	(0.2,0.7)	(0.7,0.2)	(0.7,0.7)
(0.3,3,0)	1.8	2.26	0.338	0.662
(0.3,5,0)	2.16	2.33	0.534	0.642
(0.3,8,0)	2.96	2.52	1.07	0.718
(0.7,3,0)	-0.378	-0.079	0.525	0.898
(0.7,5,0)	-0.219	-0.122	0.79	0.915
(0.7,8,0)	0.247	-0.0807	1.43	1.04
(0.3,3,1)	0.0787	0.576	-1.45	-1.09
(0.3,5,1)	0.0807	0.277	-1.6	-1.47
(0.3,8,1)	0.498	0.0809	-1.4	-1.75
(0.7,3,1)	-2.68	-2.35	-1.7	-1.29
(0.7,5,1)	-2.87	-2.74	-1.78	-1.63
(0.7,8,1)	-2.74	-3.07	-1.48	-1.87

TABLE 30. Simulation Standard Deviation for each estimation when $\alpha = 1$ and $\lambda = 0.1$.

(h_1, h_2) (x, t, d)	(0.2,0.2)	(0.2,0.7)	(0.7,0.2)	(0.7,0.7)
(0.3,3,0)	0.204	0.148	0.0462	0.036
(0.3,5,0)	0.292	0.157	0.0534	0.0346
(0.3,8,0)	0.555	0.204	0.0874	0.0384
(0.7,3,0)	0.0629	0.0685	0.0738	0.0516
(0.7,5,0)	0.074	0.0668	0.0859	0.0512
(0.7,8,0)	0.154	0.064	0.139	0.0617
(0.3,3,1)	0.553	0.247	0.341	0.164
(0.3,5,1)	0.486	0.22	0.298	0.138
(0.3,8,1)	0.308	0.2	0.178	0.129
(0.7,3,1)	0.444	0.405	0.34	0.164
(0.7,5,1)	0.591	0.357	0.293	0.144
(0.7,8,1)	0.466	0.335	0.202	0.14

TABLE 31. Simulation Standard Deviation for each estimation when $\alpha = 1$ and $\lambda = 0.9$.

(h_1, h_2) (x, t, d)	(0.2,0.2)	(0.2,0.7)	(0.7,0.2)	(0.7,0.7)
(0.3,3,0)	0.557	0.371	0.136	0.105
(0.3,5,0)	0.635	0.367	0.19	0.115
(0.3,8,0)	0.665	0.407	0.273	0.149
(0.7,3,0)	0.142	0.16	0.239	0.161
(0.7,5,0)	0.256	0.188	0.289	0.165
(0.7,8,0)	0.439	0.246	0.347	0.198
(0.3,3,1)	0.579	0.375	0.156	0.119
(0.3,5,1)	0.645	0.366	0.213	0.128
(0.3,8,1)	0.638	0.397	0.286	0.161
(0.7,3,1)	0.161	0.182	0.264	0.174
(0.7,5,1)	0.288	0.213	0.311	0.175
(0.7,8,1)	0.466	0.273	0.349	0.205

TABLE 32. Simulation squared root of mean squared errors $MSE(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda)) = \mathrm{E}(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda) - m_{x,t,d})^2$ and squared root of mean sum of squared errors $MSSE_{h_1,h_2,\lambda} = \mathrm{E}\sum_{x,t,d}(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda) - m_{x,t,d})^2$ when $\alpha = 1$ and $\lambda = 0.1$.

(h_1, h_2) (x, t, d)	(0.2,0.2)	(0.2,0.7)	(0.7,0.2)	(0.7,0.7)
(0.3,3,0)	0.925	1.09	0.114	0.251
(0.3,5,0)	1.06	0.973	0.14	0.0998
(0.3,8,0)	1.54	0.989	0.247	0.0394
(0.7,3,0)	0.462	0.323	0.0744	0.153
(0.7,5,0)	0.443	0.48	0.0925	0.0512
(0.7,8,0)	0.376	0.573	0.237	0.0896
(0.3,3,1)	1.26	1.79	0.591	0.285
(0.3,5,1)	1.28	1.45	0.486	0.147
(0.3,8,1)	1.4	1.14	0.216	0.385
(0.7,3,1)	2.32	1.37	0.559	0.213
(0.7,5,1)	2.2	1.56	0.502	0.24
(0.7,8,1)	1.7	1.85	0.335	0.533
$\sqrt{MSSE_h}$	4.82	4.26	1.21	0.864

TABLE 33. Simulation squared root of mean squared errors $MSE(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda)) = \mathrm{E}(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda) - m_{x,t,d})^2$ and squared root of mean sum of squared errors $MSSE_{h_1,h_2,\lambda} = \mathrm{E}\sum_{x,t,d}(\hat{m}(x, \phi_t, d; h_1, h_2, \lambda) - m_{x,t,d})^2$ when $\alpha = 1$ and $\lambda = 0.9$.

(h_1, h_2) (x, t, d)	(0.2,0.2)	(0.2,0.7)	(0.7,0.2)	(0.7,0.7)
(0.3,3,0)	1.88	2.29	0.364	0.67
(0.3,5,0)	2.25	2.35	0.567	0.652
(0.3,8,0)	3.03	2.55	1.1	0.733
(0.7,3,0)	0.403	0.178	0.577	0.913
(0.7,5,0)	0.336	0.225	0.841	0.929
(0.7,8,0)	0.503	0.259	1.47	1.06
(0.3,3,1)	0.584	0.687	1.46	1.1
(0.3,5,1)	0.65	0.459	1.61	1.47
(0.3,8,1)	0.809	0.405	1.43	1.76
(0.7,3,1)	2.69	2.36	1.72	1.3
(0.7,5,1)	2.88	2.75	1.8	1.64
(0.7,8,1)	2.78	3.08	1.52	1.88
$\sqrt{MSSE_h}$	6.56	6.39	4.49	4.32

4.5. Conclusion

In this chapter, we investigated nonparametric smoothing in the situation that the covariates are continuous, ordinal and nominal variables. As the previous studies we introduced in Section 1, this is the first sight on this situation to our knowledge. We proposed a NW estimator with product kernel for the three covariates. We derived the asymptotic conditional bias and asymptotic conditional variance of our estimator, which were in the same order as when doing local linear regression with two continuous covariates (Ruppert and Wand 1994, p.1357). Combining the results of the asymptotic conditional bias and variance could give the asymptotic mean conditional squared error (MSE) for estimation at $\boldsymbol{\psi}$. If we assumed h_1 and h_2 converged in the same rate, it was straightforward to show that the optimal rate of h_1 and h_2 to minimize MSE was in the order of $n^{-1/6}$. Notice the smoothing parameter λ was not involved in the leading terms of either asymptotic conditional bias or variance, and it was absorbed in the remnant terms. In this way, we could not determine its optimal rate. What we knew was that $0 < \lambda < 1$ and it was required to go to 0 when smoothing more locally. Not only did λ disappear, the asymptotic bias actually got "split" into its two components, so that the asymptotic bias for estimating m_1 did not depend at all on m_2 and vice versa. This happened because λ was forced to go to zero. Something similar happened in the asymptotic variance, but it was more subtle: when we estimated m_1, the sample size appearing the denominator was nG_1, which was basically np_1, the expected number of observations belonging to m_1. We conducted a simulation study and the results of the bias, variance, and choice of the bandwidths were in accordance with our theoretical results.

References

Aitchison, J. and C. Aitken (1976). Multivariate binary discrimination by the kernel method. *Biometrika 63*, 413420.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19 (6)*, 716–723.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics 16*, 125–127.

Bierens, H. (1983). Uniform consistency of kernel estimators of a regression function under generalized conditions. *Journal of the American Statistical Association 78*, 699707.

Bowman, A., P. Hall, and T. Titterington (1984). Cross-validation in nonparametric estimation of probabilities and probability densities. *Biometrika 71*, 341351.

Brewer, K. R. W. (1963). Ratio estimation in finite populations: some results deductible from the assumption of an underlying stochastic process. *Austral. J Statist 5*, 93–105.

Clark, R. G. and R. L. Chambers (2008). Adaptive calibration for prediction of finite population totals. *Survey Methodology 34*(2), 163–172.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplot. *J. Amer. Statist. Assoc 74*, 829–836.

Cleveland, W. S. and S. Devlin (1988). Locally weighted regression: an approach to regression analysis by local fitting. *J. Amer. Statist. Assoc 83*, 596–610.

Eubank, R. (1988). *Spline Smoothing and Nonparametric Regression.* New York: Dekker.

Fahrmeir, L. and G. Tutz (1994). *Multivariate Statistical Modeling Based on Generalized Linear Models*. Springer-Verlag.

Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc 87*, 998–1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist. 21*, 196–216.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and its Applications*. London: Chapman & Hall.

Fuller, W. A. (2009). *Sampling Statistics*. John Wiley & Sons, Inc.

Gasser, T. and H. G. Müller (1979). *Kernel estimation of regression functions. In: Lecture Notes in Mathematics, 757*. New York: Springer.

Gasser, T. and H. G. Müller (1984). Estimating regression functionsand their derivatives by the kernel method. *Scand. J. Statist 11*, 171–385.

Green, P. J. and B. W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models*. Washington, D. C.: Chapman and Hall.

Grund, B. and P. Hall (1993). On the performance of kernel estimators for high-dimensional sparse binary data. *Journal of Multivariate Analysis 44*, 321344.

Hall, P. (1981). On nonparametric multivariate binary discrimination. *Biometrika 68*, 287294.

Hall, P., Q. Li, and J. Racine (2007). Nonparametric estimation of regression functions in the presence of irrelevant regressors. *Review of Economics and Statistics 89*, 784–789.

Hall, P., J. Marron, and B. Park (1992). Smoothed cross-validation. *Probability Theory and Related Fields 92*, 120.

Hall, P., J. Racine, and Q. Li (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association 99*, 10151026.

Hall, P. and M. Wand (1988). On nonparametric discrimination using density differences. *Biometrika 75*, 541547.

Härdle, W. (1990). *Applied Nonparametric Regression.* New York: Cambridge University Press.

Herrmann, E., M. Wand, J. Engel, and T. Gasser (1995). A bandwidth selector for bivariate kernel regression. *Journal of the Royal Statistical Society, Series B 57*, 171–180.

Koralik, R. and J. Opsomer (2010). A survey analysis of the 2006 national survey of fishing, hunting, and wildlife-associated recreation wildlife-watching activities.

Li, Q. and J. Racine (2004). Cross-validated local linear nonparametric regression. *Statistica Sinica 14*, 485–512.

Li, Q. and J. Racine (2008). Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data. *Journal of Business and Economic Statistics 26*, 423–434.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology 140*, 1–55.

Müller, H.-G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *Journal of the American Statistical Association 82*, 231–238.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications 9*, 141–142.

Nascimento Silva, P. and C. Skinner (1997). Variable selection for regression estimation in finite populations. *Survey Methodology 23*, 23–32.

Ouyang, D., Q. Li, and J. Racine (2009). Nonparametric estimation of regression functions with discrete regressors. *Econometric Theory 25*, 1–42.

Racine, J. and Q. Li (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics 119*, 99–130.

Royall, R. M. (1970). On finite population sampling under certain linear regression models. *Biometrik 57*, 377–387.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics 9 (2)*, 6578.

Ruppert, D. and M. P. Wand (1994). Multivariate locally weighted least squares regression. *Ann. Statist. 22*, 1346–1370.

Särndal, C., I. Thomsen, J. M. Hoem, D. V. Lindley, O. Barndorff-Nielsen, and T. Dalenius (1978). Design-based and model-based inference in survey sampling. *Scandinavian Journal of Statistics 5, No. 1 (1978)*, 27–52.

Scott, D. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics.* New York: Springer-Verlag.

Stone, C. (1977). Consistent nonparametric regression (with discussion). *Annals of Statistics 5*, 595–645.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Annals of Statistics 8*, 1348–1360.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics 10*, 1040–1053.

Theil, H. (1961). *Economic forecasts and policy.* North-Holland Pub. Co.

Titterington, D. (1980). A comparative study of kernel-based density estimates for categorical data. *Technometrics 22*, 259268.

Wahba, G. (1990). *Spline models for observational data.* Society for Industrial and Applied Mathematics.

Wand, M. P. and M. C. Jones (1994). Multivariate plug-in bandwidth selection. *Computational Statistics 9*, 97–116.

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

Wang, L. and S. Wang (2011). Nonparametric additive model-assisted estimation for survey data. *Journal of Multivariate Analysis 102*, 1126–1140.

Wang, M. and J. van Ryzin (1981). A class of smooth estimators for discrete distributions. *Biometrika 68*, 301309.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā Series A 26*, 359–372.